HONG KONG, 13-17 may 2007

# Statistical verification of real estate estimation models

*Prof. Jozef Czaja, PhD Anna Barańska*
*Terrain Information Department*
*Faculty of Mining Surveying and Environmental Engineering*
*University of Science and Technology*
*Krakow, POLAND*

# INTRODUCTION

The selection of a model describing the market variability of real estate values in relation to their characteristic attributes is the most important stage of the real estate market analysis. In the process of modelling the real estate values, additive or multiplicative functions can be used. A real estates database, containing transaction prices and attributes of real estates, may have additive as well as multiplicative qualities.

To choose a model describing the most precisely the variability of prices in a database, it is necessary to make the estimation of parameters of additive and multiplicative model and also a complete variance analysis. The verification of statistical hypotheses concerning the value of estimated parameters of these models will be the basis for a statistical inference allowing choosing a more reliable model. Parametric tests, constructed on *T*-Student's and *F*-Snedecor distribution quantiles, will be applied to the statistical inference.

In the process of modelling the real estate market values, additive (1) or multiplicative (2) functions can be used in form:

$$w = a_0 + \sum_{k=1}^{m} g(x_k) \qquad (1)$$

$$w = a_0 \cdot a_1^{x_1} \cdot a_2^{x_2} \cdot \dots \cdot a_m^{x_m} \qquad (2)$$

where:

$w$ – unit price or value of real estate,
$x_k$ – value of attribute *k* for real estate,
$g$ – function of real estate price - attribute *k* relation,
$a_0$ – free term in the model (unit value of a real estate, for zero of all attributes),
$a_j$ – estimated model coefficients.

## ESTIMATION OF MODELS PARAMETERS

**1. Estimation of the additive model**

First of all, we will consider a special case of a system of equations in form (1), which is a linear multiple regression (linear non-only by parameters, but also by independent variables):

$$w = a_0 + \sum_{k=1}^{m} x_k \cdot a_k \qquad (3)$$

The expression (3) may be written as a matrix:

$$[W] = [X] \cdot [a] \qquad (4)$$

where:
$[W]$ - vector of dependent random variable (of a real estate value),
$[X]$ - matrix containing ones and independent variables (attributes),
$[a]$ - vector of multiple linear regression coefficients.

Solving a generalized linear model means:

- determining an unbiased estimator of vector of unknowns:

$$\hat{a} = (X^T X)^{-1} \cdot X^T C \qquad (5)$$

- determining an unbiased estimator of remainder variance (defining estimation inaccuracy of model parameters):

$$\hat{\sigma}_0^2 = \frac{C^T C - \hat{a}^T X^T C}{n - m - 1} \qquad (6)$$

- determining a covariance matrix of vector of unknowns (model parameters):

$$Cov(\hat{a}) = \hat{\sigma}_0^2 \cdot (X^T X)^{-1} \qquad (7)$$

- determining a covariance matrix of model values:

$$Cov(W) = \hat{\sigma}_0^2 \cdot X^T (X^T X)^{-1} X \qquad (8)$$

$C$ – vector of unit price or value of real estates in a given database

2. **Estimation of the multiplicative model**

For the estimation of the coefficients $a_j$ in model (2), the function (2) has to be brought to a linear form. For this purpose, we take the logarithms of both sides, using the natural logarithm, and we receive:

$$\ln w = \ln a_0 + x_1 \cdot \ln a_1 + x_2 \cdot \ln a_2 + \ldots + x_m \cdot \ln a_m \qquad (9)$$

The system of equations (9) has the features of a probabilistic model, taking on in matrix notation the following form:

$$[\ln W] = [X] \cdot [\ln a] \qquad (10)$$

where:

$[\ln W]$ - vector containing natural logarithms of real estates prices

$[X]$ - mattrix containing ones and attributes values of real estates in database,

$[\ln a]$ - vector containing estimated values of natural logarithms of the model parameters $a_j$, By applying the least squares method, we receive the following formulas for estimated parameters of a non-linear model:

$$\ln \hat{a} = \left(X^T X\right)^{-1} \cdot X^T \ln C \qquad (11)$$

To determine a covariance matrix for estimated regression parameters, we use the formula:

$$Cov(\ln \hat{a}) = \hat{\sigma}_0^2 \left(X^T X\right)^{-1} \qquad (12)$$

where $\sigma_0^2$ is the variance of the multiplicative non-linear model estimation. Its estimator is given by the following expression:

$$\hat{\sigma}_0^2 = \frac{(\ln C)^T \cdot (\ln C) - (\ln \hat{a})^T \cdot X^T \cdot (\ln C)}{n - m - 1} \qquad (13)$$

$C$ – vector of unit price or value of real estates in a given database

## EXAMINING THE SIGNIFICANCE OF MODEL PARAMETERS

Within the framework of the statistical verification of estimated models, we examine the significance of the parameter system in each model and the significance of every particular parameter. However, the basic indicator determining the quality of matching a model with data is the square of curvilinear correlation coefficient $\boldsymbol{R}^2$.

## Examining the significance of model system parameters

Verification of a parameters system significance is based on Fisher-Snedecor's statistics, using the following hypothesis

*H*$_0$: $\sum_{j=0}^{m} a_j^2 = 0$

against the alternative hypothesis

*H*$_1$: $\sum_{j=0}^{m} a_j^2 \neq 0$ .

Statistic form in this test:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} \qquad (14)$$

which, if the null hypothesis is true, has a *F*-Snedecor's distribution with (*m*, *n*-*m*-1) degrees of freedom.

$n$ - number of considered real estates in the database

$m$ - number of considered attributes

## Examining the significance of particular model parameter

Verification of particular regression parameter significance is based on *T*-Student statistics, under the null hypothesis

*H*$_0$: $a_j = 0$

against the alternative hypothesis

*H*$_1$: $a_j \neq 0$.

Statistic form in this test:

$$T = \frac{\hat{a}_j}{\sigma\left(\hat{a}_j\right)} \qquad (15)$$

which, if the null hypothesis is true, has a *T*-Student distribution with (*n*-*m*-1) degrees of freedom.

If for any of explaining variables, the statistical test does not demonstrate reasons for rejecting the null hypothesis, we eliminate this variable from the model and we reestimate the parameters.

## FORECASTING A MARKET UNIT VALUE OF A REAL ESTATE

The prediction of a real estate unit price '$w$' (market value) is determined by substituting to the evaluated model, the values of the attributes $x_j$. By converting this, we receive the forecast value for the analysed real estate:

$$w = a_0 + x_1 a_1 + x_2 a_2 + \ldots + x_m a_m$$

for additive model

$$w = \exp\left(\ln a_0 + x_1 \cdot \ln a_1 + x_2 \cdot \ln a_2 + \ldots + x_m \cdot \ln a_m\right)$$

for multiplicative model

## EXAMPLE

In the tables 1 and 2, some of model estimation results in additive (linear in consideration of independent variables and parameters or in relation to the parameters only) and multiplicative form are presented. Models were tested on two local markets of real estates of the same type (dwellings). Acquired information on transactions concern two big quarters of one city, diversified in respect of factors shaping the prices of real estates.

Last two lines in each table, include the test results of verification of parameter system significance.

**Table 1.** Results of additive model estimation

| | multiple regression | | linear model in relation to the parameters | |
|---|---|---|---|---|
| | Quarter A | Quarter B | Quarter A | Quarter B |
| $n$ | 91 | 77 | 91 | 77 |
| $u$ | 14 | 14 | 29 | 40 |
| $\sigma_0^2$ | 0,17 | 0,08 | 0,15 | 0,08 |
| $\sigma_0$ | 0,42 | 0,28 | 0,39 | 0,28 |
| $R^2$ | 0,68 | 0,85 | 0,78 | 0,91 |
| $s$ | 0,57 | 0,40 | 0,10 | 0,07 |
| $F(\alpha, u\text{-}1, n\text{-}u)$ | 1,85 | 1,86 | 1,66 | 1,72 |
| $F_{cal}$ | 12,68 | 27,52 | 7,68 | 10,44 |

where:

$n, u$ – numbers of real estates and estimated parameters,
$\sigma_0^2$ $\sigma_0$ - model remainder variance and estimation standard error,
$R^2$ - coefficient of determination,
$s$ - fraction of model parameters statistically significant,
$F_{cal}$ – calculated value of a test function.

**Table 2.** Results of multiplicative model estimation

| | Quarter A | Quarter B |
|---|---|---|
| $n$ | 91 | 77 |
| $u$ | 14 | 15 |
| $\sigma_0^2$ | 0,02 | 0,01 |
| $\sigma_0$ | 0,14 | 0,07 |
| $R^2$ | 0,68 | 0,84 |
| $s$ | 0,57 | 0,40 |
| $F(\alpha, u\text{-}1, n\text{-}u)$ | 1,85 | 1,86 |
| $F_{cal}$ | 12,49 | 23,63 |

When we eliminate from the model variables which are not statistically significant, the new estimation of parameters is made and the model is verified. After this procedure all of parameters in each model are significant ($s$ =1,00).
The results of the analysis of qualities of reduced models are presented in tables 3 and 4.

**Table 3.** Results of additive reduced model estimation

| | multiple regression | | linear model in relation to the parameters | |
|---|---|---|---|---|
| | Quarter A | Quarter B | Quarter A | Quarter B |
| $n$ | 91 | 77 | 91 | 77 |
| $u$ | 8 | 6 | | 2 |
| $\sigma_0^2$ | 0,17 | 0,22 | | 0,22 |
| $\sigma_0$ | 0,42 | 0,47 | | 0,47 |
| $R^2$ | 0,65 | 0,55 | **0,26** | 0,74 |
| $s$ | 1,00 | 1,00 | | 1,00 |
| $F(\alpha, u\text{-}1, n\text{-}u)$ | 2,13 | 2,35 | | 3,97 |
| $F_{cal}$ | 21,86 | 17,13 | | 211,64 |

In the case of a model in additive complex form for the quarter A, where the coefficient of matching is very low $R^2$=0,26, the analysis of accuracy as well as the continuation of model verification were abandoned. This model was found unfit to predict the market value of a real estate.

**Table 4.** Results of multiplicative reduced model estimation

| | Quarter A | Quarter B |
|---|---|---|
| $n$ | 91 | 77 |
| $u$ | 8 | 6 |
| $\sigma_0^2$ | 0,02 | 0,01 |
| $\sigma_0$ | 0,15 | 0,07 |
| $R^2$ | 0,64 | 0,82 |
| $s$ | 1,00 | 1,00 |
| $F(\alpha, u\text{-}1, n\text{-}u)$ | 2,13 | 2,35 |
| $F_{cal}$ | 21,53 | 23,63 |

Comparing coefficients of determination $R^2$ obtained for models in the quarter '***A***' before and after reducing parameters in models (0,68 and 0,65 in additive simple model; 0,68 and 0,64 in multiplicative model), we can notice that the reduction of number of variables is negligible for quality of model.

On the ground of obtained results, we can see that the additive form of the model illustrates better the variability of prices in relation to the qualities essential for it in the quarter ***A***. Whereas, in the quarter ***B*** the multiplicative model seems to be better matched after the elimination of parameters statistically insignificant from the models.

In all models, the system of estimated parameters shows the statistical significance. It is however noticeable that the percentage of statistically significant parameters in the model decreases considerably when the number of parameters increases (Table 1). Though, it does not cause a proportional increase of model matching. Then, complicating the model by multiplication of parameters would not necessary improve its overall quality.

## CONCLUSIONS

Recapitulating, we can say that the vicinity of local real estate markets, like in case of quarters of the same town, as well as the consideration of real estates of the same type, are not sufficient to presume that an equal form of the estimation model will be proper.

In the analyzed example, for the quarter ***A***, finally, after the elimination of statistically insignificant parameters (variables), both forms of estimation model can be applied. However, for the quarter ***B***, the multiplicative model seems to match better the market data, giving a lower estimation standard error than in case of additive model.

Thank you very much for your attention