SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2023
ročník/volume 33

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

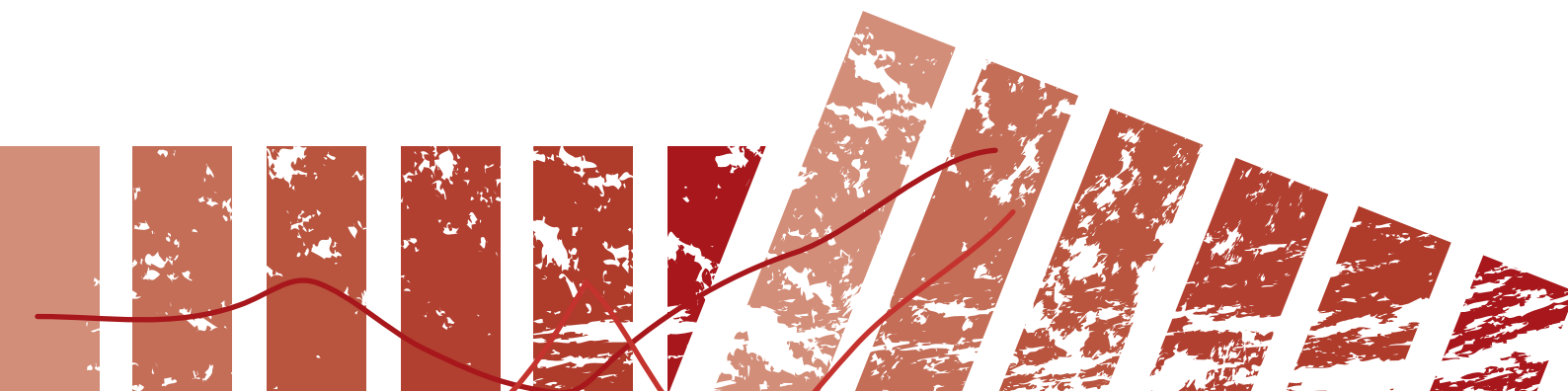
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 21 – 38

Dátum vydania/Publication date: 15. júl 2023/July 15, 2023



Peter KNÍŽAT
Statistical Office of the Slovak Republic, University of Economics in Bratislava

HEDONIC CONSUMER PRICE INDEX: DIAGNOSTICS AND ANALYSIS OF VARIANCE

HEDONICKÝ INDEX SPOTREBITEĽSKÝCH CIEN: DIAGNOSTIKA A ANALÝZA ROZPTYLU

ABSTRACT
Web scraping provides a new innovative data source that can be utilised in price statistics by National Statistical Institutes. The prices of product-offers are automatically downloaded from the internet, which allows a wider selection of representative products in the consumer basket. The other advantage of scraping online prices is that the user can obtain characteristic parameters of individual products. These parameters are used in calculating a hedonic price index that is more preferable for products with a high replacement rate. The hedonic regression assumes that the natural logarithm of the product's price can be explained by its characteristic parameters. In many previous researches, the authors estimate hedonic price indices for various products without any statistical verification of the fitted regression model. In this paper, we carry out a thorough analysis of the hedonic regression model that encompasses checking the model's diagnostics and its initial assumptions. Moreover, we use the analysis of variance to test contrasts between categories of individual characteristic parameters. The categories without any contrast are merged together and the hedonic price index is re-estimated. In the empirical study, we estimate and compare the hedonic price indices for different scenarios using observed web scraped prices from the Slovak market.

ABSTRAKT
Web scrapovanie poskytuje nový inovatívny zdroj údajov, ktorý môžu národné štatistické úrady využiť v cenovej štatistike. Ceny produktov sa automaticky sťahujú z internetu, čo umožňuje širší výber reprezentatívnych produktov v spotrebnom koši. Ďalšou výhodou scrapovania online cien je, že používateľ môže získať charakteristické parametre jednotlivých produktov. Tieto parametre sa využívajú pri výpočte hedonického cenového indexu, ktorý je vhodnejší pre produkty s vysokou mierou fluktuácie. Hedonická regresia predpokladá, že prirodzený logaritmus ceny produktu možno vysvetliť jeho charakteristickými parametrami. V mnohých predchádzajúcich výskumoch autori odhadnú hedonické cenové indexy rôznych produktov bez toho aby štatisticky vyhodnotili správnosť regresného modelu. V tomto článku vykonáme dôkladnú analýzu hedonického regresného modelu, ktorá zahŕňa kontrolu diagnostiky modelu a jeho počiatočných predpokladov. Okrem toho použijeme analýzu rozptylu na testovanie kontrastov medzi kategóriami pre jednotlivé charakteristické parametre. Kategórie bez kontrastu sa potom zlúčia a hedonický cenový index sa prepočíta. V empirickej štúdii vypočítame a porovnávame hedonické cenové indexy pre rôzne prípady, využitím web scrapovaných cien produktu zo slovenského trhu.

KEY WORDS
web scraping, online prices, consumer price index, hedonic regression, time-product dummy regression

KĽÚČOVÉ SLOVÁ
web scrapovanie, online ceny, index spotrebiteľských cien, hedonická regresia, regresia s časovo umelou premennou

1. INTRODUCTION

In the past decade, National Statistical Institutes (NSIs) consider various alternative data sources to supplement the traditional data collection of product prices. The online collection, also called web scraping, of prices is one of these alternatives. Web scraping can be automated and NSIs are able to gather all available product-offers, which can be used to select a representative sample of product items for the consumer basket. Noting that the traditional price collection only covers a limited number of product offers due to extensive requirement for human resources. The research paper [12] considers web scraping to modernise a data collection for Italian harmonised consumer price index, in the context of the European project, and evaluates its effectiveness. It states that a web scraping has a big potential as a new innovative data source for price statistics.

Through web scraping, we usually obtain daily prices of products that are collected from comparison website platforms, where multiple online retailers offer the same product. This leads to a collection of multiple daily prices for the same product, which need to be aggregated on a monthly level. In the paper [11], the author demonstrates various methods for aggregating daily prices. The empirical analysis shows that the geometric average(s) could be the most appropriate aggregation since it includes all prices of selected products and it reduces the effect of extreme price values.

Moreover, web scarping allows obtaining characteristic parameters of individual products that are assumed a significant determinant of the price of the product. A well- known method for estimating a price index, which considers characteristic parameters, is a hedonic regression. The advantage of the hedonic regression is that it can include all products, without any requirement of its price observed in every month. Additionally, new and disappearing products can also be included in the estimation of the hedonic price index that is particularly important when the replacement rate of products is substantially high. The guidelines [4] states that hedonic price indices are more appropriate when the scale of replacement rate is high.

A comprehensive guide on hedonic regression models and corresponding price indices is provided in [16]. Many other research papers, for example [5], [6] and [7] among others, are dedicated to study hedonic regression models in the context of the estimation of price indices. For example, the paper [5] concludes that a hedonic price index is more suitable, when the characteristic parameters for products are observable, than the time-product dummy price index. However, none of these shows an examination of the model's goodness of fit, statistical tests of the model's assumptions, or any detailed analysis of individual characteristic parameters.

The objective of this paper is to carry out a thorough analysis of the hedonic regression model in the context of the estimation of consumer price indices. The individual categories of characteristic parameters are evaluated by using analysis of variance, i.e., a contrast between a pairwise comparison of categories is tested by different hypothesis tests. The categories with contrast that is not statistically significant from zero, a test under the null hypothesis, are merged together and the

hedonic price index is re-estimated. In the empirical study, we estimate and compare hedonic price indices using observed web scraped prices from the Slovak market.

The paper is organised as follows. Section 2 shows a theoretical framework of the hedonic regression and its corresponding price index. It also outlines a methodology that can be used for the evaluation of the model goodness of fit and the analysis of variance for categorical characteristic parameters. Section 3 shows the empirical application of the discussed theory on the product category of refrigerators. Conclusion discusses the results, issues with hedonic regression models and a potential further research.

2. THEORETICAL FRAMEWORK

In this section, in the first part, we introduce a theoretical framework for estimating a hedonic price index that is based on defining a hedonic regression model. In the recently published guidelines by Eurostat [2], an estimation of the hedonic price index is briefly outlined but no further details of testing the model diagnostics, or a statistical significance of model parameters are provided.

In the second part, we discuss a theory of the analysis of variance (ANOVA) in the context of the hedonic regression model. Moreover, we show the diagnostic tests that are used for testing the model goodness of fit and its initial assumptions.

2.1 REGRESSION-TYPE MODELS FOR CONSUMER PRICE INDICES

We assume that the logarithm of the price of the product item i, i = 1, …, n, can be explained by its characteristic parameters, where a number of characteristic parameters z_{ik} with k = 1,…,K for each product item i is observable at the time period t, t = 0, …, T. The time window T is defined by the user and it usually covers at least one year, T = 12. The log-linear hedonic regression model can be defined as [2]:

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T} \partial^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t \qquad (1)$$

where D_i^t is a dummy variable that is assigned a value of 1 if the observed price relates to the product item i at the time period t, and 0 otherwise. The parameters ∂^0, ∂^t, and β_k are estimated through the ordinary least squares method, which yields $\hat{\partial}^0$, $\hat{\partial}^{\mathrm{t}}$, and $\hat{\beta}_k$. The regression model in Eq. (1) assumes that the random errors ε_i^t are normally distributed with mean 0 and constant variance σ^2. Noting that in order for the regression model to be estimable, we need to omit one dummy variable due to the singularity issue of the covariate matrix. Thus, the time-dummy variable D_i^t for the base period t = 0 is omitted.

The exclusion of the base period from the estimation, and the logarithmic form of the regression model, leads to a plausible interpretation of the estimated $\hat{\partial}^{\mathrm{t}}$s in the context of consumer price indices. The estimated $\hat{\partial}^{\mathrm{t}}$s are the percentage changes in average prices between the time period 0 and t, where the characteristic parameters are controlled for.

Taking the exponential of the estimated $\hat{\partial}^{\mathrm{t}}$ leads to the following expression for the hedonic price index [5]:

$$I_{Hedonic}^{0,t} = exp\left(\hat{\partial}^{t}\right) = \frac{\prod_{i\epsilon S^{t}}(p_i^t)^{\frac{1}{N^t}}}{\prod_{i\epsilon S^{0}}\left(p_i^0\right)^{\frac{1}{N^0}}} exp\left[\sum_{k=1}^{K}\hat{\beta}_k(\bar{z}_k^0 - \bar{z}_k^t)\right] \tag{2}$$

where $\hat{\beta}_k\bar{z}_k^0$ and $\hat{\beta}_k\bar{z}_k^t$ are the mean of estimated characteristic parameters effects at the time period 0 and t, respectively, where $\bar{\mathrm{z}}_{\mathrm{k}}^{0} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{0}} z_{ik}/N^0$ and $\bar{\mathrm{z}}_{\mathrm{k}}^{\mathrm{t}} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{\mathrm{t}}} z_{ik}/N^t$ are the sample averages of characteristic parameters. A detailed derivation of Eq. (2) can be found in [5]. Eq. (2) represents a hedonic price index that can be interpreted as a geometric average of price changes, between the base period 0 and the current period t, weighted by the mean effect of characteristic parameters. For the matched sample of product items between the time period 0 and t, Eq. (2) simplifies to the bilateral Jevons index; refer to [10] or [11] for further definitions.

Moreover, the hedonic regression model in Eq. (1) can be used for the missing price imputation, that is, when a product falls out of offer in a particular month but it reappears again next month, the estimated model can be utilised for the missing price prediction.

If we assume that the characteristic parameters of individual product items are unobservable, a different version of the log-linear regression model can be used. The time-product dummy (TPD) regression model is defined as [5]:

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T}\partial^t D_i^t + \sum_{i=1}^{N-1}\gamma_i D_i + \varepsilon_i^t \tag{3}$$

where D_i is a dummy variable that is assigned a value 1 if the observed price relates to the product item i, and 0 otherwise. Similarly as for the time dummy variable, a dummy variable for the arbitrary product item i is not included in order to estimate the regression model.

Similarly, the time-product dummy (TPD) price index can be expressed as [5]:

$$I_{TPD}^{0,t} = exp\left(\hat{\partial}^{t}\right) = \frac{\prod_{i\epsilon S^{t}}(p_i^t)^{\frac{1}{N^t}}}{\prod_{i\epsilon S^{0}}\left(p_i^0\right)^{\frac{1}{N^0}}} exp\left[\sum_{i=1}^{N}\left(\bar{\hat{\gamma}}_i^0 - \bar{\hat{\gamma}}_i^t\right)\right] \tag{4}$$

where $\bar{\hat{\gamma}}_i^0$ and $\bar{\hat{\gamma}}_i^t$ are the mean of estimated product items' effects at the time period 0 and t, respectively, where $\bar{\hat{\gamma}}_i^0 = \sum_{\mathrm{i}\epsilon\mathrm{S}^{0}}\hat{\gamma}_i/N^0$ and $\bar{\hat{\gamma}}_i^{\mathrm{t}} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{\mathrm{t}}}\hat{\gamma}_i/N^t$ are the sample averages of the estimated parameters $\hat{\gamma}$. A detailed derivation of Eq. (4) can be found in [5]. Again, for the matched sample of product items between the time period 0 and t, Eq. (4) simplifies to the bilateral Jevons index; refer to [10] or [11] for further definitions.

2.2 MODEL DIAGNOSTICS AND ANALYSIS OF VARIANCE

Both the hedonic regression model in Eq. (1) and the time-product dummy regression model in Eq. (3) are assumed to be of the linear form. This linearity in the

functional must be verified since its misspecification leads to incorrectly defined random errors ε_i^t that further invalidates hypothesis tests for the fitted model(s). Moreover, it can also indicate that the hedonic regression model has missing or omitted characteristic parameters, which is a phenomenon that can cause erroneous hedonic price indices.

The estimated random errors, or residuals, $\hat{\varepsilon}_i^t$ for Eq. (1) can be expressed as:

$$\hat{\varepsilon}_i^t = \ln p_i^t - \ln \hat{p}_i^t = \ln p_i^t - \left(\hat{\partial}^0 + \sum_{t=1}^{T} \hat{\partial}^t D_i^t + \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right) \tag{5}$$

A similar expression can be shown for Eq. (3). To verify the aforementioned assumptions of the fitted model, i.e., a normal distribution of the estimated residuals and the assumption for a constant variance, we can use standard diagnostic tests that are estimable by any statistical software in the regression procedure. We use the SAS software, where the *proc glm* procedure displays, in a graphical format, the following diagnostic tests:

Table No. 1: Diagnostic tests

Diagnostic test (plot)	Correctly specified model	Mis-specified model
Predicted vs observed responses	The values are close to the diagonal line.	The values that are far from the diagonal line → the predicted response is far from the observed response, i.e., the residual is large.
Residuals and predicted residual	The values show no pattern → a set of independent and identically distributed random variables.	The values follow a curve → missing characteristic parameters. The values are 'fan shaped' → a non-constant variance. The values are not randomly scattered → a correlation between residuals (autocorrelation in responses).
Studentized residuals versus the leverage	The values are in the range ±2.	The values that exceed ±2 can be considered outliers.
Residuals vs quantiles	The values are close to the diagonal line.	The few values fall on a diagonal line → outliers. The left end of is below / above the diagonal line → long / short tail in the left. Similarly, for the right end.

Source: SAS, Author's construction

The manuscript [16] provides a comprehensive study of hedonic regression models when the functional form of Eq. (1) is misspecified, particularly, due to the missing characteristic parameters in the model.

Furthermore, we introduce a concept of the analysis of variance (ANOVA) that is a general method for studying contrasts, or differences, between treatments of the

categorical variable for the given response variable. In general, the ANOVA functional form can be expressed as the general linear model. To formulate the terminology, we have a number of observed prices, N_c, in each category (treatment) c = 1,…, C, for the given characteristic (of the categorical format) parameter, and the linear ANOVA model for the i^{th} logarithmic price in the c^{th} category can be defined as [1]:

$$ln\, p_{ic}^{t} = \mu + \alpha_c + \varepsilon_{ic}^{t} \tag{6}$$

where μ is the overall mean across all N observed prices and $N = \sum N_c$. The parameters α_c represent the specific effects which are departures from the overall mean specific to each category c. The errors ε_{ic}^{t} are the unexplained variation specific to the i^{th} observation within category c and are assumed to be normally distributed with zero mean and a constant variance σ^2. To be able to identify individual categories uniquely, the constraint $\sum_c \alpha_c = 0$ has to be satisfied.

It follows that Eq. (6) can be rewritten in the form of the general linear model, where the design matrix X_{ij} is defined as containing p scalar independent variables coded 0's or 1's for the response category memberships. It implies that Eq. (6) can be re-expressed by setting $\beta_0 = \mu$, $\beta_2 = \alpha_1$, and so on to $\beta_c = \alpha_c$. Hence, it follows that Eq. (6) has the equivalent formulation as:

$$ln\, p_{ic}^{t} = \sum_{j=0}^{c+1} X_{(ic)j}\beta_j + \varepsilon_{ic}^{t} \tag{7}$$

The design matrix $X_{(ic)j}$ in Eq. (7) is equivalent to Eq. (1) for each characteristic parameter z_{ik} that is of the categorical format. Eq. (7) includes all independent variables in the model, including the time dummy variables, which serve as control variables in the estimation. Noting that the parameter β_j, usually for the last category c, is not defined, also called a controlled category, to maintain a non-singularity of the design matrix. It follows that the null hypothesis test for testing the statistical significance for each category is defined as:

$$H_0: \{\beta_1 - \beta_c = 0\ and\ \beta_2 - \beta_c = 0\ and\ \ldots and\ \beta_{c-1} - \beta_c = 0\} \tag{8}$$

The F statistics is used for testing the hypothesis; a detailed derivation of the hypothesis test is provided in [1]. Noting that the linear difference of β_js to the overall mean of the characteristic parameter, instead of the controlled category, can also be tested.

Moreover, we are interested in comparing the contrast between each category, not only for the controlled category. The main reason is that the categories for which the response is shown statistically indifferentiable can be merged into one category. This could lead to a computational efficiency when dealing with a large number of characteristic parameters that have multiple categories.

The null hypothesis test for the pairwise comparison of contrasts between categories can be defined as follows:

$$H_0: \{\beta_i - \beta_k = 0\} \quad for\ i \neq k \tag{9}$$

The SAS procedure *proc glm*, through the statement *LSMeans*, carries out the pairwise comparison; details can be found in [14], with a theory of the hypothesis test discussed in [1]. Similarly, the contrast between a specific category versus multiple other categories[1] can also be tested, which, in our case, is needed when we require to merge more categories into another specific category.

Various other multiple comparison tests for the pairwise comparison were proposed in the context of ANOVA. These tests are relevant for different types of experimental design. In the case of the hedonic model, an experimental design is completely randomized, i.e., we assume that prices are assigned to the categories completely at random, with unequal sample sizes of product items within each category.

The Tukey method, first proposed in 1953 by Tukey in [17], which is considered more conservative among other multiple comparison tests, is the most suitable in our case. The Tukey's confidence interval for the difference between categories i and k at the α% significance level is defined as [1]:

$$\beta_i - \beta_k \pm \left(\frac{q_{v,n-v,\alpha}}{\sqrt{2}}\right)\sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_k}\right)} \tag{10}$$

where $q_{v,n-v,\alpha}$ is a critical point at the α% significance level from the *Studentised range distribution* and r_i and r_k are sample sizes of category i and k, respectively; a detailed derivation is shown in [1]. If the confidence interval in Eq. (10) contains zero, the difference between categories β_i and β_k is not significant at the α% significance level. The original Tukey's method was developed for the equal sample sizes in categories. Hayter in [9] shows that the same form of interval in Eq. (10) can be used for unequal sample sizes in individual categories, and that the overall confidence level is then at least 100(1 − α)%.

3. RESULTS

In the empirical study, we use web scraped data to demonstrate an application of the theoretical framework outlined in the previous section. Daily web scraped prices of the product category refrigerators (ECOICOP 5-digit: 05.3.1.1, refer to [13]) are available from 01 December 2019 until 31 December 2020. The prices of product items, including its corresponding characteristics parameters, were scraped from the website https://www.heureka.sk/. The aggregation of daily prices is carried out as in [10], i.e., daily prices of product items are geometrically averaged on a monthly level.

Table No. 2 shows a list of the characteristic parameters that are used in the estimation of Eq. (1).

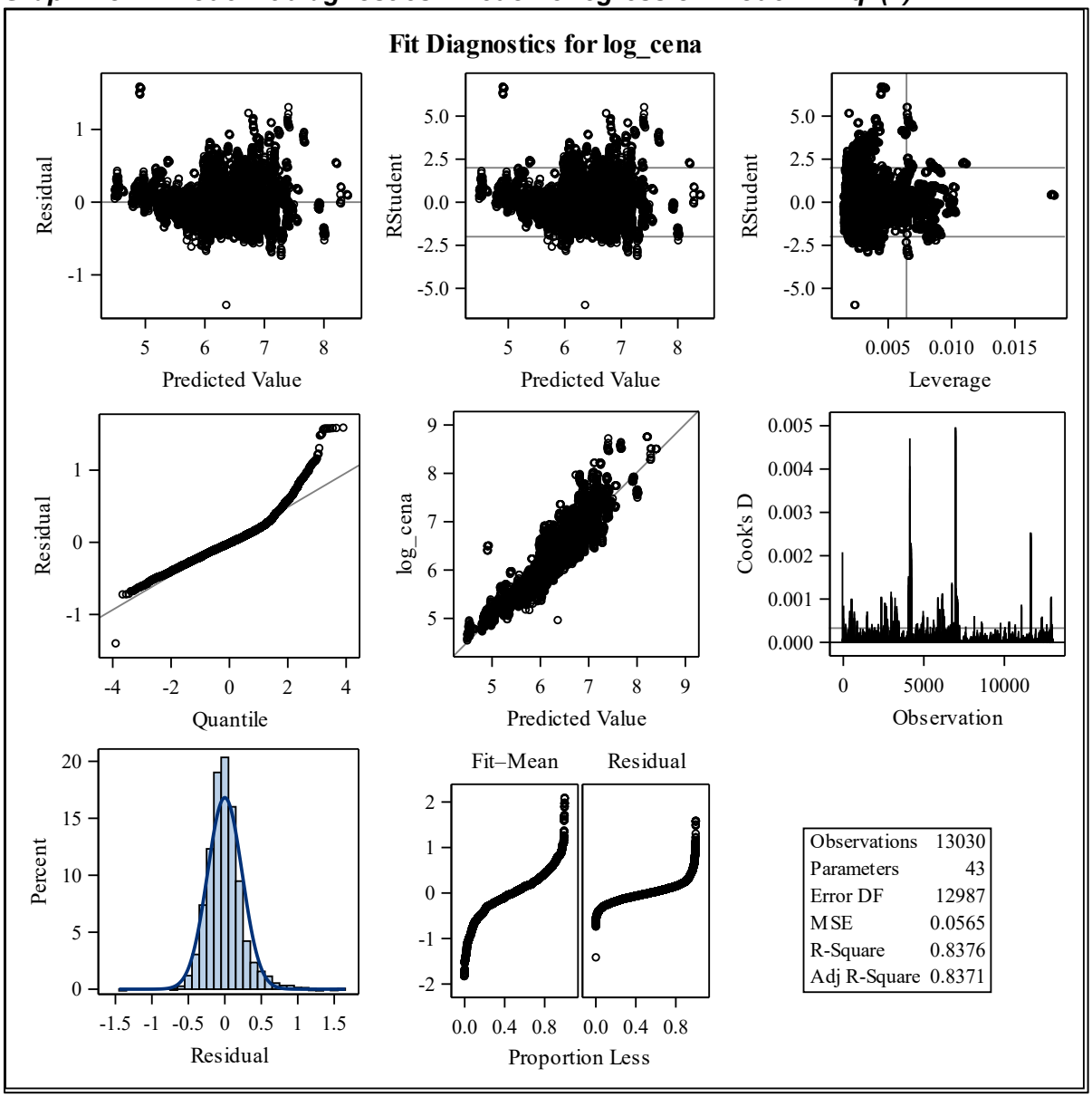
[1] It has to be correctly defined, which depends on the number categories tested, in the contrast statement in the SAS GLM procedure.

Table No. 2: The characteristic parameters

Parameter	Type (no. of categories)
Producer (výrobca)	Categorical (20)
Placement type (spôsob umiestnenia)	Categorical (2)
Noise (hlučnosť)	Numerical
Design (prevedenie)	Categorical (4)
Electricity consumption per year (spotreba energie za rok)	Numerical
Net volume of the fridge (čistý objem chladničky)	Numerical
Net volume of the freezer (čistý objem mrazničky)	Numerical
Height (výška)	Numerical
Width (šírka)	Numerical
Depth (hĺbka)	Numerical

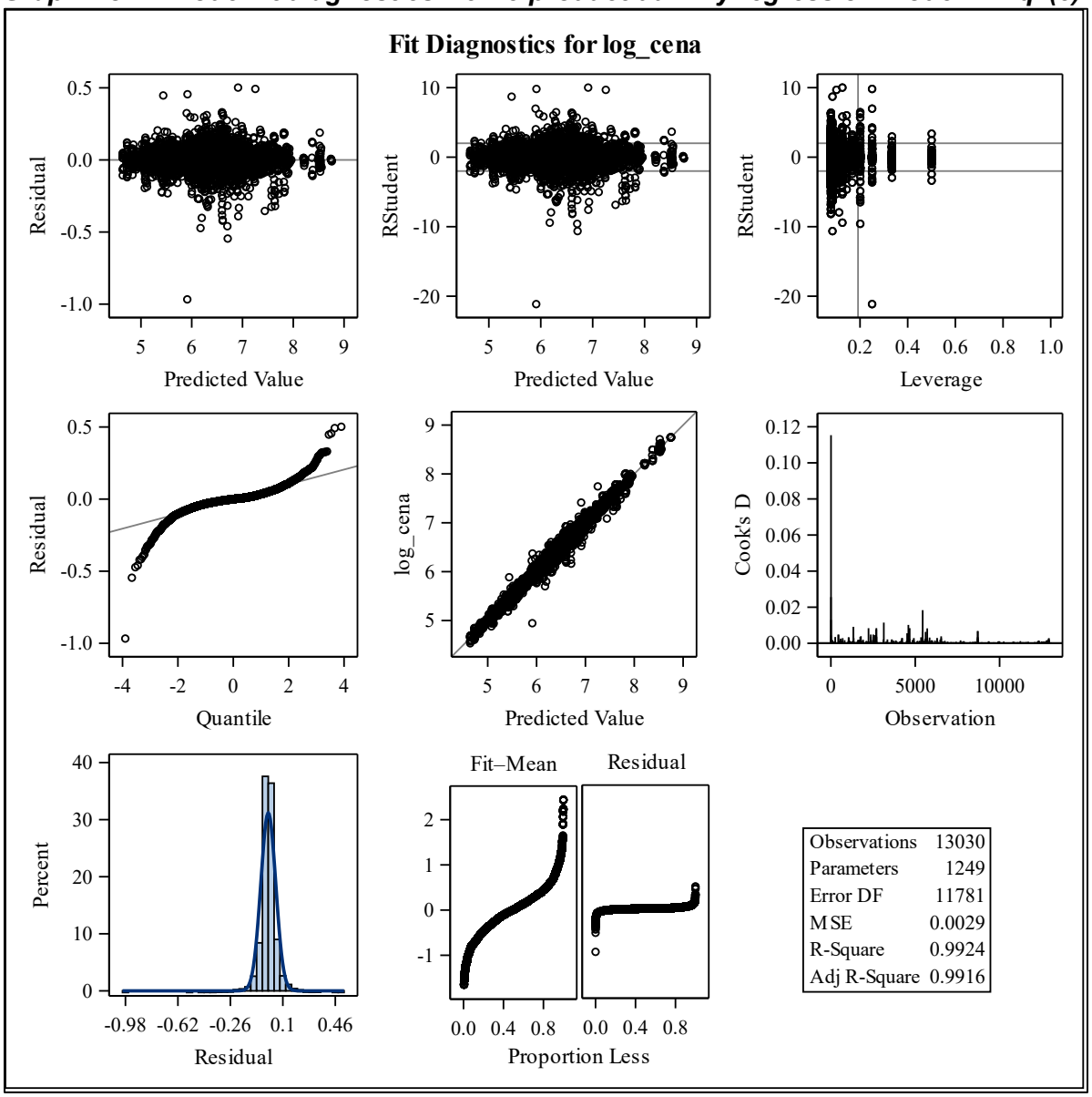
Source: SO SR, Author's construction

Graph No. 1: Model fit diagnostics – hedonic regression model in Eq. (1)



Source: SAS, Author's construction (Note: log_cena means log_price)

Graph No. 2: Model fit diagnostics – time-product dummy regression model in Eq. (3)



Source: SAS, Author's construction (Note: log_cena means log_price)

The estimation of Eqs. (1) and (3) is carried out in the SAS software, using the *proc GLM* procedure. The overall hypothesis test, using F statistics, show that both regression models are statistically significant at the 5% level. Additionally, all parameters, displayed in Table No. 2, in Eq. (1) are statistically significant at the 5% level. The full results are available from the author on request.

The graphs show the diagnostic tests, displayed in Table No. 1, for the fitted regression models in Eqs. (1) and (3), respectively.

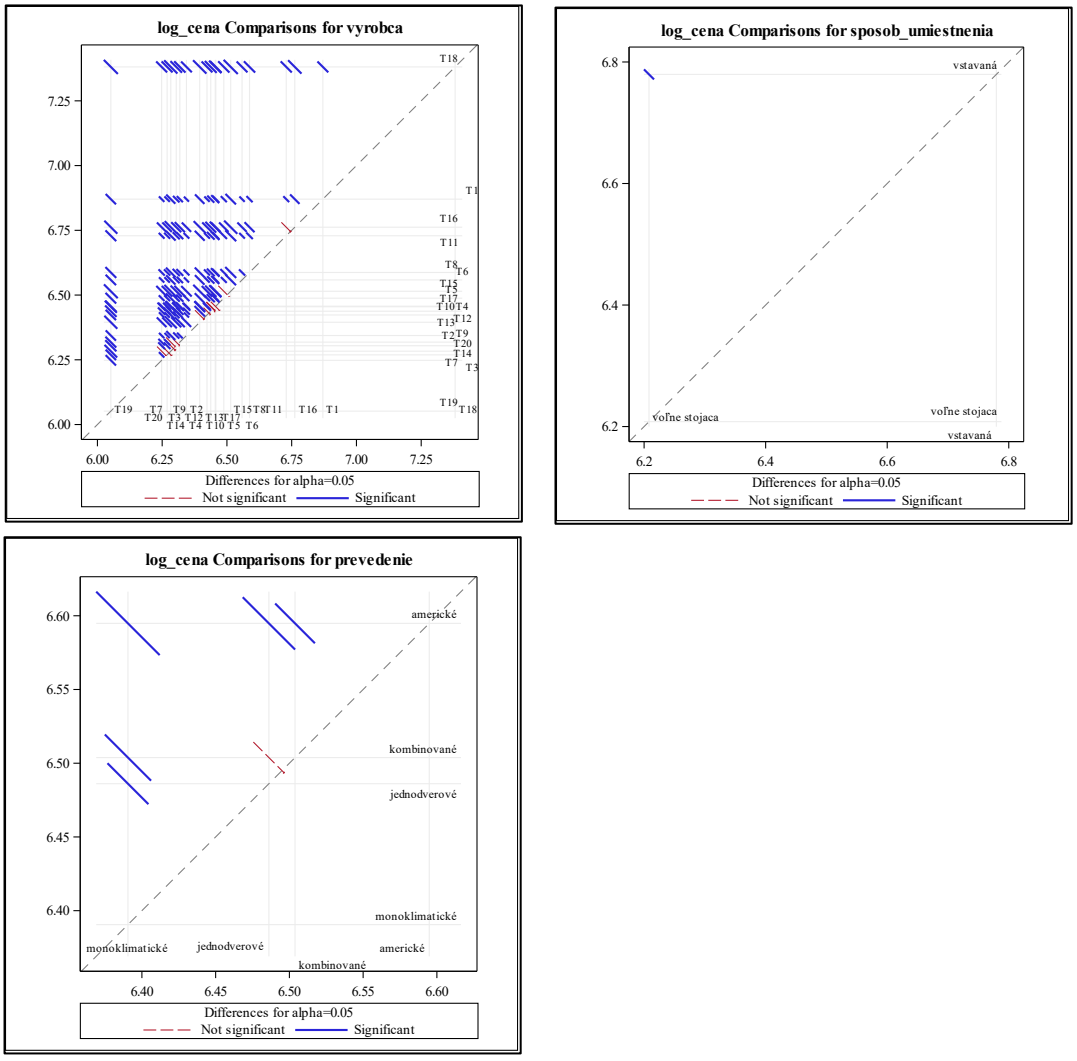
Graph No. 1 shows that the functional form of Eq. (1) is mis-specified. The plot of the residual vs predicted values shows a non-random distribution and a 'megaphone-like' shape that indicates a correlation in random errors. This leads to a conclusion that the random errors are not normally distributed and do not have a constant variance. Moreover, the graph of RStudent vs predicted values has many points beyond the limit(s) of ±2 that shows a presence of outliers in the observed prices that can be

inadvertently affecting the fitted model. From the second row of Graph n. 1, we can infer that the large prices cause a violation of the distributional properties of the random errors, which is seen in both the residual vs the quantile plot, the right end is far off the diagonal line, and the observed (log_cena) vs the predicted values plot.

Similar conclusions, for the presence of outliers in the observed prices, can be drawn for the fitted regression model in Eq. (3). However, the plot of residuals vs predicted values shows somewhat random distributions of values, which does not indicate a correlation among random errors, or a violation of the normal distribution assumption. In the third row, a distribution of residuals also resembles a normal curve, albeit with long tails caused by many outliers. Hence, we can infer that the fitted regression model is specified correctly, but the observed prices should be cleaned from outliers on both sides of the distribution.

We proceed to carry out a pairwise comparison of individual categories for each characteristic parameter. Noting that we have three categorical parameters, refer to Table No. 2. Graph No. 3 shows a significance level (at α = 5%) for the pairwise comparison.

Graph No. 3: Hypothesis test for the pairwise comparison of categories



Source: SAS, Author's construction (Note: log_cena means log_price)

The red line(s), in individual plots, indicate that the difference between categories[2] for the given characteristic parameter is not significantly different from zero, refer to Eq. (9). For categories that show non-significant differences, we create a single category. However, we consider the significance level for the difference at α = 10%. The following table shows the categories that are merged into one category.

Table No. 3: Merged categories – pairwise comparison

Control (main) category	Merged category
T12	T17
T12	T4
T14	T20
T14	T9
T15	T5

Source: Author's construction

Noting that for the merger of multiple categories, we carry out an additional contrast hypothesis test for in-between differences, for example, under the null hypothesis $\beta_{T12} = \beta_{T17} = \beta_{T4}$; refer to the SAS code in Annex B, particularly, the statement contrast.

In the follow-up analysis, we proceed to test the hypothesis of the pairwise comparison using the Tukey' multiple comparison method in Eq. (10). The results, which we display only for pairwise categories that are not significantly different from zero, are shown in Table No. 4:

Table No. 4: Confidence intervals – Tukey method

Comparisons significant at the 0.05 level are indicated by ***.			
vyrobca Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
T16 - T18	-0.085394	-0.177286	0.006497
T5 - T10	-0.006769	-0.048698	0.035159
T1 - T10	-0.022936	-0.062997	0.017125
T1 - T5	-0.016167	-0.054711	0.022377
T8 - T17	-0.056017	-0.113843	0.001808
T8 - T13	0.007288	-0.059478	0.074055
T13 - T17	-0.063306	-0.137796	0.011185
T14 - T15	-0.029089	-0.104795	0.046616
T3 - T15	-0.055178	-0.127308	0.016951
T3 - T14	-0.026089	-0.068946	0.016767
T3 - T4	0.025384	-0.010586	0.061355
T12 - T19	0.012626	-0.065001	0.090253

Source: SAS, Author's construction

Similarly based on the Tukey's method, we merged the categories with non-significant differences. Noting that the confidence interval that contains zero for the pairwise comparison indicates that the difference between these categories is not significant from zero. Table No. 5 shows the merged categories.

[2] *For the translation of individual categories refer to Annex A.*

Table No. 5: Merged categories – Tukey method

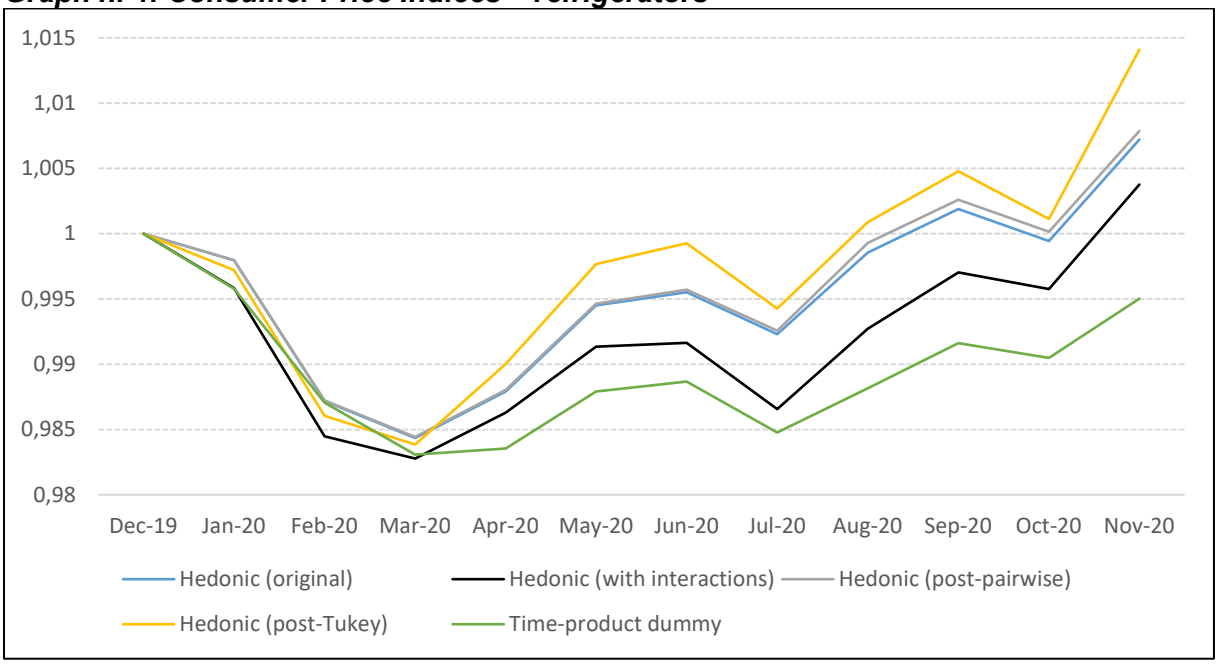
Control (main) category	Merged category
T16	T18
T1	T10
T1	T5
T8	T17
T8	T13
T3	T15
T3	T4
T3	T14
T12	T19

Source: Author's construction

Noting that after the merge of the categories in Table No. 5, the null hypothesis for the Tukey's test is rejected, at α = 5%, for all pairwise comparisons.

In the final stage of the analysis, we show the estimation of consumer price indices for all scenarios. The hedonic consumer price index is estimated for four different scenarios: original data, original data with interactions, post-pairwise and post-Tukey's merge of categories, respectively, along with the time-product dummy (TPD) consumer price index. The following graph shows the results.

Graph n. 4: Consumer Price Indices – refrigerators



Source: SOSR, Author's construction

Graph No 4 shows that all indices have the same trend. The TPD index is the lower bound in most of time periods. Adding the effect of interactions between characteristic parameters in Eq. (1) shifts the hedonic index down, towards the TPD index. This could indicate that missing parameters in the model might cause upward drift in the hedonic index. The hedonic index, after merging categories with non-significant differences, moves upwards. Noting that relatively big differences between individual price indices, for example, almost 2% between the lowest and largest index value, is solely due to

the specification of models, i.e., the number of parameters, explanatory variables, in the model.

All results have been generated in SAS software and are available from the author on request.

In general, based on the results, we can conclude that the number of characteristic parameters in the model plays an important role in the context of the estimated price indices. Therefore, it is important that all the relevant parameters of the particular product are collected and included in the estimation of the model. Moreover, the diagnostic tests show that extreme price values can cause a violation of the initial model's assumptions that can result in the erroneous price indices. The exclusion of extreme prices should be considered.

4. CONCLUSION

At the outset of the paper, we demonstrate a theoretical framework of hedonic and time-product dummy regression models. Additionally, a concept for diagnostic tests of the fitted regression model is outlined, including the analysis of variance that, in general, tests for the (average price) difference between categories of each characteristic (categorical) parameter. The diagnostic tests are crucial for testing the assumptions and the model's goodness of fit. An incorrect functional form of the regression model leads to errors in the estimated parameters, particularly, in hedonic consumer price indices. Moreover, the ANOVA might indicate that some categories of the characteristic parameters can be merged, which might lead to a minimisation of the computational time in practice.

It is important to consider online prices in the calculation of consumer price indices since the economic activity shows that consumers purchase some products online. In the empirical study, we use web scraping, which is considered a new data source, for collecting prices of products. The online daily prices are aggregated by using a geometric average on a monthly level.

Moreover, in the empirical analysis, using online prices for the product category of refrigerators from the Slovak market, we show that the functional form of the fitted hedonic regression model is misspecified. This is shown in the inspection of the diagnostic tests for the fitted model. The outliers play a role in the violation of the model assumptions, in particular, the assumptions of the normal distribution of random errors and its constant variance are not met. Hence, we are unable to confirm a suitability of the hedonic consumer price index for this case.

To improve the model's performance, we might consider excluding the extreme prices on both end of the distribution. On the other hand, these excluded product items, if economically important, would not be reflected in the estimation of the hedonic consumer price index. This shows that a disadvantage of web scraped data is that they do not contain any information on sales.

We show that characteristic parameters, and its corresponding categories, are statistically significant in determining the log-price of the product. Since some product categories, with a high replacement rate, and availability of big pool of product items require more complex formulas to capture the average price changes across time more

accurately, we recommend that hedonic price indices are considered by National Statistical Institutes for price statistics. Nevertheless, as demonstrated in the paper, NSIs must conduct a thorough analysis of hedonic regression models before its implementation in the production environment, either for estimating consumer price indices or for missing price imputations.

Further research, which can be part of the ongoing project of the modernisation of data sources for price statistics at the Statistical Office of the Slovak Republic, should concern the analysis of splicing methods that are used for avoiding the revision problem in consumer price indices when using multilateral methods.

ACKNOWLEDGEMENTS

This work was conducted as part of the European project *Dynamic Pricing Model* (311071AA56) that is supported by the European Union and co-financed through the European Regional Development Fund. The author would like to acknowledge a support of the grant by the Grant Agency of the Slovak Republic VEGA 1/0047/23 „The importance of spatial spillover effects in the context of the EU's greener and carbon-free Europe priority“. The author would also like to express an appreciation to the Heureka management for allowing the Statistical Office of the Slovak Republic to scrap daily product prices from their website platform.

BIBLIOGRAPHY

[1] DEAN, A., VOSS, D.: Design and Analysis of Experiments, Springer Texts in Statistics, 1999, Springer-Verlag New York Inc.
[2] EUROSTAT, Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Manuals and Guidelines, 2022, Luxembourg: Publication Office of the European Union. ISBN 978-92-76-44354-4.
[3] EUROPEAN COMISSION, EUROSTAT: Practical guidelines on web scraping for the HICP, Harmonised Indices of Consumer Prices, Directorate C: Macro-economic statistics, Unit C-4: Price statistics, Purchasing Power Parities, Housing statistics, November 2020.
[4] ILO/IMF/OECD/UNECE/EUROSTAT/THE WORLD BANK. Consumer Price Index Manual: Theory and Practice, 2004. ILO Publications, Geneva.
[5] DE HAAN, J., HENDRIKS, R.: Online data, fixed effects the construction of high-frequency price indexes. In: Paper presented at the Economic Measurement Group Workshop, 2013, 28-29 November 2013, Sydney, Australia.
[6] DE HAAN, J., KRSINICH, F.: Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes. In: Journal of Business & Economic Statistics 32, 2014, pp. 341-358.
[7] DE HAAN, J.: Hedonic Prices Indexes: A Comparison of Imputation, Time Dummy and ‘Re-Pricing’ Methods. In: Jahrbücher f. Nationalökonomie u. Statistik, Lucius & Lucius, Stuttgart, 2010, Bd. (Vol.) 230/6, pp. 772-791.
[8] GLASER-OPITZOVÁ, H.: Nové zdroje údajov pre cenovú štatistiku a metódy ich spracovania. In: Slovenská štatistika a demografia, 2019, roč. 29, č.4, str. 49 – 66.
[9] HAYTER, A. J.: A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. In: The Annals of Statistics, 1984, vol. 12, no. 1, pp. 61-75.
[10] KNÍŽAT, P., GLASER-OPITZOVÁ, H.: Index spotrebiteľských cien z webscrapovaných údajov: analýza vybranej produktovej skupiny. In: Slovenská štatistika a demografia, 2023, roč. 33, č.1, pp. 37 – 49.

[11] KNÍŽAT, P.: Web scraped data in consumer price indices. In: Statistical Journal of the IAOS, 2023, vol. 39, no.1, pp. 203-212.
[12] POLIDORO, F., GIONNINI, R., LO CONTE, R., MOSCA, S. and ROSSETTI, F.: Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. In: Statistical Journal of the IAOS, 2015, vol. 31, no. 2, pp. 165-176.
[13] RAMON - Reference and Management of Nomenclatures: Europa - RAMON - Classification Detail List [cit. 2022-12-05].
[14] SAS INSTITUTE INC.: SAS/STAT® 13.1 User's Guide, The GLM Procedure, 2013, Cary, NC: SAS Institute Inc.
[15] Tovarová skupina chladničky: https://lednice.heureka.sk/ [cit. 2022-12-05].
[16] TRIPLETT, J. E.: Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. Organization for Economic Co-operation and Development, 2006, Paris.
[17] TUKEY, J. W.: The problem of multiple comparisons. Dittoed manuscript of 396 pages, 1953, Department of Statistics, Princeton University.

RESUME
The Statistical Office of the Slovak Republic examines the use of new data sources for official price statistics. One of the alternative data sources for collecting prices of products is the online environment. Web scraping is an automated way of online price collection. The advantage of web scraping is that it does not require vast human resources that are needed in the traditional data collection from physical stores. Moreover, it also enables collecting characteristic parameters of individual products that can be used in the estimation of the hedonic regression model. The hedonic regression model can be used for the estimation of the hedonic price index, or for the imputation of missing prices. The advantage of the hedonic price index is that it captures the changes of all available products in the time window, including the new and disappearing products. In this paper, we show the estimation of the hedonic regression model and carry out various statistical tests to verify the model's specifications. In addition, we analyse the statistical significance of individual characteristic parameters, including a detailed analysis of variance for the categorical variables. Based on the analysis, we propose adjusting the categories for individual characteristic parameters that can reduce a calculation time when estimating the hedonic consumer price index for big data in practice. In the empirical study, we show the application, and compare hedonic consumer price indices for different scenarios, for the product category of refrigerators. In paper is part of the project that deals with the modernization of price statistics from the perspective of different data sources at the Statistical Office of the Slovak Republic.

RESUMÉ
Štatistický úrad SR skúma využitie nových zdrojov údajov pre oficiálnu cenovú štatistiku. Jedným z alternatívnych zdrojov dát na zber cien produktov je online prostredie. Web scraping je automatizovaný spôsob zberania cien z internetu. Výhodou web scrapingu je, že nevyžaduje veľa ľudských zdrojov, ktoré sú potrebné pri tradičnom zbere dát z obchodov. Okrem toho umožňuje aj zber charakteristických parametrov jednotlivých produktov, ktoré je možné použiť pri odhade hedonického regresného modelu. Model hedonickej regresie možno použiť na výpočet hedonického cenového indexu alebo na imputáciu chýbajúcich cien. Výhodou hedonického cenového indexu je, že zachytáva zmeny cien všetkých dostupných produktov v časovom okne, vrátane nových a tzv. odchádzajúcich produktov. V článku ukážeme

odhad hedonického regresného modelu a vykonáme rôzne štatistické testy na overenie správnosti modelu. Okrem toho analyzujeme štatistickú významnosť jednotlivých charakteristických parametrov, vrátane podrobnej analýzy rozptylu pre kategorické premenné. Na základe analýzy navrhneme upraviť kategórie jednotlivých charakteristických parametrov, čoho dôsledkom môže byť napr. skrátenie výpočtového času pri odhade hedonického indexu spotrebiteľských cien pre veľké dáta. V empirickej štúdii ukážeme aplikáciu výpočtu hedonických indexov spotrebiteľských cien a ich výsledky porovnáme pre rôzne prípady, pri výpočte použijeme ceny pre produktovú kategóriu chladničky. Príspevok je súčasťou projektu, ktorý sa zaoberá modernizáciou cenových štatistík z pohľadu využitia nových zdrojov údajov v Štatistickom úrade SR.

CURRICULUM VITAE

Peter Knížat MSc. *is an external PhD student at the Faculty of Economic Informatics, University of Economics in Bratislava. His main research interests are in spatial regressions and functional data; his publications can be found at: https://orcid.org/0000-0001-5100-1319. He teaches practical exercises for: Support of decision-making processes and Multiple attribute decision-making, including its application in the statistical software R. He works as a statistician at the Directorate of General Methodology, Registers and Coordination of the National Statistical System, Statistical Office of the Slovak Republic (SO SR), where he is responsible for proposing a statistical methodology for big data analysis. Prior to working at SO SR, he worked in an international bank as a senior risk and portfolio manager, where he lead the development of Basel internal risk-based models used in internal credit risk assessment processes and in the calculation of the bank's regulatory capital, and expected credit loss models used in the International Financial Regulatory Standards (IFRS) 9 reporting.*

CONTACT

peter.knizat@statistics.sk
peter.knizat@euba.sk

ANNEX A

Acronym	Producer (vyrobca)
T1	LIEBHERR
T2	GORENJE
T3	WHIRLPOOL
T4	ELECTROLUX
T5	SAMSUNG
T6	BOSCH
T7	BEKO
T8	AEG
T9	CANDY
T10	LG
T11	SIEMENS
T12	AMICA
T13	PHILCO
T14	HISENSE
T15	LORD
T16	HAIER
T17	CONCEPT
T18	MIELE
T19	INDESIT
T20	ZANUSSI

Acronym	Placement Type (sposob_umiestnenia)
vstavana	built-in
volne stojaca	free-standing

Acronym	Design (prevedenie)
americke	american
kombinovane	combined
jednodverove	one-door
monoklimaticke	monoclimatic

ANNEX B

```
/* Hedonic Regression index - original data */
ODS GRAPHICS ON;
TITLE; TITLE1 "Hedonic Model - ANOVA";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";

ods output
parameterestimates=FE_parameters_h;
proc glm data=unmatched_index_2
          PLOTS(maxpoints=none)=(DIAGNOSTICS RESIDUALS);
class vyrobca sposob_umiestnenia prevedenie cum_time (REF=FIRST);
model log_cena = cum_time vyrobca sposob_umiestnenia prevedenie
```

```
spotr_energie_rok objem_chladnicka vyska sirka hlbka objem_mraznicka hlucnost /
SS3 solution;
        LSMEANS vyrobca / PDIFF;
         LSMEANS sposob_umiestnenia / PDIFF;
         LSMEANS prevedenie / PDIFF;

         MEANS vyrobca / TUKEY CLDIFF ALPHA=0.05;
         MEANS sposob_umiestnenia / TUKEY CLDIFF ALPHA=0.05;
         MEANS prevedenie / TUKEY CLDIFF ALPHA=0.05;

         /* contrasts after pairwise */
        contrast "AMICA vs CONCEPT/ELECTROLUX"
                    vyrobca  0 0 0 -1 0 0 0 0 1,
                    vyrobca  0 0 0 -1 0 0 0 0 0.5 0 0 0 0 0 0.5;
         contrast "HISENSE vs ZANUSSI/CANDY"
                    vyrobca  0 0 0 0 0 -1 0 0 0 0 0 0 1,
                    vyrobca  0 0 0 0 0 -1 0 0 0 0 0 0 0.5 0 0 0 0 0 0 0.5;
run;

TITLE; FOOTNOTE;
ODS GRAPHICS OFF;

/* TPD Regression index - original data */
ODS GRAPHICS ON;
TITLE; TITLE1 "TPD Model";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";

ods output
parameterestimates=FE_parameters_tpd;
proc glm data=unmatched_index_2
          PLOTS(maxpoints=none)=(DIAGNOSTICS RESIDUALS);
class  cum_time (REF=FIRST) id;
model log_cena = cum_time id / SS3 solution;
run;

TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```