Jan Magnus, Anatoly Peresetsky

The price of Moscow apartments[1]

We present a simple hedonic model for apartment prices in Moscow in the year 2003. Based on some 15,000 observations we estimate the model and use the estimates for prediction. Pretest issues are explicitly taken into account.

Key words: Hedonic prices, Moscow, pretesting.

1. Introduction

In this paper we attempt to explain and predict the price of apartments in Moscow in the year 2003. The price of an apartment is hypothesized to depend on certain characteristics, such as size and location. We thus follow the 'hedonic method, ' which goes back to Haas (1922) and was made popular in studies by Griliches (1961) on car prices and Chow (1967) on computer prices; see also Lancaster (1966) for the economic theory on which the hedonic method is based.

The development of asking prices in Moscow is given in Figure 1 over a sixteen-year period. After the fall of the Berlin Wall on 9 November 1989 and the 'August Coup' in 1991, the Soviet Union was dissolved and Boris Yeltsin became president of Russia. The ruble suffered two serious crises in the 1990s, first on 1 1 October 1994 ('Black Tuesday') and then on Monday 17 August 1998. The effects are visible in the figure. After 2000, property values increased dramatically. The asking price for an average apartment was $ 652/sqm in June 2000. Six years later, in June 2006, the asking price for the same apartment had increased to $ 4072/sqm. Currently Moscow is one of the most expensive property markets in the world (after Monaco and London), and it has the largest concentration of billionaires after New York and London. Prestigious projects, such as the 64-storey Federation Tower at Krasnopresnenskaya Embankment are under construction.

Against such a volatile background the estimation and prediction of property prices is a hazardous exercise. We do, however, have access to a very large and unique data set (about 15,000 observations), which gives us some hope of gaining insight into this market.

If we had been interested in the development of property prices over time, then we should have taken into account the state of the economy, the financial market, and expectations. Since we are primarily interested in the relative prices at a specific point in time, the hedonic method seems appropriate, at least as a benchmark.

The purpose of this paper is threefold. First, we model and estimate the price of an apartment as a function of characteristics as advertised on websites. This is done using one half of the sample. Second, we use the other half of the sample to test whether the estimated model can be used in predicting the prices. Given the idiosyncracies of property markets in general and the

[1] Acknowledgements: We thank Dmitry Danilov, Arthur van Soest, Marina Velikova, and Marno Verbeek for helpful comments, Chris Muris and Hongying Sun for valuable research assistance, and G. Malginov and G. Sternik for providing the data underlying Figure 1. E-mail addresses: Magnus: magnus@uvt.nl; Peresetsky: peresetsky@cemi.rssi.ru

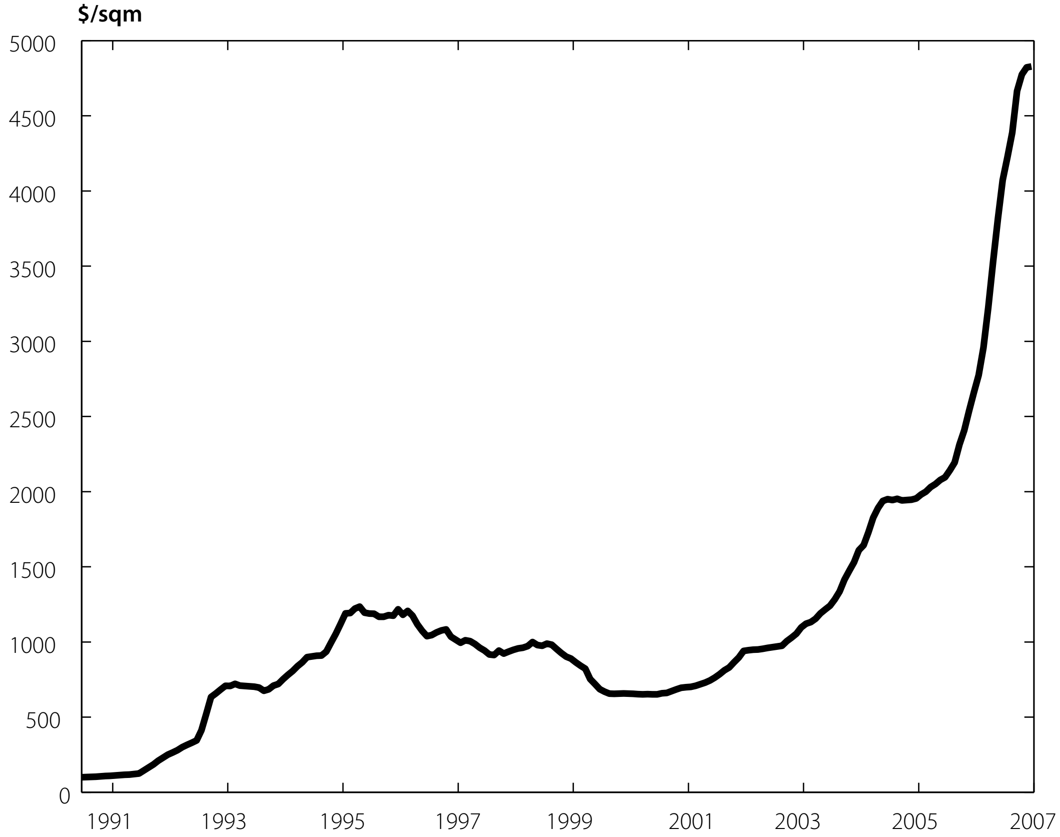

Figure 1: Asking prices of Moscow apartments, 1990–2006 [2]

volatile Moscow market in particular, it would be naive to believe that property prices can be determined by a simple model where by necessity many important details have been omitted. Nevertheless, the predictions are reasonable and useful. In addition to these two aims, we also wish to take a serious view towards pretesting and the fact that essentially all empirical work ignores the possible dangers of not accounting for it.

In real estate economics, the hedonic method is used to account for the problem that buildings are heterogeneous, so that it is difficult to estimate the demand for buildings generically. The hedonic method assumes that a house or apartment can be decomposed into characteristics such as number and size of the rooms, distance to the city center, environmental quality (air pollution, water pollution, noise), environmental amenities (aesthetic views or proximity to recreational sites), and the like.

Applications of the hedonic method to real estate indices include studies in the USA, most prominently Haas (1922) in Minnesota, 1916–1919; Bailey, Muth, and Nourse (1963) in St. Louis, 1937–1959; Witte, Sumka, and Erekson (1979), based on theory developed in Rosen (1974),

[2] The figure is based on monthly observations from June 1990 until December 2006; see Malginov and Sternik (2006) for the development until December 2005. The data for 2006 were kindly provided by G. Sternik.

in North Carolina, 1972; Milton, Gressel, and Mulkey (1984) in Florida; and Mills and Simenauer (1996) in US regions, 1986–1992.

Among the European studies we mention applications by Bender, Gacem, and Hoesli (1994) in Geneva, 1978–1992; Lansink and Thijssen (1998) in The Netherlands, 1970–1988; Maurer, Pitzer, and Sebastian (2004) in Paris, 1990–1999; and Van Soest and Verbeek (2010) in Moscow, 1994 and 1996.

Most of the studies — with the notable exception of Maurer, Pitzer, and Sebastian (2004) — analyze only a few hundred transactions. In contrast, we have about 15,000 observations.

The plan of this paper is as follows. In Section 2 we describe the data. In Section 3 we estimate the model and discuss the results in some detail. We then use the estimated model for prediction (Section 4) and assess the strengths and limitations of the predictions. In Section 5 we discuss pretest issues. Section 6 concludes.

2. The data

The data were collected by students of the New Economic School in Moscow in January – February 2003 as part of a course project in econometrics. The raw data consist of 16,115 apartments (flats) containing one, two, three, or four rooms. The typical apartment is a two-room apartment (43% of our sample), while one-room (25%) and three-room (28%) apartments are also common. The four-room apartments (4%) have a small market, and — according to Moscow estate agents — are priced idiosyncratically. Since our interest lies primarily in modeling standard mass-produced apartments, we have omitted the four-room apartments from our sample. Then 15,476 apartments remain.

The data on these one-, two-, and three-room apartments were collected by 66 students, divided into 19 groups. Each group had the task to collect data on one apartment category (1-2-3 rooms) only. Groups working on the same category coordinated their work (typically by Moscow regions or metro lines), so that apartments could not be counted twice.

The source of the data are websites of the most commonly used real estate agents in Moscow in 2003:

www.appartment.ru, www.astet.ru, www.babilon.ru, www.kdo.ru, www.estate.msk.ru,
www.kont.ru, www.mian.ru, www.miel.ru, www.novostroy.ru, www.orsn.ru,
www.realty.ru, www.trigon.ru,

and the 'Fili' real estate agency. Three typical entries are exemplified in Table 1, where we see two three-room apartments and one one-room apartment advertised. The nearest metro station is provided, the exact address ('D' stands for Dom (house, building); 1/17 indicates a building at a corner: 1 is the number at the main street (Trofimova), while 17 is the number of the same building at the cross street). The floor number 4/8 means that the apartment is on the fourth floor of an eight-floor building. Tot/Liv/Kit gives the area (in m^2) of the total space, the living space, and the kitchen space, respectively.

All prices are asking prices, since realized sales prices are not publicly available. In Moscow (and in Saint Petersburg) most of the apartments are priced in US dollars; in other Russian cities mostly in rubles. The data contain both newly built and old apartments. The variables in our data set are as follows.

Table 1

Three advertised apartments

# Rooms	3	3	1
Metro	Aviamotornaya	Avtozavodskaya	Akademicheskaya
Address	Sinichkina ul., 2nd, D. 4	Trofimovaul., D. 1/17	Shvernikaul., D. 7
Distance to metro	5 publ. transp.	10 by foot	10 by foot
Price	$ 75,000	$ 86,000	$ 39,000
Floor Number	5/5	4/8	4/12
Type of building	brick	brick	panel
Tot/Liv/Kit	73/49/9	83/54/9	35/14/10

price: asking price of the apartment (in 1000 US$) as advertised, used as a proxy to the realized sales price which is not available.

space variables:

- *totsp*: total area of the apartment (in m^2).
- *livsp*: area of the living space (in m^2), that is, living room, bedrooms, dining room.
- *kitsp*: area of the kitchen (in m^2). Only typical apartments with a separate kitchen are considered.

In addition to *livsp* and *kitsp* an apartment will have additional space (*addsp*) for a bathroom, toilet, hall, and the like. By definition,

$$totsp = livsp + kitsp + addsp.$$

rooms: number of rooms (1, 2, or 3).

distance variables:

- *dist*: distance to the center of Moscow (in km). Moscow has a radial-circle structure with a well-defined center (near Red Square). The distance is measured (using a city map) from the nearest metro station to the city center.
- *metrdist*: distance to the nearest metro station (in minutes), either by foot or by public transport (bus, trolley, tram, or dolmush); see the dummy variable *walk*.

dummy variables:

- *walk*: 1 if the apartment is within walking distance from the metro; 0 otherwise. This dummy is to be combined with *metrdist* in order to measure the distance (in traveling time) from the apartment to the nearest metro station. For example, if the advertisement says 'five minutes from metro by foot, ' then *metrdist* = 5 and *walk* = 1, and when it says 'ten minutes from metro by public transport, ' then *metrdist* = 10 and *walk* = 0.
- *brick*: 1 if the apartment is a 'brick' building; 0 otherwise. The term 'brick' includes apartment buildings made of *kirpichniy* (brick) or *monolitniy zhelezobeton* (cast in-situ reinforced concrete),

but not buildings made of *panelniy* (panel construction) or *blochniy* (breeze block: a light concrete building block made with cinder aggregate). The dummy is a proxy for the quality of the building: brick is better than no-brick.

- *tel*: 1 if the apartment possesses a regular city telephone line; 0 otherwise. Typically, *tel* = 1 unless the building is newly constructed in an area with poor infrastructure. Therefore, *tel* serves as a proxy for the development of the infrastructure. If there is no telephone, then this is a disadvantage, because it may take one or two years to get connected. Of course, with the appearance of mobile phones the presence of the regular city line has become less important.
- *balc*: 1 if the apartment has at least one balcony or loggia; 0 otherwise. Ground floor apartments typically do not have balconies for reasons of security.
- *floor*: 1 if the apartment is not located on the ground or upper floor; 0 otherwise. In Moscow, apartments on the ground or upper floor are less popular (luxury penthouses excluded), because they typically do not have a balcony. Also, the upper floor may be noisy from elevator machinery, may suffer water leakage when the roof is damaged, and may be too hot in the summer if the roof is poorly insulated. The ground floor is considered more dangerous, may be noisy because of the entrance door and elevator, may smell badly if isolation from the basement is inadequate, the floor may be cold, and there may be lack of privacy.

Before the data can be used in an analysis they need to be screened and cleaned. There are a number of obvious misprints which need to be deleted. For example, *totsp* = 100,000 is not likely, *floor* < 0 is impossible, and so is the combination (*walk* = 0 and *metrdist* = 0) which would imply traveling zero minutes by public transport to the nearest metro station. In addition we impose the following restrictions:

$20 \le price \le 250$. If the price is less than $ 20,000 then there must be a special situation. Say the advertised price is $ 18,000. Then it could be that the actual price is $ 36,000 and the buyer is supposed to pay 50% now and 50% next year. Another possibility is the agent tells you: 'Yes, $ 18,000 is the price. But you see, the owner is a very old lady and needs money. So we sign a contract in which you pay money now, but only take possession of the apartment after the old lady has died. ' If the price is higher than $ 250,000, then it must relate to a 'luxury' apartment which represents a very different segment of the housing market.

$25 \le totsp \le 150$. If the total space of a one-room apartment is less than 25, this probably indicates a 'dormitory' or 'hotel' apartment, which we do not consider.

$10 \le livsp \le 100$. One can hardly imagine a one-room apartment with a room space less than 10 m^2.

$5 \le kitsp \le 25$. A kitchen space of less than 5 m^2 is not usual, and it indicates a 'dormitory' or 'hotel' apartment or a misprint.

$6 \le addsp \le 45$. It is hard to imagine a hall and bathroom (combined with toilet) of less than 6 m^2.

$0.4 \le psqm \le 3.0$, where $psqm := price/totsp$ denotes the price per square meter. In January/February 2003 a price of less than $ 400 per m^2 is suspiciously low, and might indicate prepayment, an elderly owner, a dormitory, or some other special situation. Prices higher than $ 3000 are related to 'luxury' apartments.

$9 \le livsp/rooms \le 30$ indicates that the average space per room should neither be too small nor too big. If the average is more than 30 then the apartment is most likely one with 'free planning' (where the new owners need to construct the inner walls themselves).

Imposing these restrictions implies that we must delete 754 data, so that our data set reduces further from 15,476 to 14,722. These are the data we will work with in our analysis.

The students who collected the data are serious and intelligent master's students. They were carefully supervised during the data collection process. Nevertheless, there may have been some cheating in one or two groups, typically by using data from a previous year and multiplying all prices by a fixed amount. We will see later how we can control for possible cheating.

A preliminary data analysis yields summary statistics tabulated in Table 2, based on 3533 one-room apartments, 6785 two-room apartments, and 4404 three-room apartments — in total 14,722 apartments. We see that the apartments do not differ much in distance from a metro station, distance from the center of Moscow, in price per square meter, and in the quality of the building. This is in agreement with common sense because a typical Moscow building usually contains many different types of apartments. Also, the average space of the kitchen is almost the same for all apartments. However, *totsp*, *livsp*, *addsp* (and of course *price*) differ considerably. The variable *floor*, not reported in the table, is smaller for one-room apartments, which could mean that old houses have more one-room apartments and fewer floors, or that there is high turnover of one-room ground-floor apartments.

Table 2

Summary statistics

rooms		price	totsp	livsp	kitsp	addsp	psqm	dist	metrdist
1	Mean	38.3	37.3	19.3	8.2	9.8	1.03	11.9	9.4
	Median	36.0	37.0	19.0	8.0	9.0	1.00	12.5	10.0
2	Mean	58.6	52.1	31.3	8.0	12.8	1.11	9.9	8.4
	Median	50.0	52.0	31.0	8.0	12.0	1.00	10.5	10.0
3	Mean	76.4	72.1	46.6	8.7	16.9	1.04	10.1	8.7
	Median	68.0	72.6	45.0	8.5	16.0	0.97	10.5	10.0

3. Estimation

In order to select the appropriate model and estimate the parameters, we choose log (*price*) rather than *price* as our dependent variable, because this variable is much better approximated by a normal distribution. We also distinguish between focus variables and auxiliary variables.

The focus variables are in the model because we want them to be, either because the associated estimate is of central importance to us or because economic theory or common sense indicates that this variable must be in the model. In contrast, auxiliary variables are not of direct interest; they only appear because we have an a priori belief that they might improve the properties of the focus estimates. The a priori distinction between focus variables and auxiliary variables implies that variables may be kept in the model even when the associated *t*-ratios indicate otherwise (are 'small'), and it also allows us to properly account for distortions in pretest bias and variance as we shall see later.

We have selected eight focus variables and three auxiliary variables, as follows:

focus variables: We consider *constant* as a focus variable. The space of an apartment is represented by log (*totsp*) and the ratio *kitsp/totsp*. The distance is represented by log (*dist*), and three variables indicating together the distance to the nearest metro station: *metrdist* × *walk*, *metrdist* × (1 − *walk*), and *walk*. The variable *floor* is a focus variable because we wish to test whether this variable has an impact or not.

auxiliary variables: We consider the existence of a balcony (*balc*) as an auxiliary variable. Since balconies do not appear on the ground floor, we also include a cross-term *balc* × *floor*. Finally, *group 18* is a dummy for a specific group of students who we suspect of cheating.

The unrestricted model includes all eleven variables, while the restricted model only includes the eight focus variables. Model selection takes place only over the auxiliary variables.

Before we select our preferred model and estimate our parameters, we split the sample in two halves, because we wish to test the prediction power of our selected model and we wish to do so on a new set of data. So we randomly split the sample, resulting in 7395 observations for estimation (sample 1) and 7327 observations for analyzing our predictions (sample 2).

3.1. Model selection

All estimation results are thus based on sample 1 and they are summarized in Table 3.

We first estimate the unrestricted model. Based on the *t*-values of the three auxiliary parameters (7.5, −4.8, and −1.4, respectively), we decide to drop the *group 18* variable but keep *balc* and *balc* × *floor*. It makes no difference whether we perform general-to-specific or specific-to-general or some other selection procedure — the selected model is always the same. The estimates of the selected model are given in the second column of Table 3 and we see that the parameter estimates are hardly affected.

Columns 3 and 4 in the table are given for comparison only. The restricted model, although rejected in favor of the selected model, still produces no big changes in the estimates of the focus variables. The small 'naive' model is less good, although it too gives credible estimates of the included parameters.

All estimates have signs that correspond to our a priori beliefs. According to common belief, the price of an apartment depends primarily on size and location. Indeed, location is important: both the distance to the center and the distance to the nearest metro station are significant regressors. Since we lack information on the proximity of trees, air quality, the view, and other location variables, we cannot be more precise on the effect of location.

While location is important, size is more important. If we drop log (totsp) from our regression, then R2 reduces from 0.781 to 0.352. It is not true that the price per square meter is constant. In fact, larger apartments are not only more expensive but also relatively more expensive: the coefficient of log (totsp) is significantly larger than one. This is due to the fact that larger apartments are relatively sparse, but also to the fact that larger apartments are often more luxurious.

Table 3

Regression results

Variable	Unrestricted	Selected	Restricted	Naive
constant	– 0.053	– 0.060	– 0.032	—
	(0.044)	(0.043)	(0.043)	—
log (*totsp*)	1.080	1.082	1.083	1.094
	(0.009)	(0.009)	(0.009)	(0.002)
kitsp/totsp	0.904	0.891	0.936	—
	(0.058)	(0.057)	(0.057)	—
log (*dist*)	–0.203	–0.203	–0.199	–0.180
	(0.004)	(0.004)	(0.004)	(0.004)
metrdist × *walk*	–0.004	–0.004	–0.004	—
	(0.001)	(0.001)	(0.001)	—
metrdist × (1 – *walk*)	–0.007	–0.007	–0.007	—
	(0.001)	(0.001)	(0.001)	—
walk	0.053	0.053	0.052	0.093
	(0.011)	(0.011)	(0.011)	(0.005)
floor	0.093	0.093	0.058	—
	(0.010)	(0.010)	(0.005)	—
balc	0.069	0.069	—	—
	(0.009)	(0.009)	—	—
balc × *floor*	–0.054	–0.054	—	—
	(0.011)	(0.011)	—	—
group 18	–0.014	—	—	—
	(0.010)	—	—	—
$\hat{\sigma}^2$	0.033	0.033	0.033	0.036
R^2	0.781	0.781	0.779	0.760

3.2. Has cheating taken place?

It is a lot of work for the students to collect the data and it is possible that some cheating has taken place. In the previous year a similar data collection exercise had taken place, though on a smaller scale. It would have been quite easy to take last year's data and multiply all prices by a fixed amount. And this would have been hard to detect. We are suspicious of one group in particular, namely group 18. This group was one of the groups collecting data on one-room apartments, and they obtained data on 710 apartments (364 in sample 1). We see from Table 3 that the coefficient is not significant (*t*-value is –1.4) and we conclude from this that there is insufficient evidence to decide that cheating has taken place.

Of course, on may argue — since group 18 only collected data on one-room apartments — that we should include apartments dummies. If we include a dummy *R1*, which takes the value 1 if the apartment is a one-room apartment and zero otherwise, and a dummy *R2*, similarly defined for two-room apartments, then the estimate for *group 18* changes to –0.006 (0.011), which is also not significantly different from zero.

Hence we proceed as if the students have been reliable and that no cheating has taken place.

3.3. Optimal apartment plan

The estimate of the coefficient of *kitsp/totsp* is 0.891, suggesting that kitchens at the moment are too small. In our sample the kitchen occupies 15.9% of a typical apartment. The larger the kitchen relative to the total apartment, the more expensive is the apartment. Up to a point of course. Thus, how big should the kitchen be?

To see what would be optimal we include the variable $(kitsp/totsp)^2$ in our regression. In general, in a relationship

$$y=\alpha+\beta_1 x+\beta_2 x^2+\dots,$$

where $\beta_2 < 0$, the value of *x* which maximizes y is given by $x^* = -\beta_1/(2\beta_2)$. Letting $x = kitsp/totsp$, we find $\hat{\beta}_1 = 3.636$ and $\hat{\beta}_2 = -7.411$, and hence $x^* = 0.245$. This suggests that the kitchen should occupy 24.5% in the apartment in order to maximize the price. In fact, the new series of mass-produced apartment blocks in Moscow almost achieve this optimum.

3.4. Distance to center

The typical Moscow apartment has two rooms and there is a shortage of one-room apartments. A consequence of the high demand for one-room apartments could be that their price decreases slower with distance from the center than is the case with two-room apartments. To investigate this hypothesis, we replace the constant term by three dummies *R1*, *R2*, and *R3*, and we replace log(*dist*) by log(*dist*) × *R1*, log(dist) × *R2*, and log(*dist*) × *R3*. In our estimated model, a 1% increase in distance from the center leads to a 0.20 percent decrease in price. If we distinguish three categories, the estimated elasticities are –0.16, –0.21, and –0.19, respectively, and the differences are statistically significant. This confirms our hypothesis.

We see that the price of three-room apartments also decreases slower with distance from the center than is the case with two-room apartments. This is not due to a shortage of three-room apartments. The reason is probably that owners of three-room apartments are relatively rich and have cars, so that distance is less important. Also, three-room apartments are often occupied by families with children who may prefer a cleaner location further from the center, while singles in one-room apartments prefer the center.

3.5. Metro distance

Recall that the variable *metrdist* indicates the travel time (in minutes) to the nearest metro station, be it on foot or by public transport. Table 3 shows that

$$\log(price)=\begin{cases}\ldots-0{,}007\times metrdist, & \text{if } walk=0,\\ \ldots+0{,}053-0{,}004\times metrdist, & \text{if } walk=1.\end{cases}$$

Hence, in comparison to an apartment next to a metro station, an apartment five, ten, twenty, or thirty minutes away by foot reduces the price by 2.0%, 3.9%, 7.7%, 11.3%, respectively. And, if the nearest metro station is only reachable by public transport, it reduces the price by 8.6%, 11.6%, 17.6%, 23.1%, respectively.

More detailed analysis shows that buyers of three-room apartments find the distance to the metro less important than buyers of one- or two-room apartments. Probably they are rich enough to go by car.

3.6. Relevance of floor

The estimated coefficient for the *floor* variable is 0.093, showing a preference for apartments that are not on the ground floor or on the top floor. The impact of the *floor* variable on one-room apartment prices is much smaller than for larger apartments, perhaps because one-room apartments are typically bought by relatively poor young families who may be less sensitive to 'bad' floors.

4. Prediction

We do not expect that apartment prices are predicted accurately by our selected model. Property prices depend on many variables and in particular the location variables are incomplete in our data set. The predictions of our model should therefore only be seen as a base value which buyer, seller, and estate agent may use as a starting point of valuation.

We consider the selected model and the estimated coefficients obtained from sample 1. We now use the second half of our data set (sample 2) in order to produce predictions. For each of the apartments in our sample we thus know the values of the relevant regressors, so that we can calculate the predicted price, which can then be confronted with the observed price. Letting p denote the observed price and $\hat{p}$ the predicted price, we plot in Figure 2 the predicted $\log(\hat{p})$ against the observed $\log(p)$ for each of the 7327 observations in sample 2. The discreteness of the data shows in the vertical lines of the picture: the asking prices are discrete, but the predictions are continuous. The point cloud is slightly curved, indicating that the model underestimates prices at the top end. A possible reason is that the more expensive apartments are distinguished by features not measured in our data set: a view of a park or river, respectable neighborhood, or jacuzzi bathroom.

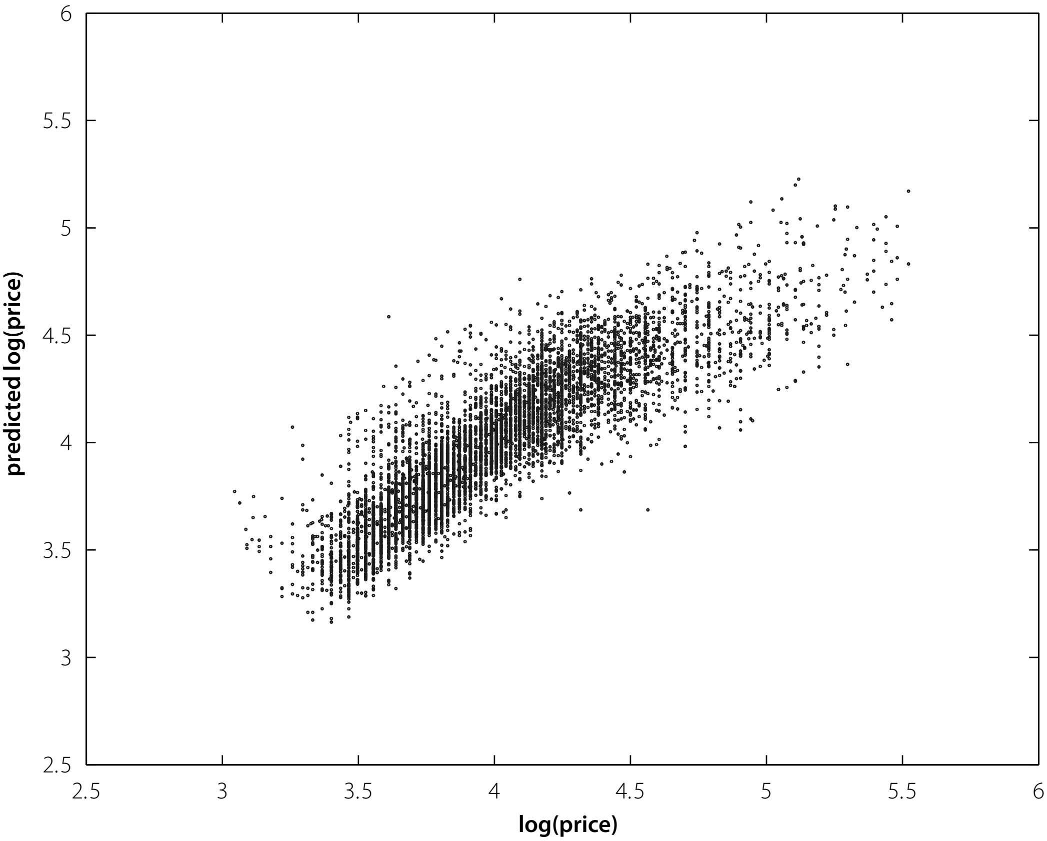

Figure 2: Prediction scatter of log ($\hat{p}$) against log (p)

The relative prediction error $(\hat{p} - p)/p$ is plotted in Figure 3. As expected, the histogram is skewed to the right. Only 28.6% of the predicted prices deviate less than 5% from the actual prices; 52.2% deviate less than 10%; and 85.6% deviate less than 25%. Furthermore, 6.5% of the predictions are more than 25% below the actual prices, and 7.8% are more than 25% above the actual prices. The predictions provide useful benchmarks for the valuation of an apartment, but they also confirm that property prices are determined by more factors than the ones observed in our sample.

5. The effects of pretesting

In econometrics we typically use the same data for both model selection and estimation (or prediction). Standard statistical theory is therefore not directly applicable, because the properties of the estimates depend not only on the stochastic nature of the selected model, but also on the way this model was selected. The relevant 'pretest theory' required to deal with this problem (at least in part) was developed in Magnus and Durbin (1999) and Danilov and Magnus (2004a, b). In this section we investigate what effect, if any, pretesting has on the distribution of the estimates presented before.

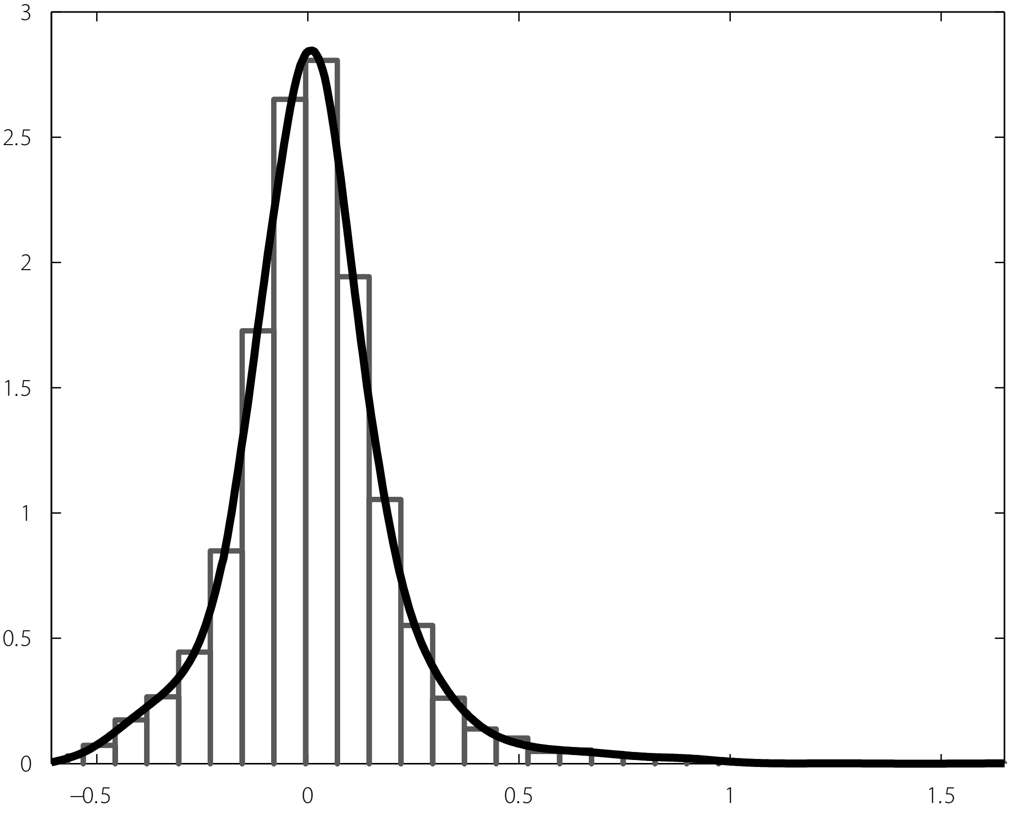

Figure 3: Histogram of relative predicted prices $(\hat{p} - p)/p$

Consider the standard linear regression model

$$y = X\beta + Z\gamma + \varepsilon,$$

where $y(n\times 1)$ is the vector of observations, $X(n\times k)$ and $Z(n\times m)$ are matrices of nonrandom regressors, $\varepsilon\,(n\times 1)$ is a random vector of unobservable disturbances, and $\beta\,(k\times 1)$ and $\gamma\,(m\times 1)$ are unknown nonrandom parameter vectors. We assume that $k \geq 1$, $m \geq 1$, $n - k - m \geq 1$, that the design matrix $(X:Z)$ has full column-rank $k + m$, and that the disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i. i. d. $N(0, \sigma^2)$.

The reason for distinguishing between *X* and *Z* is that *X* contains explanatory variables ('focus' regressors) that we want in the model on theoretical or other grounds, while Z contains additional explanatory variables ('auxiliary' regressors) of which we are less certain and whose role is only to improve the properties of the estimates of β. In our case we have $k = 8$ focus regressors and $m = 3$ auxiliary regressors.

We define the matrices

$$M := I_n - X(X'X)^{-1}X' \quad \text{and} \quad Q := (X'X)^{-1}X'Z(Z'MZ)^{-1/2},$$

and the normalized parameter vector $\theta = (Z'MZ)^{1/2}\gamma$. The least-squares (LS) estimators of β and γ are $b_u = b_r - Q\hat{\theta}$ and $\hat{\gamma} = (Z'MZ)^{-1/2}\hat{\theta}$, where $b_r = (X'X)^{-1}X'y$ and $\hat{\theta} = (Z'MZ)^{-1/2}Z'My$.

The subscripts '*u*' and '*r*' denote 'unrestricted' and 'restricted' (with $\gamma = 0$) respectively. Notice that $\hat{\theta} \sim N(\theta, \sigma^2 I_m)$ and that b_r and $\hat{\theta}$ are independently distributed.

Let S_i be an $m \times r_i$ selection matrix of rank r_i ($0 \le r_i \le m$), so that $S_i' = (I_{r_i} : 0)$ or a column-permutation thereof. The equation $S_i' \gamma = 0$ thus selects a subset of the γ's to be equal to zero. Following Danilov and Magnus (2004a), the LS estimators of β and γ under the restriction $S_i' \gamma = 0$ are then given by

$$b_{(i)} = b_r - QW_i\hat{\theta}, \qquad c_{(i)} = (Z'MZ)^{-1/2} W_i \hat{\theta},$$

where

$$W_i := I_m - (Z'MZ)^{-1/2} S_i (S_i'(Z'MZ)^{-1} S_i)^{-1} S_i' (Z'MZ)^{-1/2}$$

is a symmetric idempotent $m \times m$ matrix of rank $m - r_i$. (If $r_i = 0$ then $W_i = I_m$.) The distribution of $b_{(i)}$ is given by

$$b_{(i)} \sim \mathrm{N}(\beta + Q(I_m - W_i)\theta, \quad \sigma^2((X'X)^{-1} + QW_iQ')).$$

There are 2^m (in our case eight) different models to consider, one for each subset of $\gamma_1, \ldots, \gamma_m$ set equal to zero. A *pretest* estimator of β is obtained by first selecting one of these models (using *t*- or *F*-tests or other model selection criteria), and then estimating β in the selected model. We assume that the model selection is based exclusively on the residuals from the restricted model, that is, on My. This assumption is satisfied in all standard cases. Thus, a pretest estimator of β can be written as $b = \sum_i \lambda_i b_{(i)}$, where the weights satisfy

$$\lambda_i = \lambda_i(My), \qquad \lambda_i \ge 0, \qquad \sum_i \lambda_i = 1,$$

the sum is taken over all 2^m models, and all λ_i are 0 except one which is 1. The pretest estimator of β is equal to $b = b_r - QW\hat{\theta}$, where $W = \sum_i \lambda_i W_i$. Notice that W is a random matrix, because the $\{\lambda_i\}$ are random.

Let $\eta := \theta/\sigma$ and $\hat{\eta} := \hat{\theta}/\sigma$. Also, let $V := \sigma^2 (X'X)^{-1}$ denote the variance in the restricted model, with diagonal elements υ_{jj} ($j = 1, \ldots, k$). The equivalence theorem for estimation (Danilov and Magnus, 2004a, Theorem 1) then implies that, for $j = 1, \ldots, k$, the bias, variance, and mean-squared error of the components of b are given by

$$\mathrm{bias}(b_j) := \mathrm{E}(b_j - \beta_j) = -\sqrt{\upsilon_{jj} q_{0j}^2} \cdot q_j' \mathrm{E}(W\hat{\eta} - \eta),$$
$$\mathrm{var}(b_j) := \upsilon_{jj}(1 + q_{0j}^2 \cdot q_j' \mathrm{var}(W\hat{\eta}) q_j),$$

and hence that

$$\mathrm{MSE}(b_j) = \upsilon_{jj}(1 + q_{0j}^2 \cdot q_j' \mathrm{MSE}(W\hat{\eta}) q_j),$$

where

$$q_{0j}^2 := \frac{e_j'QQ'e_j}{e_j'(X'X)^{-1}e_j}, \qquad q_j := \frac{Q'e_j}{\sqrt{e_j'QQ'e_j}},$$

and e_j denotes the j-th unit vector. The properties of the complicated pretest estimator thus depend critically on the properties of the less complicated estimator $W\hat{\eta}$ of η. In particular, different model selection procedures may lead to different effects of pretesting, and this works exclusively through $W\hat{\eta}$.

In what follows we shall assume that σ^2 is known, namely equal to $\hat{\sigma}^2$ in the unrestricted model. The makes the analysis somewhat easier and with the large number of observations that we work with the effect is negligible; see Danilov (2005).

It is convenient to scale the auxiliary regressors z_1, z_2, and z_3 by $z_i^* := z_i \Big/ \sqrt{z_i' M z_i}$ for $i = 1, 2, 3$. This has no effect on the analysis. Then,

$$Z^{*\prime} M Z^* = \begin{pmatrix} 1.000 & 0.809 & -0.012 \\ 0.809 & 1.000 & 0.001 \\ -0.012 & 0.001 & 1.000 \end{pmatrix}.$$

The first two auxiliary regressors are strongly correlated, but the third auxiliary regressor (*group 18*) is almost uncorrelated with the first two. If the correlation is strong, then MSE($W\hat{\eta}$) is typically large. So we would expect a stronger effect of pretesting for the first two auxiliary variables. Here the effects are small because the t-values (and the number of observations) are large.

Table 4

Statistics for the auxiliary variables

Variable	$\hat{\gamma}$	t_γ	$\hat{\gamma}^*$	$\hat{\theta}^*$	$\hat{\eta}^* = \hat{\theta}^* / \hat{\sigma}$
balc	0.069	7.541	2.338	1.403	7.707
balc × *floor*	–0.054	–4.839	–1.500	–0.275	–1.510
group 18	–0.014	–1.352	–0.246	–0.266	–1.463

In Table 4 we present the t-values and associated statistics in the unrestricted model for the three auxiliary regressors. Of course, $\hat{\gamma}^*$ and $\hat{\gamma}$ are not the same, because of the rescaling, but their corresponding t-values are the same. The values of $\hat{\theta}$ and $\hat{\eta}$ (not reported) are close to (but not the same as) the values of $\hat{\theta}^*$ and $\hat{\eta}^*$. Since η is the unobserved 'theoretical' t-ratio, a comparison of t_γ and $\hat{\eta}^*$ is of interest. If $Z^{*\prime} M Z^*$ were equal to the identity matrix, then $t_\gamma = \hat{\eta}^*$ and hence there is little surprise that this equality almost holds for the third auxiliary variable *group 18*.

The effects of pretesting are summarized in Table 5 for the case of general-to-specific (backward) model selection. The first two columns repeat the estimates b of the focus parameters β and their reported standard errors $\sqrt{\widetilde{\mathrm{var}}(b)}$ from Table 3. The estimates themselves do not change

Table 5

Effects of pretesting on the focus variables

Variable	b	$\sqrt{\widetilde{\text{var}}(b)}$	q_0^2	bias	RMSE	RMSE*
constant	–0.053	0.044	0.027	–0.0037	0.044	0.045
log (*totsp*)	1.080	0.009	0.029	0.0012	0.009	0.009
kitsp/totsp	0.904	0.058	0.037	–0.0075	0.057	0.058
log(*dist*)	–0.203	0.004	0.023	–0.0002	0.004	0.004
metrdist × *walk*	–0.004	0.001	0.001	0.0001	0.001	0.001
metrdist × (1 – *walk*)	–0.007	0.001	0.001	–0.0000	0.001	0.001
walk	0.053	0.011	0.001	–0.0007	0.011	0.011
floor	0.093	0.010	2.770	0.0002	0.010	0.013

whether or not we take pretesting into account — only their properties change. Hence, var (*b*) is not the correct variance when pretesting is taken into account.

The values of q_0^2 are important because they are multiplication factors. A small value of q_0^2 implies a small difference between the reported and the actual variance and MSE, thus resulting in a negligible effect of pretesting. In our case, q_0^2 is quite small for all regressors except *floor*. Thus, if any pretest effect would occur, it would occur through this variable.

The pretest bias, variance, and mean-squared error are all functions of the parameter vector η. This vector is unknown, but an estimate $\hat{\eta}^*$ is available. This estimate is unbiased, but its variance is constant, since $\hat{\eta}^* \sim \mathrm{N}(\eta, I_m)$. Hence the variance does not converge to zero when the sample increases. In columns 4 and 5 we present the bias and root mean-squared error, calculated at $\eta = \hat{\eta}^*$. Of course, η will not be equal to $\hat{\eta}^*$. In the last column of Table 5 we therefore present RMSE*, a 95% upper bound of the RMSE calculated as the maximum in a three-dimensional probability box around $\hat{\eta}^*$.

We conclude from the results in Table 5 that the effect of pretesting in this case is very small, due to the strong *t*-ratios and the low values of q_0^2. The estimates of all eight focus variables are therefore unbiasedly estimated and have the correct standard errors. The only possible exception is the variable *floor*, where the root mean-squared error may be 37% higher than reported.

In the case of specific-to-general (forward) model selection, the same results hold in essence. The root mean-squared error of the coefficient of the variable *floor* may now be 44% higher than reported, in line with the simulations of Danilov and Magnus (2004a).

6. Conclusions

We are well aware of the limitations of the hedonic method. In particular, the method captures only the buyer's willingness to pay for perceived differences in attributes and their direct consequences. Thus, if people are not aware of the linkages between the attribute and benefits to them or their property, the value will not be reflected in property prices. The method also assumes that people have the opportunity to select the combination of features they prefer, given their income. The property market may however be affected by outside influences, like taxes, interest rates, or other factors. Despite these possible problems we found that our results are remarkably robust in the sense that different model specifications lead to essentially the same conclusions.

The main determinant of the apartment price in Moscow is its size (*totsp*), not its location. The elasticity of the price with respect to size is large than one, so that larger apartments are not only more expensive but also relatively more expensive. A possible explanation is that in the observed segment of the market ($totsp < 150$), larger apartments often correspond to more luxurious apartments, and that this is the reason they are relatively more expensive.

There exists an 'optimal' apartment plan where the kitchen occupies 24.5% of the total apartment area — much higher than the 2003 ratio (15.9%). Interestingly, the new series of mass-produced apartment blocks in Moscow almost achieve this optimum.

Since Moscow has a radial structure, it is to be expected that the distance to the center is important for the apartment price. We found that a 1% increase in distance from the center leads to a 0.20% decrease in price. The price of one-room apartments decreases slower than the price of larger apartments due to a shortage of these apartments.

Moscow has an effective and reliable metro system, which also has a radial structure and is the principal provider of public transport. Proximity to a metro station is therefore important. Compared to an apartment next to a metro station, an apartment ten minutes away by foot decreases the price by 3.9%, while the same apartment ten minutes away by public transport decreases the price by 11.6%. The distance to a metro station is less important for owners of three-room apartments — they often travel by car.

Muscovites dislike ground-floor or top-floor apartments. Such apartments are on average 9.3% less expensive. This effect is much smaller for one-room apartments than for larger apartments, possibly because one-room apartments are typically bought by relatively poor young families.

One can not expect that apartment prices would be predicted accurately by a model such as ours, which takes only the most basic properties of an apartment into account. Property prices depend on many other factors to which we have no access: location ecology, a good view, respectable neighborhood, quality of construction and decoration, and more. Still, the relative prediction error is less than 25% for 85.6% of the apartments, which provides a useful prediction tool for buyers, sellers, and estate agents.

The effects of pretesting are surprisingly small in our analysis. This is due in part to the large number of data, but also to specific characteristics of the data (low values of q_0^2). The model selection procedure thus has very little effect on the properties of our estimates and predictions. For practical purposes the estimates and predictions may therefore be considered unbiased with the correct estimated precisions.

References

Bailey, M. J., R. F. Muth, and H. O. Nourse (1963). A regression method for real estate price index construction, Journal of the American Statistical Association, 58, 933–942.

Bender, A. R., B. Gacem, and M. Hoesli (1994). Construction d'indices immoboliers selon l'approche hédoniste, Finanzmarkt und Portfolio Management, 8, 522–534.

Chow, G. C. (1967). Technological change and the demand for computers, The American Economic Review, 57, 1117–1130.

Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known, The Econometrics Journal, 8, 277–291.

Danilov, D. and J. R. Magnus (2004a). On the harm that ignoring pretesting can cause, Journal of Econometrics, 122, 27–46.

Danilov, D. and J. R. Magnus (2004b). Forecast accuracy after pretesting with an application to the stock market, Journal of Forecasting, 23, 251–274.

Griliches, Z. (1961). Hedonic price indices for automobiles: An econometric analysis of quality change, in: The Price Statistics of the Federal Government, National Bureau for Economic Research, General Series No. 73, New York, 137–196.

Haas, G. C. (1922). Sales prices as a basis for farm land appraisal, Technical Bulletin No. 9, The University of Minnesota Agricultural Experiment Station, St. Paul.

Lancaster, K. J. (1966). A new approach to consumer theory, Journal of Political Economy, 74, 132–157.

Lansink, A. O. and G. Thijssen (1998). Testing among functional forms: An extension of the generalized Box-Cox formulation, Applied Economics, 30, 1001–1010.

Magnus, J. R. and J. Durbin (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest, Econometrica, 67, 639–643.

Malginov, G. and G. Sternik (2006). The housing market of Moscow region, Section 4.9 in: Russian Economy in 2005: Trends and Outlooks, Institute for the Economy in Transition, Issue 27, Moscow, 429–448.

Maurer, R., M. Pitzer, and S. Sebastian (2004). Hedonic price indices for the Paris house market, Allgemeines Statistisches Archiv, 88, 303–326.

Mills, E. S. and R. Simenauer (1996). New hedonic estimates of regional constant quality housing prices, Journal of Urban Economics, 39, 209–215.

Milton, J. W., J. Gressel, and D. Mulkey (1984). Hedonic amenity and functional form specification, Land Economics, 60, 378–388.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition, Journal of Political Economy, 82, 34–55.

van Soest, A. and M. Verbeek (2010). Empirical Applications I, Econometric Exercises, Volume 4, Cambridge University Press, New York.

Witte, A. D., H. J. Sumka, and H. Erekson (1979). An estimate of a structural hedonic price model of the housing market: An application of Rosen's theory of implicit markets, Econometrica, 47, 1151–1174.