Centre for Transport Studies
STOCKHOLM

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

Shiva Habibi-KTH
Marcus Sundberg-KTH
Anders Karlström-KTH

CTS Working Paper 2013:13

Abstract

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' specifications. Most of the studies in the field, take an inference-driven approach to select best models to estimate relevant parameters and project the results to the future, whereas we do take a prediction-driven approach. We use feature (variable) selection and cross-validation algorithms to improve predictive performance of models. These methods have been extensively used in other fields such as marketing but are scarce studies employing them in the choice modeling field. Additionally, we introduce four different prediction questions or loss-functions: overall prediction (log-likelihood), brand market share, ethanol (E85)/brand market share, and total share of ethanol cars and the predicted results of these models are compared. The results show that 'best' models prediction depend different prediction questions to answer. Also, they indicate that log-likelihood does not perform accurately when the objective is to predict a sub-section of population such as total share of E85 cars.

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

Centre for Transport Studies
SE-100 44 Stockholm
Sweden
www.cts.kth.se

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 *

Abstract

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' specifications. Most of the studies in the field, take an inference-driven approach to select best models to estimate relevant parameters and project the results to the future, whereas we do take a prediction-driven approach. We use feature (variable) selection and cross-validation algorithms to improve predictive performance of models. These methods have been extensively used in other fields such as marketing but are scarce studies employing them in the choice modeling field. Additionally, we introduce four different prediction questions or loss-functions: overall prediction (log-likelihood), brand market share, ethanol (E85)/brand market share, and total share of ethanol cars and the predicted results of these models are compared. The results show that 'best' models prediction depend different prediction questions to answer. Also, they indicate that log-likelihood does not perform accurately when the objective is to predict a sub-section of population such as total share of E85 cars.

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

*Corresponding author: shivah@kth.se

1 Introduction

The composition of the car fleet with respect to age, fuel consumption and fuel types has a great impact on environment. Recently, several different policies have been implemented to affect car fleets composition. Therefore, building models that forecast future composition of car fleet more reliably is very important. Additionally, these models are used to predict and evaluate the changes in the fleet under influence of these policies to provide decision makers with a technical support. In this paper we analyze forecasting problem and focus at building a discrete choice model to predict accurately, the demand for new cars in the Swedish car fleet in short-term future.

There are two epistemological approaches to modeling: *absolute* and *pragmatic* (Keane and Wolpin, 2007). "Absolutists" assume that there exists a true model which generates observed data. Therefore, they validate the models by testing their fitness to the data. On the other hand, "pragmatics" claim that there is no such a true model and all the models are simplifications of the real behavior of a system. Therefore, no *best* model that holds for all cases and parallel 'best' models (even seemingly contradictory) can exist and they only outperform each other according to the problems that are to be solved. One of the basic criteria is that the model should result in accurate prediction.

In the field of choice modeling, as a common practice, the absolutist approach is taken by using *statistical inference.* In statistical inference, it is assumed that an unknown true probability distribution exists from which observed data has been generated and the objective is to derive the properties of this unknown distribution from observed data. The process of finding existing pattern in the data with the objective of drawing conclusions about an unknown value (e.g. mean) in the population is called *inference.* The inferred conclusions are to validate and support the existing theories or to be in consistent with a priori knowledge. Examples of common statistical inference are hypothesis testing *(e.g. t-test)*, estimation *(e.g. log-likelihood)*and model fitness *(e.g. log-likelihood ratio).* These theory-driven restrictions on models may prevent accurate predictions results. In contrast to statistical inference, the objective of statistical prediction is to project information from data to other unknown population over time. Here, understanding data or supporting any theory is not necessary.

The objective of this study is to build a car-type choice model that gives the accurate predicted results. Therefore, we take pragmatic approach such that *predictive performance* of the model is the objective of problem solving and searching for the *best* model. Since we have large number of data related to cars attributes, we use *feature (variable) selection* [1] algorithm to obtain best predictive models. Feature selection is a search algorithm to reduce the size of data by finding the subset of variables which are useful for the analysis of interest which in our case is to select the model that generates the highest predictive performance. We use *cross-validation* to select the best model. Cross-validation (CV) is a model selection 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 validation the estimated results. A single data split is called simple validation or *hold-out* validation, and averaging over several splits is called a *cross-validation* . Various splitting strategies are found in the literature which lead to various versions of CV.

There are a vast number of methods for model selection in statistical literature including Akaike information criterion (AIC), Bayesian information criterion (BIC)[2] and CV among others. AIC and BIC penalize goodness of fit caused by model complexity which is measured by number of parameters in the model. This penalty is to avoid *over-fitting*[3]. Over-fitting occurs when a model is too fitted to the available data that looses its generality to be applied on another independent data. An over-fitted model generally has poor predictive performance. CV is a method raised specifically to fix over-fitting problem AIC is used when the objective of model selection is prediction while BIC is used to find the best structure for modeling.

The problem of optimistic outcomes of a model as a result of estimating and validating on the same data, was first addressed by Larson, 1931 (Arlot and Celisse, 2010). Cross-validation was raised to fix this problem by evaluating the generality of the

[1] ”variables” are the raw input variables and ”features” are variables constructed for the input variables. ”variable” and ”feature” can be used interchangeably when there is no impact on the selection algorithms (Guyon and Elisseeff, 2003)

[2] The AIC is $-2LL + 2K$, where LL is the value of log-likelihood and K is the number of parameters. The BIC is $-2LL + \log(N)K$, where N is sample size

[3] BIC penalizes over-fitting stronger than AIC

results of a statistical model to an independent data. (Hills, 1966; Lachenbruch and Mickey, 1968; Mosteller, F. and Tukey, 1968; Stone, 1974 ; Allen, 1974; Geisser, 1975). The idea of using CV for model selection was discussed by Efron and Morris, 1973 and Geisser, 1975. Since then, cross validation is widely used to for model selection due to its simplicity and universality. Its only assumption is that the data is independently identically distributed (i.e. i.i.d.) which can also be relaxed (Arlot and Celisse, 2010). In choice modeling literature, few studies exist that focus on prediction-driven approach to provide prediction results. McFadden et al., 1977 considers the observations before introduction of San Francisco Bay Area Rapid Transit (BART) as the estimation sample and observations after BART introduction, as the validation sample and compare the predictions with actual results. This study is unique in that there are few before and after comparison of prediction of new products in the literature. The same data are used in Train, 1979 to compare the predictive ability of complex models measured by higher number variables that are learned from data with simple models with fewer variables that are based on background knowledge. The results show that overall prediction results of the complex model are more accurate. Hensher and Ton, 2000 compare predictive performance of nested logit (NL) and artificial neural network (ANN) for commuter mode choice. In more methodological studies, Keane and Wolpin (2007) investigate the ability of nonrandom holdout sample in prediction the impacts of introduced policy and Huang et al. (2012) propose two new prediction-driven approaches to discrete choice modeling. In the car type choice modeling application, employing a multinomial logit 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 few studies focusing on evaluating prediction 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 bootstrapping method to measure the effect of the forecast error. Mohammadian and Miller (2002) compare the predictive potential of nested logit (NL) with artificial neural network (ANN) in car-type choice application. While cross-validation is used by Mohammadian and Miller, 2002 in the training process of neural network, interestingly, there is no study that apply cross-validation method in logit models and to the best of our knowledge, no study use 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 on car type choice which provides the highest predictive performance. We use Swedish car fleet data (2006-2008) to find a robust model to predict the demand for new cars while future alternatives (supply) are unknown and misspecified. Considering the changes in the supply of each year (future alternatives), using the same year supply 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 a given year. Therefore, we use out-of-sample prediction where the validation sample is the data of consecutive year. Finally, the results of this paper show as pragmatics argue, the 'best' models for prediction differ significantly according to the prediction question to answer.

This paper is structure as follows: in the next section the data is described, methodology is discussed in detail,in the section 3 and following to that models specification and results will be presented, in last section, main conclusions and future research are discussed.

2 Data

For the results presented in this paper we merge two different data sources for the years of interest, namely 2006, 2007 and 2008. The first data source is the car register that contains all passenger cars in the Swedish fleet and some characteristics of each car. The second data source contains very detailed information about all car makes/models/versions, including price, that were available on the Swedish market these years. In this section we describe each of these data sources and finally how the two are merged.

2.1 Observations from the car register

The car register contains all passenger cars that are owned privately or by a company. In this paper, we only focus on this segment. In addition to information specific to the registration of the car (e.g. first registration date and date for last status change), some main car characteristics are stored in the register such as brand, model name, vehicle year, fuel type, weight, power and body type. The age, gender and home municipality of the owner are also given in the registry data. The vehicle year is defined based on a combination of three attributes; model year, production year and first registration date because all three attributes are not available for all observations. Vehicle year is equal to model year if it is available, otherwise, the production year of the car and if this is not available either then it is equal to the year of first registration date. Since we are interested in new cars these observations need to be selected. For this purpose cars that are registered for the first time a given year but that are actually older should be excluded. We consider that a car has been bought new in 2006 if the first registration date is equal to 2006 and the vehicle year is equal to 2006 or 2007. We define new cars for 2007 in the same way. Imported cars are not included in any case. With this information there are 107,771 observations in 2006, 116,566 in 2007 and 83,609 in 2008. This definition of a newly bought car is slightly different from the one used in the official statistics that also counts older cars in. We choose to exclude these so that we can have a more accurate idea about the price paid for the car.

Table 1 reports the number and share of ethanol cars sold in 2006, 2007 and 2008. comparing 2006 to 2007 demand, the share of sold petrol cars decreased with 20% mainly in favor of diesel cars but also ethanol cars. This changes in 2008 mainly towards ethanol cars with 10% increase. The share of electic-hybrid cars and gas cars remain almost the same.

Fuel Type	**2006**		**2007**		**2008**	
	Number	Share	Number	Share	Number	Share
Petrol	83416	77.4	67011	57.5	35912	45.2
Diesel	18650	17.3	38118	32.7	26957	33.9
El-hybrid	475	0.4	586	0.5	690	0.9
Ethanol	5107	4.7	10739	9.2	15721	19.8
Gas	69	0.1	112	0.1	151	0.2

Table 1: Observations by fuel type in 2006, 2007 and 2008

2.2 Car alternatives available on the market

Some interesting attributes of the chosen cars such as price, fuel consumption and CO2 emission are missing in the car register. In order to impute this information as well as defining the choice set we use an additional data source provided by a consultant company, Ynnor, containing detailed information about all cars available on the Swedish market on the make/model/version level of detail.

In the remainder of the paper we denote this data source as supply. For 2006, 2007 and 2008 there are 2320, 2679 and 2981 cars available, respectively, corresponding to 45 different makes. Table 2 shows the share of available ethanol cars in 2006, 2007 and 2008. Since there is an increase in the number of cars available on the market from 2006 to 2007 one can note that the number of petrol cars increases but the share decreases continuously in favor of diesel cars an ethanol cars.

Fuel Type	**2006**		**2007**		**2008**	
	Number	Share	Number	Share	Number	Share
Petrol	1579	68.0	1748	65.2	1806	60.5
Diesel	703	30.3	863	32.7	1035	34.7
El-hybrid	11	0.5	13	0.5	13	0.4
Ethanol	16	0.7	44	1.6	109	3.6
Gas	11	0.5	11	0.4	12	0.4

Table 2: Cars available in the market by fuel type for 2006, 2007 and 2008

2.3 Data merging

As described above we have on the one hand the demand data from the car register where the characteristics of the chosen cars are crudely defined. On the other hand, the alternatives in the supply data are defined at a very detailed level. When merging these two data sources to impute missing information several alternatives may correspond to the same observation. The matching is done so that observation and alternatives have the same make, model, vehicle year and fuel type which are observable from demand data as well. The resulting data set contains 103,155, 116,344 & 79,435 observations and

398, 397 & 401 aggregated alternatives in 2006, 2007 and 2008, respectively.

3 Methodology

For the goal of increasing the performance of the predictive models, we use hold-out validation and feature selection methods. As stated before, CV avoids over-fitting by setting a side part of data for validation which is testing sub-set and the rest of the data is used for estimating which is training sub-set. In our case study due to the unknown and misspecified future alternatives (supply), using cross validation on a given year, will lead to a model that is over-fitted to the supply of that year; therefore we use hold-out validation in which validation is done on the data of consecutive year and it is not random. The idea of hold-out sample is to choose a non-random sample as validation set which is significantly different from estimation sample along the prediction question of interest. The robustness of this method is discussed in Keane and Wolpin, 2007 as non-random hold-out sample. They argue that although, common cross validation methods use separate estimation and validation methods, samples are still within the pattern of the same data which can not be confidentially generalized to beyond the support of the data. This is also a valid issue in our case where supply changes significantly over time. Another problem usually faced in car-type choice modeling is the large number of variables for the cars in the supply. Additionally, brand specific constants, variables' transformation and/or interactions also increase number of variables to be included in model. This problem motivates use of feature selection method. Feature selection is a search algorithm for returning new feature subsets out of the space of features such that the selected model optimizes the selection criterion which in our study is the predictive performance. Some of the main benefits provided by feature selection include: reducing the dimensionality of the feature space which reduces storage requirements and training time, removing of redundant or noisy data, better data understanding and model interpretability and improving performance of predictor. There are different algorithms for feature selection. We employ wrappers method in which a selection criterion is used to score subsets of variables based on the given criterion function. They are usually criticized to be computationally intensive (Guyon and Elisseeff, 2003). The selection criterion function is called *loss function* which is usually a specific measure of predictive error for models' fit. The objective of validation is to evaluate the predictive performance of a model over sub-set of features and optimize the loss function. The validation method is used as an accuracy selection method which computes the loss function (selection criterion) for each candidate feature subset. Usually, an exhaustive comparison of the loss function value at all 2^n possible subsets of an n-feature data is not practical. Therefore, feature selection method needs a search algorithm. Greedy search algorithms are among the most popular ones. The name greedy is due to the fact that the former decision to include or exclude variables are not re-evaluated when next decisions are taken. These algorithms are computationally advantageous and robust against over-fitting (Guyon and Elisseeff, 2003). There include two variants of these algorithms: Sequential forward selection (SFS) and sequential backward selection (elimination) (SBS or SBE). Sequential forward selection (SFS) sequentially adds features to the empty sets until any further addition does not decrease the value of loss function while sequential backward selection (elimination) (SBS or SBE) sequentially removes features from this set until any further removal does not increase the value of loss function. We use wrapper and sequential forward selection as feature selection and search algorithm, respectively.

3.1 Statistical framework

3.1.1 General notation

We have at hand a set of observations $\xi = \xi_1, .., \xi_n \in \Xi$ with common distribution F called sample. ξ is the realization of a random vector Ξ with unknown joint distribution function, $F_\Xi(\xi)$, or in other words, $F_\Xi(\xi)$ is a true data generating distribution which has generated sample ξ. Each observation ξ_i consists of $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ which are explanatory and response variables, respectively, and $\Xi = \mathcal{X} \times \mathcal{Y}$.

A statistical inference is to make statements about the unknown distribution function $F_\Xi(\xi)$, based on the observed sample ξ. Let s_0 be a quality of $F_\Xi(\xi)$ and $\hat{s}$ be the estimate of s_0, then the function $\hat{s}(\xi)$ is called an estimator. A *loss function* $l(\hat{s}, s_0)$, where $l : \mathbb{S} \times \mathbb{S} \longmapsto \mathbb{R}, s \in \mathbb{S}$ measures the discrepancy obtained by estimating s_0 with $\hat{s}$.

The risk of an estimator, $R(\hat{s})$, is expected value of its loss:

$$R(\hat{s}) = \mathbb{E}_{\xi \sim F}\left[l\left(\hat{s}(\xi), s_0\right)\right] \tag{1}$$

Both s_0 and $F_\Xi(\xi)$ are unknown, therefore, the risk is not known as well and should be estimated. In prediction we aim at predicting a quantity of interest $Y \in \mathcal{Y}$ given an explanatory variable $X \in \mathcal{X}$. $p \in \mathbb{S}$ maps $\mathcal{X}$ to $\mathcal{Y}$, and the loss function $l(p;(x,y))$ measures the discrepancy between y and its predicted value $p(x)$. The loss is minimal for $p(x) = y$.

Several loss functions can be chosen for a given statistical problem. Some of the popular loss functions are as follows:

- *Negative log-likelihood*, where $l(\hat{s}(\xi), s_0) = -\ln(\hat{s}(\xi))$ aiming at estimating quality s of F.
- *Squared error*, where $l(\hat{s}(\xi), s_0) = (\hat{s}(\xi) - s_0)^2$. When the squared error is used as a loss function, then the risk is called mean squared error (MSE) and the square root of it is called, root mean squared error (RMSE).

3.1.2 Feature selection

Each x_i consists of m input variables $x_{i,k},\ \ k = 1, ..., m$. Define $v^{(j)} := (x_{i,k})_{k \in I^{(j)}}$ where $I^{(j)}$ is a subset of $1, ..., m$, our objective is to find $v^* \in \mathbb{V}$ that minimizes loss of prediction as a criterion function:

$$v^* := \arg\min_{v^{(j)}} l(p(v^{(j)}), y^{(j)}) \tag{2}$$

To evaluate criterion function, we use hold-out validation. Training sample is used to estimate the estimator and validation sample is used to estimate the risk of this estimator (See (1)). Hold-out validation (or CV in general) selects the model with the smallest estimated risk. The hold-out estimator is generally defined as (Arlot and Celisse, 2010):

$$l^{HO}(\xi; I^{(t)}) := \frac{1}{\left|I^{(h)}\right|} \sum_{i \in \xi^{(h)}} l(\hat{s}(\xi^{(t)}); \xi_i)\ \ , \tag{3}$$

where $I^{(t)}$ is subset of $1, ..., n$ with its complement $I^{(h)}$, $\xi^{(t)} := (\xi_i)_{i \in I^{(t)}}$ is the training set and $\xi^{(h)} := (\xi_i)_{i \in I^{(h)}}$ is the validation set. re-arranging (3) in feature-selection frame work with respect to $p(v^j)$ results in:

$$l^{HO}(v^{(t,j)}; I^{(t)}) := \frac{1}{\left|I^{(h)}\right|} \sum_{i \in \xi^{(h)}} l(p(v^{(t,j)}), (v^{(t,j)}, y^{(j)})) \ , \tag{4}$$

where $v^{(t,j)} := (x_{i,k})_{i \in I^{(t)}, k \in I^{(j)}}$.

4 Model specification and results

We use registry data from 2006-2008 and our objective is to build a model to predict year 2008 market share based on data on years 2006 and 2007. We are interested in these year series due to the large changes happening in both supply and demand side including the introduction of purchasing subsidy of 10,000 SEK for clean cars in year 2007 and the boost in the number of ethanol cars introduced in the market in these years as can be seen in table 2.

Table 3 shows the car attributes used for modeling. These variables together with quadratic form of non-dummy variables contain a set of 85 variables used for sequential feature selection in the same order presented in table 3. For estimation on the training data, we use multinomial logit model (MNL) with linear-in-parameter specifications. As discussed in section 2, alternatives available in the supply are not observed in the demand data and we have to use aggregated form of alternatives. We correct aggregation of alternatives as follows (Ben-Akiva and Lerman (1985)):

$$V_i = \bar{V_i} + \mu \ln n_i + \mu \ln[\frac{1}{n_i} \sum_{l \in L_i} \exp^{\frac{1}{\mu}(V_l - \bar{V_i})}] \tag{5}$$

where,

V_i, deterministic utility of aggregate alternative,

$V_i = \frac{1}{m_i} \sum_{l \in L_i} V_l$, average of disaggregate alternatives' deterministic utilities,

L_i, set of disaggregate alternatives corresponding to aggregate alternative i,

n_i, number of disaggregate alternatives in the L_i,

Table 3: Description of cars attributes

Attribute	**Description**
log(n)	Number of sub-alternatives constituting each aggregated alternative
Brand Specific	Constant
Origin Specific	Constant
Cabriolet	Dummy for cabriolets
Copue	Dummy for Coupes
Hatch-backs	Dummy for Hatch -backs
Minibuss	Dummy for minibuss
Minivan	Dummy for minivan
MPV	Dummy for MPVs
Sedan	Dummy for Hatch -backs
SUV	Dummy for Hatch -backs
GAS	Dummy for gas cars
E85	Dummy for ethanol-hybrid cars
El	Dummy for electrical-hybrid cars
Diesel	Dummy for diesel cars
Price	Purchase Price in 1000,000 SEK
Tax	Vehicle circulation tax in 1000 SEK[1]
Fuel-consumption	liter per 100 km
CO2	gr per 10 m
Tank-volume	in liter
Weight	kg
Power	kw
Clean	Dummy for clean cars
Fuel-cost	Fuel cost per 100 km in 1000 SEK[2]
Lux	Dummy for luxury car (purchase price over 800,000 SEK)
Weight/power	kg/kw

[1] vehicle circulation tax= base tax(360 SEK) + CO2 component (20 SEK/gr of CO2 emission for conventional, 10 SEK/gr of CO2 emission for alternative fuels. For diesel cars, tax of conventional car tax is multiplied by 3.15. 1 USD is approx 6.45 SEK in February, 2013.

[2] Fuel cost = fuel price (SEK/lit) * fuel consumption (lit/100km). For the hybrid vehicles the minimum cost of running on different fuels is used.

μ, nesting parameter.

The second term of the equation is the measure for the size and the third term is the measure for the heterogeneity. We treat $\log(n)$ as one of the variables and estimate its parameter, μ. We do not consider the correction for heterogeneity in this study.

We introduce four different loss functions as well as two types of training sets: cross-section (i.e. 2007) and time-series (i.e. 2006-2007). This is motivated by the fact that time-series data include variation in both supply and demand in successive years, therefore it is assumed to give more robust prediction. Finally, we will end up 8 different loss function that will give us 8 different loss functions are as follows:

1. Negative log-likelihood,
   (a) $-\sum_i(\log(\hat{P}_{i,2008}|x_{i,k,2007}, \hat{\beta}_{2007}))$
   (b) $-\sum_i(\log(\hat{P}_{i,2008}|x_{i,k,2006-2007}, \hat{\beta}_{2006-2007}))$
2. Root mean square error for brands market share
   (a) $\sqrt{\frac{1}{|brands|}\sum_{b\in brands}((\hat{P}_{b,2008}|x_{i,k,2007}, \hat{\beta}_{2007}) - Sh_{b,2008})^2}$
   (b) $\sqrt{\frac{1}{|brands|}\sum_{b\in brands}((\hat{P}_{b,2008}|x_{i,k,2006-2007}, \hat{\beta}_{2006-2007}) - Sh_{b,2008})^2}$
3. Root mean square error for ethanol/brand market share [4]
   (a) $\sqrt{\frac{1}{|E85/brands|}\sum_{e\in E85/brands}((\hat{P}_{e,2008}|x_{i,k,2007}, \hat{\beta}_{2007}) - Sh_{e,2008})^2}$
   (b) $\sqrt{\frac{1}{|E85/brands|}\sum_{e\in E85/brands}((\hat{P}_{e,2008}|x_{i,k,2006-2007}, \hat{\beta}_{2006-2007}) - Sh_{e,2008})^2}$
4. Total share of ethanol cars
   (a) $((\hat{P}_{E85,2008}|x_{i,k,2007}, \hat{\beta}_{2007}) - Sh_{E85,2008})$
   (b) $((\hat{P}_{E85,2008}|x_{i,k,2006-2007}, \hat{\beta}_{2006-2007}) - Sh_{E85,2008})$

   where,

   brands and *E85/brands* denote the set of available brands and the ethanol

[4] The reason that we consider ethanol cars and not clean cars, is that based on definition of clean cars in Sweden we also need to have CO_2 emission and fuel consumption to identify clean cars whereas this data is missing in the demand and to acquire this data, we need to include the average of relevant attributes from supply side which brings uncertainty in market share of the clean cars.

cars within each brand, respectively,
$\hat{\beta}$ denotes estimated parameter on relevant training set,
Sh denotes different market share, and,
$\hat{P}$ denotes estimated probability of considered market share.

Based on what discussed here and in section 3, procedure of searching for the best models are summarized as follow:

1. variables are sequentially added to the empty subset, $I^{(j)}$
2. training and testing sets are generated, which are $\xi^{(t,yr)}$, $\xi^{(h,yr)}$, respectively. yr denotes year.
3. a discrete choice model of make/model/fuel-type is estimated on the training set, i.e. $\hat{\beta}_{yr}$
4. The value of pertinent loss-function is calculated and compared to the value from the previous iteration.
   These steps are repeated until adding more variables to the set do not reduce the loss-function value.

Table 6 in Appendix presents the final results of selected variables based on different loss-functions. It shows that totally different 'best' models are acquired by introducing different loss-functions. As can be seen, $\log(n)$ is only included when the loss-function is log-likelihood indicating the fact that each alternative aggregated by make/model/fuel-type contains different versions of cars, is important when the purpose of prediction is to predict the market share of each aggregated alternative as in log-likelihood function. Comparing the entire set of models estimated on pooled data with that of ones estimated on the cross-section data, show less variables are included in the former models except than the models with RMSE of brand share as loss function. Additionally, models with LL as their loss-functions contain more variables among others for each data set. In other words these models describe data better. Price variable is only included in models with LL as loss function, it has expected negative sign and significant. But it shows less sensitivity with pooled data. Also, quadratic form of the price is included in cross-section data set in LL as loss function with positive sign.

Unlike price, tax is not included in any of the models. However, quadratic form of it gets included in models with RMSE of E85/brand shares in both datasets. But it has positive sign in cross-sectional data set. Fuel cost is included significantly and with expected sign, in the model estimated on pooled data and total share of E85 as loss function, also the one estimated on cross-section data and RMSE of E85/brand share as loss function.

Finally, we might not able to motivate inclusion or exclusion of every variables or their signs as we would with the standard logit models, specially when it comes to the large number of brand specific constants and other correlated variables such as fuel cost and fuel consumption, or tax and price with fuel dummies, but as described in section 1 the objective of this study is not to find the most relevant variables or to explain data as good as possible but to include the ones which give higher prediction results regardless of their ability to explain data or significance of the relevant parameters.

Results presented in tables 4 and 5 compare the predictive performance of different acquired models. In these tables the RMSE of brand and E85/brand market shares of models with respective loss functions are compared with the ones with LL as loss function for cross-section and time-series data. In all cases the results given by LL are higher. The values of RMSE for E85/brand market share for LL loss function are 1.86 and 1.96 for cross-section and time-series data, respectively presented in table 4. These values are both more than their corresponding values for RMSE of E85/brand, which are 0.37 and 0.43 respectively. Total share of E85 cars with the perspective loss-functions are the same as its actual number which is 19.89. The corresponding values from models with LL as a loss function equivalent to 17.36 & 18.43 in cross-section and time-series data, respectively. The same trend can be observed in table 5. Comparing LL with RMSE of brand market share as loss functions, also shows the superiority of models acquired by the respective RMSE as loss function to the ones with LL as loss function, with the value of RMSE 0.65 and 0.35 less than 1.52 and 1.59 respectively for for cross-section and time-series data. From the presented results, it can not be concluded that which of the models estimated on time-series or cross-section data perform better regarding prediction. Models estimated on time-series are superior in predicting total share of ethanol cars with LL as loss function and RMSE of brand market share while the models acquired from cross-section data give better predicted results in both market shares of brand and E85/brand with LL as loss function and market share of E85/brand with respective loss function. This results are not in accordance with our hypothesis that time-series data will increase the predictive performance. The reason to these results could be the fact that supply of 2007 is more similar to the that of 2008 than supply of 2006 which suggests the probable over-fitting of the time-series models to the supply of two successive years.

		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	–	0,09	0,02	–
CHEVROLET	0,02	0,01	0,09	–	0	0	–
CITROEN	0,17	1,21	1,22	–	0,98	0,86	–
FORD	3,04	2,55	3,13	–	2,28	2,72	–
PEUGEOT	1,16	2,16	1,44	–	2,1	1,7	–
RENAULT	0,76	1,87	0,52	–	2,33	0,68	–
SAAB	3,27	4	2,9	–	4,62	2,59	–
SEAT	0,25	0,22	0,12	–	0,33	0,19	–
SKODA	1,33	1,73	1,37	–	2,12	1,71	–
VOLKSWAGEN	1,95	1,39	1,6	–	1,68	1,24	–
VOLVO	7,83	2,1	7,8	–	1,9	7,93	–
Total	**19,89**	17,36	20,2	**19,89**	18.43	19,64	**19,89**
RMSE		**1,83**	**0,37**		**1.96**	**0,43**	

Table 4: Predicted results for market share of E85 cars

5 Conclusion and future work

The objective of this study is to find a robust model for prediction of car-type choice in short-run future while taking the pragmatic prediction-driven view and applying hold-out validation together with feature selection to find the best prediction model. We estimate MNL model on a given year, t (i.e. 2006 or 2006-2007) and validate the estimated results to the successive year $t + 1$ (i.e. 2008). The year $t + 1$ is called hold-out sample and this method is called hold-out validation. This prevents over-fitting of the models to the supply of a specific year. Feature selection is an automatic way of select-

		Estimation on 2007		**Estimation on 2006-2007**	
	Loss function:	LL	RMSE brand share	LL	RMSE brand share
Brand	**Actual share**	**Predicted share**		**Predicted share**	
ALFA ROMEO	0,04	0.05	0.04	0.08	0.05
AUDI	4,38	4.73	4.68	5.39	4
BENTLEY	0,01	0	0	0	0
BMW	4,23	6.47	4.03	6.51	3.9
CADILLAC	0,12	0.16	0.03	0.16	0.11
CHEVROLET	0,58	0.39	0.26	0.52	0.35
CHRYSLER	0,08	0.08	1.02	0.2	0.24
CITRON	4,10	6.25	6.54	5.13	4.9
DODGE	0,02	0.15	0.17	0.15	0.07
FERRARI	0,03	0.01	0	0.02	0
FIAT	0,46	0.43	0.24	0.36	0.2
FORD	5,83	5.65	6.02	5.06	6.47
HONDA	2,20	2.02	2.21	1.91	1.67
HUMMER	0,00	0	0.07	0	0
HYUNDAI	5,66	6.44	4.51	5.87	4.89
JAGUAR	0,16	0.06	0.05	0.09	0.07
JEEP	0,06	0.05	0.08	0.16	0.13
KIA	3,15	2.52	3.74	2.05	4.13
KOE	0,00	0	0	0	0
LAMBORGHINI	0,01	0.01	0	0	0
LAND ROVER	0,07	0.03	0.12	0.05	0.09
LEXUS	0,21	0.19	0.16	0.23	0.27
LOTUS	0,00	0	0	0	0
MASERATI	0,03	0	0	0.01	0
MAZDA	1,62	2.18	1.92	2.43	1.91
MERCEDES	1,95	2.51	1.91	3.08	1.93
MINI	0,39	0.32	0.34	0.32	0.32
MITSUBISHI	1,14	0.78	1.27	0.8	1.38
MORGAN	0,00	0	0	0	0
NISSAN	2,49	1.72	1.69	1.98	1.68
OPEL	3,71	3.25	3.57	3.3	3.79
PEUGEOT	6,13	9.06	6.99	8.1	6.82
PORSCHE	0,12	0.05	0.05	0.15	0.14
RENAULT	1,83	6.14	2.96	6.45	3.51
ROLLS_ROYS	0,00	0	0	0	0
SAAB	4,32	5.25	3.53	6.73	3.93
SEAT	1,15	0.89	0.7	0.99	0.75
SKODA	6,26	5.46	6.45	7.05	6.18
SMART	0,06	0.02	0.05	0.01	0.02
SSANGYONG	0,05	0.04	0.06	0.07	0.04
SUBARU	1,03	0.79	1.12	0.67	1.19
SUZUKI	0,20	0.57	0.76	0.32	0.64
TOYOTA	10,61	7.2	8.23	6.14	8.6
VOLKSWAGEN	9,62	9.33	9.1	8.51	10.17
VOLVO	15,91	8.75	15.31	8.95	15.43
Total	100	100	99.98	100	99.97
RMSE	**-**	**1.52**	**0.65**	**1.59**	**0.35**

Table 5: Predicted results for market share of brands

ing variables to be included in the model. This method is very useful in car type choice application since there exists a large number of car attributes to select among and also due to their correlations and interactions, selecting them based only on priori knowledge will not be accurate.

Table 6 shows that different prediction questions or their consequent loss function result in different 'best' models which validates the approach of pragmatics that alternative models may coexist for different purposes unlike the absolutist assumption of existing a 'true' moedl. Tables 4 and 5 show that in all cases, the best models resulted from particular prediction question outperform the ones with log-likelihood as their loss-function. The explanation to this is that LL assigns the same weights on all alternatives and observations and gives the overall prediction while in prediction we are specifically interested in a sub-section of the data. Considering the fact that log-likelihood is a robust estimator, these results indicate the objective of selecting the models with good predictive performance is not the same as the objective of selecting the estimation of relevant variables as in log-likelihood function.

$\log(n)$ is not included in any of the best prediction models. Thus, simple aggregation of alternatives lead to the best predicted results which implies that unlike inference (estimation), the fact that each aggregate alternative consists of several sub-alternatives do not matter in prediction. Stating intuitively, our result show that to avoid over-fitting to a supply of a given year, using time series data could result in a more generalizable results.

With applying the methods presented in this paper, we got closer to the objective of accurate prediction of car fleet. Yet, There are still other issues that need to be addressed and considered. Right now, it is probable that these results might be over-fitted to the actual supply of the year 2008. The 'best' models chosen here need to be tested by using possible future scenarios, like supply of another year (e.g. 2009) to gain enough confidence in the accuracy of their prediction. Moreover, the results presented in this study, show the importance of time-series particularity which verifies the sensitivity of models to the supply and therefore modeling of the supply side could be considered in future studies.

Additionally, more aggregated alternatives (e.g. the choice of ethanol vs. non-ethanol cars) can also be considered to investigate of predictive power of LL as loss-function while alternatives represents the sub-section. Finally, We used forward selection as a search algorithm which is more efficient computationally than backward elimination. However, backward elimination might lead to the better models since variables are evaluated in the presence of other variables. So, it could be interesting to compare these results from that of SBE if it is computationally practicable.

Appendix

Table 6: Variables selected in each model

		Estimation on 2006-2007 / Validation on 2008				**Estimation on 2007 / Validation on 2008**			
	loss function:	LL	RMSE Brand share	RMSE E85/brand share	E85 total share	LL	RMSE Brand share	RMSE E85/brand share	E85 total share
No	**Variable**								
1	logN	0.95	0.00	0.00	0.00	0.86	0.00	0.00	0.00
		(305.89)	(0.00)	(0.00)	(0.00)	(159.97)	(0.00)	(0.00)	(0.00)
2	Bsc_ALF	-0.74	-1.79	-0.98	0.00	-0.97	-2.17	-1.49	0.00
		(-4.30)	(-18.42)	(-10.17)	(0.00)	(-4.27)	(-12.57)	(-4.44)	(0.00)
3	Bsc_AUD	-0.57	1.61	0.00	0.00	-0.55	-0.53	-1.69	0.00
		(-48.11)	(0.00)	(0.00)	(0.00)	(-32.71)	(-29.53)	(-90.00)	(0.00)
4	Bsc_BEN	2.34	0.00	6.42	0.00	0.00	0.00	3.17	0.00
		(7.19)	(0.00)	(19.55)	(0.00)	(0.00)	(0.00)	(4.99)	(0.00)
5	Bsc_BMW	0.00	2.00	0.00	0.00	0.00	0.00	-1.61	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-79.04)	(0.00)
6	Bsc_CAD	-1.67	0.00	-3.53	0.00	-1.20	-4.50	-4.42	0.00
		(-11.23)	(0.00)	(-23.83)	(0.00)	(-3.70)	(-18.43)	(-18.24)	(0.00)
7	Bsc_CHE	0.00	1.25	-2.65	0.00	0.00	-2.53	0.00	0.00
		(0.00)	(19.22)	(-67.08)	(0.00)	(0.00)	(-35.87)	(0.00)	(0.00)
8	Bsc_CHR	0.00	1.44	-1.28	0.00	0.00	0.00	-1.57	0.00
		(0.00)	(22.69)	(-35.11)	(0.00)	(0.00)	(0.00)	(-25.23)	(0.00)
9	Bsc_CIT	-0.71	-0.43	-0.94	0.00	-0.14	-0.16	-0.18	0.00
		(-62.57)	(-39.53)	(-77.13)	(0.00)	(-8.45)	(-11.65)	(-10.88)	(0.00)
10	Bsc_DOD	-0.57	0.00	-2.73	0.00	-0.46	-2.44	-2.62	0.00
		(-8.40)	(0.00)	(-43.79)	(0.00)	(-4.74)	(-34.13)	(-36.06)	(0.00)
11	Bsc_FER	0.00	0.00	9.43	0.00	2.97	0.00	0.92	0.00
		(0.00)	(0.00)	(48.88)	(0.00)	(7.42)	(0.00)	(0.97)	(0.00)
12	Bsc_FIA	-0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(-2.19)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
13	Bsc_FOR	0.18	3.55	0.00	0.00	0.85	0.00	0.00	0.00
		(6.32)	(65.45)	(0.00)	(0.00)	(16.29)	(0.00)	(0.00)	(0.00)
14	Bsc_HON	1.37	0.29	0.48	0.00	1.25	0.70	-0.71	0.00
		(73.21)	(14.87)	(26.16)	(0.00)	(48.00)	(31.05)	(-33.91)	(0.00)
15	Bsc_HUM	-2.56	0.00	-2.65	0.00	-1.31	0.00	-2.56	0.00
		(-3.82)	(0.00)	(-3.50)	(0.00)	(-2.69)	(0.00)	(-4.55)	(0.00)
16	Bsc_HYU	0.00	0.00	0.00	0.00	0.00	0.00	-1.63	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-77.42)	(0.00)
17	Bsc_JAG	0.00	1.85	0.00	0.00	0.00	2.47	0.00	0.00
		(0.00)	(10.24)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
18	Bsc_JEE	-0.21	0.85	-1.82	0.00	-0.66	-2.89	-2.28	0.00
		(-3.48)	(11.45)	(-32.55)	(0.00)	(-6.83)	(-33.94)	(-26.26)	(0.00)
19	Bsc_KIA	-0.79	0.00	0.08	0.00	-0.71	0.00	-1.09	0.00
		(-48.96)	(0.00)	(3.82)	(0.00)	(-36.60)	(0.00)	(-56.97)	(0.00)
20	Bsc_KNG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
21	Bsc_LAM	-0.46	0.00	5.94	0.00	0.00	0.00	0.00	0.00
		(-1.30)	(0.00)	(18.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
22	Bsc_LAN	0.00	3.75	0.00	0.00	0.00	5.21	3.33	0.00
		(0.00)	(20.06)	(0.00)	(0.00)	(0.00)	(0.00)	(20.09)	(0.00)
23	Bsc_LEX	-0.46	-2.79	0.00	0.00	-0.77	-2.23	-2.71	0.00
		(-11.41)	(-70.61)	(0.00)	(0.00)	(-11.66)	(-35.49)	(-41.66)	(0.00)

Continued on next page

Table 6 – continued from previous page

		Estimation on 2006-2007 /				Estimation on 2007 /			
				Validation on 2008				Validation on 2008	
	loss function:	LL	RMSE	RMSE	E85	LL	RMSE	RMSE	E85
			Brand	E85/brand	total		Brand	E85/brand	total
			share	share	share		share	share	share
No	**Variable**								
24	Bsc_LOT	-1.91	0.00	-5.59	0.00	-2.82	-1.82	-1.60	0.00
		(-4.11)	(0.00)	(-11.22)	(0.00)	(-1.68)	(-2.72)	(-1.56)	(0.00)
25	Bsc_MAS	0.00	-0.14	3.72	0.00	1.36	0.00	0.00	0.00
		(0.00)	(0.00)	(9.44)	(0.00)	(5.12)	(0.00)	(0.00)	(0.00)
26	Bsc_MAZ	0.31	0.10	0.02	0.00	0.23	0.00	-0.81	0.21
		(15.44)	(4.64)	(1.15)	(0.00)	(7.64)	(0.00)	(-31.69)	(9.30)
27	Bsc_MER	-0.95	1.75	0.00	0.00	-0.72	-0.81	-1.52	0.00
		(-59.78)	(0.00)	(0.00)	(0.00)	(-30.26)	(-34.89)	(-64.84)	(0.00)
28	Bsc_MIN	0.37	4.35	0.00	0.00	0.00	4.99	2.28	0.00
		(5.40)	(24.26)	(0.00)	(0.00)	(0.00)	(0.00)	(16.02)	(0.00)
29	Bsc_MIT	0.69	0.00	0.00	0.00	0.34	0.05	-0.74	0.00
		(31.15)	(0.00)	(0.00)	(0.00)	(10.09)	(1.60)	(-23.94)	(0.00)
30	Bsc_MOR	-2.17	-0.23	-5.14	0.00	-1.50	0.00	0.00	0.00
		(-6.12)	(-0.99)	(-15.31)	(0.00)	(-3.74)	(0.00)	(0.00)	(0.00)
31	Bsc_NIS	0.00	0.18	0.00	0.00	0.00	0.04	-0.73	0.00
		(0.00)	(8.11)	(0.00)	(0.00)	(0.00)	(1.50)	(-28.45)	(0.00)
32	Bsc_OPE	-0.98	2.17	0.06	0.00	-0.88	-0.48	-0.96	0.00
		(-82.78)	(0.00)	(4.74)	(0.00)	(-48.79)	(-26.77)	(-57.33)	(0.00)
33	Bsc_PEU	-0.37	0.00	-0.40	0.00	0.00	0.00	0.00	0.00
		(-41.46)	(0.00)	(-40.38)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
34	Bsc_POR	-1.45	0.24	-1.65	0.00	-1.18	-2.95	-3.23	0.00
		(-24.19)	(0.00)	(-29.64)	(0.00)	(-9.19)	(-23.87)	(-25.76)	(0.00)
35	Bsc_REN	-0.92	-0.19	-1.05	0.00	-0.47	-0.53	-0.18	0.00
		(-76.15)	(-17.10)	(-83.92)	(0.00)	(-22.59)	(-30.56)	(-8.58)	(0.00)
36	Bsc_RSR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
37	Bsc_SAA	-0.05	0.00	0.19	0.00	-0.22	0.00	0.00	0.00
		(-4.06)	(0.00)	(17.33)	(0.00)	(-11.71)	(0.00)	(0.00)	(0.00)
38	Bsc_SEA	-2.38	-2.16	-2.57	0.00	-2.53	-2.26	-2.30	0.00
		(-89.61)	(-81.57)	(-95.00)	(0.00)	(-61.81)	(-56.44)	(-56.82)	(0.00)
39	Bsc_SKO	0.00	0.40	0.25	0.00	-0.26	0.54	0.00	1.30
		(0.00)	(40.59)	(21.95)	(0.00)	(-15.68)	(39.36)	(0.00)	(97.81)
40	Bsc_SMA	-4.62	0.00	0.00	0.00	-4.61	-2.99	-3.10	0.00
		(-34.58)	(0.00)	(0.00)	(0.00)	(-30.91)	(-20.26)	(-19.80)	(0.00)
41	Bsc_SSY	-2.07	-2.25	-0.79	0.00	-2.25	-2.15	-2.04	0.00
		(-20.37)	(-22.33)	(-7.54)	(0.00)	(-21.18)	(-20.95)	(-19.65)	(0.00)
42	Bsc_SUB	0.00	-0.13	0.29	0.00	0.34	0.00	-0.79	0.00
		(0.00)	(-5.47)	(12.44)	(0.00)	(9.71)	(0.00)	(-23.84)	(0.00)
43	Bsc_SUZ	-0.10	0.00	-0.74	0.00	0.00	0.00	-0.71	0.00
		(-3.46)	(0.00)	(-26.65)	(0.00)	(0.00)	(0.00)	(-19.62)	(0.00)
44	Bsc_TOY	1.41	1.48	1.39	0.00	1.50	1.49	0.00	0.00
		(92.97)	(91.21)	(95.38)	(0.00)	(64.59)	(80.13)	(0.00)	(0.00)
45	Bsc_VWA	0.00	2.80	0.55	0.00	0.21	0.00	0.00	0.00
		(0.00)	(0.00)	(48.50)	(0.00)	(14.39)	(0.00)	(0.00)	(0.00)
-	Bsc_VOL	-	-	-	-	-	-	-	-
46	Osc_AME	-1.55	-3.86	-0.73	0.00	-2.13	-0.51	-0.87	0.00
		(-59.26)	(-72.07)	(-65.02)	(0.00)	(-42.21)	(-37.13)	(-61.44)	(0.00)
47	Osc_BRI	-2.42	-5.67	-2.69	0.00	-2.42	-6.60	-4.42	0.00
		(-47.78)	(-32.56)	(-77.66)	(0.00)	(-54.14)	(0.00)	(-34.07)	(0.00)
48	Osc_CZE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Continued on next page

Table 6 – continued from previous page

		Estimation on 2006-2007 / Validation on 2008				Estimation on 2007 / Validation on 2008			
	loss function:	LL	RMSE Brand share	RMSE E85/brand share	E85 total share	LL	RMSE Brand share	RMSE E85/brand share	E85 total share
No	**Variable**								
49	Osc_FRE	0.00	0.00	0.00	0.00	-0.34	0.00	-0.49	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(-24.13)	(0.00)	(-34.82)	(0.00)
50	Osc_GER	0.00	-2.88	-1.12	0.00	0.00	-0.22	0.00	0.00
		(0.00)	(0.00)	(-123.95)	(0.00)	(0.00)	(-20.52)	(0.00)	(0.00)
51	Osc_ITA	-2.55	-2.57	-3.35	0.00	-2.66	-2.68	-3.12	0.00
		(-16.43)	(-59.58)	(-73.32)	(0.00)	(-44.83)	(-48.90)	(-54.56)	(0.00)
52	Osc_JPN	-1.31	-1.06	-1.44	0.00	-1.30	-1.08	-0.21	0.00
		(-96.03)	(-70.83)	(-101.81)	(0.00)	(-58.91)	(-66.74)	(-15.72)	(0.00)
53	Osc_KOR	0.00	0.00	-1.57	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(-105.30)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
54	Osc_SPA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
55	Cab	-0.92	-1.25	0.00	0.00	-1.57	-1.51	-1.42	0.00
		(-42.85)	(-77.42)	(0.00)	(0.00)	(-50.25)	(-61.73)	(-52.79)	(0.00)
56	Coupe	-1.28	0.00	0.00	-0.62	-1.44	0.00	0.00	-0.64
		(-58.28)	(0.00)	(0.00)	(-42.41)	(-56.51)	(0.00)	(0.00)	(-34.28)
57	Hatch	-0.39	0.00	0.00	0.00	-0.74	0.00	0.00	0.00
		(-45.52)	(0.00)	(0.00)	(0.00)	(-58.83)	(0.00)	(0.00)	(0.00)
58	Minibuss	-2.93	-2.22	0.00	0.00	-1.96	-2.41	0.00	0.00
		(-45.16)	(-35.70)	(0.00)	(0.00)	(-21.06)	(-26.84)	(0.00)	(0.00)
59	Minivan	-1.55	0.00	0.00	0.00	-1.78	0.00	0.00	0.00
		(-75.79)	(0.00)	(0.00)	(0.00)	(-56.18)	(0.00)	(0.00)	(0.00)
60	MPV	-1.95	0.00	-1.70	0.00	-1.43	0.00	-0.95	-1.49
		(-60.20)	(0.00)	(-52.11)	(0.00)	(-36.85)	(0.00)	(-23.96)	(-38.80)
61	Sedan	-1.42	0.00	0.00	0.00	-1.69	0.00	0.00	0.00
		(-124.56)	(0.00)	(0.00)	(0.00)	(-97.94)	(0.00)	(0.00)	(0.00)
62	SUV	-0.31	0.00	0.28	0.00	-0.38	0.00	0.00	0.00
		(-25.70)	(0.00)	(24.54)	(0.00)	(-23.50)	(0.00)	(0.00)	(0.00)
63	Gas	-2.23	0.00	-3.48	0.00	-2.60	0.00	-5.89	0.00
		(-28.37)	(0.00)	(-44.22)	(0.00)	(-26.48)	(0.00)	(-58.61)	(0.00)
64	E85	0.39	0.58	-0.96	1.34	0.15	0.45	-1.21	1.30
		(20.77)	(65.36)	(-47.09)	(158.68)	(5.36)	(40.95)	(-12.23)	(124.80)
65	El	-1.08	0.00	0.00	0.00	0.00	0.00	-3.37	0.00
		(-30.45)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-53.20)	(0.00)
66	Diesel	0.00	0.00	0.00	0.00	0.00	0.00	-4.44	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-50.56)	(0.00)
67	Price	-0.51	0.00	0.00	0.00	-6.55	0.00	0.00	0.00
		(-8.38)	(0.00)	(0.00)	(0.00)	(-29.87)	(0.00)	(0.00)	(0.00)
68	Tax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
69	Fuel consumption	0.00	0.00	22.43	0.00	1.98	0.01	26.79	0.00
		(0.00)	(0.00)	(82.50)	(0.00)	(10.42)	(0.00)	(9.73)	(0.00)
70	CO2	0.00	0.00	-5.28	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(-40.46)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
71	Tank_volume	0.00	13.48	0.00	0.01	0.00	0.00	0.00	0.00
		(0.00)	(98.57)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
72	weight	3.81	0.00	-1.87	0.00	0.00	0.00	12.04	0.00
		(42.06)	(0.00)	(-88.64)	(0.00)	(0.00)	(0.00)	(33.56)	(0.00)
73	Power	0.00	0.00	0.00	0.00	-0.30	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(-6.10)	(0.00)	(0.00)	(0.00)
74	Clean	1.12	0.00	0.09	0.00	1.08	0.00	0.00	0.00
		(67.88)	(0.00)	(5.08)	(0.00)	(54.31)	(0.00)	(0.00)	(0.00)

Continued on next page

Table 6 – continued from previous page

		Estimation on 2006-2007 / Validation on 2008				Estimation on 2007 / Validation on 2008			
	loss function:	LL	RMSE Brand share	RMSE E85/brand share	E85 total share	LL	RMSE Brand share	RMSE E85/brand share	E85 total share
No	**Variable**								
75	Fuel_cost	0.00	0.00	0.00	-0.11 (-	0.00	0.00	-4.02	0.00
		(0.00)	(0.00)	(0.00)	132.24)	(0.00)	(0.00)	(-17.71)	(0.00)
76	Lux	-0.51	-2.03	-1.10	0.00	0.00	0.53	-2.18	0.00
		(-8.44)	(-46.77)	(-24.04)	(0.00)	(0.00)	(5.67)	(-19.74)	(0.00)
77	Weight/Power	0.00	-1.94	0.00	0.00	-2.64	-11.49	-6.41	0.11
		(0.00)	(-142.12)	(0.00)	(0.00)	(-46.16)	(-85.20)	(-23.63)	(10.40)
78	sq_Price	0.00	0.00	0.00	0.00	3.19	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(23.30)	(0.00)	(0.00)	(0.00)
79	sq_Tax	0.00	0.00	-0.05	0.00	0.00	0.00	0.02	0.00
		(0.00)	(0.00)	(-81.65)	(0.00)	(0.00)	(0.00)	(10.91)	(0.00)
80	sq_Fuel consumption	0.00	0.00	0.00	0.00	1.43	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(4.57)	(0.00)	(0.00)	(0.00)
81	sq_CO2	0.00	0.00	1.49	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(41.51)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
82	sq_Tank_volume	2.11	-13.92	1.52	0.00	0.00	0.00	-0.29	0.00
		(53.95)	(-117.58)	(36.90)	(0.00)	(0.00)	(0.00)	(-4.40)	(0.00)
83	sq_weight	-1.89	0.00	0.00	0.00	0.00	0.00	-3.11	0.00
		(-62.47)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-24.86)	(0.00)
84	sq_Power	0.00	0.00	0.00	0.00	0.00	-0.57	-0.04	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(-125.24)	(-4.19)	(0.00)
85	sq_Fuel_cost	0.00	0.00	-0.11	0.00	-0.03	0.00	0.04	0.00
		(0.00)	(0.00)	(-81.45)	(0.00)	(-12.94)	(0.00)	(19.53)	(0.00)
86	sq_Weight/Power	0.00	0.00	-0.62	0.00	0.00	3.02	1.31	0.00
		(0.00)	(0.00)	(-86.87)	(0.00)	(0.00)	(58.11)	(13.28)	(0.00)

References

Allen, D. M. (1974). The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction, *Technometrics* **16**(1): 125–127.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys* **4**: 40–79.

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.

Hills, M. (1966). Allocation Rules and their Error Rates, *Journal of the Royal Statistical Society. Series B (Methodological)* **28**(1): 1–31.

Huang, Z., Zhao, H. and Zhu, D. (2012). Two New Prediction-Driven Approaches to Discrete Choice Prediction, *ACM Transactions on Management Information Systems (TMIS)* **3**(2): 9.

Keane, M. P. and Wolpin, K. I. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior*, *International Economic Review* **48**(4): 1351–1378.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis, *Technometrics* **10**(1): 1–11.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation.

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

McFadden, D., Talvitie, A., Cosslett, S., Hasan, I., Johnson, M., Reid, F. and Train, K. (1977). *Demand model estimation and validation*, Vol. 5, Institute of Transportation Studies.

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.

Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics, *in* E. Lindzey, G. and Aronson (ed.), *Handbook of Social Psychology*, chapter 10.

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