dr inż. Łukasz Mach
Narodowy Bank Polski, Oddział Okręgowy Opole
Wydział Statystyczno-Dewizowy

Logit modelling as a tool supporting decision making in the real estate market

Abstract: The article includes a logit model construction process to support the process of decision making in real estate market. The model of logistic regression which has been elaborated, shall define the probability of transactions in the real estate market and it will indicate statistic variables which influence the demand significantly. The process of decision making (based on logit models) - substantially prepared and correctly executed - is a key determinant having influence on improving the competitiveness of companies, especially during the global economic crisis.

Keyword: logit modelling, residential real estate market, decision making process

1. Introduction

Essentially prepared and properly conducted decision making process is the key factor, which influences the competitiveness of companies. Well prepared decision making process enables to minimize the risk of making wrong decisions and helps to enhance the competitive advantage of companies. During the economic crisis, especially companies, which operate in the prone to crisis industries should consider the proper decision making process. Taking into consideration economic theory, the main factors, which influences on competitiveness of economies there are investments implied mainly by housing market. The growth dynamics can be parameterized by the volume of sold residential real estates.

The aim of the article is researching the secondary real estate market in Poland by building three qualitative decision making models. The aim of the above mentioned is estimation of probability of selling real estates and indication (for every single model) diagnostic variables, which characterize particular residential real estates and influence on selling them. Building three different models was a result of analysis of residential real estate on the homogenous markets i.e. the markets homogenous according to the assumptions.

The process of grouping residential real estate markets in provinces for homogenous groups was conducted by applying cluster analysis (by using inter alia macroeconomic data of

economic and business data)[1]. The research process, which aims at building quality models describing probability of selling residential real estate based on the model of logistic regression (with the use of bidding data and selling data from third quarter of 2013 in capital of provinces[2]).

2. The research process- grouping provinces into homogenous area

The research process was conducted in two main stages. In the first stage, taking into consideration socio-economic factors which affect substantially on the price of square meter of residential, Polish provinces were grouped into three homogenous groups. Determinants describing residential real estate and their influence on selling apartments were specified in the second stage by applying logistic regression. In the process of grouping, by applying cluster analysis technique, the key factor was a choice of discriminating variables, which were used for creation of homogenous provinces groups. The assumption, that socio-economic variables would be determinants used in grouping particular districts, was made. The recognition of the factors, which mainly influence on the Polish districts development in the socio-economic dimension, was made with the use of the factor analysis. The observation matrix (16 provinces, 18 variables) was the basis for conducting factor analysis.

Diagnostics variables included: X1- new entities of national economy for 1000 population; X2- entities of national economy, which finished their activity for 1000 population; X3- number of micro enterprises for 10000 population; X4- number of small enterprises for 10000 population; X5- number of medium-sized enterprises for 10000 population; X6- number of large enterprises for 10000 population; X7- Gross domestic product per 1 person; X8- average monthly gross salary; X9- the average pension outside agricultural social security system; X10- retail per 1 inhabitant; X11- average monthly expenses per one person for use of apartment; X12 – average monthly expenses per one person per equipment of apartment; X13- average monthly income per person in household; X14- number of rooms for 1000 population; X15- the average usable area per one inhabitant; X 16- apartments for 1000 population; X 17- apartments for 1000 marriages; X18- new apartments in new residential areas (with the right of usage in the whole building or in particular parts) or in non-residential areas for 1000 population.

[1] data from Central Statistical Office
[2] data from National Bank of Poland

Implementation and assumptions referring to the correctness of the calculation in the research process were described in the previous thesis of the author [compare Mach 2012, pages 106-116]. In the above mentioned thesis diagnostic variables, which substantially affect for the price of the square meter of residential, were specified. In the above mentioned thesis, the socio-economic variables, which affect on the price of square meter of residential were specified by applying factor analysis and multiple regression. The factor analysis proved to be an effective tool enabling for detection of hidden development factors. Use of the multiple regression enabled for identification of factors substantially influencing on the price of square meter of residential. Parameterization of the researched relations, which specify development of particular provinces and indication of their affection for formation of the square meter's price of residential, can help enterprises, which operate in real estate industry as a prerequisite decreasing the risk of decisions, which were made in the managing process.

Finally, the variables, which substantially affected on the price of square meter of residential were economic and business variables including: number of micro enterprises for 10000 population, number of small enterprises for 10000 population, number of medium-sized enterprises for 10000 population, number of large enterprises for 10000 population, Gross domestic product per 1 person, average monthly gross salary, retail per 1 inhabitant, average monthly expenses per one person for use of the apartment, average monthly income per person in household. Diagnostic variables defined as above, were subsequently used as input variables applied in cluster analysis aimed at grouping Polish provinces for homogenous provinces. Ward's Algorithm was used as a agglomeration method and the distance measure was Euclidean distance, while applying clustering analysis. Requirements referring to application of clustering analysis and formal record can be found in thesis of the following authors [Aczel, A.D., 2000, s. 849-916; Panek, T. 2009, s.105-169; Witkowska. D., 2002, s.80-90]. Results of applying cluster analysis are depicted at the illustration 1.

Binding distance between16500 and 48000 was made as an assumption during the division of the Polish provinces into three homogenous groups. For the reason of the article groups were defined as follows:

- Group 1- provinces with small local real estate markets;
- Group 2- provinces with medium and big local real estate markets;
- Group 3- provinces playing vital role in the growth of the residential real estate market.

The following provinces were classified to the group 1: lubelskie, podkarpackie, podlaskie, świętokrzyskie, warmińsko-mazurskie, lubuskie, opolskie, zachodnio-pomorskie, kujawsko-

pomorskie. To the group 2: łódzkie, pomorskie, dolnośląskie, małopolskie, wielkopolskie, śląskie. To the group 3 only mazowieckie province.

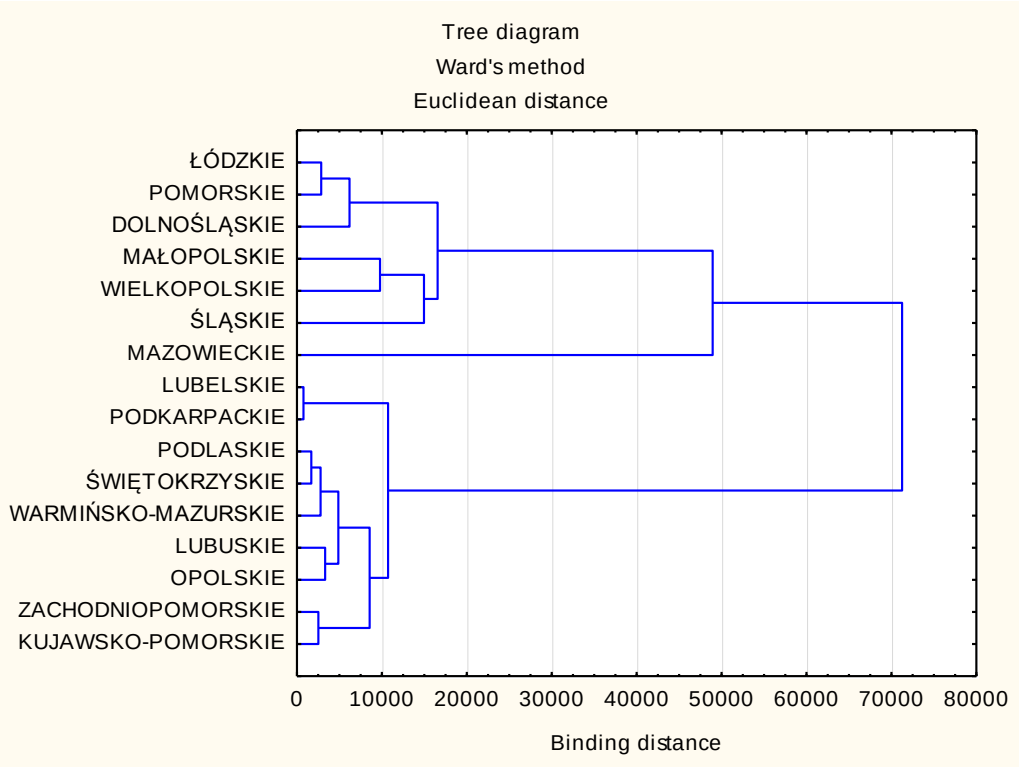

Illustration 1. The result of applying cluster analysis in order to distinguish similar provinces

3. The research process – application of logistic regression in specified groups of provinces

The first activity during the creation of logit models was checking whether outliers exist. Subsequently the aim of the research was defined: building a logistic model, which parameterizes the probability of selling the apartment according to qualities by which an apartment is characterized. Input logistic regression model was created for all groups of provinces and then defined in the formula 1:

$$P(Y=1/x_1,x_2,...,x_9)=\frac{e^{a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+a_5x_5+a_6x_6+a_7x_7+a_8x_8+a_9x_9}}{1+e^{a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+a_5x_5+a_6x_6+a_7x_7+a_8x_8+a_9x_9}} \quad (1)$$

where:

x_1 - the number of floors in the building (building with 5 floors at maximum, building with more than 5 floors);

x_2 - number of rooms in the apartment (one-room apartments, two-room apartments or three-room apartments);

x_3 - kind of the kitchen in the apartment (dark, bright, annex);

x_4 - estimation of location (bad location, average, good);

x_5 - estimation of apartment' s location in the building (not very favourable, average, favourable);

x_6 - floor, at which room is located (first floor, ground floor or the top floor, middle floor);

x_7 - the area of the apartment (apartment up to 40 m^2; apartments with area from 40 m^2 to 80 m^2, apartments with the area of more than 80 m^2);

x_8 - standard of interior completion (high standard, average, low);

x_9 - building technology (traditional, traditional improved, monolithic, prefabricated, wooden, steel frame);

$a_0, a_1, a_2, ..., a_9$ – structural parameters of the model.

After defining formal logit model (compare formula 1) the estimation procedure was conducted three times in sequence, then verification of parameters of structural models for provinces from group 1, group 2 and group 3 and results interpretation was made. Requirements referring to application of logistic regression and formal record can be found in thesis of the following authors [Dittmann, P., 2004, s.137-138; Maddala, G.S., 2008, s. 371-382].

3.1. Use of logistic regression for provinces from group 1- estimation, model verification and interpretation of result

Nine provinces were classified to the first group. The smallest one is opolskie province and the biggest one lubelskie province.[3] Shortly characterizing secondary residential real estate market we can state as follows:

- There were 4881 records in the first group, where condition of the real estate market was described with 449 selling transactions. In order to precisely classify cases in the created logistic model, the file with 899 was created (with sustainable structure of offers and transactions, where accordingly 450 offers and 449 transactions were located);

[3] Data included the actual population of the particular provinces (stated on 31th December of 2012)

- In the first group of provinces, the lowest price for the square meter was 2174 PLN and the highest 6527 PLN;
- 72 % of apartments were located at the buildings with maximum 5 floors;
- 74% of apartments had 2 or 3 rooms;
- 80% of apartments had bright kitchen;
- 39% of apartments were located on the ground floor or on the top floor.

In table 1 preliminary estimation results of logit model were presented with the use of Quasi-Newton method. The aim of the above mentioned was estimation of probability of selling residential real estate located in provinces from group one.

Table 1. Preliminary results of estimation logit model (own study)

	Const	X_1	$X_{2.}$	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Estimation	-2,07	-0,21	-0,35	0,87	-0,31	-0,27	-0,16	0,69	0,19	0,50
Standard deviation	0,13	0,12	0,17	0,37	0,29	0,15	0,13	0,12	0,13	0,11
t(110)	-15,51	-1,75	-2,06	2,37	-1,08	-1,81	-1,23	5,58	1,42	4,38
Standard p	0,00	0,08	0,04	0,02	0,28	0,07	0,22	0,00	0,16	0,00

Analyzing presented results, variables, which affect substantially for variable Y are $x_2, x_{3,} x_6, x_7, x_9$. After subsequent rejecting of variables, which are irrelevant statistically, the final version of logit model was expressed in the formula 2:

$$P(Y) = \frac{e^{-0,03-0,57x_2-0,34x_6+0,57x_7+0,70x_9}}{1+e^{-0,03-0,57x_2-0,34x_6+0,57x_7+0,70x_9}} \quad (2)$$

where:

x_2 - the number of rooms in apartment (apartments with more than 3 rooms);

x_6 - floor, where the apartment is located (ground floor or the top floor);

x_7 - the area of the apartment (apartments to 40 m^2);

x_9 - building technology (prefabricated);

Summary of the estimation process for logistic regression model are presented in table 2. In this table variables list is presented, coefficients for structural parameters, their estimation errors and the marginal error.

Table 2. Results of estimation process for group 1 of provinces significant variables (own study)

	Coefficient	*Standard deviation*	*z*	*The marginal effect*
const	-0,0372132	0,123053	-0,3024	
x2	-0,568865	0,205495	-2,7683	-0,140339
x6	-0,348923	0,144945	-2,4073	-0,0870095
x7	0,56663	0,175294	3,2325	0,139488
X9	0,698333	0,148452	4,7041	0,172085

The statistical value of p for the whole model was lower than 0,05, what confirms the relevance of the model comparing to the model with intercept only. This datum confirms of model creation purposefulness because implies the statement, that created model implements something new.Also interpretation of log-likelihood, which is the measure of fitting the whole model was made. This logarithm is calculated with the use of statistics -2 log with maximum likelihood of created model and in model with the intercept only (in the created model accordingly 1024,57 and 1098,18). At the base of the above mentioned values pseudo R^2 was calculated and it equals 0,06.

Interpreting the results (compare table 2) we can draw the following conclusions:

- probability of selling apartment with more than 3 rooms is about 0,14 lower than 1, 2 or 3-room apartments;
- probability of selling an apartment located at the ground floor or the top floor is lower about 0,09 comparing to apartments located at the other floors;
- probability of selling an apartment with less than 40 square meters is 0,14 times higher than selling apartment with more than 40 square meters;
- probability of selling apartment located in the building, which was built in prefabricated technology is about 0,17 higher than buildings built in other technologies.

In table 3 both correct and incorrect classified cases for the model were presented. The odds ratio, which equals 2,59 was calculated (ratio of quotient of correctly classified cases to quotient incorrectly classified). The value higher than unity indicates, that this classification is better than this, which will be conducted by random.

Table 3. Table relevancy (own study)

	Expected 0	Expected 1	Percent accuracy
0	168	194	46,40884
1	109	326	74,94253

3.2 The use of logistic regression for provinces in group 2- estimation, verification of models and interpretation of results

Six provinces were classified to the second group. The smallest one is pomorskie province and the biggest one śląskie province.[4] Shortly characterizing secondary residential real estate market we can state as follows.

- There were 7220 records in the second group, where condition of the real estate market was described with 415 selling transactions. In order to precisely classify cases in the created logistic model, the file with 830 records was created (with sustainable structure of offers and transactions, where accordingly 415 offers and 415 transactions were located);
- In the second group of provinces, the lowest price per square meter was 2500 PLN and the highest 12068 PLN;
- 72 % of apartments were located in buildings with maximum 5 floors;
- 77,5% of apartments had 2 or 3 rooms;
- 70% of apartments had bright kitchen;
- 41% were located on the ground floor or on the top floor;
- 34% of apartments were built in prefabricated technology.

In table 4 were shown results of the estimation of logit model with applying Quasi- Newton method. Preliminary results of logit model estiomation (own study)

Table 4. Preliminary results of logit model estiomation (own study)

	Const	X_1	$X_{2.}$	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Estimation	0,11	-0,23	-0,77	-0,42	-0,08	-0,08	0,10	0,72	1,42	0,48
Standard deviation	0,21	0,20	0,26	0,43	0,62	0,24	0,19	0,21	0,36	0,19
t(110)	0,50	-1,15	-2,92	-0,96	-0,13	-0,35	0,51	3,35	3,94	2,57
Standard p	0,62	0,25	0,00	0,34	0,89	0,73	0,61	0,00	0,00	0,01

[4] Data included the actual population of the particular provinces (stated on 31th December of 2012)

Analysing the results we can notice that variables x_2, x_7, x_8, x_9 .substantially affect for variable Y.After subsequent rejecting of variables, which are irrelevant statistically, the final version of the logit model was expressed in the formula 5

$$P(Y) = \frac{e^{-0,27-0,68x_2+0,71x_7+1,54x_8+0,68x_9}}{1+e^{-0,27-0,68x_2+0,71x_7+1,54x_8+0,68x_9}} \quad (5)$$

where:

x_2 - number of rooms in the apartment (apartments with more than 3 rooms);

x_6 - floor, where the apartment is located (ground floor or the highest floor);

x_7 - area of the apartment (floors to 40 m^2);

x_8 - standard of apartment's interior finishing (low);

x_9 - building technology (prefabricated).

Summation of the estimation process of logistic regression model was depicted in table 5. In table 5, list of variables, structural parameters coefficients, estimation error and the marginal effect are presented.

Table 5. Estimation process results for relevant variables (own study)

	Coefficient	*Standard deviation*	*z*	*The marginal effect*
const	-0,26751	0,107728	-2,4832	
X2	-0,680145	0,259629	-2,6197	-0,168284
X7	0,710262	0,195656	3,6302	0,169132
X8	1,54336	0,358479	4,3053	0,316343
X9	0,683677	0,166316	4,1107	0,165505

Likewise in the previous model, the statistical value of p for the model was under 0,05 what confirms the relevance of the model comparing to the model with intercept only. Pseudo R^2 was at the level of 0,08. Interpreting results we can withdraw the following conclusions:

- probability of selling an apartment with more than 3 rooms is lower about 0,17 comparing to one-room apartment, two-room apartment or three-room apartment;

- probability of selling an apartment with less than 40 square meters was 0,17 times higher than the apartment with more than 40 square meters;
- probability of selling an apartment with low interior finishing is 0,32 times higher than apartment with medium and high standard;
- probability of selling an apartment located in the building built with the use of prefabricated technology is about 0,17 higher than apartments build in another technologies.

In the table 6 both correct and incorrect classified cases for the model were presented. The odds ratio, which equals 2,66 was calculated (ratio of quotient of correctly classified cases to quotient incorrectly classified). The value higher than unity indicates, that this classification is better than this, which will be conducted by random.

Table 6. Table relevancy (own study)

	Expected 0	Expected 1	Percent accuracy
0	166	123	57,43945
1	132	260	66,32653

3.3. The use of logistic regression for provinces classified to group 3- estimation,

After applying cluster analysis only one province i.e. mazowieckie was classified to the third group. Defining the secondary real estate market in the third group, we can state as follows:

- 5922 of records in third group and 333 of transactions. Taking into consideration precise classification of cases during the process of building logistic model the file with 666 records was created, with accordingly 333 offers and transactions;
- In the third group of provinces, the lowest price per square meter was 3401 PLN and the highest 15000 PLN[5];
- 54% of apartments where located in the buildings with maximum 5 floors;
- 75,6% of apartments had 2 or 3 rooms;
- 60% of apartments had bright kitchen;
- 36,8% of apartments were located on the ground floor or on the top floor;
- 25% of apartments were built with the use of prefabricated technology.

In table 7 were shown results of the estimation of logit model with applying Quasi- Newton methods.

[5] Maximum value is the result of the assumption

Table 7. Preliminary results of logit model estiomation (own study)

	Const	X_1	$X_{2.}$	X_3	X_6	X_7	X_8	X_9
Estimation	-0,08	0,38	-0,12	-0,19	-0,33	0,43	0,27	-0,02
Standard deviation	0,16	0,18	0,28	0,36	0,18	0,22	0,26	0,20
t(110)	-0,53	2,12	-0,44	-0,52	-1,89	1,99	1,07	-0,11
Standard p	0,59	0,03	0,66	0,60	0,06	0,05	0,29	0,91

Analysing the results we can notice that x_1, x_7 substantially affect for variable Y. After subsequent rejecting of variables, which are irrelevant statistically, the final version of logit model was expressed in the formula 6:

$$P(Y) = \frac{e^{0,11-0,40x_1+0,48x_7}}{1+e^{0,11-0,40x_1+0,48x_7}} \tag{6}$$

where

x_1 - number of the floors in the building (building up to 5 floors);

x_7 - the area of the apartment (apartment up to 40 m^2).

Summing up the estimation process of the logistic regression model was show in table 8. In this table the list of variables, coefficients for the structural parameters, estimation errors and the marginal effect were presented.

Table 8. Estimation process results for relevant variables (own study)

	Coefficient	*Standard deviation*	*z*	*Marginal effect*
const	0,111986	0,125753	0,8905	
X1	-0,402567	0,157786	-2,5514	-0,100295
X7	0,479893	0,190048	2,5251	0,118881

The statistical value of p for the model was 0,00078 what confirms the relevance of the model comparing to the model with intercept only. What is more it is the proof , that model brings new conclusions. Pseudo R^2 was at the level of 0,02.

Interpreting data we can draw the following conclusions.

- probability of selling an apartament located in the building with maximum 5 floors is about 0,10 lower than in the building with more than 5 floors;
- probability of selling an apartament with less than 40 square meters is 0,12 times higher than selling apartments with more than 40 square meters.

In the table 9 based both correct and incorrect classified cases for the model were presented. The odds ratio was calculated (ratio of quotient of correctly classified cases to quotient incorrectly classified) as 1,44.

Table 9. Table relevancy (own study)

	Expected 0	Expected 1	The percentage of correctness
0	162	171	48,64865
1	132	201	60,36036

Summary

The first stage during attemption of estimating the probability of selling residential real estates was the division of Polish regions for relatively homogenous areas.The criterion of division was defined by the set of economic and business variables, which affect significantly for the price of square meter in the residential real estates.During the research three groups of provinces were specified (group 1, group 2 and group 3). Subsequently, basing on the application of logistic regression tools, the attempt of estimation of these three models was made. Models define the probability of selling residential real estate according to the qualities by which are characterized.In the first group of provinces, from 9 diagnostic variables, 4 variables were substantial for estimation the probability of selling: number of rooms (apartments with more than three rooms); the floor, where the apartment is located (ground floor or the top level); area of the apartment (floors till 40 m^2), standard of the apartment (low), building technology (prefabricated). In the second group, the substantial variables turned out to be number of rooms (apartments with more than three rooms); the floor, where the apartment is located (ground floor or the top floor); area of the apartment (floors up to 40 m^2), standard of the apartment (low), building technology (prefabricated). Only two variables i.e. the number of floors (buildings with maximum 5 floors), area of the apartment (floors up to 40 square meters) in the logit model were classified to third group of provinces. Each of the three logit functions (compare formula 1,2,3) enables to define the probability of selling residential stocks depending on individual qualities. Moreover

calculated marginal effects, give interpretative possibilities referring to residential stocks within provinces included to the particular group. For every estimated logit model table relevancy was appointed and the value of pseudo R^2 was calculated. Taking into consideration results from table relevancy we can state, that correctness of classification of particular models is estimated accordingly at the level of 62,0%; 62,5% and 54,5%. The disturbing fact, that should be noticed, is the very low value of coefficient pseudo R^2. The process of improvement of input data quality, which characterize secondary residential real estate market, should be continuation of the conducted research. Improvement of data quality would be at the same time improvement of measurers, which describe quality of the models (inter alia pseudo R^2). Improvement of data quality could be reached through building logistic models considering data from longer period.

Literature

1. Aczel, A.D., Statystyka w zarządzaniu, Wydawnictwo Naukowe PWN, Warszawa, 2000.
2. Dittmann P., Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie, Oficyna Ekonomiczna, Kraków, 2004.
3. Mach, Ł., *Determinanty ekonomiczno-gospodarcze oraz ich wpływ na rozwój rynku nieruchomości mieszkaniowych*, Ekonometria, 4(38)/2012, ISSN 1507-3866, s. 106-116.
4. Maddala G.S., Ekonometria, Wydawnictwo Naukowe PWN, Warszawa 2008.
5. Panek, T., Statystyczne metody wielowymiarowej analizy porównawczej, Oficyna Wydawnicza SGH w Warszawie, Warszawa, 2009.
6. Witkowska D., Sztuczne sieci neuronowe i metody statystyczne, Wydawnictwo C.H. Beck, Warszawa, 2002.