Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

63rd ISI World Statistics Congress (WSC) / IFC-sponsored Invited Paper Sessions (IPS) related to new developments in central bank statistics

July 2021, The Hague, Netherlands, virtual format

New hedonic quality adjustment method using sparse estimation[1]

Sahoko Furuta and Yoshiyuki Kurachi,
Bank of Japan

[1] This presentation was prepared for the WSC. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

BANK OF JAPAN

New Hedonic Quality Adjustment Method using Sparse Estimation

Sahoko Furuta

Bank of Japan

Research and Statistics Department

Introduction

- ✓ The hedonic estimation generally has issues with multicollinearity and the omitted variable bias. This leads to a low estimation accuracy and a large estimation burden in practice.

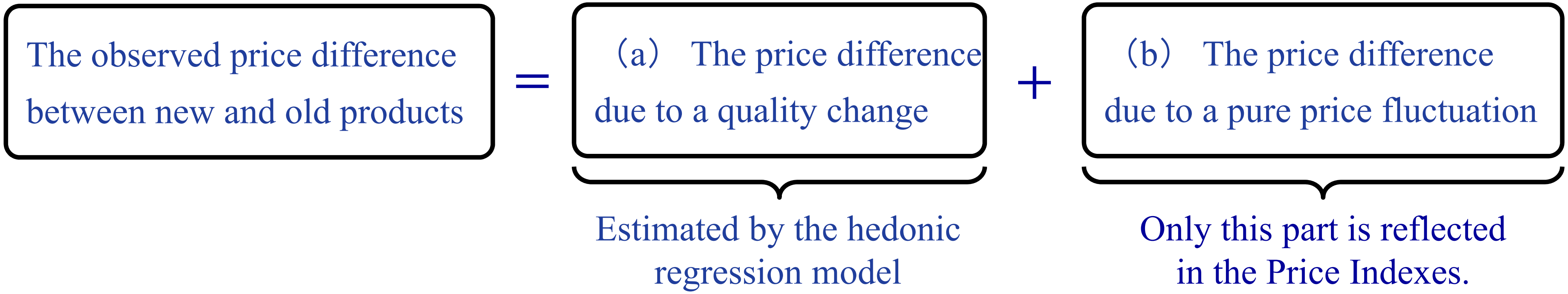
- ✓ To overcome these problems, we introduce new estimation method using "sparse estimation" as a way to automatically select the meaningful variables from a large number of candidates.

- ✓ The new method brings three benefits;
  1. A significant increase in the number of variables in the model
  2. An improvement in fit of the model to actual prices
  3. A reduction of the over-estimation in quality improvements due to the omitted variable bias

1. Motivations

What is Hedonic Quality Adjustment?

- ✓ The Bank of Japan applies **the hedonic quality adjust method** in the compilation of the Price Indexes to eliminate the effect of products' quality changes.
- ✓ When a product turnover occurs, the observed price difference between new and old products is decomposed into (a) the difference due to a quality change and (b) the difference due to a pure price fluctuation, which is called **quality adjustment**.



The observed price difference between new and old products = （a） The price difference due to a quality change + （b） The price difference due to a pure price fluctuation

Estimated by the hedonic regression model

Only this part is reflected in the Price Indexes.

- ✓ In the hedonic method, the relationship between **product quality** and **price** is statistically regressed with a large amount of data. This method is not only highly objective, but also applicable to various changes in characteristics of products.

Overview of Conventional method

✓ Given the non-linear relationship between the price and characteristic of a product, the hedonic regression model often has both of linear parts and non-linear parts by the Box-Cox transformed term.

$$y_i^{(\lambda_0)} = \beta_0 + \sum_{k=1}^{p_d} \beta_{dk} x_{dk,i} + \sum_{j=1}^{p_c} \beta_{cj} x_{cj,i}{}^{(\lambda_j)}$$

y_i: theoretical price, $x_{cj,i}$: continuous variable, $x_{dk,i}$: dummy variable,
β_0: constant term, β_{cj}: coefficient on a continuous variable,
β_{dk}: coefficient on a dummy variable,
λ_0: Box-Cox parameter for theoretical price,
λ_j: Box-Cox parameter for a continuous variable,
p_c: number of continuous variables, p_d: number of dummy variables

Issues of Conventional method

Accuracy of estimation

- Multicollinearity
- The omitted variables bias

These problems are likely to arise when the characteristics of the products are highly correlated. They disturb accurate estimation of the parameters.

Burden of estimation

Repeating estimation while changing the set of the variables (excluding variable that cause multicollinearity and including the meaningful variables) to obtain good results.

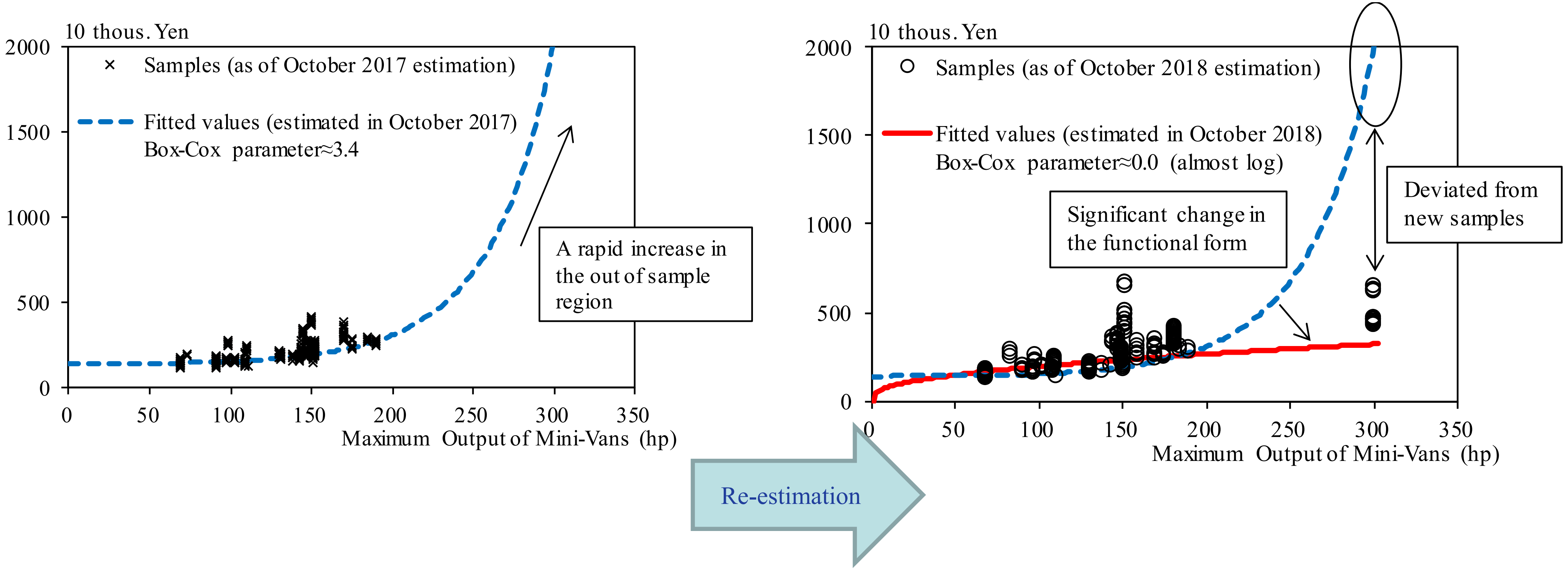
Accuracy of estimation (1)

日本銀行 BANK OF JAPAN

- ✓ Estimated parameters on variables may become unstable due to the problem of multicollinearity and the omitted variables bias.
- ✓ Multicollinearity refers to a state in which there is a high inter-connection among the variables. Multicollinearity makes it difficult to identify price effects of variables, and it may also cause the omitted variables bias through the variable selection based on the statistical significance. As a result, the parameters are not estimated accurately.
- ✓ It is known that these problems can be more serious as the model has more complex functional form to deal with the non-linear effects of price determining characteristics.

Accuracy of estimation (2)

- ✓ A distorted functional form has a problem, called "overfitting."
- ✓ The model may give quite poor estimates for the new products (i.e. out-of-sample).



Burden of estimation

✓ As mentioned, the model with complex functional form may be suffered by the problem of multicollinearity and the omitted variables bias.

✓ Then, a slight change in sample and regressors often leads to a quite different estimation result in each re-estimation. Discontinuity in the estimates is highly problematic in practice.

⇒We have to repeat estimation with changing the set of the variables each time until obtaining a better and acceptable result.

✓ This problem is serious in the estimation of "passenger car", where there are many candidate variables and they are highly correlated.

2. New Method using Sparse Estimation

Sparse Estimation (1)

✓ Sparse estimation has a property that select the meaningful variables from a large number of candidates and gives zero coefficients to the rest of the variables ("Sparsity"). It can perform "variable selection" and "coefficient estimation" at the same time and can automatically derive a stable and well fitted model.

✓ The new estimation method proposed in this study employs an Adaptive Elastic Net (AEN), which enjoys two desirable properties;

1. "Group Effect" that gives robustness for multicollinearity
2. "Oracle Property" that ensures the adequacy of variable selection and estimated coefficients.

Sparse Estimation (2)

- ✓ For example, Lasso, a typical sparse estimation, estimates $\boldsymbol{\beta}$, by minimizing loss function: the sum of the squared errors and the regularization term ($\boldsymbol{L}_1$norm of $\boldsymbol{\beta}$).
- ✓ Lasso has similar loss function with Ridge, but differs in that it has sparsity.

Lasso

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}|^2+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

Ridge

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}|^2+\lambda\sum_{i=1}^{p}{\beta_j}^2\right)$$

$\lambda > 0$: regularization parameter (It selects relatively smaller number of variables if λ is large)

Sparse Estimation (3)

- ✓ In the bivariate model, $\boldsymbol{\beta}$ is derived from the intersection of the contour line of the sum of squared error and the constraint.
- ✓ Lasso gives $\boldsymbol{\beta}$ at the corners of rhombus of the constraint, and then one coefficient is estimated to be exactly zero.

Lasso

$$\underset{\beta_1,\beta_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \beta_1 X_{1,i} - \beta_2 X_{2,i}\right)^2$$

$$\text{s.t. } |\beta_1| + |\beta_2| \leq s$$

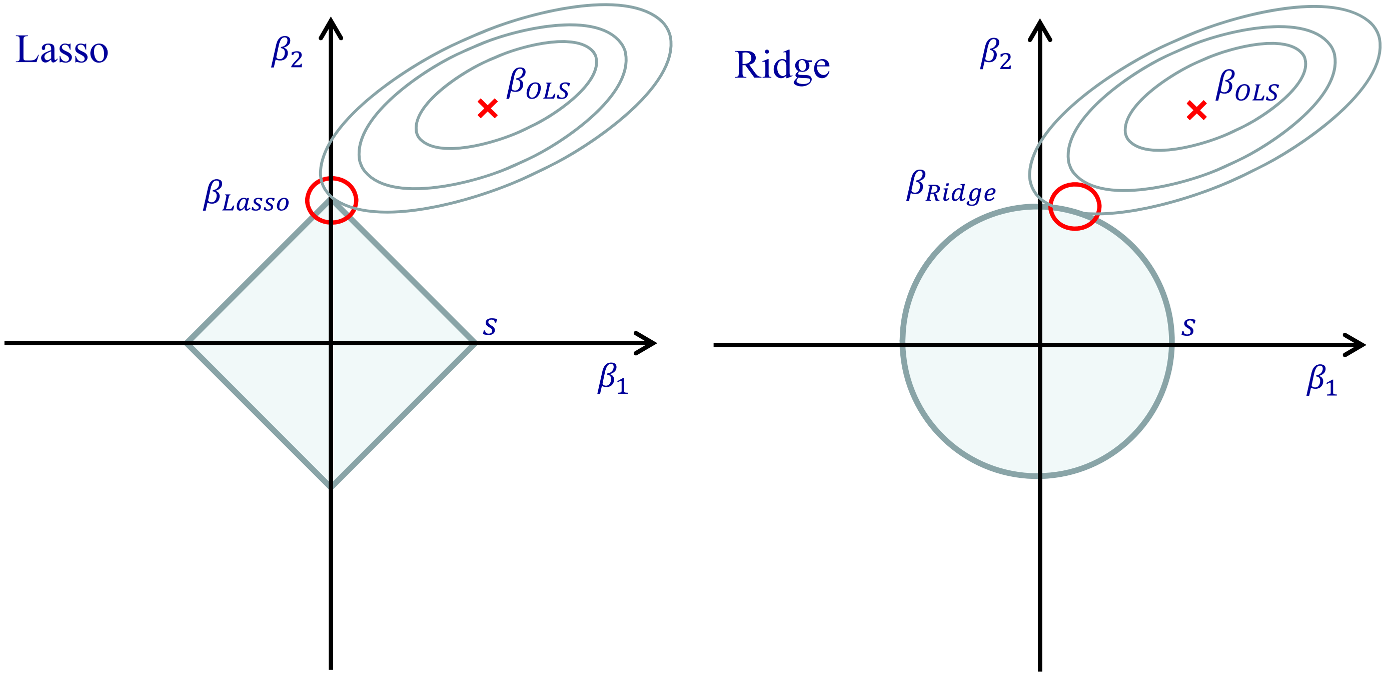
$s > 0$: 1−1 corresponding to λ

Ridge

$$\underset{\beta_1,\beta_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \beta_1 X_{1,i} - \beta_2 X_{2,i}\right)^2$$

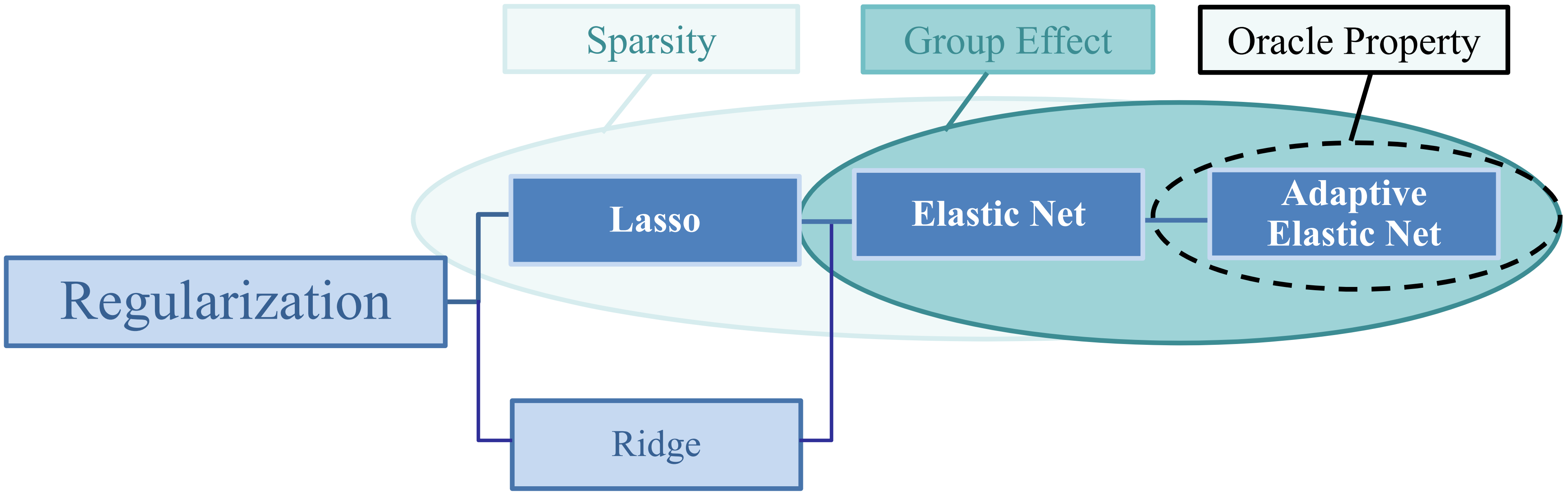
$$\text{s.t. } {\beta_1}^2 + {\beta_2}^2 \leq s^2$$

$s > 0$: 1-1 corresponding to λ



Adaptive Elastic Net (1)

- ✓ AEN can be interpreted as the combination of the Lasso and the Ridge.
- ✓ It has "group effect" and "oracle property."



Adaptive Elastic Net (2): Group Effect

- ✓ For Lasso, the results of variable selection are known to be unstable in data has strong multicollinearity.
- ✓ A typical method to overcome this problem is the "Elastic Net (EN)."
- ✓ The robustness of EN for multicollinearity is called "group effect". It is a property that gives similar coefficients on variables when the correlation between them is high.

$$\hat{\boldsymbol{\beta}}(EN) = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\boldsymbol{\beta}}{\operatorname{argmin}}|\boldsymbol{Y} - \boldsymbol{X\beta}|^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|\right\}$$

$\lambda_2 > 0$: L_2 norm regularization parameters
$\lambda_1 > 0$: L_1 norm regularization parameters
n: number of observations

Adaptive Elastic Net (3): Oracle Property

✓ The "oracle property" is known as a property that asymptotically guarantees the appropriateness of both the "variable selection" and the "coefficient estimation".

When $\boldsymbol{\beta}^*$ is the true coefficient, the estimator $\widehat{\boldsymbol{\beta}}$ satisfies the following;

(1) Variable Selection Consistency

$$\lim_{n\to\infty} P\left(\hat{\beta}_j = 0\right) = 1 \qquad with\ \beta_j{}^* = 0$$

(2) Asymptotic Normality of the Non-zero Coefficients

$$\lim_{n\to\infty} \frac{\left(\hat{\beta}_j - \beta_j{}^*\right)}{\sigma\left(\hat{\beta}_j\right)} \sim \mathrm{N}(0{,}1) \quad with\ \beta_j{}^* \neq 0$$

$\sigma^2\left(\hat{\beta}_j\right)$: asymptotic variance of estimator

Adaptive Elastic Net (4)

- ✓ We employ AEN as a new estimation method for hedonic regression model.

- ✓ The AEN estimation is performed in two stages. At the first stage, we estimate the coefficients with EN. Then, EN is performed again to impose greater penalties for variables with small absolute values of the coefficients.

$$\widehat{\boldsymbol{\beta}}(AEN) = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(|\boldsymbol{Y} - \boldsymbol{X\beta}|^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 + {\lambda_1}^* \sum_{j=1}^{p} \widehat{w}_j |\beta_j|\right)\right\}$$

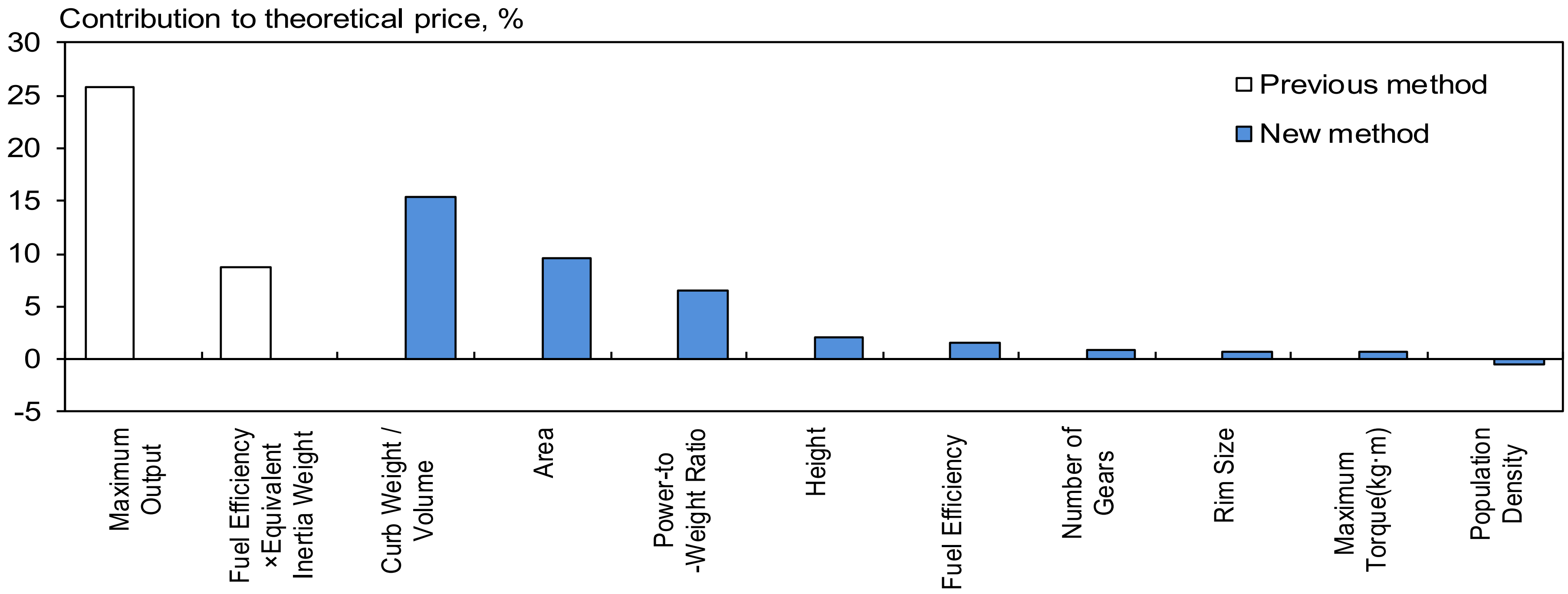
$$\widehat{w}_j = \left(|\hat{\beta}_j(EN)|\right)^{-\gamma}$$

${\lambda_1}^* > 0$: L_1 norm regularization parameters (2nd stage)
$\widehat{w}_j > 0$: adaptive weight, $\gamma > 0$: adaptive parameter
(Larger γ imposes larger penalties corresponding to the absolute value of the coefficient)

3. Estimation Results

Continuous variables in the model

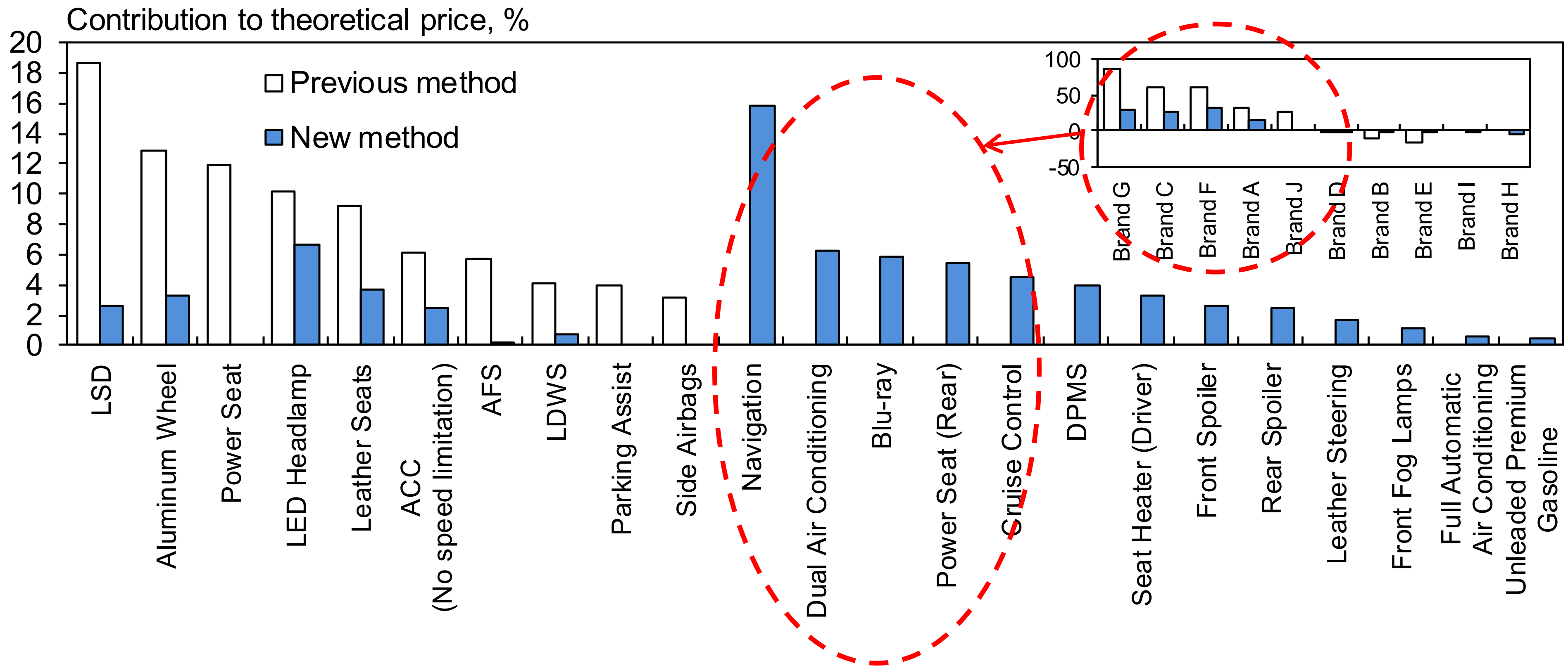
- ✓ We apply new and previous hedonic regression models to passenger cars in Japan and compare those results.
- ✓ The number of continuous variables in the regression models increases and this is accompanied by a reduction in dependence on just a few specific variables.



Note: Bar charts indicate the rates of change in theoretical price due to one unit increase in variables where all variables of a product are set at sample means.

Dummy variables in the model

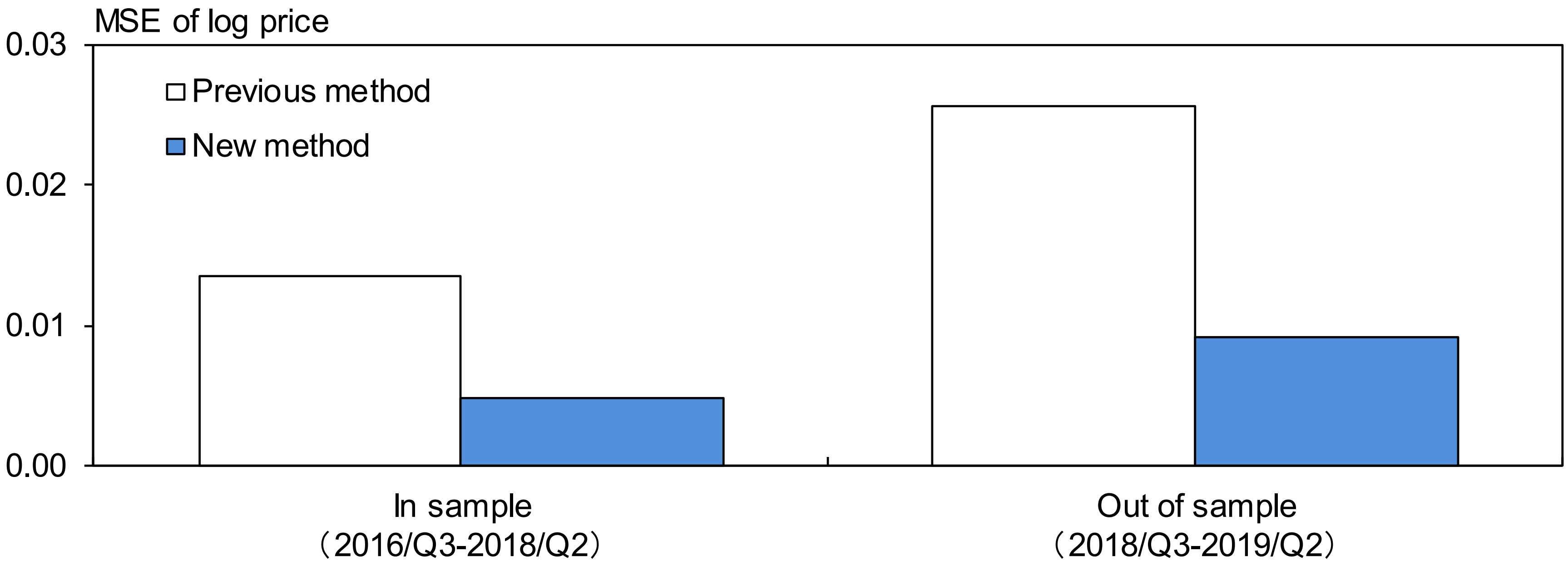
✓ As a result of the increased number of characteristics, the new regression model reduces its reliance on manufacturer dummies (control variables).



Note: Bar charts indicate the rates of change in theoretical price due to one unit increase in variables where all variables of a product are set at sample means.

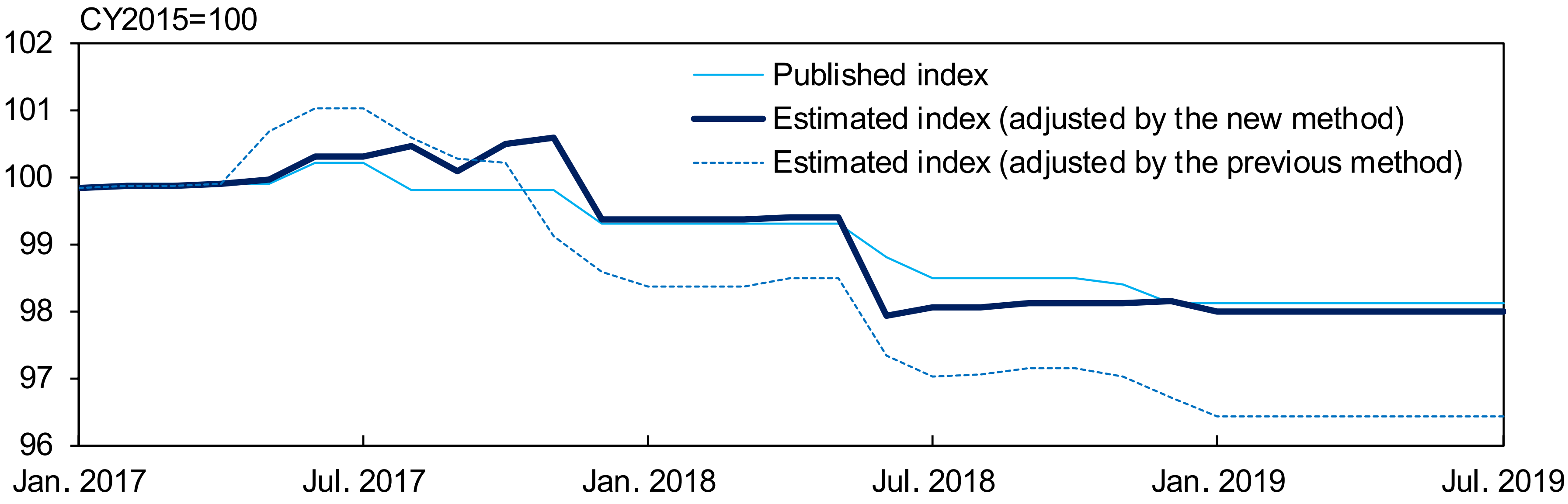
Fit of the model

- ✓ The fit (mean squared errors) of regression models to actual price improves in the new estimation method for both in-sample and out-of-sample period.
- ✓ Since the quality adjustment is generally applied to products, released after the estimation, the improvement in the out-of-sample fit implies an increase in the usefulness of the hedonic quality adjustment method in practice.



Estimated Price Index

- ✓ The estimated price index of "standard passenger cars (gasoline cars)" in the PPI, which is retrospectively calculated by applying the new hedonic estimation method to all quality adjustments, shows similar developments to the published price index.
- ✓ On the other hand, the previous method highlights the risk of over-estimating the rate of quality improvement as it shows an excessive decline in the price.



4. Conclusion

Conclusion

✓ The new estimation method using "sparse estimation"

1. mitigates the problems of omitted variables and multicollinearity significantly.
2. improves estimation accuracy and reduces estimation burden.
3. possibly improves the accuracy of the price index.

✓ The proposed method can supports effective use of big data for price statistics as it can automatically build a good performance model by extracting all necessary information even with the large dataset.

New Hedonic Quality Adjustment Method using Sparse Estimation

Sahoko Furuta[1]*, Yoshiyuki Kurachi[2]

[1] Bank of Japan, Tokyo, JAPAN, sahoko.furuta@boj.or.jp
[2] Bank of Japan, Tokyo, JAPAN, yoshiyuki.kurachi@boj.or.jp

Abstract:
In the application of the hedonic quality adjustment method to the price index, multicollinearity and omitted variable bias arise as practical issues. This paper proposes the new hedonic quality adjustment method using "sparse estimation" in order to overcome these problems. The new method deals with the problems by ensuring two properties: the "Grouped Effect" that gives robustness for multicollinearity and the "Oracle Property" that provides the appropriate variable selection and the asymptotically unbiased estimators. We perform an empirical analysis applying the new method to the producer price index of passenger cars in Japan. The result shows that, compared with the conventional standard estimation method, the new method brings the following three benefits; 1) a significant increase in the number of variables in the regression model, 2) an improvement in fit of the model to actual prices, and 3) the reduction of the overestimation in quality improvements due to the omitted variable bias. These points suggest that the proposed method is likely to improve the accuracy of the price index while enhancing the usefulness of the hedonic quality adjustment method. We expect that this method supports effective use of big data for price statistics through automatically building a good performance model by extracting all necessary information even with the large dataset.

Keywords:
Price index; Quality adjustment; Hedonic regression model; Sparse estimation; Adaptive Elastic Net

1. Introduction:

Given the price index indicates "pure" price changes of the product over time, it is essential to adjust the price difference attributable to quality differences between old and new products in response to the renewal of products in the market. The hedonic quality adjustment method is one of the quality adjustment methods for the price index. It extracts a quality change by using the regression model which estimates the relationship between characteristics and prices while assuming the quality of a product can be represented by the accumulation of individual characteristics.

The hedonic quality adjustment method has two main advantages; 1) it can objectively evaluate the quality changes of products based on data and statistical methods rather than on the subjective judgement, and 2) even if there are various changes in characteristics of products, it can comprehensively evaluate the effects of these changes on the product prices. Therefore, the hedonic approach has been applied in the compilation of the consumer price index (CPI) and the producer price index (PPI) in many countries.

However, there are some issues for applying the hedonic quality adjustment method in practice. First, if the characteristics of the products are highly correlated, the problem of multicollinearity on the explanatory variables is likely to arise, and it may cause the omitted variables bias through the variable selection based on the statistical significance.

Furthermore, it is known that the problems of multicollinearity and omitted variable bias can be more serious as the model has more complex functional form to deal with the non-linear effects of price determining characteristics.

In this paper, we attempt to overcome these problems by introducing the new estimation method employing "sparse estimation" in the estimation of the hedonic regression model.

2. Methodology:

Taking into account the non-linear relationship between the price and characteristics of a product, the Bank of Japan (BOJ) previously employed the following hedonic regression model with the Box-Cox transformed term[1], and it was estimated by using the ordinary least squares (OLS) method, in the compilation of the PPI and export/import price index.

$$y_i^{(\lambda_0)} = \beta_0 + \sum_{j=1}^{p_c} \beta_{cj} x_{cj,i}^{(\lambda_j)} + \sum_{k=1}^{p_d} \beta_{dk} x_{dk,i}\,, \tag{1}$$

where y_i: theoretical price, $x_{cj,i}$: continuous variable, $x_{dk,i}$: dummy variable,
β_0: constant term, β_{cj}: coefficient on a continuous variable,
β_{dk}: coefficient on a dummy variable,
λ_0: Box-Cox parameter for theoretical price,
λ_j: Box-Cox parameter for a continuous variable,
p_c: number of continuous variables, p_d: number of dummy variables.

However, there are some issues with this method in that the parameters are not stable due to multicollinearity and omitted variables, which can cause bias in the parameters when the explanatory variables (characteristics in the hedonic regression model) are highly correlated. In particular, it is known that the omitted variable bias becomes more severe on complex functional forms, and it poses a risk of generating downward bias in the price index because of an overestimation of the rate of quality improvement.

To deal with the aforementioned issues, we introduced the new estimation method with "adaptive elastic net: AEN", a type of sparse estimation proposed in Zou and Zhang (2009). Sparse estimation performs variable selection and coefficient estimation at the same time under the property called "sparsity". This method has an advantage over the previous one (equation (1)) in that it can automatically derive a more stable and fitted model. In addition, the AEN incorporates the L_1 norm (sum of absolute values) and the L_2 norm (sum of squares) of coefficients as regularization terms in the two-stage estimation of coefficients (see equations (3)-(5) below). Then it enjoys two desirable properties: the "Grouped Effect" that gives robustness for multicollinearity and the "Oracle Property" that ensures the adequacy of variable selection and coefficients (Zou, 2006).

Given these properties, the new estimation method selects variables and a functional form simultaneously by extracting variables from the quadratic multivariate regression model with interaction terms, shown as equation (2), in the AEN estimation. Note that this regression model is to incorporate interaction effects among characteristics of a product while maintaining the non-linear relationship between price and characteristic in the regression model.

$$Y_i \equiv \log y_i\,,$$

[1] The Box-Cox transformation of a variable x with the Box-Cox parameter (λ) is as follows (Box and Cox, 1964).

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log x & (\lambda = 0) \end{cases}$$

$$Y_i = \hat{\beta}_{00} + \sum_{j=1}^{p} \hat{\beta}_{0j} x_{j,i} + \sum_{j=1}^{p} \hat{\beta}_{jj} {x_{j,i}}^2 + \sum_{k>j\geq 1} \hat{\beta}_{jk} x_{j,i} x_{k,i} \,, \tag{2}$$

where

$$\widehat{\boldsymbol{\beta}} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(|\boldsymbol{Y} - \boldsymbol{X\beta}|^2 + \lambda_2 \sum_{k\geq j\geq 0} {\beta_{jk}}^2 + {\lambda_1}^* \sum_{k\geq j\geq 0} \widehat{w}_{jk}\left|\beta_{jk}\right|\right)\right\}, \tag{3}$$

$$\widehat{w}_{jk} = \left(\left|{\hat{\beta}_{jk}}^{1st}\right|\right)^{-\gamma}, \tag{4}$$

$$\widehat{\boldsymbol{\beta}}^{1st} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(|\boldsymbol{Y} - \boldsymbol{X\beta}|^2 + \lambda_2 \sum_{k\geq j\geq 0} {\beta_{jk}}^2 + \lambda_1 \sum_{k\geq j\geq 0} \left|\beta_{jk}\right|\right)\right\}, \tag{5}$$

y_i: theoretical price, $x_{j,i}$: explanatory variable, $\widehat{\beta}_{jk}$: coefficient on $x_{j,i}x_{k,i}$,
p: number of candidate explanatory variables, n: number of samples in dataset,
$\lambda_1 > 0$: L_1 norm regularization parameter (1st stage),
${\lambda_1}^* > 0$: L_1 norm regularization parameter (2nd stage),
$\lambda_2 > 0$: L_2 norm regularization parameter,
$\gamma > 0$: adaptive parameter, $\widehat{w}_{jk} > 0$: adaptive weight.

3. Result:

In this section, we apply new and previous hedonic regression models to passenger cars in Japan and compare those results.

Chart 1 shows the rate of change in theoretical price due to one standard deviation increase in continuous variables where a hypothetical data with all variables are set at the mean value over the sample period. It is clear that the number of explanatory variables in the regression models increases and this is accompanied by a reduction in dependence on just a few specific variables. For instance, regarding the driving performance of passenger cars, in addition to the maximum output, which is solely selected in the previous model, the new estimation method enables the incorporation of characteristics related to the acceleration performance into the model, such as number of gears and maximum torque.

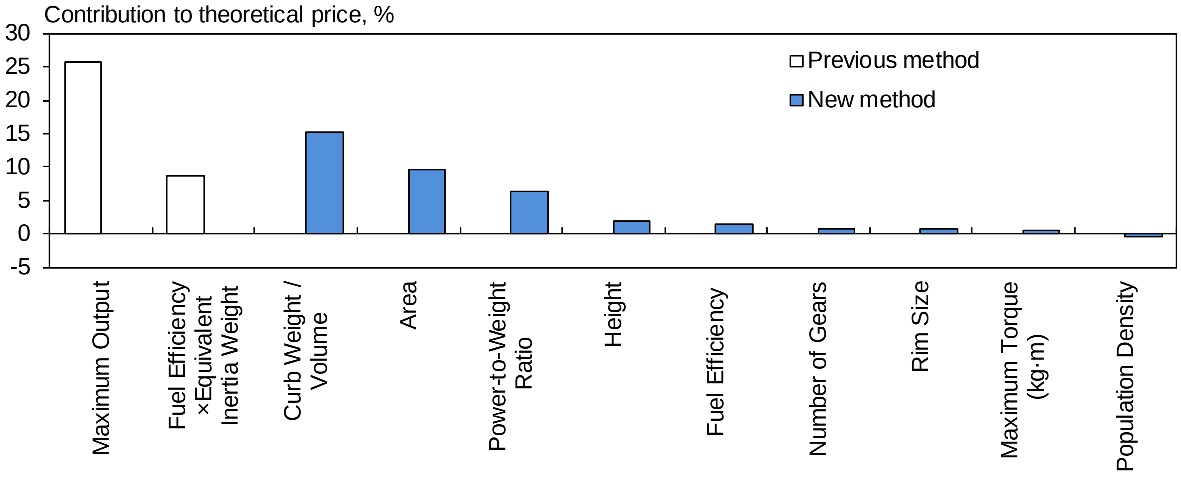


Chart 1: Estimated Effects of Characteristics on Price of Passenger Cars

As for the estimation accuracy of these models, we calculate the mean squared errors for both in-sample and out-of-sample period (Chart 2). We can find that the fit of regression models to actual price generally improves in the new estimation method for both in-sample and out-of-sample data. Since the quality adjustment is generally applied to products that is released after the estimation, the improvement in the out-of-sample fit implies an increase in the usefulness of the hedonic quality adjustment method in practice.

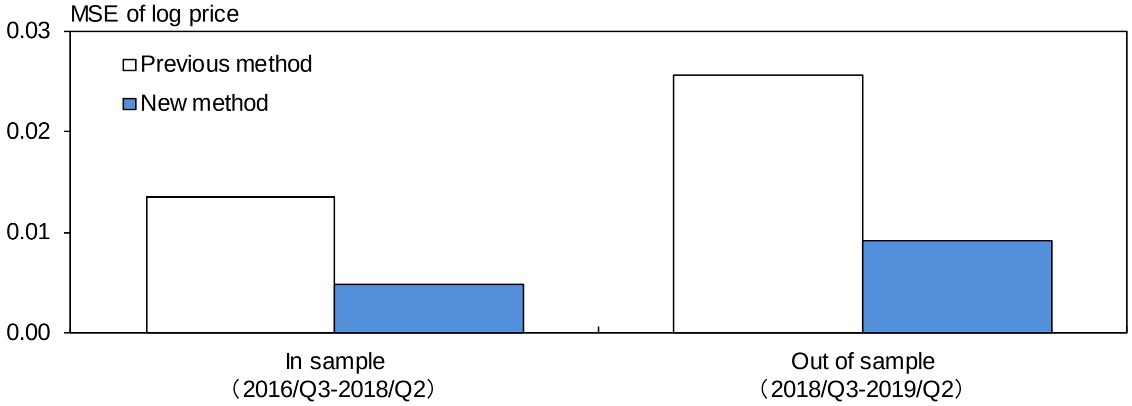


Chart 2: Fit of Hedonic Regression Models by Estimation Method

In fact, looking at the price index of "standard passenger cars (gasoline cars)" in the PPI, it can be seen that the estimated index, which is retrospectively calculated by applying the new hedonic estimation method to all quality adjustments, shows similar developments to the published price index[2] (Chart 3). On the other hand, the estimated index by the previous method highlights the risk of over-estimating the rate of quality improvement as it shows an excessive decline in the price. These observations suggest that an increase in the number of explanatory variables under the new method contributes to the accurate estimation of quality improvement rates in practice.

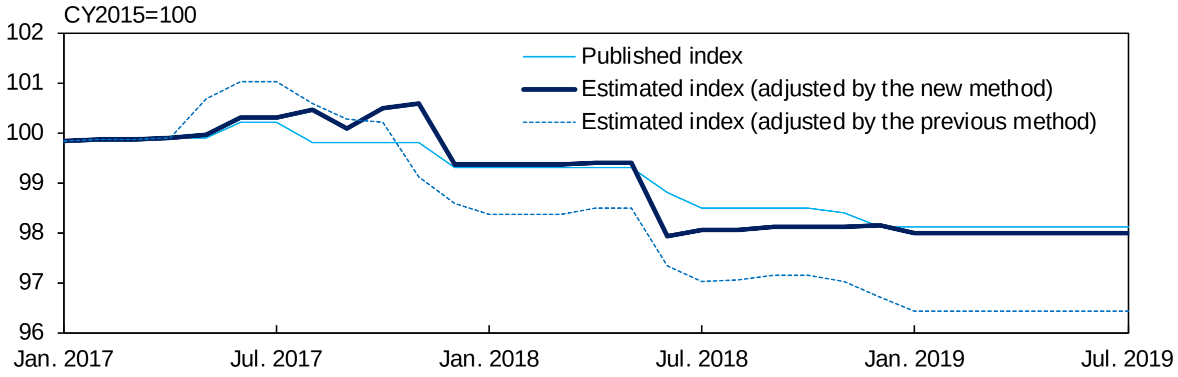


Chart 3: Estimated Price Index by New and Previous Methods

4. Discussion and Conclusion:

This paper introduces the new estimation method for the hedonic quality adjustment to overcome the problems due to multicollinearity and omitted variables. The AEN, which is employed in the new method, provides two desirable properties: the "Grouped Effect" that gives robustness for multicollinearity and the "Oracle Property" that ensures the adequacy of variable selection and asymptotic unbiasedness of coefficients.

[2] In the compilation of the PPI, the BOJ choose the most appropriate one among various quality adjustment methods including the hedonic quality adjustment method based mainly on review of an estimated quality improvement with a respondent firm. If the estimates by the hedonic quality adjustment method cannot pass this review, the BOJ applies other quality adjustment methods such as the production cost method.

The empirical analysis for passenger car prices in Japan suggests that the new method using the AEN potentially offers following benefits: 1) a significant increase in the number of variables incorporated in the model, 2) an improvement in fit especially for the out-of-sample period, and 3) less omitted variable bias which reduces the risk of over-estimation of the quality improvement rate. Therefore, the new method is expected to make the hedonic quality adjustment more accurate and more applicable for various sample replacements.

As mentioned above, since the hedonic method is based on data and statistical methods, it has strength in its objectiveness and applicability for a quality change accompanied by developments in a wide range of product's characteristics. The increased usability of the hedonic regression model will lead to the more accurate price index. Moreover, the proposed estimation method is a highly efficient as it can automatically build a good performance model by extracting all necessary information even with the large dataset. We expect that this method supports effective use of big data for price statistics.

References:

1. Box, G. E. P. & Cox, D. R. (1964). An Analysis of Transformations, *Journal of the Royal Statistics Society Series B*, **26**, 211-252.
2. Furuta, S., Hatayama, Y., Kawakami, A., & Oh, Y. (2021). New Hedonic Quality Adjustment Method using Sparse Estimation, Bank of Japan Working Paper Series, Forthcoming.
3. Triplett, J. E. (2006). *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*, OECD Publishing.
4. Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418-1429.
5. Zou, H. & Zhang, H. H. (2009). On the Adaptive Elastic-Net with a Diverging Number of Parameters, *The Annals of Statistics*, **37**(4), 1733-1751.