Please cite this article as:
Andrzej Grzybowski, Asking price model for used car market - incorporating prior information, Scientific Research of the Institute of Mathematics and Computer Science, 2005, Volume 4, Issue 1, pages 50-57.
The website: http://www.amcm.pcz.pl/

ASKING PRICE MODEL FOR USED CAR MARKET - INCORPORATING PRIOR INFORMATION

Andrzej Grzybowski

Institute of Mathematics and Computer Science, Czestochowa University of Technology

Abstract. The paper is devoted to the problem of a choice between various regression models for asking price of used cars. Different models can be obtained with the help of different estimators which can be adopted during the process of identification of model parameters. We compare models obtained using prior information with the ones obtained with the help of ignoring the information LS-estimator.

Introduction

When one wants to apply the linear regression model to a set of data, there are various methods one could use to estimate the regression coefficients. The most popular one is called the method of least squares (LS). This method, however, has its well-known weaknesses, see e.g. Berger [1], Grzybowski [2]. Among others, the usual least-squares estimator does not incorporate prior information and so, to use this information we need some alternative such as Bayes, robust Bayes or minimax estimators, to mention at least the most popular ones - see e.g. [1, 2]. However, the optimal performance of the estimators depends on the problem formulation and various assumptions about the model. In actual usage it is often difficult to decide what description of the problem is most appropriate, and consequently, which estimator is most suitable. In this paper we focus on such situations and emphasise the role of cross-validation simulations in choosing the estimator when the prediction is the main purpose of the model building. The presented idea will be adopted to built a model for asking price in a given segment of used car market.

1. Problem formulation - assumptions and the data

Suppose we are interested in finding out pricing behavior for used car market. Although there is a substantial amount of knowledge to be acquired through experience on pricing strategies we want to built a model explaining the behavior.

An important marketplace is an Internet platform. For our analysis we have data collected from August, 2005 announcements in http://moto.money.pl/autokomis.

Data for 113 automobiles were collected, in particular the asking price for a car (*AskPr*), information on manufacturer (*CM*), the model type (*MT*), the model year (*MY*), the car age (*CA*), the odometer reading (*OR*), the engine cubic capacity (*EC*), the engine type (*ET*). The objective is to develop a model to understand the determinants of asking prices for used automobiles. We also want to verify whether or not the information of regression parameters for a model built for one market can be adopted in estimation for similar, yet competing markets.

We assume the considered model has the linear form: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}$, where $\mathbf{Y}$ is an n-dimensional vector of observations of the dependent variable (asking price), $\mathbf{X}$ is a given nonstochastic (n x k) matrix with the rank k, β is a k-dimensional vector of unknown regression coefficients, $\mathbf{Z}$ is an n-dimensional vector of random disturbances.

Typical model building consists, roughly speaking, of two main stages; model simplification achieved by determination „insignificant" explanatory variables and setting their coordinates equal to zero and then application of a good estimation procedure to obtain the remaining ones. During the first stage, which is outside of the scope of this paper, we determined the variables *CA*, *OR* and *EC* as significant ones. It also appears that the data do not provide sufficient evidence to reject the assumption about the linear form of the model, so we decide to build a model given by the following formula:

$$AskPr = \beta_1 + \beta_2 CA + \beta_3 OR + \beta_4 EC + Z$$

Now we want to select, if possible, procedure which is most appropriate for the estimation of the regression parameter $\beta = (\beta_1,...,\beta_4)^T$ in our problem.

2. Estimators under consideration

In the paper we build our models with the help of the estimator $\mathbf{b}^{\mathbf{EB}}$, empirical linear bayes with respect to any distribution with the mean vector ϑ and covariance matrix Δ, and the robust generalised bayes estimator (RB-estimator) $\mathbf{b}^{RB}$ developed by Berger in [1]. We compare the models with the one obtained with help of LS estimator.

The estimator $\mathbf{b}^{\mathbf{EB}}$ is given by the following formula:

$$\mathbf{b}^{\mathbf{EB}}(\mathbf{Y}) = \mathbf{C}(\Delta,\Sigma)\mathbf{X}^T\Sigma^{-1}\mathbf{Y} + \mathbf{C}(\Delta,\Sigma)\Delta^{-1}\vartheta \tag{1}$$

where Δ is a given positive definite ($k \times k$) matrix and $\mathbf{C}(\Delta,\Sigma) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X}+\Delta^{-1})^{-1}$. Σ is a positive definite ($n \times n$) matrix - a given consistent estimate for the covariance matrix of $\mathbf{Z} = (Z_1,...,Z_n)^T$.

The robust generalized bayes estimator $\mathbf{b}^{RB}$ suggested for use by Berger in [1] is given by:

$$\mathbf{b}^{RB}(\mathbf{Y}) = \mathbf{b}^{LS}(\mathbf{Y}) - h_k(\| \mathbf{b}^{LS}(\mathbf{Y}) - \vartheta \|_1^2)\mathbf{V}(\mathbf{V} + \Delta)^{-1}(\mathbf{b}^{LS}(\mathbf{Y}) - \vartheta) \qquad (2)$$

where

$$\mathbf{V} = (\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X})^{-1} \text{ and } \| \mathbf{b}^{LS}(\mathbf{Y}) - \vartheta \|_1 = (\mathbf{b}^{LS}(\mathbf{Y}) - \vartheta)^{\mathrm{T}}(\mathbf{V} + \mathbf{A})^{-1}(\mathbf{b}^{LS}(\mathbf{Y}) - \vartheta)$$

The function $h_k()$ is a certain increasing function which can be reasonably approximated by the following formula:

$$h_k(a^2) \cong \min\left\{1, \frac{k+1}{a^2}\right\}$$

The point ϑ appearing in both above definitions can be thought of as a prior guess for the regression coefficient, the matrix Δ can describe our uncertainty connected with the guess, for a discussion see e.g. Grzybowski [2].

3. Cross-validation techniques

This section provides an introduction to a variety of CV heuristics. In the sequel „to select a model" means „to select an estimator which lead to a given estimates of the model coefficients".

Hold out set CV. „Hold out set" CV (hereafter referred to as HOS-CV) is perhaps the most obvious form of CV. The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the estimation (or training) set and the validation (or testing) set. The estimator fits a function using the estimation set only. Then on the base of the obtained model the output values are predicted for the data in the validation set. The errors it makes are used to evaluate the model.

Multifold CV. HOS-CV is often sufficient when abundant data are available. However, when it is not known whether or not the available amount of data is sufficient to ensure proper training and still provide an adequate amount of data for testing, it is well known that HOS-CV can be sensitive to user-set parameters that determine the size of the validation set (and hence, as well, the size of the training set) as well as the choice of split (i.e., a particular distribution of the available records into either the training set or the validation set). Let N_t and $N_v = N\text{-}N_t$ denote the sizes of the training set and validation set, respectively. Given choice of N_v and N_t, one direct method to reduce the sensitivity of HOS-CV on the choice of split is to repeat the HOS-CV procedure a number of times and average the results. We refer to this method generically as „multifold" CV. Multifold CV can be implemented in several ways. One way is „Leave-Many-Out" CV (LMO-CV) Note that there are $C(N, N_v)$ ways to select a validation set of size N_v, where

$$C(k,n) = k!/[k!(n\text{-}k)!]$$

gives the number of combinations available by choosing k things out of n that are available.

Here are three methods of implementing LMO-CV:

1) „Exhaustive" LMO-CV (ELMO-CV): average over all $C(N, N_v)$ splits.
2) „Disjoint set" LMO-CV (DLMO-CV): split the data into $[N/N_v]$ disjoint validation subsets of size N_v i.e., take the „floor" of N/N_v, rounding it down to the nearest integer.
3) „Monte-Carlo" LMO-CV (MCLMO-CV): select each validation subset by drawing N_v records at random without replacement (see Shao [3]).

ELMO-CV can be prohibitively expensive for even modest amounts of data, as $C(N,N_v)$ grows large quickly with N for $N_v > 1$. DLMO-CV is the implementation of multifold CV very commonly used in literature (for instance, it is essentially equivalent to an implementation known as „V-fold CV", where $V = [N/N_v]$ gives the number of disjoint validation sets, or, „folds"). MCLMO-CV was suggested by Shao [3] and has some appealing (and possibly surprising) properties. MCLMO-CV is similar to the bootstrap in that a particular estimation set could be generated more than once. However, note that MCLMO-CV differs from the conventional bootstrap, where training sets are obtained by sampling from the available data at random *with* replacement. Both ELMO-CV and MCLMO-CV average over many more validation sets than DLMO-CV, in general, because they generate many more splits than does DLMO-CV.

Leave-One Out CV. All of the multifold methods involve choice of N_v. Setting $N_v = 1$ yields the popular and well-studied „Leave-One-Out" CV (LOO-CV). LOO-CV may be considered as a special case of LMO-CV. Perhaps LOO is an appealing alternative to LMO-CV to many practitioners because:

- it eliminates the need to select the value of M,
- it limits the number of splits (to a total of N), and
- it results in the maximum number of records (i.e., N-1) allotted to each training set.

However, these reasons can be outweighed by the much better performance that can result by using LMO-CV, as is discussed e.g. in Plutowski, Sakata, White [4].
In our examples we make use of the two most effective cross-validation methods: LOO-CV and MCLMO-CV.

4. Results

We use the above described techniques LOO-CV and MCLMO-CV to select best asking price model. The comparison of the *EB*-, *RB*- and *LS*-estimators is made for two loss functions measuring ex post forecast accuracy: the distance between the forecast and the actually observed value of dependent variable

$$L_1(y, \hat{y}) = |y - \hat{y}|$$

and for the quadratic loss function given by

$$L_2(y, \hat{y}) = (y - \hat{y})^2$$

The first model we develop is based on data gathered for Skoda-Fabia model type. The regression parameters obtained with *LS*-estimator will be a prior information for next models built for Opel Astra, VV Golf, Renault Megane and Toyota Corolla. We denote the models *SF*-, *OA*-, *VVG*-, *RM*-, *TC*-, respectively.
Adopting usual *LS*-estimator we obtain the following *SF*-model:

$$AskPr = 24414.4 - 2413.7\, CA + 7.17 OR - 0.0011 EC$$

Thus the prior information is given by $\vartheta = (24414.4, -2413.7, 7.17, -0.0011)$. The uncertainty connected with the information is given by the diagonal matrix $\Delta = [24414.4^2, 2413.7^2, 7.17^2, 0.0011^2)]$. The elements on the principal diagonal are equal to squares of corresponding regression coefficients - this may reflect our belief that the sign of a given coefficient in all models are the same and an appropriate coefficient should not be much greater (not greater than, say, three times). More thorough analysis of the description of the uncertainty can be found in Grzybowski [2, 5].
Now we build *TC*-model. To compare our estimators first we use the LOO-CV technique. The data set contains 20 records devoted to Toyota. So the presented results are an average loss obtained for 20 problems. Next the MCLMO-CV method with $N_v = 5$ is adopted. The random split of the data set was generated 50 times. Tables 1 and 2 show results obtained in this case. Apart from the mean values of loss functions L_2 and L_1 we also compute the value of a statistic

$$U = \frac{\overline{L}_i^{\mathbf{b}1} - \overline{L}_i^{\mathbf{b}2}}{\sqrt{S_i^2(\mathbf{b}1) + S_i^2(\mathbf{b}2)}} \sqrt{n}$$

to test the null hypothesis that the mean values of loss (L_1 or L_2) for given estimators are actually equal. Here the symbols $\overline{L}_i^{\mathbf{b}j}$ and $S_i^2(\mathbf{b}j)$ stand for the average loss L_i, i = 1,2, and its variance, computed for estimator $\mathbf{b}j$, j = 1,2, respectively, n is a number of *all* predictions made during the simulations.

Table 1

The comparison of *EB* and *LS* estimators for *TC*-model

CV method	*EB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	4462	3.24×10^7	4744	3.35×10^7	–53	–1670
LOO	3864	1.96×10^7	4655	2.81×10^7	–63	–7060

Table 2

The comparison of *RB* and *LS* estimators for *TC*-model

CV method	*RB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	4614	$3.28\text{x}10^7$	4744	$3.35\text{x}10^7$	–25	–1084
LOO	4243	$2.34\text{x}10^7$	4655	$2.81\text{x}10^7$	–32	–3775

As we can see the results of cross validation show that both incorporating prior information estimators are better than the *LS*-estimator - no matter what the loss function is. On the other hand we can also notice that both estimators, *EB* and *RB*, have very similar performance, compare Table 1 and Table 2. The following tables present results obtained for the remaining markets (in all cases prior information is as previously described).

Table 3

The comparison of *EB* and *LS* estimators for *VVG*-model

CV method	*EB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	4118	$2.84\text{x}10^7$	4943	$4.07\text{x}10^7$	–151	–18800
LOO	8512	$9.15\text{x}10^7$	8516	$9.64\text{x}10^7$	–0.16	–1723

Table 4

The comparison of *RB* and *LS* estimators for *VVG*-model

CV method	*RB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	4251	$3.05\text{x}10^7$	4943	$4.07\text{x}10^7$	–125	–15400
LOO	8512	$9.15\text{x}10^7$	8516	$9.64\text{x}10^7$	–0.16	–1723

Table 5

The comparison of *EB* and *LS* estimators for *RM*-model

CV method	*EB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	6357	$7.41\text{x}10^7$	6753	$9.09\text{x}10^7$	–55	–16100
LOO	8626	$9.38\text{x}10^7$	8976	$1.21\text{x}10^8$	–15	–8600

Table 6

The comparison of *RB* and *LS* estimators for *RM*-model

CV method	*RB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	6709	$8.70\text{x}10^7$	6753	$9.09\text{x}10^7$	–6.13	–3700
LOO	8626	$1.13\text{x}10^8$	8976	$1.21\text{x}10^8$	–15	–2256

Table 7

The comparison of *EB* and *LS* estimators for *OA*-model

CV method	*EB* estimator		*LS* estimator		*U* statistic	
	L_1	L_2	L_1	L_2	L_1	L_2
LMO	3111.72	1.28×10^7	3275.25	1.49×10^7	–57.6	–4970
LOO	3036.01	1.28×10^7	3192.71	1.37×10^7	–13.92	–906

The results of the comparison of *RB* and *LS* estimators for *OA*-model are exactly the same as for *EB* estimator presented in Table 7, so we omit the next table. Such a phenomenon can be observed quite often, especially when the prior information ϑ is close to the actual value of the regression parameter β. In such a case both estimators are the same, see the formulae (1) and (2). From the results presented above we can judge that the prior information is precise enough to use EB-estimator which leads to a little bit less loss. Then the models for particular markets are as follows:

TC-model:

$$AskPr = 50241.6 - 6166.6\,CA + 2.79OR - 0.0012EC$$

VVG-model:

$$AskPr = 33288.1 - 5009.7\,CA + 14.6OR - 0.0011EC$$

RM-model:

$$AskPr = 30561.2 - 9197.46\,CA + 21.8OR - 0.0012EC$$

OA-model:

$$AskPr = 31726.5 - 3964.5\,CA + 8.32OR - 0.0011EC$$

Finally, we should once again estimate *SF*-model - this time with the help of *EB*-estimator. First, we should choose a prior information. We omit here the discussion of this problem, but it appears that each possible prior vector ϑ, taken from any of the considered markets, leads to similar improvement in prediction precision. Assuming ϑ equals the regression parameter in *VVG*-model we obtain the following:

$$AskPr = 24133.1 - 2487.5CA + 7.5OR - 0.0011EC$$

5. Final remarks

In the paper we present the usage of the cross validation techniques in order to choose models for various used automobile markets. However we should stress that the criteria of performance of the compared estimators are based on the prediction error and thus if the principal aim of the estimation is not the prediction, the results may be misleading. Such a situation occurs e.g. when we want to investigate the relationship between response and some of explanatory variables rather than to predict future values of the response. It is well known from decision theory that the prediction and estimation task are usually not exchangeable in that sense,

that the estimators better for one task does not have to be better for another. On the other hand computer simulations show that estimators incorporating prior information are even better as tools for the other task, see e.g. Grzybowski [2].

References

[1] Berger J.O., Statistical Decision Theory and Bayesian Analysis, Springer Verlag, New York 1985.

[2] Grzybowski A., Metody wykorzystania informacji a priori w estymacji parametrów regresji, Wydawnictwo Politechniki Częstochowskiej, seria monografie 89, Częstochowa 2002.

[3] Shao J., Linear model selection by cross-validation, Journal of the American Statistical Association. 1993, 88, 422, 486-494.

[4] Plutowski M.E., Sakata S., White H., Cross-validation estimates integrated mean squared error, C.L. Giles, S.J. Hanson, J.D. Cowan (eds.), Advances in neural information processing systems 6, Morgan Kaufmann Publishers, San Mateo, CA 1994.

[5] Grzybowski A., Analiza funkcji ryzyka estymatorów opartych na pewnym opisie niepewności, Modelowanie Preferencji a Ryzyko, praca zbiorowa pod redakcją T. Trzaskalika, 2002.

[6] Birkes D., Dodge Y., Alternative Methods of Regression, Wiley & Sons, New York 1993.

[7] Frees E.W., Data Analysis Using Regression Models - the Business Perspective, Prentice-Hall Inc., New Jersey 1996.