The model for housing unit price estimation in Moscow

Foreword

The current practice offers two base approaches to estimation of residential property value: expert opinions and mathematical models (based on the regression analysis and the neural networks).

Expert opinions are most common, as they give account for recent trends in the market behavior and other price factors difficult for formalized estimation. However, this method is rather cost intensive and inclined to produce discrepant results because of the biased nature of expert opinions.

Typically, when there is a need to rapidly evaluate a rather larger number of property units the preference is given to statistical methods. The goal of this method is to identify and evaluate in quantitative terms a group of factors capable to affect the property prices. The quality of the produced model is manifested in its capacity to approach expert opinions about the property value.

This research has been undertaken with an objective for estimation residential property of each household included into the database of the household survey conducted by the Moscow City Statistics Committee (Mosstat) in 2002. To reach this goal we build up a hedonic regression model for estimation of Moscow housing prices using the real data on housing for sale in March, 2003. This dataset was provided by Moscow realtors.

Overview of related researches

Main efforts on building up regression models for residential property appraisal are taken by real estate companies, which are concerned with production of preliminary estimates of housing unit prices for their clients. Unluckily, this information is only for the internal use, which makes impossible the fair assessment of the quality of these researches.

For a rather long time students of the New Economic School guided by P. Katyshev and A. Peresetsky have been working on the residential property appraisal model. Findings of this research are easily accessible.

However, most of researches use statistical methods of appraisal under which the investigation of the housing market is tied up to a fixed time period, which makes impossible the production of accurate estimates in cases when trends (prevailing factors) are changing.

Description of variables

There are literally hundreds of potential housing characteristics that could be included on the right hand side. While theory is not much of a guide, experience from many studies suggests that, whatever the purpose, a full dataset would include the following[1]:

- Rooms, in the aggregate, and by type (bedrooms, bathrooms, etc.)
- Floor area of the unit

[1] Malpezzi, Stepeh. 2002. *Hedonic Pricing models: A Selective and Applied Review.* The Center for Urban Land Economics Research.

- Structure type (single family, attached or detached, if multifamily the number of units in the structure, number of floors)
- Type of heating and cooling systems
- Age of the unit
- Other structural features, such as presence of basements, fireplaces, garages, etc.
- Major categories of structural materials, and quality of finish
- Neighborhood variables, perhaps an overall neighborhood rating, quality of schools, socioeconomic characteristics of the neighborhood
- Distance to the central business district, and perhaps to sub-centers of employment; access to shopping, schools and other important amenities.
- Among characteristics of the tenant that affect prices: length of tenure (especially for renters), whether utilities are included in rent; and possibly racial or ethnic characteristics (if these are hypothesized to affect the price per unit of housing services faced by the occupant)
- Date of data collection (especially if the data are collected over a period of months or years).

The regression model for estimating Moscow housing unit prices is based on the supply statistics of the Moscow secondary housing market dated by March 2003 and provided by A. Sapozhnikov, Russian multi-listing system (RMLS). This database contains more than 19 thousand offers and covers more than 40 housing property criteria, from which only those criteria that were simultaneously included into the RMLS inquiry form and the secondary housing market database were selected[2]. This was done in an effort to receive estimations that will come most closely to the real housing prices (that is more accurate and less biased).

Dependent variable:
Value of a housing unit, thou. dollars. As is evident from earlier investigations, models that estimate total value of a dwelling produce more accurate estimations than models that use price per square meter as the dependant variable. We use the seller's bid price rather than the actually paid purchase price for dwelling as the dependent variable. We have to do this because nowadays there is no well-developed system of collection of information about really closed transactions. Occasionally the bid price may substantially vary from the purchase price, but generally this difference is not large according to Moscow real estate experts and brokers (See Fig. 1).

[2] For the full inventory of the housing property criteria see the RLMS inquiry form, Standard Annual Form #1 (approved by the RF State Statistics Committee, resolution #31 from 04/17/02) and Standard Quarterly Form #1-B (approved by the RF State Statistics Committee, resolution #2 from 01/15/02)

Fig 1. Average per square meter price of secondary housing in Moscow

Source: *Dengy* weekly, № 35 (440), 08.09 – 14.09.2003

Descriptive variables:

Living space, sq. meters. Presumably, there is the direct relationship between the cost and size of a dwelling (the larger a dwelling is, the higher is its cost).

Subsidiary space, sq. meters. This indicator shows the floorspace of subsidiary premises of a dwelling including a kitchen, corridors, a bathroom, a toilet, balconies and loggias. Similarly to the previous indicator, this one also directly affects the cost of a dwelling.

Number of rooms. From earlier investigations of housing price factors we know that there is a variety of mechanisms for pricing dwellings in accordance with the number of rooms. To put it differently, the estimated cost of housing criteria is not a constant value; it may vary with the type of dwelling. In our case this indicator gave rise to four more dummy variables useful for more accurate evaluation of correlations.

Construction period. Most professional appraisers consider the age of a dwelling as a major price factor. Regretfully, the database provided by A. Sapozhnikov, does not contain this indicator. However, considering its importance for the analysis we decided not to exclude it but try to estimate using the available supply statistics of the secondary housing market as follows[3]:

- Time period 1(before 1965)

  Building materials - brick; number of stories – no more than 10
- Time period 2 (1965 – 1990)
  - Building materials - brick; number of stories – not less than 12
  - Building materials - blocks; number of stories – any
  - Building materials – panels; number of stories – no more than 16
- Time period 3 (from 1990 – to present)

  Building materials – panels; number of stories – not less than 17

[3] See N.N. Nozdrina, A.Yu.Sapozhnikov, G.M. Sternik, S.G. Sternik, *Quality Grouping of Moscow Housing* - http://www.realtymarket.org

Building materials – monolith; number of stories – any.

We can not use variables "Building materials" and "Number of stories" separately in the regression since the variable "Number of stores" is not included into Mosstat survey. Doing so will create a systematic bias in imputed data on value of currently occupied units.

The qualitative variable "Period of construction" was then converted into three dummy variables to include into the regression model. Typically, the older the dwelling is, the lower is its price.

New housing, that is, apartments in recently constructed buildings, made up a separate group. As a rule, such apartments need finishing, and therefore a respective variable should be subject to adjustment with a negative coefficient. In other words, with all other terms equal the need of finishing will depreciate the value of a dwelling.

Telephone connection. The connection of a dwelling to the telephone line brings the price of a dwelling to increase.

Building materials. From earlier researches we know that the building material can substantially affect the price of a dwelling. To include this factor into the model a respective dummy variable was used with value 0 assigned to brick buildings, and 1 – to the rest[4]. Accordingly, it is anticipated that a respective variable would be subject to adjustment with a negative coefficient, because with all other terms equal apartments in brick buildings will cost higher than the rest.

Location. Location may also significantly change the price of a dwelling. Typically housing located in the downtown is priced higher than housing in the outskirts. This factor is included into the model as a dummy variables showing the location of a dwelling. Distribution of housing by this factor was made in accordance with Moscow administrative districts (9, excluding Zelenograd since there is no data on housing supply there).

Estimation method

There is no strong theoretical basis for choosing the correct functional form of a hedonic regression. Follain and Malpezzi[5] (1980), for example, tested a linear functional form as well as a log-linear (also known as semi-log) specification. But they found the log-linear form had a number of advantages over the linear form, detailed below.

The log-linear form is written:

$$\ln R = \beta_0 + \beta_1 S + \varepsilon$$

where ln R is the natural log of dwelling value, S are structural, neighborhood, locational, and other characteristics of the dwelling, β_0 and β_i and ε are the hedonic regression coefficients and error term, respectively.

The log-linear form has five things to recommend it. First, the semi-log model allows for variation in the dollar value of a particular characteristic so that the price of one component depends in part on the house's other characteristics. For example, with the linear model, the value added by a third bathroom to a one-bedroom house is the same as it adds to a five-bedroom house. This seems unlikely. The semi-log model allows the value added to vary proportionally with the size and quality of the home.

[4] We studied a variety of methods of this factor evaluation (building material) and found the suggested one the best as it makes it possible to receive the most accurate regression equation.

[5] Follain, James R., and Stephen Malpezzi. 1980. *Dissecting Housing Value and Rent.* Washington D.C.: The Urban Institute.

Second, the coefficients of a semi-log model have a simple and appealing interpretation. The coefficient can be interpreted as approximately the percentage change in the rent or value given a unit change in the independent variable. For example, if the coefficient of a variable representing central air conditioning is .219, then adding it to a structure adds about 22 percent to its value or its rent. Actually, the percentage interpretation is an approximation, and it is not necessarily accurate for dummy variables. Halvorsen and Palmquist[6] show that a much better approximation of the percentage change is given by e^b-1, where b is the estimated coefficient and e is the base of natural logarithms. So a better approximation is that central air will add exp (.219) – 1 = 24 percent.

Third, the semi-log form often mitigates the common statistical problem known as heteroskedasticity, or changing variance of the error term. Fourth, semi-log models are computationally simple, and so well suited to examples. The one hazard endemic to the semilog form is that the anti-log of the predicted log house price does not give an unbiased estimate of predicted price. This can, however, be fixed with an adjustment. Finally, it is possible to build specification flexibility into the right-hand side, using dummy (or indicator) variables, splines and the like (of which more shortly). This allows us a fair amount of flexibility in estimation, even with the semi-log form.

We selected to use the following hedonic regression model:

$$\ln(Value) = b_0 + b_1 \cdot \ln(Liv\_sp) + b_2 \cdot \ln(Rst\_sp) + c \cdot D + k \cdot D \cdot \ln(Tot\_sp) + \varepsilon$$

where D – dummy variables vector (number of rooms, building materials, construction period, administrative district),

c, k – vector coefficients,

ε – random term.

The model was estimated by OLS (ordinary least square) technique including the White heteroskedasticity correction.

Estimation results

Estimation results are shown in Table 1. All variables that had coefficients with a statistically insignificant bias from 0 were excluded from the regression model. Additionally, a selection of dummy variables (Tsentralny Administrative District; construction period – after 1990; number of rooms – 1) was also excluded from the equation in an effort to avoid multicollinearity problems.

Statistical parameters of the model are good: all coefficients of variables are significant and have the expected value; the equation in general is significant (F-test for the equation insignificance was rejected).

A high value of the determination coefficient R^2 (88%) signifies that the model has good forecasting capacities.

Supposing we would like to estimate the cost of a two-room apartment, the total size of which is 54 sq. meters including 30 sq. meters of living space. The apartment is located in the Central administrative district and is not connected to a telephone line. Under the suggested model, the cost of this apartment will approach to 83.5 thou US dollars.

From the second column of Table 1 is evident that the telephone connection raises the cost of housing unit by 1.8%. In our case this will raise the cost of the apartment by 1.5 thou US dollars.

[6] Halvorsen, Robert, and Raymond Palmquist. 1980. The Interpretation of Dummy Variable in Semilogarithmic Regressions. *American Economic Review*, 70, June, pp. 474-5.

The need of finishing of a newly-built unit decreases its cost by 3.77% with all other terms equal. In other words, if this apartment was in a newly-built house, its cost would go down by 3.1 thou US dollars.

The building material remains a rather significant price factor. For example, units in buildings other than brick will cost 3.29% less than similar units in brick buildings.

Most high-priced dwellings are located in the downtown. Thus dwellings of similar type but located in other districts will be less costly. As an illustration, the difference in prices between the Southern East District and downtown is 30.56% and the South District and downtown – 27.46%. Turning back to our case, we see that the cost of an apartment in the South-East district averages 58 thou US dollars, while a similar apartment in the South district costs by 2.5 thou more (near 61 thou US dollars).

If we compare prices for two similar units with one of them in a building constructed before 1965, and the second – in a building constructed after 1990, the difference will be 25.77%. Typically units in buildings constructed between 1965 – 1990 cost 2.45% less than similar units in buildings constructed after 1990, but 23.32% higher than units in buildings constructed before 1965.

The relationship of the unit price and size is also ruled by some auxiliary factors such as the number of rooms and the construction period. Table 2 examines various situations and estimates the relationship between the size and cost of a unit. In our case if a buyer wants to purchase an apartment by 10 sq. meters larger he/she will have to pay by $\frac{10}{30} \cdot 0.75 = 0.25$ percent, or 20.85 thou US dollars, more.

Table 1. Model estimation outcomes

	Log (Value of housing)	%
Const	8,598328	8,60
	(123,89)	
Log (Living space)	0,468594	0,47
	(25,54)	
Log (Rest space)	0,335178	0,34
	(35,52)	
Rooms = 2	-1,034431	-64,46
	(11,91)	
Rooms = 3	-1,165722	-68,83
	(11,85)	
Rooms = 4	-1,770713	-82,98
	(10,83)	
Construction period: before 1965	-0,298038	-25,77
	(8,27)	
Construction period: 1965 - 1990	-0,02485	-2,45
	(7,53)	
Presense of telephone	0,017882	1,80
	(2,86)	
Newly constructed	-0,038466	-3,77
	(3,58)	
Material of construction: other then brick	-0,033498	-3,29
	(28,22)	
East	-0,31659	-27,14
	(52,62)	
West	-0,142693	-13,30
	(20,53)	
North	-0,264949	-23,28
	(41,65)	
North - East	-0,281143	-24,51
	(45,71)	
North - West	-0,273518	-23,93
	(40,86)	
South - East	-0,364642	-30,56
	(58,39)	
South - West	-0,21134	-19,05
	(30,80)	
South	-0,321	-27,46
	(54,37)	
Interaction terms		
Room = 2 * Log (Total space)	0,278451	0,28
	(11,66)	
Room = 3 * Log (Total space)	0,297199	0,30
	(11,44)	
Room = 4 * Log (Total space)	0,433396	0,43
	(11,30)	
Construction before 1965 * Log (Total space)	0,055026	0,06
	(5,87)	
R ^2	0.880755	
R ^2 adj	0,880605	
F stat	5882,003	
Prob F	0	
Number of obs	17543	

All coefficients are extremely significant (Absolute value of t-stat. in parentheses)
The "Center", "Construction period: after 1990" and “Rooms = 1” were not included into regression to avoid multicollinearity.
Due to heteroskedasticity the White procedure is applied.

Table 2. An increase in dwelling value due to 1% increase in living space.

		Number of rooms			
		1	2	3	4+
Construction period	Before 1965	0.52%	0.80%	0.82%	0.96%
	1965 – 1990	0.47%	0.75%	0.77%	0.90%
	After 1990	0.47%	0.75%	0.77%	0.90%

Conclusions

The regression analysis carried out in the course of this research appeared to be rather useful not only for accomplishment of the key research goal – to estimate housing prices in accordance with a set of parameters, but for examination of the pricing mechanism of the secondary housing market. The ultimate findings of this research showing the relationship of housing prices with various factors may be of particular use for decision-making by real brokers, developers and constructors.

The inconvenience of the suggested method consists in the need to regularly repeat these estimations in accordance with changes in the secondary housing market demand. Another difficulty is the selection of variables used for estimation purposes: a larger number of variables lowers the forecasting potential of the model, while a small number of them is exposed to the risk of ignoring significant factors. In our case the selection of variables was guided by the specific goals of this research and was limited to a set variables used in official household surve.

The suggested model has a good forecasting potential and thus can be used for simulation the behavior of housing prices on the market. From the standpoint of forecasting and statistical quality this model appears to be the best in the group of simple and linear regression models.