The Application of Partial Least Squares Method in Hedonic Modelling

Anna Król

Abstract The theory of hedonic models states that it is possible to precisely describe the price of a heterogeneous commodity by a set of its characteristics. However, the quality of the obtained results depends on the completeness of the set of significant attributes of the commodity used for estimation, as well as the statistical correctness of the model. In cases where the set of explanatory variables is numerous, and when the problem of multicollinearity occurs, it is not possible to use standard methods of estimation (such as OLS). The aim of this article is to examine the usefulness of the partial least squares method (PLS) in the estimation of hedonic models with a large number of correlated characteristics. It is shown that the use of PLS can yield better results than the alternative solution - the removal of problematic variables from the data set. Empirical research was carried out for selected groups of durable commodities. The databases were created using a tool for collecting data from web pages developed by the author.

Anna Król
Wrocław University of Economics
✉ anna.krol@ue.wroc.pl

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/05
ISSN 2363-9881

1 Introduction

The theory of hedonic models states that it is possible to precisely describe the price of a heterogeneous commodity by a set of its characteristics. The assumption that consumers derive utility from attributes of goods, rather than the good itself, led to the conception in which the price of a commodity is determined as an aggregate of values estimated for each significant characteristic of this commodity (so called hedonic prices). However, in cases of some goods the number of characteristics which significantly influence its price can be quite large. For smaller data sets this might cause a problem with the degrees of freedom, if standard methods of model estimation (such as OLS) are applied. Moreover, the bigger the number of prediction variables the more probable that they will be correlated to one another, which might cause problems induced by multicollinearity, especially when a large number of explanatory variables is binary, as it is sometimes the case in hedonic models, where variables often represent the existence of specific features in the given variant of a good, as well as its brand name or model.

The presence of correlated predictors in the data set used for the estimation of a regression model can cause a variety of problems. One of them, especially consequential in hedonic modelling, is that the individual coefficients estimates might not be precise. Multicollinearity can increase the variance of the regression coefficients, making them unstable (see Greene, 2008; Maddala and Lahiri, 2009). This influences the correctness of the interpretation of hedonic prices, and the ability of a hedonic model to fulfill one of its main goals, which is to provide estimates on the market prices of unobservable characteristics. The common solution to the multicollinearity issue is to remove highly correlated predictors in question, or linearly combine the problematic variables. However, apart from the fact that this might lead to a loss of information and omitted variable biases, it might also result in a model without important characteristics from the theoretical (or practical) point of view.

The aim of this article is to examine the usefulness of the partial least squares method (PLS) in the estimation of a hedonic model with a large number of correlated binary characteristics. It is shown that the use of PLS can yield acceptable results in comparison to some alternative solutions, such as ignoring the problem of multicollinearity, and removal of problematic variables from the data set. Empirical research was carried out for the data set of over 10 000 used hatchback cars offered for sale in Poland in May 2015.

2 Hedonic models theory

The foundations of hedonic methods are formed by the so-called hedonic hypothesis which states that heterogeneous commodities are characterized by a set of significant attributes (characteristics) (see Dziechciarz, 2004; Triplett, 2006). The relationship between the price of commodity (*PRICE*) and the set of its characteristics (X) described by a certain function f is called hedonic regression and may be described in the following general notation:

$$PRICE = f(X;\beta;\varepsilon), \tag{1}$$

where ε is the error term of the model. One of the main application areas of hedonic regression is to provide estimates of prices of individual characteristics (so-called hedonic prices or implicit prices). This feature of hedonic models is particularly interesting in cases of attributes whose prices are not directly observable on the market (such as for example product brand names). Usually empirical research presents the *a priori* assumption of the functional form f of the hedonic regression. The commonly used approaches are linear, exponential, double-log, logarithmic or a mixture of the above, depending on which fits the data best (see Brachinger, 2002). The standard estimation technique in such cases is the ordinary least squares method, or, in the presence of heteroskedasticity of error term, which is quite common in cross-sectional models, the weighted least squares method. However, if multicollinearity distorts the results, an alternative estimation approach is necessary.

3 Partial least squares method

Partial least squares (PLS) regression is a technique that reduces the number of predictors to a smaller set of uncorrelated latent components, and performs least squares regression on these components, instead of on the original data (see Hastie et al, 2009; Helland, 1990). It is commonly used in cases when predictors are correlated, or when there are more predictors than observations in the data set. The PLS method is similar to principal component regression, in which the principal components obtained for the predictors matrix X are used as regressors on the response Y. The orthogonality of principal components eliminates the multicollinearity problem. However there is no guarantee that the chosen latent components will be relevant for explaining the dependent

variable. In contrast, PLS aims at finding uncorrelated linear transformations of the original predictor variables which have a high covariance with the response variable (see Abdi, 2003). Several algorithms have been developed and implemented in various statistical programs for PLS regression estimation. In the presented research the R package *plsdepot* has been used (see Sanchez, 2012). The algorithm involves a few simple steps performed iteratively. The latent components are calculated as the weighted sums of predictors:

$$T = XW, \tag{2}$$

where $T = [t_1 \;\; t_2 \;\; \ldots \;\; t_k]$ is a matrix of k latent components and $W = [w_1 \;\; w_2 \;\; \ldots \;\; w_k]$ is a matrix of k projection weights. The projection weights are computed so that each of them maximizes the covariance between response Y and the corresponding components, and normalized so that $\|w_i\| = 1$. Moreover, the extracted components are orthogonal. Using latent components, both X and Y are decomposed by means of OLS:

$$X = TP^T + \xi, \tag{3}$$

$$Y = US^T + \zeta, \tag{4}$$

with $U = TD$, where D is a diagonal matrix, $P = [p_1 \;\; p_2 \ldots \;\; p_k]$ a matrix of k loadings of X, $S = [s_1 \;\; s_2 \ldots \;\; s_k]$ a matrix of k loadings of Y and ξ, ζ are the error terms.

Finally, the vector of parameters estimates may be obtained using the following formula:

$$\hat{B} = WS^T. \tag{5}$$

In order to determine the appropriate number of components in a PLS model, a cross-validation procedure is commonly used. In the presented research the leave-one-group-out cross-validation method has been applied. In this procedure the data set is randomly split into 10 segments of approximately equal size. Then, the observations in one of the groups are omitted and left outside as a test set. The other nine segments are used as learning set to estimate a model and predict the observations in the test segment. The procedure is repeated for each of the 10 segments. In the end the prediction error sum of squares (PRESS) is calculated using the following formula:

$$PRESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i}), \tag{6}$$

where $\hat{Y}_{i,-i}$ is the fitted value for the omitted observation. The presented above procedure is applied for each component k, and so called cross-validated R-squared (Q_k^2) is calculated to assess the predictive ability of potential models:

$$Q_k^2 = 1 - \frac{PRESS_k}{RSS}, \tag{7}$$

where $PRESS_k$ is the prediction error sum of squares using the k^{th} component and RSS the residual sum of squares. A component k is considered significant if Q_k^2 is greater than 0,0975. The cumulative cross-validated R-squared ($cumQ_k^2$) measures the predictive power of a model with components from 1 to k. The point in which adding another component increases the $cumQ_k^2$ only marginally marks the optimal PLS model specification.

4 Hedonic models for used hatchback cars

The main goal of the presented research was to obtain the hedonic model which can provide interpretable estimates of all significant characteristics, and can be used for their valuation (for example to acquire evaluations of brand premiums). The sources of data were used hatchback cars sales offers available on one of the biggest advertising site for individual second-hand vehicle sellers in Poland. Using this internet service each advertiser, for the price of a small fee, can place a sale offer, which is then stored in databases and visible for potential buyers on the web site. The databases were created using a tool for collecting data from web pages developed by the author. The tool comprises mutually related PHP scripts integrated with a SQL database, and is using the PHP framework *Simple HTML DOM Parser* written by S.C. Chen[1]. Its objective is to enter specific offers published on the web site (determined by the user), to collect the data necessary for the research, and to store it in the SQL database. Afterwards, another PHP script transfers the data in the user desired format to the research computer, ready to be processed and analysed.

[1] Available at: `http://simplehtmldom.sourceforge.net/` (accesed: 12.10.2015).

Table 1 Descriptive statistics

Variable name	Min	Max	Average	Std. deviation	Structure [%]
PRICE	15 882	700	209 000	13 538	
AGE	10.35	1	25	4.30	
CAPACITY	1 569.40	599	5 700	345.63	
MILLAGE	150 730	1 000	420 000	64 820	
PETROL					56.23
DIESEL					43.77
LPG					6.73
METALLIC					68.88
STANDARD					31.12
DOOR_3					29.01
DOOR_5					70.99

4.1 Data set description

The data set comprises 10 423 hatchback cars of 27 different makes offered for sale in Poland in May 2015 on the secondary market. Each offer in the data set is described by price[2] (PRICE [PLN]) and the following main characteristics: car age (AGE [years]), engine cubic capacity (CAPACITY [cm^2]), mileage (MILAGE [km]), fuel type (2 categories: PETROL, DIESEL), liquefied petroleum gas installation (LPG), paint type (2 categories: METALLIC, STANDARD), and number of doors (2 categories: DOOR_3, DOOR_5). Table 1 presents basic descriptive statistics for the PRICE variable and the main attributes of cars. In addition, the set of dummy variables was used to capture all other attributes. The variables used relate car history (3 dummies: registered in Poland (REGIST), serviced in authorized car service center (SERV), never crashed (UNCRASH)), presence of accessory features (16 dummies: aluminum wheels (ALU), ABS, ASR, ESP, electric mirrors (EL_MIRR), xenon headlights (XENON), air conditioning (AIR), CD, on-board computer (COMP), multifunction steering wheel (STEER), rain sensor (RAIN), parking aid system (PARK), leather seats (LEATHER), navigation system (NAVI), bluetooth (BLUE), tinted windows (TINTED)), as well as car brand names (27 makes).

The data set was split into two parts: A training set used for the models estimation, and a test set of 500 randomly chosen observations for the validation of the obtained models.

[2] The main disadvantage of the used the data source is the fact that the collected prices are offer prices, and not transaction prices.

4.2 Estimation results

In the first step of the research the log-linear model (with dependent variable lnPRICE), which provided the best fit to the data, was estimated using the ordinary least squares method on the group of 9 923 observations from the training set. Since the presence of heteroskedasticity was detected (using White's test), the model was re-estimated using White's weighted least squares method (WLS). The procedure involves OLS estimation of the model of interest, followed by an auxiliary regression to generate an estimate of the error variance. Afterwards weighted least squares estimation is performed, using the reciprocal of the estimated variance as weights (see White, 1980). For the reference brand (which was left outside of the model) the variable VOLKSWAGEN was chosen, as the one with the largest number of representatives in the data set. Upon model estimation three brands turned out to be insignificantly different from the reference brand: HONDA, SMART and TOYOTA. Therefore, they were removed from the data set, and the model was once again estimated. The results are presented in Table 2 and Table 3 in column 1 (WLS (1)). All the variables in the model WLS (1) are highly statistically significant (p-values smaller than 0.01), except for the variable VOLVO (significance on the level 0.05). The goodness-of-fit of the model measured by R^2 statistic is on a satisfactory level (around 87%).

Problematic for the model interpretation is however the estimate for the variable DIESEL. It is highly significant, but it is also negative. This result is contrary to the *a priori* expectations. Diesel cars are on average more expensive, which is caused by higher production costs, and also are more valued by the buyers due to lower operation costs (cheaper fuel and lower fuel consumption). Moreover, investigation of the correlation matrix (Table 4) shows that the DIESEL variable is significantly and positively correlated with variable lnPRICE. One can as well observe correlations between DIESEL, AGE, MILAGE and CAPACITY variables, which combined with the fact that the model has a great number of dummy variables (45), was probably the source of the problem.

In the first attempt to deal with the issues induced by multicollinearity, three additional models were estimated using WLS: A model without the problematic DIESEL variable (WLS (2)), a model without the AGE variable (WLS (3)), and model without the CAPACITY variable (WLS (4)). The results are presented in Tables 2 and 3 in columns 2, 3 and 4 respectively. The model WLS (2) does not

Table 2 Comparison of hedonic models for used hatchback cars (dependent variable: lnPRICE)

	(1) WLS	(2) WLS	(3) WLS	(4) WLS	(5) PLS
constant	9.997***[b]	10.02***	8.974***	10.31***	9.954
AGE	−0.1226***	−0.1198***		−0.1192***	−0.1105
CAPACITY	0.0003301***	0.0002945***	5.236e-05***		0.0002355
MILAGE	−7.281e-07***	−9.096e-07***	−4.454e-06***	−4.124e-07***	−1.736e-06
DIESEL	−0.05632***		0.1959***	0.004309	0.069731
LPG	0.03083***	0.05700***	0.1224***	0.05233***	0.15664
METALLIC	0.02255***	0.02286***	−0.01955**	0.02821***	0.03763
DOOR_5	0.1288***	0.1285***	0.2164***	0.1481***	0.1422
REGIST (51.3%)[a]	0.06796***	0.06185***	0.06098***	0.06861***	0.07548
SERV (41.4%)	0.02119***	0.02773***	0.07835***	0.02322***	0.03719
UNCRASH (59.6%)	0.05097***	0.05367***	0.03973***	0.05130***	0.06393
ALU (47.9%)	0.06558***	0.06717***	0.03084***	0.08692***	0.05840
ABS (89.8%)	0.05611***	0.05806***	0.2627***	0.06520***	0.08164
ASR (35.7%)	0.01737***	0.01760***	0.04588***	0.02905***	0.01681
ESP (35.9%)	0.04794***	0.05002***	0.1402***	0.06419***	0.07119
EL_MIRR (72.1%)	0.03448***	0.03671***	0.04846***	0.05899***	0.01983
XENON (6.8%)	0.06050***	0.07058***	0.09640***	0.1124***	0.08557
AIR (80.8%)	0.08230***	0.08152***	0.2362***	0.1026***	0.10569
CD (77.5%)	0.03097***	0.02971***	0.1637***	0.02736***	0.01946
COMP (56.7%)	0.04511***	0.04461***	0.1246***	0.05300***	0.04823
STEER (38.9%)	0.03795***	0.03535***	0.1139***	0.03721***	0.01922
RAIN (17%)	0.02628***	0.02624***	0.05072***	0.04249***	0.03283
PARK (15.6%)	0.04945***	0.04892***	0.1156***	0.05706***	0.03415
LEATHER (8.1%)	0.03571***	0.03806***	0.03023*	0.06702***	0.04353
NAVI (7.4%)	0.04978***	0.04718***	0.1276***	0.06279***	0.05543
BLUE (10.6%)	0.03827***	0.03793***	0.1790***	0.03858***	0.05160
TINTED (17.3%)	0.01706***	0.01740***	0.04107***	0.02620***	0.02771

[a] In parenthesis the percentage of cars with given feature is given.
[b] Stars indicate the significance level: '***' means significance on the level 0.01, '**' on the level 0.05, '*' on the level 0.1.

cause problems with the interpretation of coefficients, but at the same time fails to provide the estimate of an important hedonic price for the type of fuel. At the same time, none of the used statistics for models comparison (R^2, Schwarz information criterion SC) support the action of variable deletion. The removal of the AGE variable from the model (WLS (3)) leads to a positive estimate for the DIESEL characteristic, however its value is quite high. It conveys that the diesel car is on average over 21% more expensive in comparison to similar cars which run on petrol fuel and are similar in all other respects. Moreover, one can observe substantial increases in the estimates of a few other variables. For

Table 3 Comparison of hedonic models for used hatchback cars (continuation of Table 2)

	(1) WLS	(2) WLS	(3) WLS	(4) WLS	(5) PLS
ALFAROMEO (1.1%)[a]	−0.2781***[b]	−0.2729***	−0.2742***	−0.2673***	−0.1321
AUDI (4.9%)	0.2030***	0.2017***	0.1555***	0.2331***	0.3394
BMW (1.7%)	0.09640***	0.1030***	0.1306***	0.1603***	0.2751
CHEVROLET (0.5%)	−0.4803***	−0.4759***	−0.2535***	−0.4883***	−0.4763
CHRYSLER (0.2%)	−0.4190***	−0.4074***	−0.4727***	−0.2691***	−0.2431
CITROEN (4.6%)	−0.3094***	−0.3206***	−0.2935***	−0.3389***	−0.2151
DAEWOO (0.7%)	−0.6222***	−0.6323***	−0.8104***	−0.7079***	−0.5813
FIAT (6.0%)	−0.3473***	−0.3525***	−0.2802***	−0.3697***	−0.2892
FORD (7.8%)	−0.3013***	−0.3042***	−0.2371***	−0.2926***	−0.2058
HONDA (2.9%)					0.1788
HYUNDAI (1.4%)	−0.3223***	−0.3266***	−0.1971***	−0.3310***	−0.3029
KIA (1.1%)	−0.2462***	−0.2454***	−0.09278***	−0.2505***	−0.1962
MAZDA (3.1%)	−0.2305***	−0.2332***	−0.2443***	−0.2221***	−0.0898
MERCEDES (2.6%)	−0.1256***	−0.1285***	−0.2685***	−0.09775***	0.0085
MINI (0.5%)	0.2592***	0.2597***	0.3077***	0.2577***	0.4321
MITSUBISHI (1.0%)	−0.3648***	−0.3641***	−0.3339***	−0.3630***	−0.2778
NISSAN (3.0%)	−0.2937***	−0.2903***	−0.2881***	−0.3046***	−0.1932
OPEL (12.0%)	−0.1966***	−0.1983***	−0.2006***	−0.2091***	−0.0749
PEUGEOT (5.3%)	−0.3029***	−0.3096***	−0.2427***	−0.3128***	−0.2268
RENAULT (8.4%)	−0.4053***	−0.4102***	−0.4548***	−0.4213***	−0.2703
SEAT (5.7%)	−0.1108***	−0.1126***	−0.06946***	−0.08465***	−0.0107
SKODA (3.9%)	−0.1707***	−0.1680***	0.03678*	−0.1526***	−0.0926
SMART (0.5%)					0.0027
SUZUKI (1.4%)	−0.1975***	−0.1928***	−0.1003***	−0.2227***	−0.1035
TOYOTA (5.4%)					0.1096
VOLVO (0.4%)	0.1098**	0.09152*	0.1850***	0.1492***	0.1748
Adjusted R^2	0.8719	0.8716	0.6973	0.8574	0.8621
SC	41 052.42	42 917.45	41 655.97	43 176.23	
MAPE (training set)	2.053	2.061	3.383	2.132	2.124
MAPE (test set)	2.149	2.155	3.451	2.191	2.165

[a] In parenthesis the percentage of cars of given brand is given.
[b] Stars indicate the significance level: '***' means significance on the level 0.01, '**' on the level 0.05, '*' on the level 0.1.

Table 4 Correlation matrix

	lnPRICE	AGE	DIESEL	CAPACITY	MILEAGE
lnPRICE	1.0000	−0.8387	0.2158	0.2118	−0.4503
AGE		1.0000	−0.1747	0.0614	0.6045
DIESEL			1.0000	0.4289	0.2314
CAPACITY				1.0000	0.3436
MILEAGE					1.0000

example, the price premium for the cars equipped with ABS increases to 26% (in comparison to less than 6% in model WLS(1)), and for cars with air conditioning to over 24% for this feature (only about 8% in the initial model). Such instabilities might be related to the omitted variable biases caused by the removal of the significant variable AGE from the model. Finally, the model WLS (4) estimated excluding the CAPACITY variable, provides a positive albeit highly statistically insignificant (p-value = 0.4697) estimate of the DIESEL coefficient. To sum up, the classic remedy for the problem with correlated variables, consisting in the deletion of problematic variables, proved to be unsatisfactory in the presented research. Such a solution might be advisable in hedonic modelling in cases when the correlated variables provide very similar information about the price of the commodity. In such situations the loss of information caused by the removal of one of the characteristics is minimal, and the possibility of omitted variable biases is reduced. The research described in the article Dziechciarz-Duda and Król (2015) could serve as an example, where out of two highly correlated characteristics describing the price of a tablet device - vertical screen resolution and horizontal screen resolution - one could be removed without much consideration. Both attributes carry very similar information about the quality of the screen, and therefore are likely to perform equally good in the model interchangeably. In such cases a linear combination of both variables may be used as well (for example the sum of vertical and horizontal resolutions), however this leads to a more problematic interpretation of the obtained coefficient.

As an alternative approach, an attempt to apply the partial least squares method was made. Using the cross-validation procedure, the number of necessary latent components was established. Figure 1 presents the cumulative cross-validated R-squared measures, and favours a model with four components. The predictive power of models with additional components (more than 5) increases only marginally, so further expansion of the latent component matrix would not bring the expected improvement.

Subsequently, the iterative procedure described in Sect. 3 was performed. The final results of estimation are presented in Table 2 and Table 3 in column 5 (PLS (5)). Elimination of the multicollinearity problem by introducing the orthogonal latent components allowed to obtain the model with all the coefficients in accordance with the expectations, including the parameter estimate for variable DIESEL. The diesel car is on average around 7% more expensive in comparison to cars which run on petrol fuel, ceteris paribus. Each additional feature of a car positively influences its price. For example, the price of a car

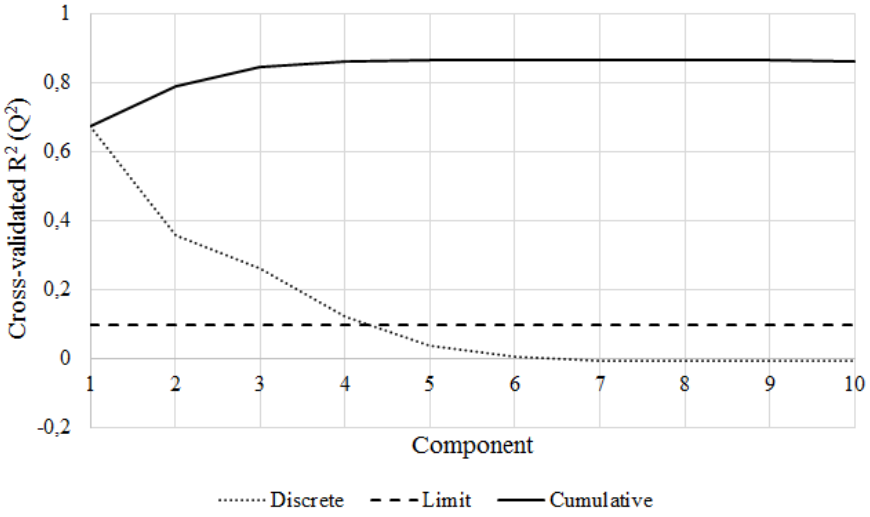

Fig. 1 Cross-validation results

equipped with air conditioning is on average about 11% higher, and xenon headlights increase the price by 8%. Moreover, the hedonic model allows for a calculation of brand name premiums. In the group of producers with higher brand equities than the reference brand (VOLKWAGEN) are AUDI, BMW, HONDA, MERCEDES, MINI, SMART, TOYOTA and VOLVO. In turn, the least valued brands are CHEVROLET, DAEWOO and HYUNDAY. The estimated hedonic model may be used as well for pricing the cars which are intended for sale. In terms of predictive power all the models performed similarly. The mean average percentage errors (MAPE), both in training set and in test set, were within limits of 2%-2.2%, and in case of the model WLS (3) on a slightly higher level around 3.5%.

5 Conclusions

The aim of this article was to examine the usefulness of partial least squares method (PLS) in estimation of hedonic models with a large number of correlated characteristics. It is shown that the use of PLS can yield better results than the alternative solution - the removal of problematic variables from the data set.

Therefore, partial least squares regression might be an alternative to OLS/WLS methods of hedonic models estimation in cases of multicollinearity, especially when the deletion of correlated variables is problematic from the theoretical or practical point of view, or when omitted variable biases occur. By applying the PLS method it was possible to obtain a hedonic model for used hatchback cars with values of coefficients which comply with the expectations and are easy to interpret. Moreover, the goodness-of-fit and predictive power of the model estimated with partial least squares were only slightly inferior in comparison to the initial WLS model, and somewhat better in comparison to the models from which problematic variables were removed.

Acknowledgements The study was conducted in the framework of the research project entitled *The Application of Hedonic Methods in Quality-Adjusted Price Indices* (*Zastosowanie metod hedonicznych do uwzględniania różnic jakości dóbr we wskaźnikach dynamiki cen*). The project has been financed by the National Science Centre on the basis of decision no. DEC-2013/09/N/HS4/03645.

References

Abdi H (2003) Partial least squares (pls) regression. In: Lewis-Beck M, Bryman A, Futing T (eds) Encyclopedia of Social Sciences Research Methods, Sage, Thousand Oaks, pp 792–795

Brachinger HW (2002) Statistical theory of hedonic price indices. DQE Working Papers 1, Department of Quantitative Economics, University of Freiburg/Fribourg, Switzerland

Dziechciarz J (2004) Regresja hedoniczna. próba wskazania obszarów stosowalności. In: Zeliaś A (ed) Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych, Wyd. AE, Kraków, pp 163–175

Dziechciarz-Duda M, Król A (2015) Zastosowanie analizy unfolding i regresji hedonicznej do oceny preferencji konsumentów. Taksonomia 25(385):90–98, DOI 10.15611/pn.2015.385.10

Greene WH (2008) Econometric Analysis, 6th ed. Prentice Hall, New Jersey

Hastie T, Tibishirani R, Friedman J (2009) The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York

Helland I (1990) Partial least squares regression and statistical models. Scandinavian Journal of Statistics 17(2):97–114

Maddala G, Lahiri K (2009) Introduction to econometrics. 4th edition. Wiley, Chichester

Sanchez G (2012) R package `plsdepot`. URL `https://cran.r-project.org/web/packages/plsdepot/index.html`

Triplett J (2006) Handbook on hedonic indexes and quality adjustments in price indexes. Oecd publishing, OECD Directorate for Science, Technology and Industry, Paris

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838, DOI 10.2307/1912934