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

日本銀行 BANK OF JAPAN

✓ 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

日本銀行 BANK OF JAPAN

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

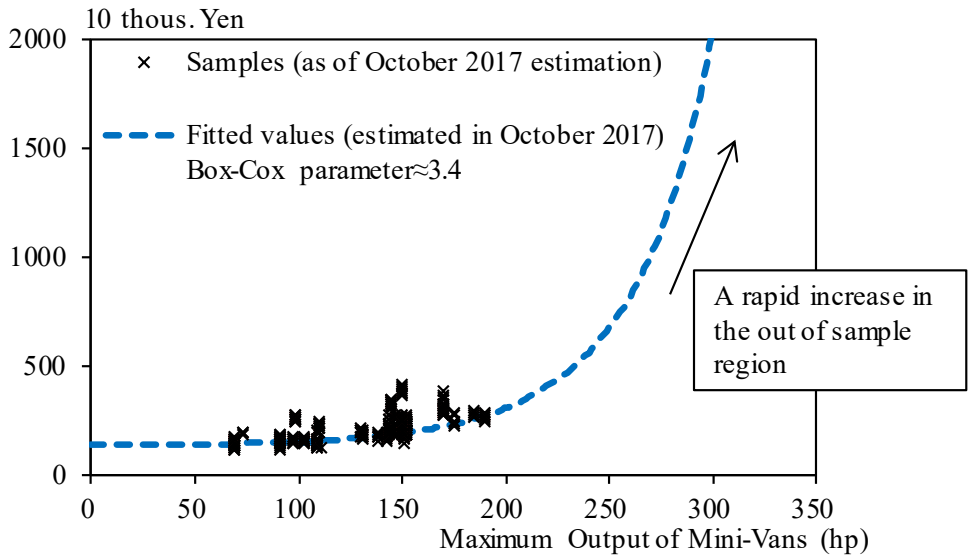
Repeating estimation while changing the set of the variables (excluding variable that cause multicollinearity and including the meaningful variables) to obtain good results.

Accuracy of estimation (1)

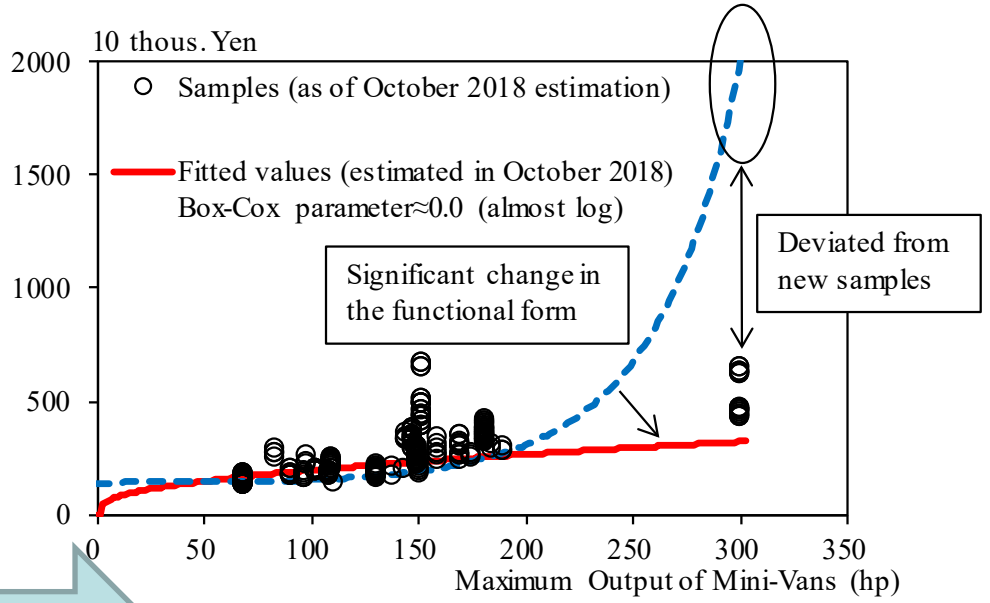
- ✓ Estimated parameters on variables may become unstable due to the problem of multicollinearity and the omitted variables bias.
- ✓ Multicollinearity refers to a state in which there is a high inter-connection among the variables. Multicollinearity makes it difficult to identify price effects of variables, 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).



Re-estimation



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 (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

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

s.t. $|\beta_1| + |\beta_2| \leq s$

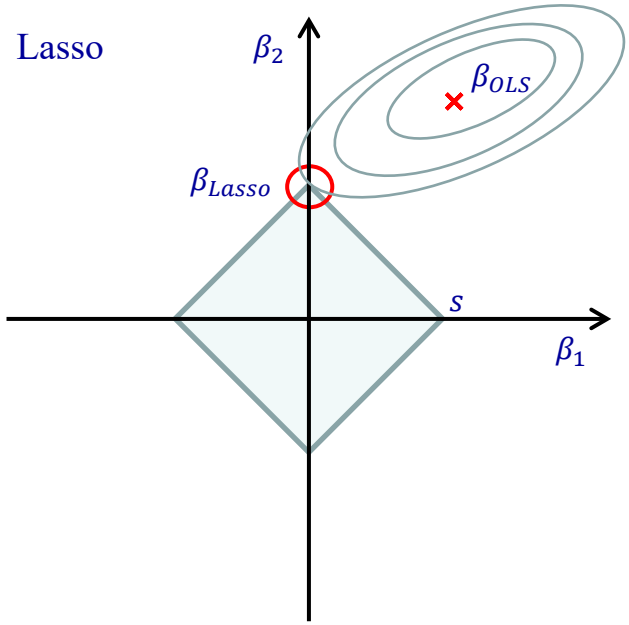
$s > 0$: 1−1 corresponding to λ

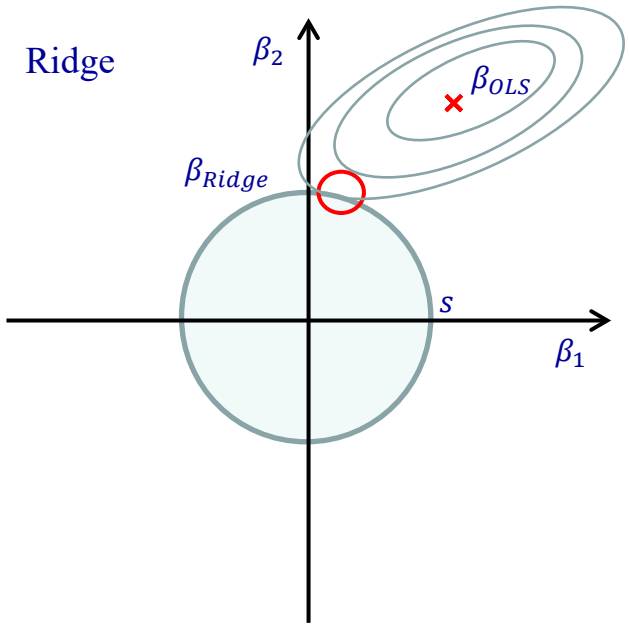
Ridge

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

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

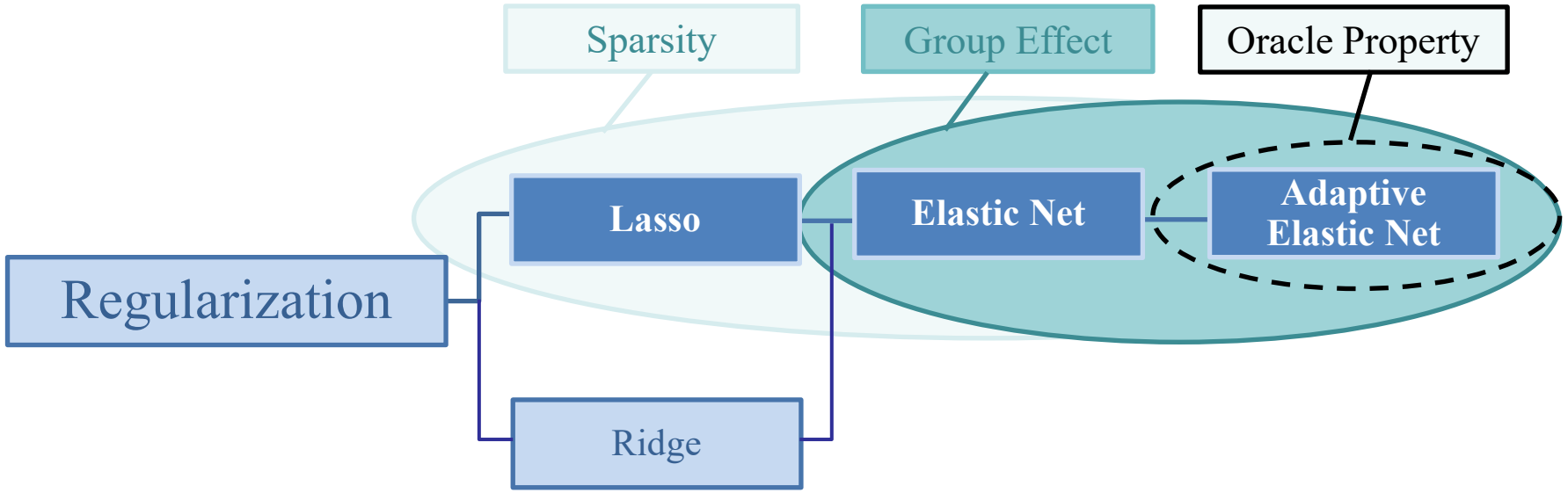
$s > 0$: 1-1 corresponding to λ





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.

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

$\lambda_2 > 0$: 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 $\hat{\boldsymbol{\beta}}$ 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.

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

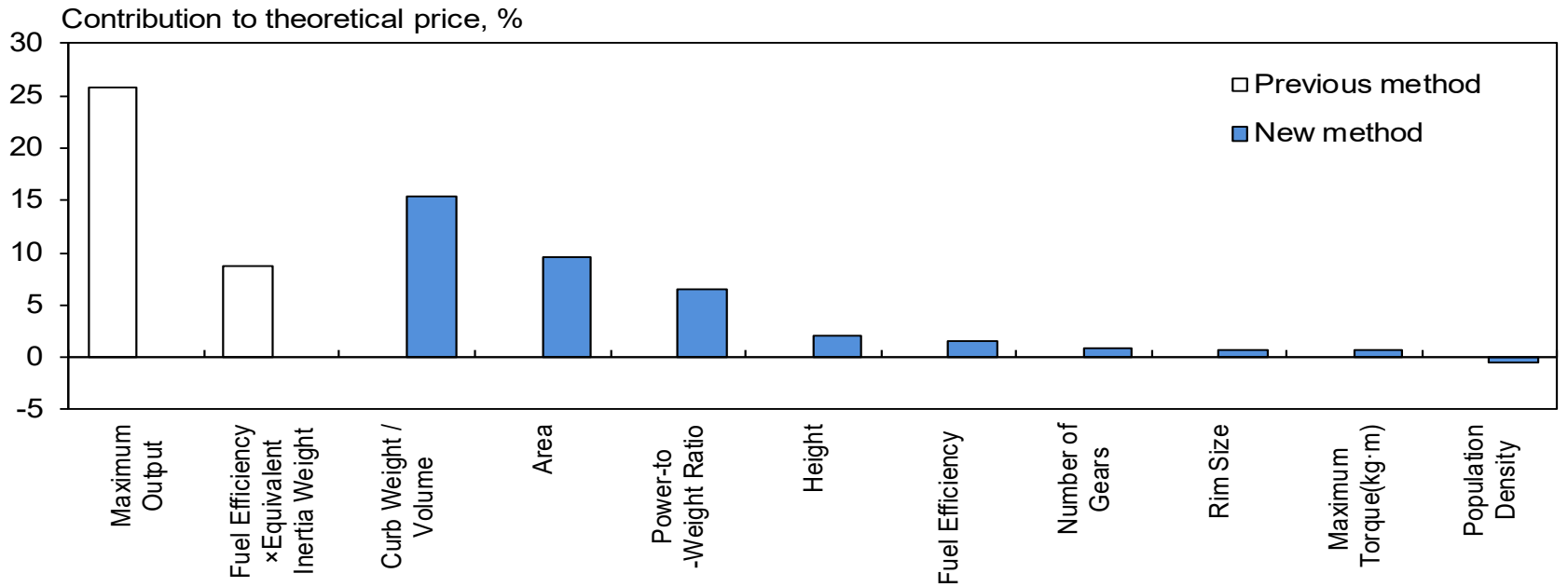
$$\widehat{w}_j = \left(\left|\hat{\beta}_j(EN)\right|\right)^{-\gamma}$$

${\lambda_1}^* > 0$: 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

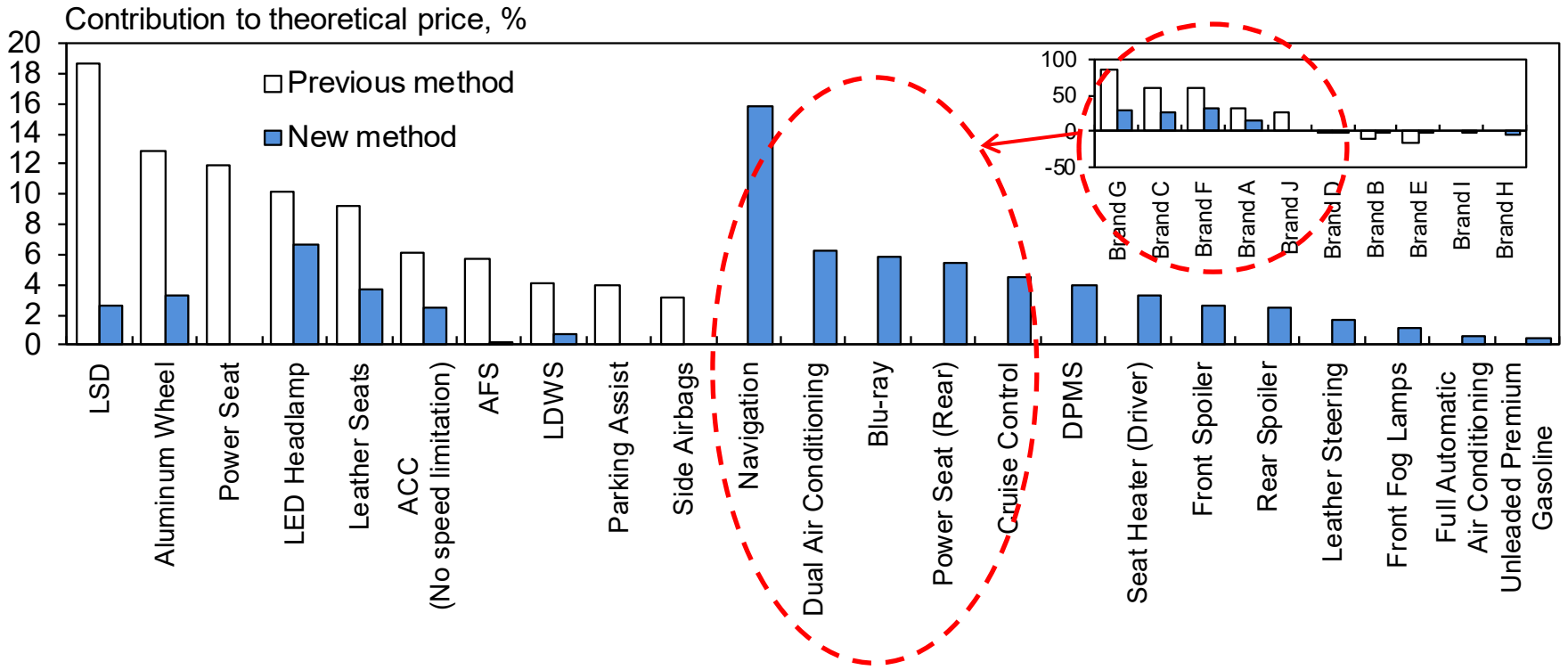
- ✓ We apply new and previous hedonic regression models to passenger cars in Japan and compare those results.
- ✓ The number of continuous variables in the regression models increases and this is accompanied by a reduction in dependence on just a few specific variables.



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

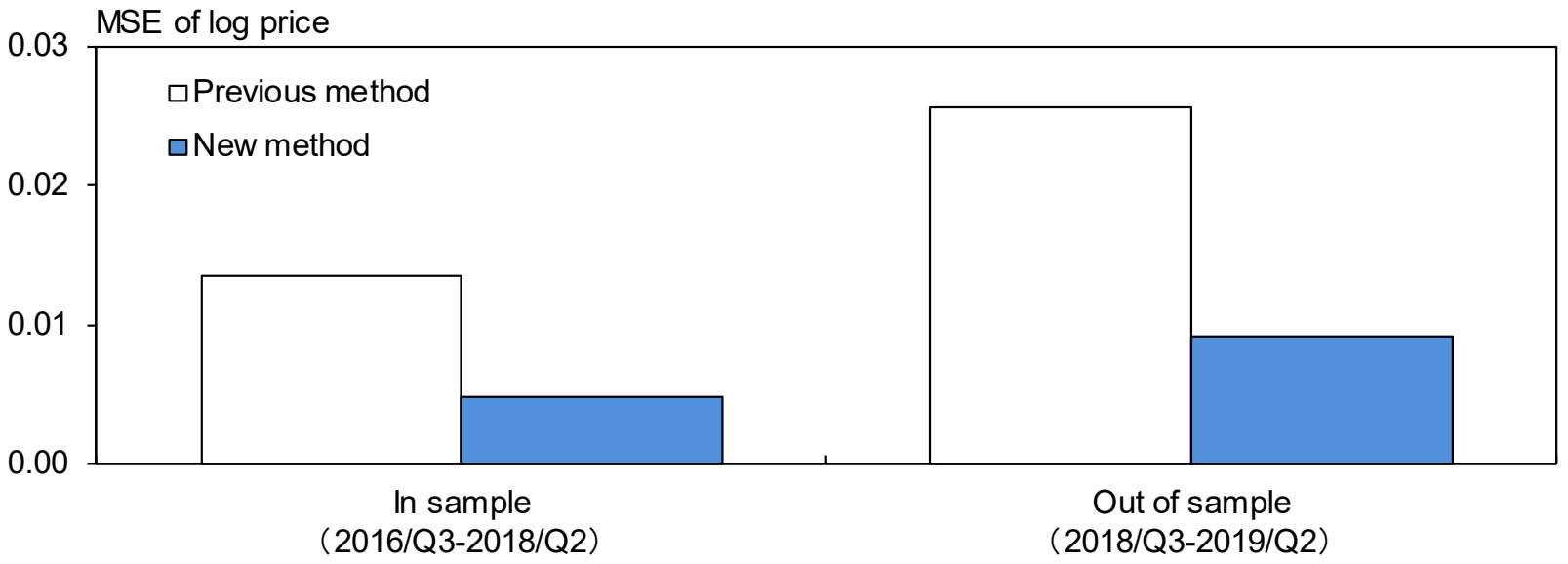
✓ As a result of the increased number of characteristics, the new regression model reduces its reliance on manufacturer dummies (control variables).



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.

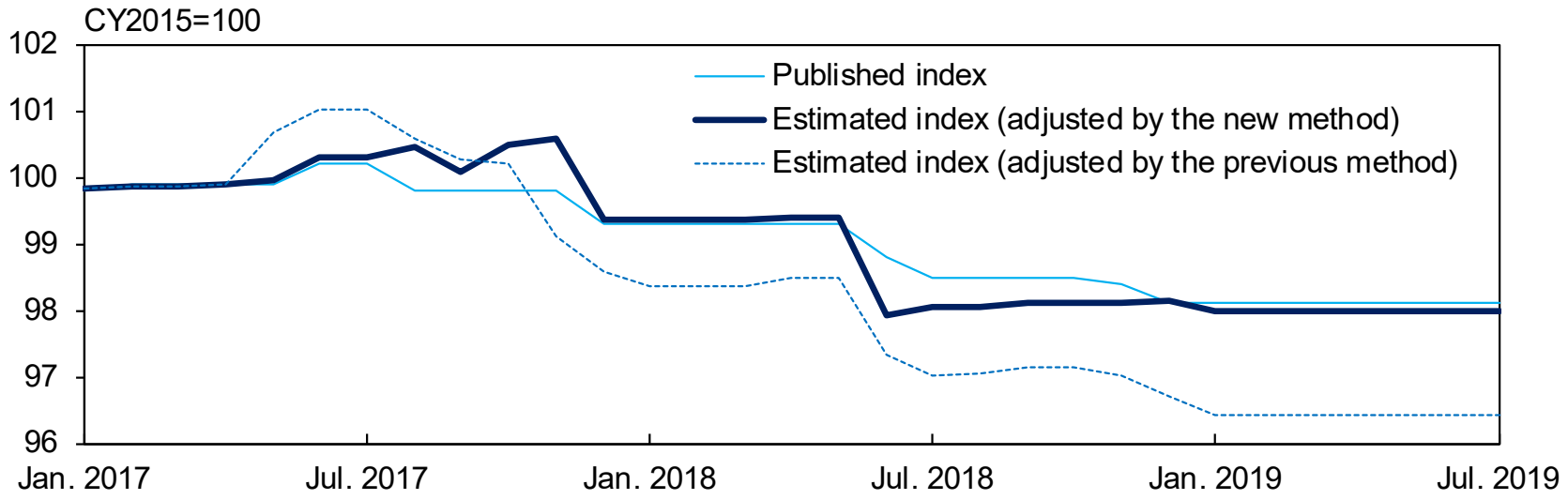
Fit of the model

- ✓ 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.



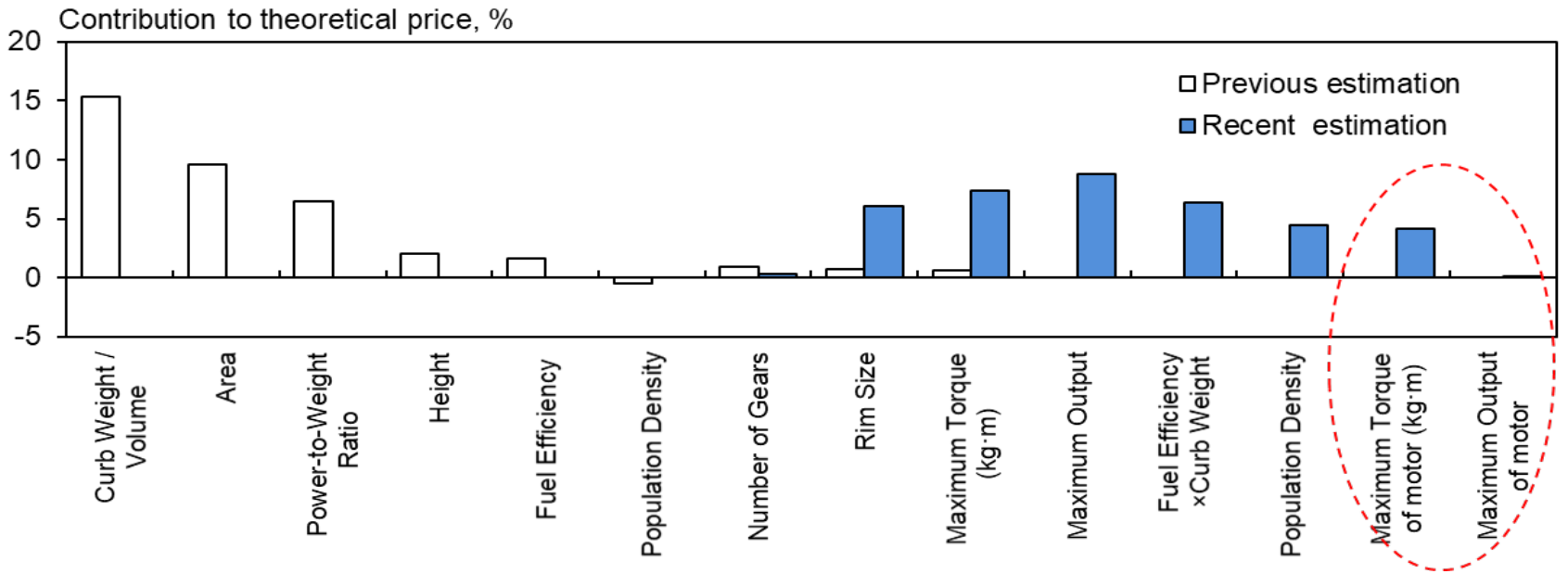
Estimated Price Index

- ✓ 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.



Recent Estimation Results

- ✓ In recent years, passenger car quality has changed greatly following the move to electric cars from gasoline cars in the market.
- ✓ The new method worked well and some features related to electric motors were adopted in recent estimation. It allowed for a more accurate evaluation of passenger car quality in this market situation.



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.

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 automatically build a good performance model by extracting all necessary information even with the large dataset (e.g. 1500 samples and 100 candidate variables in the dataset of passenger car), and it can be expected that this method supports effective use of big data for price statistics.