Centraal Bureau voor de Statistiek

New dwellings;

output price indices building costs, 2015=100

Method description

summary   *This is a description of the Output price index of building costs of new dwellings. It describes the method of calculation of the index.*

keywords   *Method, price index, building costs, newly build, dwellings*

Introduction

This document describes the methodology of the output price index (O-PINW) for newly built dwellings. It pays special attention to among other things the regression model and the outlier detection method.

For a better understanding of the background of the index, we start with some general remarks about the data on which the index is based. The subsequent section elaborates on the regression model and describes the variables and formulas.

1. General remarks

The O-PINW is based on information from municipal administrations (the number of building licences granted). Building licences are issued on a project basis, i.e. more than one dwelling may be constructed with one licence. The observed prices refer to complete projects. Prices per dwelling are calculated by dividing the project price by the number of dwellings in the project. The price index is calculated using a regression model, also known as a hedonic model. An average price of a basic dwelling in the project is estimated every quarter, on the basis of price determining variables. This estimated price is then compared with the estimated average price in the reference year (see section 3.4.2), resulting in a price index.

The regression model is based on five variables that partly determine the price of a dwelling, namely:
- Volume of the dwelling (in cubic metres),

- Number of dwellings in the project,

- Category of commissioning party,

- Type of soil, and

- Market, i.e. owner-occupied or rental.

2. Method description

After these general remarks, we shall give a more detailed explanation of the regression model and the variables used. We start with a step-by-step description of the calculation of the index, which serves as a guideline for the method description.

2.1 The calculation – step by step

To determine the average values of the variables in the base year[1] and the corresponding reference price in the reference year[2], the following steps are carried out once:

---

[1] The base year is the year for which the values of the variables of the (average) basic dwelling are determined, in this case 2016.

[2] The reference year is the year for which the price index (arithmetic) is set at 100, in this case 2015.

A1. The regression model is calculated unweighted to determine the residues. Based on these residues the outliers are determined.

A2. The outliers are removed from the dataset.

A3. Based on the number of dwellings named in a building permit the new dataset (without the outliers) is weighted and the average values of the variables are calculated for the base year.

A4. Again the regression model is calculated unweighted on the new dataset. This yields regression coefficients per variable.

A5. The regression coefficients calculated in step A4 are linked to the average values of the variables of the base year (step A3). The exponent of the sum of these pairs is the estimated price for a particular quarter of the reference year.

A6. The arithmetic mean of the four estimated quarter prices of the reference year is the reference price.

To calculate a price index, the next steps are carried out quarterly (steps B1 and B2 are the same as steps A1 and A2, and steps B3 and B4 are the same as A4 and A5):

B1. The regression model is calculated unweighted to determine the residues. Based on these residues the outliers are determined.

B2. The outliers are removed from the dataset.

B3. Again the regression model is calculated unweighted on the new dataset (without the outliers). This yields regression coefficients per variable.

B4. The regression coefficients calculated in step B3 are linked to the average values of the variables of the base year (step A3). The exponent of the sum of these pairs is the estimated price for a particular quarter of the reference year.

B5. The estimated price divided by the estimated price of the reference year (i.e. the reference price of step A6) and multiplied by 100 gives the price index figure.

2.2 The variables

The structure of the variables described in section 2 and how these are used in de regression model are described in this subsection. The dependent variable in the regression model is the logarithm of the building costs per dwelling.

The model contains five independent variables. The first variable is the volume of the dwelling. This is the result in m^3 of the total volume of the project divided by the number of dwellings in the project. Just as for the dependent variable, for this variable too the logarithm is used in the model. The number of dwellings in a project is the second independent variable. This variable is also converted into logarithms. The other three variables are changed into dummy variables. In other words, category of commissioning party, type of soil, and market sector of the dwelling (either owner-occupied or rental) are broken down into a number of variables, analogous to the categories of the variables. The variable 'category of commissioning party' is divided into three categories, namely: DO1 (government and housing corporations), DO2 (commercial developers

and corporate investors) and DO3 (private persons and others). The variable 'type of soil' is not available in the dataset of the building permits. To determine this, the most common soil type per municipality is established. This information is linked to the building permit through the municipality code. The variable is split into five categories: DG1 (sand, Wadden Sea, hills, and dunes), DG2 (fenland), DG3 (river areas), DG4 (sea clay soil) and DG5 (tidal areas, Zeeland and enclosed see inlets). For the market sector of the dwelling (H = rental sector, E = owner-occupied) only one category is taken into account, because if one is not valid the other automatically is. For the three latter variables, which are split into dummies, one dummy is not included as it serves as a reference. We have chosen to use the most frequent category as the reference. As a result, the regression model contains the variables DO1, DO2, DG2 to DG5 and H/E.

Lastly, every quarter a constant is also added to the regression model. This constant is the coefficient of the quarter dummy which has value 1 for every record of the current quarter and value 0 for the records of the other quarters. Because the regression model contains a constant per quarter it is not necessary (and impossible) to include a general constant for all records of a year. Therefore, in the SPSS syntax the regression model is specified as a model without a constant.

2.3 Outlier detection method (step A1, A2, B1 and B2)

Before the estimation of the regression coefficients, the outliers are deleted from the dataset. The method used to do this is based on the residues of the regression model. The projects with the largest residues are defined as outliers. These include 5 percent of all projects of which 2.5 percent with the largest positive residues and 2.5 percent with the largest negative residues. To determine the residues the regression model is applied to the complete dataset. After deleting the outliers the model is estimated again (see section 3.5).

2.4 The reference year and base year

Because of the base shift, the average values of the variables in the base year (2016) and the reference price for the reference year (2015) are determined.

2.4.1 Basic dwelling in base year (step A3)

The average values of the variables are calculated from the weighted dataset of the base year without the outliers. The dataset is weighted by the number of dwellings known for a building permit. The average values refer to the whole base year and together form the average basic dwelling. Note that these are the average values of the regression variables, so the mean of the logarithm of the volume, the logarithm of the number of dwellings, and the dummy variables. The quarterly estimated regression coefficients are linked to those of the basic dwelling. The mean of the quarter dummy is 1 for the current quarter and 0 for the other quarters.

2.4.2 Reference price in the reference year (steps A4 to A6)

The reference price is the arithmetic mean of the estimated prices of the four quarters of the reference year. This means that the calculation in reference year is carried out completely until an estimated price for every quarter is available. The mean of these four quarters is the reference price. In formula:

(1) $$\hat{\bar{P}}_r = \left(\hat{P}_{1r} + \hat{P}_{2r} + \hat{P}_{3r} + \hat{P}_{4r} \right)/4$$

where

$\hat{\bar{P}}_r$ is the reference price, i.e. the mean estimated price of the reference year

$\hat{P}_{kr}$ is the estimated price per quarter in the reference year

For the explanation of the calculation of $\hat{P}_{kr}$ see section 3.5, as this calculation is the same as the calculation of the estimated price in any quarter of any year $\hat{P}_{kj}$.

2.5 The calculation (steps B3 to B5)

Once the reference price of the basic dwelling has been estimated (steps A1 to A6) and the outliers have been deleted (steps B1 and B2), the regression model can be calculated every quarter. In brief this means that the regression coefficients of the explanatory variables are determined (formula 2). These regression coefficients are then linked to the mean values of the base year (formula 3). The exponent of the summed coefficient is the estimated building costs of the current quarter. Lastly, the estimated building costs are compared with the reference price in the reference year, resulting in an index (formula 4).

Translating this into formulas:

$$(2)\ \ln(P_{kj}) = c_{kj} + \beta_{1kj}\ln(x_{1kj}) + \beta_{2kj}\ln(x_{2kj}) + \beta_{3kj}x_{3kj} + \beta_{4kj}x_{4kj} + \beta_{5kj}x_{5kj} + \beta_{6kj}x_{6kj} + \beta_{7kj}x_{7kj} + \beta_{8kj}x_{8kj} + \beta_{9kj}x_{9kj} + \varepsilon_{kj}$$

where

P_{kj} is the building costs per dwelling of quarter k of year j

c_{kj} is the coefficient of the constant

β_{xkj} is the regression coefficient of variable x for quarter k of year j

x_1 is the variable Volume

x_2 is the variable Number of dwellings

x_3 x_4 x_5 and x_6 are dummy variables of categories 2, 3, 4 and 5 of the variable Type of soil

x_7 en x_8 are the dummy variables of categories 1 and 2 of the variable Category of commissioning party

x_9 is the dummy variable of the category H/E of the variable Market sector of dwelling

ε_{kj} is disturbance.

The regression model gives unweighted estimates for the coefficients of the variables per quarter for one year at the same time. This means that every quarter the regression coefficients are calculated separately on the basis of the building permits issued in a certain quarter[3].

The regression coefficients are linked to the mean values of the variables of the base year. Because the logarithm of the building costs is calculated, the exponent of the summed coefficients is multiplied by the mean values of the basic dwelling.

---

[3] There is a distinction between year and month under *review* and year and month of *calculation*. The numbers of building permits issued are regular reported to Statistics Netherlands by the municipalities in a different month than the month in which they were issued. Therefore, since 2001, the building permits are backdated, i.e. they are used in the calculation for the quarter they apply to instead of the quarter they were reported.

$$\text{(3) } \hat{P}_{kj} = \exp\left(\begin{array}{l} c_{kj} * \bar{c}_b + \hat{\beta}_{1kj} * \bar{x}_{1b} + \hat{\beta}_{2kj} * \bar{x}_{2b} + \hat{\beta}_{3kj} * \bar{x}_{3b} + \hat{\beta}_{4kj} * \bar{x}_{4b} + \\ \hat{\beta}_{5kj} * \bar{x}_{5b} + \hat{\beta}_{6kj} * \bar{x}_{6b} + \hat{\beta}_{7kj} * \bar{x}_{7b} + \hat{\beta}_{8kj} * \bar{x}_{8b} + \hat{\beta}_{9kj} * \bar{x}_{9b} \end{array} \right)$$

where

$\hat{P}_{kj}$ is the estimated price for quarter k of year j

$\bar{x}_{xb}$ is the weighted mean value of variable x in base year b

$\hat{\beta}_{xkj}$ are the regression coefficients per variable.

To calculate a price index per quarter, the estimated prices per quarter $\hat{P}_{kj}$ are divided by the reference price $\bar{P}_r$. Multiplied by 100, this gives a price index figure.

$$\text{(4) } I_{kj} = 100 * \left(\hat{P}_{kj} / \hat{\bar{P}}_r \right)$$

where

$\hat{\bar{P}}_r$ is the reference price, i.e. the mean estimated price for the reference year

I_{kj} is the price index figure for quarter k of year j.