HOX Product Description

General

This report has been updated 2011-03-16

Type of index: Transactions-based price index for single family houses and for housing cooperative apartments (condominiums) in Sweden

Name of Index: Nasdaq OMX Valueguard-KTH Housing Index™

Metropolitan areas: Stockholm, Gothenburg and Malmö

Composite indexes: Medium Cities

Period: 2005-

Base period: January 2005

Base period index number: 100

Update period: The index is calculated monthly

Type of homes: Single family houses and cooperative apartments

Publisher: Valueguard Index Sweden AB

Index developers: Professor Mats Wilhelmsson (Center for Banking and Finance, KTH), Dr. Han-Suck Song (Center for Banking and Finance, Department of Real Estate and Construction Management, KTH), Jacob Winstrand, Lars-Erik Ericson, Valueguard Index Sweden AB.

Patent: U.S. Provisional Patent Application Atty Docket No. 0245-010

Transactions data supplier: Mäklarstatistik AB

Complementary data supplier: Lantmäteriet

Market coverage: About 60 percent of all sales. The market coverage is higher in large cities e.g. it is about 70 percent in Stockholm).

Contract date: Unique to this database is that each observation contains information on both the contract date and date of transfer of the deed. As the difference is normally 1-3 months it is important that the correct date of sale is used. Not using the contract date means that the index will be lagged in time and thus obsolete.

Regression Construction

Background

Essentially three different types of transaction-based methods can be used in the study of apartment price fluctuations. The first is a simple average price index based on arithmetic mean or median sales prices, the second is the so-called repeated-sales price index, and the third is the hedonic price index.

In addition to indexes based on transactions, there also exist appraisal-based indexes, and indexes that combine transaction-based and appraisal-based methods.

Construction of index

The Nasdaq OMX Valueguard-KTH Housing Index™ is a hedonic index. This means that the index controls for variations over time in a large number of home and local area characteristics: the index is a constant-quality index.

The hedonic regression model

By using the data collected, we construct a cross-sectional and time-series hedonic equation of cooperative apartment prices. A hedonic equation is a regression of prices against attributes that determine these prices and time. The interpretation of regression coefficients is as estimates of the implicit (hedonic) prices of these attributes, hence, the willingness-to-pay for the attribute in question. The hedonic price equation is expressed as

$$Y_{i,t} = \beta_0 + X_{i,t}\beta_1 + TD_t\beta_{2,t} + \varepsilon_{i,t} \qquad i = 1,\ldots,N \text{ and } t = 1,\ldots,T$$

where Y is the dependent variable, that is the (natural logarithm) of transaction prices, β_1 is a vector of all parameters (regression coefficients) associated with matrix X, the matrix of observations on the explanatory variables. We decompose X into, for example, structural apartment and property attributes, as well as neighborhood attributes.

The matrix TD with subscript t includes a time dummy vector for each time period. N is equal to the number of observations and T is equal to the number of periods. The stochastic term ε is assumed to have a constant variance and normal distribution.

Explanatory variables: The hedonic regression model includes a large number of explanatory variables: living area (square meters), number of rooms, monthly fee, floor, number of floors in the multi-family property, elevator and balcony. For the single family house variables are used such as plot area, type of building and the assessed standard of the house (for taxation purposes). Furthermore, the data consists of the geographical location based on the x and y coordinates as well as data on administrative parish.

From these variables, a number of new variables are generated: first floor, top floor, building year (see below), different type of geographical variables (see below for examples).

The age of the property is a proxy for the quality of the property and the apartment, but instead of using a continuous variable, we have constructed seven different dummy variables. Our hypothesis is that relatively new and very old apartments have the highest prices, while the apartments built during the "Million programme" have the prices lowest.

Dummy variable	Year of construction	Explanation
Byear1	Before 1900	
Byear2	1900-1939	Before Second War II
Byear3	1940-1959	Post-war period
Byear4	1960-1975	The "Million programme" era
Byear5	1976-1990	Period with high construction subsidies
Byear6 (Base variable in regressions below)	After 1990	Abolishment of the subsidy system
New	If year of construction = year of sale	

Access to information concerning neighborhood characteristics is scarce. By estimating the distance to the centre of the city, a price gradient can be estimated. Naturally, distance is supposed to have a negative effect on price. We analyze whether the CBD accessibility affects the price differently in different geographical directions. Thus, besides distance, we also divide the city into four quadrants (northwest, northeast, southwest and southeast). Three of these have been included in the model. Moreover, we also test if the price gradient is different in different geographical directions. In order to reduce spatial dependency further, longitude and latitude coordinate attributes are included in the model. We also use a dummy variable for the inner city. Some of the spatial variables are interacted with the apartment attributes.

Furthermore, the hedonic model includes a large number of dummy variables concerning sub-markets. The submarkets are defined as the administrative parish. Parish and distance variables are included in order to reduce omitted variables bias and to mitigate spatial dependence.

Time dummies representing sales during a specific month are included in the hedonic price equation. We use contract dates instead of the transfer dates.

Handling outliers and leverage

The presence of measurement errors, leverage and outliers can be a serious problem in the estimation of real estate price indexes. In order o detect and mitigate the problem, a three-step testing procedure has been utilized. The first two steps is implemented in the statistical program Stata (Robust regression) and described in their technical handbook.

In the first step, the statistical measure Cook's distance is estimated in order to test whether an individual observation have a substantial effect on the predicted price (expected value). It is measured as the absolute value of the difference in expected value with and without an individual observation included in the estimation. Observations with a Cook's distance larger than the critical value will be dropped from the estimation. The critical value has been decided by a grid search maximizing the adjusted coefficient of determination.

In the next step, the absolute errors (observed price minus predicted price) has been used in order to down weight observations with large errors. The regression parameters have been estimated by using two different iteration processes (Huber and biweighting). The method can be seen as a WLS (weighted least square) method where observation with a high leverage and outliers is handled. If the errors are not normally distributed, WLS is more efficient than OLS.

The third step is an evaluation process in which the preferred estimation process (in this case handling leverage and outliers) is determined. In the testing procedure, a traditional OLS (ordinary least square) model including all observations is compared to an OLS model excluding observation with a high leverage. The (naïve) exclusion is based on the 1st and the 99th percentile on each independent variable and observation with a lower value than the 1st percentile or higher than the 99th percentile is excluded. Moreover, the two OLS models are compared to the WLS model, described above, using an out-of-sample prediction test. The out-of-sample test uses the first 80 percent of the observations in order to predict the price of the last 20 percent. The sample is a random sample with equal probability. This procedure has been carried out 10 times; hence, a new sample has been estimated 10 times and all parameters have been estimated and a price predicted. The model with the lowest RMSE (root mean square error) has been chosen as estimation method.

Functional form: We have used a log-linear specification indicated by the Box-Cox tests. That is, all continuous variables are transformed to natural logarithms. The dependent variable *Y* is equal to ln(*price*) and *X* is equal to ln(*independent variables*). Test procedures have been defined as follows (Box-Cox):

$$\frac{P_i^{\lambda}-1}{\lambda} = \beta_0 + \sum_1^k \beta_k \left(\frac{Z_i^{\gamma}-1}{\gamma} \right) + \varepsilon_i$$

if	$\lambda=\gamma=0$	$\Rightarrow$	linear
	$\lambda=\gamma=1$	$\Rightarrow$	log-linear
	$\lambda=1$ och $\gamma=0$	$\Rightarrow$	semi log-linear

λ=0 och γ=1 $\Rightarrow$ inverse semi log-linear

Multicollinearity: To investigate the existence of multicollinearity, we compute variance inflationary factors (VIF) for individual coefficients. The principle that VIF larger than 10 suggests multicollinearity is used. The reason for why the distance from centre variable has a high VIF is that the submarket dummies and the coordinates are included in the model specification. The other explanatory variables do not seem to suffer from multicollinearity problems.

Interpretation of regression results: We estimate hedonic price equations in which the dependent variable price appears in logarithmic form, with a number of monthly time dummy variables as independent variables. Therefore the parameter estimates concerning the dummy variables have a percentage interpretation: when multiplied by 100, the estimated parameter on a dummy variable is interpreted as the approximate percentage change in the dependent variable. In a hedonic price index application, the exponential of the estimated time dummy parameters might therefore be interpreted as the approximate rate of growth in the mean price with respect to the mean price at the beginning.

When the estimated dummy parameters indicates a large change in the dependent variable, the exact percentage difference should be computed as

100*[exp(b) – 1].

An even more exact formula is given by

p = [exp(b – 1/2V(b)) – 1]*100

where V(b) is an estimate of the variance of b. When computing the index, this latter formula is used.

Goodness-of-fit: Goodness-of-fit (explanation power) measures how much of the variation in price that can be explained by the variation in the independent variables. The model for condominium prices in the city of Stockholm during the period 2005-2009 shows an explanation power of around 85-90%. This may be regarded as a very high for this type of models. Definition of the explanation degree is defined in accordance with:

R^2=SSE/SST

where SSE is equal to sum of square residuals and SST is equals to the total variation in Y.

$$\sum \left(y_i - \bar{y} \right)^2 = \mathrm{SST}$$

$$\sum \left(\hat{y}_i - \bar{y} \right)^2 = \mathrm{SSE}$$

Variance and standard deviation: The variance and standard deviation in individual estimates is equal to *Var*(b) and *se*(b).

Test whether price changes are statistically significant: Coefficient estimates of the time variables (the time dummy variables) represent the estimate of price change.

We perform hypothesis testing to test whether the point estimate at time t is significantly different from the point estimate at the previous time period *t*-1. Hypothesis testing is performed on monthly, quarterly and annual changes. We perform hypothesis testing (one-sided) at 5% significance level to determine whether the price change is significantly different from 0.

The test statistic is:

$$\frac{b_t - b_{t-1}}{Var(b_t - b_{t-1})}$$

where

$$\sqrt{Var(b_t - b_{t-1})} = \sqrt{Var(b_t) + Var(b_{t-1}) - 2Cov(b_t, b_{t-1})}.$$

Omitted variable bias

If an unobservable characteristic increases over time, it can be mistakenly taken as a price appreciation when it is actually the result of the change in characteristics. If the omitted variable is positively correlated with the time dummies and with price, our estimated parameters concerning time will be upwardly biased. If the omitted variable is negatively related to price, the bias will be negative

Spatial dependence

We have utilized Moran's *I* in order to test for spatial correlation. Spatial dependency bears some resemblance to temporal dependency. A general spatial model incorporates a spatial structure into both the dependent variable and the error term.

$$Y_{i,t} = \beta_0 + \rho W Y_{i,t} + X_{i,t}\beta_1 + TD_t\beta_{2,t} + \varepsilon_{i,t}$$
$$\varepsilon_{i,t} = \lambda W Y_{i,t} + \eta_{i,t}$$

The parameter ρ is the coefficient of the spatially lagged dependent variable and the parameter λ is the coefficient of the spatial autoregressive structure of the error. *W* is equal to the spatial weight matrix. The special structure is a spatial weight matrix that is defined by how much a nearby (in space) observation should influence the averaging procedure. The spatial weight matrix is usually defined by inverse square distance between observations or by the nearest neighbors. We will use

the latter definition. If parameter ρ is equal to zero, a spatial error model (SEM) is estimated. If λ is equal to zero, a spatial autoregressive (SAR) model is estimated.

Composite index construction

A number of different methods can be used to develop a composite index from a number of regional index series. In order to calculate a composite index, the included regional indexes need to be weighted. The weights are based on both the stock and values of apartments in the different regions. The used housing stocks are figures from Statistic Sweden (SCB) with one year lag.

Not all regions are included in the composite index. Instead the index for condominiums include the 20 largest regions. The included regional indexes in the composite index are based on number of the sales in the region. The index for single family houses include the labour market regions of the largest cities and of larger regional centras in Sweden, as defined by the Swedish Agency for Economic and Regional Growth.

Monte-Carlo simulations have been utilized in order to evaluate different weighting methods. The used weighting procedure is similar to the procedure used for constructing value-weighted stock market indexes, for example, the OMX Stockholm/30.

The used value-weighting procedure can be described as in the equation below. Assume that we have 20 different regions. The weight (V) concerning region 1 at time t is defined as:

$$V_{1,t} = \frac{(\bar{P}_{1,t-12} S_{1,t-12})}{\sum_{i=1}^{20} \bar{P}_{i,t-12} S_{i,t-12}}$$

where S is equal to housing stock and P is equal to housing price. The housing prices are obtained from the transactions data. The average price in region 1 multiplied with the stock in region 1 is divided by the total value of the stock in all 20 regions. If the total value rise in one region, because housing stock and/or housing prices increases, its weight will increase, everything else equal, and, thereby, the region will be more important in the overall composite index. In the equation above, the housing stock and price is lagged twelve months. However due to data availability the actual lag can be six to 18 months. When the weights are estimated, the composite index is estimated as:

$$Composite\ index_t = \sum_{i=1}^{20} V_{i,t} * Index_{i,t}$$

where subscript i represents the 20 largest regions in Sweden based on number of sales, and t is period (month).

Annual revision

The weights are adjusted in July each year when the index-value for June is released. The average prices are calculated for the period July-June, in order to match the data concerning housing stock. At the same time the choice of regions used for calculating the index can be changed, based on number of the sales in each region during the weighting-period. With the new weights a different index-series will be generated, which will be adjusted to the most recent published value using a divisor:

$$Adjusted\ index\ value\ t_n = Generated\ index\ value\ t_n \Big/ \frac{\text{New index value}(t_0)}{\text{Old index value }(t_0)},$$

where (t0) marks the time for the last published index value before the change of weights.

Appendix 1 Explanatory Variables (examples)

Variable	Definition
Price	SEK
Living area	Square meters
Rooms	Number of rooms
Fee	Monthly fee: SEK*
Balc	Dummy: Balcony*
First	Dummy: First floor*
Top	Dummy: Top floor*
Plot area	Plot area**
Assement value	Assessed value for taxation purpose*
Byear1	Dummy: Apartment produced before 1900
Byear2	Dummy: Apartment produced during 1900-1939
Byear3	Dummy: Apartment produced during 1940-1959
Byear4	Dummy: Apartment produced during 1960-1975
Byear5	Dummy: Apartment produced during 1976-1990
Byear6	Dummy: Apartment produced during After 1990
New	Dummy: New building
Elev	Dummy: Elevator*
Distance	Meters to city
NE	Dummy: Northeast
NW	Dummy: Northwest
SW	Dummy: Southwest

* Condominiums only

** Single family houses only

Besides the variables above, several interaction-variables are included in the model specification.