Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytetu Szczecińskiego

Folia Oeconomica Stetinensia
DOI: 10.2478/v10031-009-0017-3

THE COX PROPORTIONAL HAZARDS MODEL IN THE ANALYSIS OF PROPERTY TRANSACTIONS

Iwona Foryś, Ph.D.

Faculty of Economics and Management
University of Szczecin
Mickiewicza 64, 71-101 Szczecin
e-mail: forys@wneiz.pl

Received 20 April 2010, Accepted 8 June 2010

Abstract

The main aim of the paper is to present the methodological basis of housing transactions which uses duration analysis and Cox' proportional hazards model. The study will take into consideration single episodes with one starting point (the purchase of a flat on the resale market) and one final point (selling the flat to another owner). The analysis has covered over 1016 transactions concerning the sale and purchase of cooperative flats in 2000-2009, including repeated transactions. For each of the transactions the date, the initial and final prices as well as the basic parameters of a flat have been given. For each of the transactions the date, the initial and final prices as well as the basic parameters of a flat have been given. The study will verify the hypothesis of the high probability of the existing owner to have further *'lasting right'* to the flat, i.e. the correlation within times of event (sell flat) and the basic parameters of a flat.

Keywords: duration analysis, housing mobility, the Cox' proportional hazard model.

JEL classification: R31.

Introduction

Usually research into the property market turnout focuses on determining the apartment price on the basis of a property's features. In the constructed models the price is the resultant of its elementary attributes such as location, size or standard. There are no analyses available, however, that link the incidence of concluded deals (the probability that an apartment will be sold) with the apartment attributes. Having information about the date of the apartment purchase on the secondary market and the date of its further resale as well as about its features, we can analyze the probability of an event (the sale of the apartment) to happen and the correlation between the time the apartment owner used it and a group of variables independent of that time.

The aim of this study is to test the usefulness of the Cox proportional hazards model in evaluating the span of time between the purchase and sale of an apartment in connection with its attributes. Then, the classical analysis of phenomena occurring on the apartment market can be supplemented with actuarial methods and techniques that have been rarely used in property market analysis. Actuarial methods, which in economics are called *a duration analysis* or *a transition analysis*, in demography, medicine and biology – *a survival analysis*, in technical science – *a reliability or failure time analysis*, in sociology are defined as *an event history analysis*. The latter term (the event history analysis) or the duration analysis most adequately refer to the proposed field of application and employs statistical methods used to analyze the spans of time among a finite number of states or events that can occur in a randomly chosen moment[1].

The study covered 1016 free market sale transactions of cooperative apartments that took place between 4 May 2000 and 31 December 2009 in Spółdzielnia Mieszkaniowa in Stargard Szczeciński, including resale transactions (8.56%). The author verified a hypothesis of high likelihood of the same person's ownership right to continue further as well as of the relation between the time when an event (the sale of an apartment) occurred and the attributes of the apartment itself.

1. The elements of actuarial theories in the housing market analyses

The possible subject for considerations is that part of the housing resource that can be sold according to the current legal regulations. It concerns both the primary and the secondary market where apartments can be resold. The event history analysis is then an instrument aid-

ing studies on the transition between purchase and sale (transferring rights to another subject) on the housing market in connection with the duration of a given state. Methods (models) used in practice depend on '*the nature of the examined process and the pool of available statistical data*'[2]. Actuarial methods make it possible to analyze the duration of phenomena of asymmetrical and incomplete distribution. They include parametric, non-parametric and semi parametric models. Also, due to the number of states in a given process methods can relate to individual episodes (one initial state and one final state or, in competitive risk models, many final states) or many episodes (many initial and many final states). The subject of this study is the period of time between the start of the observation and the event which ends the observation, but first of all its likelihood in subsequent units of time. If the event does not happen by the end of the observation, the observation is terminated (a censored observation). Most commonly it is right censoring because of the time of termination.

In the adopted area of applications single episodes with one initial state (purchase of an apartment on the secondary market) and one final state (sale of the apartment to another owner) and right censoring (as for 31 December 2009) are taken into account. The time of an event incidence *t* is a random variable of non-negative values which can be described by means of a cumulative density function *F(t)*, a density function *f(t),* a survival function *S(t), a* hazard function *h(t)* of randomly chosen non-negative values and a cumulative hazard function o *H(t)* as well as a plausibility function *(L)*[3].

The measure of probability that in time $\langle 0;t \rangle$ the sale of an apartment will take place is a cumulative density function of a random variable *t* (continuous and non-negative) defined by the following formula:

$$F(t) = P(T \le t) = \int_0^t f(z)dz, \tag{1}$$

where $F(t) \in < 0;1 >$. A probability density function:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}, \Delta t > 0, \tag{2}$$

allows to estimate the empirical distribution of events in the assumed duration intervals. The function of probability that by the time t the episode ending event has not happened and the process is being continued is described as the following survival function:

$$S(t) = P(T > t) = exp\left(-\int_0^t h(z)dz \right). \tag{3}$$

The transition intensity (or hazard) rate is a hazard function described as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}, \Delta t > 0, \tag{4}$$

that provides information about failure levels. It is a characteristics of a given unit, the estimated conditioned probability (probability expressed in time units) of the event incidence in an infinitely narrow span of time $(t; t + \Delta t)$ assuming that the event ad not occurred until the interval started. The cumulative hazard function is described by the following formula:

$$H(t) = \int_0^t h(z)dz, \tag{5}$$

while the plausibility function used for single episodes is described by:

$$L = \prod_k h(t_k)^{\delta_k} \cdot S(t_k), \tag{6}$$

where δ_i – a censoring indicator of value 1 if the event occurred in the time *t* or of value 0 when information has been censored.

Popular procedures of estimating theoretical survival function are grounded on the method of least squares and on the method of least weighted squares. They are also based on fitting one of typical distributions of the exponential survival, hazard, Weibull or Gompertz functions to the empirical distribution[4].

One of the commonly used methods of estimating the survival (duration) function that do not require arbitrarily defined time variable intervals is the Kaplan-Meier method. It is based on the fact that the evaluation of probability is a product of consequent conditioned probabilities estimated individually for continuous duration intervals[5]. It should be noted that the more precise time measures are (from transition years to months or even dates), the more effective function estimators can be achieved. Similarly, the condition to obtain non-burdened estimators of survival and hazard functions is a minimum size sample that does not exceed 30 observations[6]. What is more, the introduction of a 0-1 dummy variable to the trial makes it possible to eliminate the burden resulting from selecting individual units for this trial (the variable plays the role of a cohort variable).

Duration can be analyzed with many additional factors in view and by means of regression models: parametric (e.g. the linear regression model, the log-normal regression model, the exponential model) and non-parametric (the Cox proportional hazards model). In the above models for every group distinguished due to its feature that is independent of duration (location, flat size, etc.) the survival function is estimated and pairs of the obtained functions are compared by means of non-parametric tests (survival times do not have normal

distribution). A zero hypothesis is verified that there are no general differences among many survival functions described as:

$$H_0: S_1(t) = S_2(t) = S_3(t) = ... = S_k(t) \tag{7}$$

for all *k* groups.

The most popular tests used in calculation software are: Gehan generalized Wilcoxon test, Cox-Mantel test, log-rank test, Peto and Peto generalized Wilcoxon test, F Cox test or Mantel-Haenszel test.

Supplementary to the analysis in groups is testing the event incidence trends in subsequent trials ordered according to an assumed criterion (e.g. the age of a building where the apartment is located). The impact of many features on the expected duration of an unknown survival function can be measured by means of semi parametric models, including the Cox proportional hazards model[7]:

$$h(t : x_1, x_2, ..., x_n) = h_0(t) \cdot e^{\sum_{i=1}^{n} a_i x_i} = h_0(t)\, exp(a_1 x_1 + a_2 x_2 + ... + a_n x_n) \tag{8}$$

where:

$h(t : x_1, x_2, ..., x_n)$ – the first element of the model, parametrically non-specified time function t, resultative hazard of given *n* concomitant variables $x_1, x_2, ..., x_n$ and an adequate survival time;

$h_0(t)$ – hazard function for which all the variables equal zero (base hazard);

$exp(a_1 x_1 + a_2 x_2 + ... + a_n x_n)$ – the second element of the model – a specified exponential function;

$a_1, a_2, ..., a_n$ – model coefficients;

t – observation time.

The base hazard function is a *t* time-dependent function while the second element of the Cox model depends only on time-independent variables x_i. The exponential form of this element gives non-negative values of the estimated risk. To apply the Cox model the following assumptions must be met:

- the exact shape of the dependence between the hazard ratio and time is not known;
- there is no sound theoretical basis to use parametric models;
- there is difficulty in adjusting a variable function to duration;
- the influence of variables on duration is being examined;

- the value and direction of the influence of time-dependent variables are interesting by virtue of the study aim;
- the Cox model is used as a basic scenario.

Calculating bilaterally a logarithm of the formula (8) the Cox model can be described in a linear form, which is easier to estimate:

$$\ln h(t : x_1, x_2, ..., x_n) = \ln h_0(t) + \sum_{i=1}^{n} a_i x_i \qquad (9)$$

With different values of the function $\ln h_0(t)$ we can obtain different forms of the model. The elementary method of estimating the model coefficients is the partial plausibility method, while in a popular Statistica packet the Cox model coefficients are estimated by means of the highest plausibility method[8].

Another great advantage of the Cox model is the opportunity to use many variables, both qualitative as well as quantitative.

2. Estimating the parameters of a theoretical function of apartment ownership duration

To start with, a survival function, a probability density function and a hazard function were determined for all the concluded transactions. The duration is a span of time between the time of an apartment purchase (initial state) by a present owner and the date of its sale (final state). Those of the transactions that did not end with reselling the apartment required right censoring on 31 December 2009. It is then assumed that the apartment will be sold in the future (the event will take place, but not within the time of the observation), although the researcher is not able to record the exact date of the transaction.

In order to avoid the arbitrary choice of the number of intervals and to obtain estimators that do not depend on data grouping the duration curve was estimated by means of the Kaplan-Meier method (see Figure 1). The Chart shows several phases of decrease that start on the time axis after 377 days (more than a year), then after 1 596 days (more than four years) and after 2 276 days (i.e. after over six years). Then the curve declines less sharply and its shape is determined by censored observations.

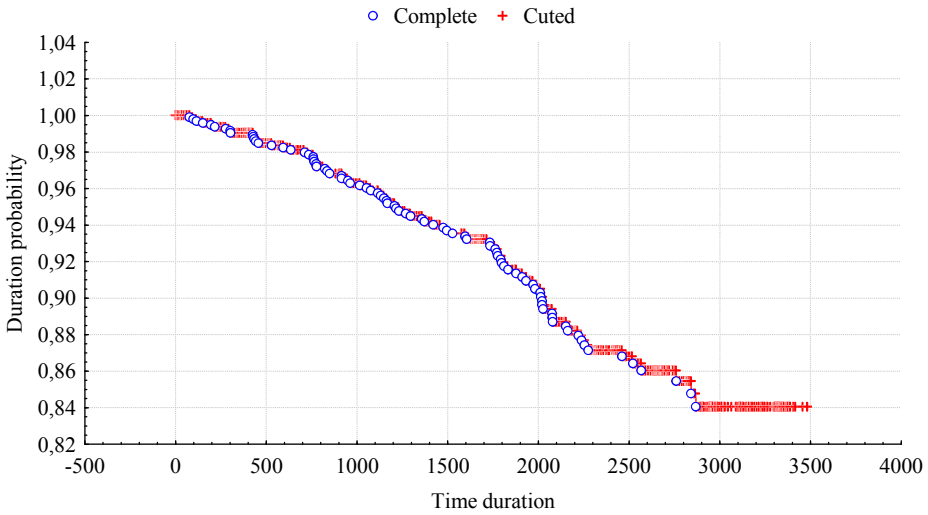

Fig. 1. Kaplan-Meier estimate of the duration function for the housing transactions (N=1016)
Source: own study on the basic of data from SM in Stargard Szczeciński.

As a second step the author introduced a variable that grouped the apartment location in one of six cooperative housing estates and she defined basic descriptive statistics for each group (see Table 1).

Table 1. Duration time descriptive statistics for the local cooperative district and sum of point Mantel statistics

District	Median	Mean	Standard deviation	Number of non truncated observations	Number of truncated observations	Sum of point
Zachód	1 695	1 640.85	969.65	24	292	2 128
Kluczewo	1 856	1 690.63	908.90	16	131	-1 230
St. Miasto	1 736	1 636.14	985.52	30	288	-3 980
Chopina	1 794.5	1 649.60	1 008.99	8	72	-1 036
Pyrzyckie	1 777	1 701.22	997.01	4	103	3 574
Letnie	1 852	1 843.56	921.02	4	44	544

Source: own study on the basic of data from SM in Stargard Szczeciński.

The obtained values of chi-square statistics (5.9881) at a relevance level of $p<0.05$ for the Mantel test, that compares selected survival curves, allow for the hypothesis that the graphs of these curves differ. For each housing estate duration functions were determined, which confirmed graphically the differences in the courses of the observed phenomenon in individual groups (the probability of selling an apartment in a given housing estate). The lowest probability was observed in Osiedle Pyrzyckie and the highest – in Stare Miasto (see Figure 2).

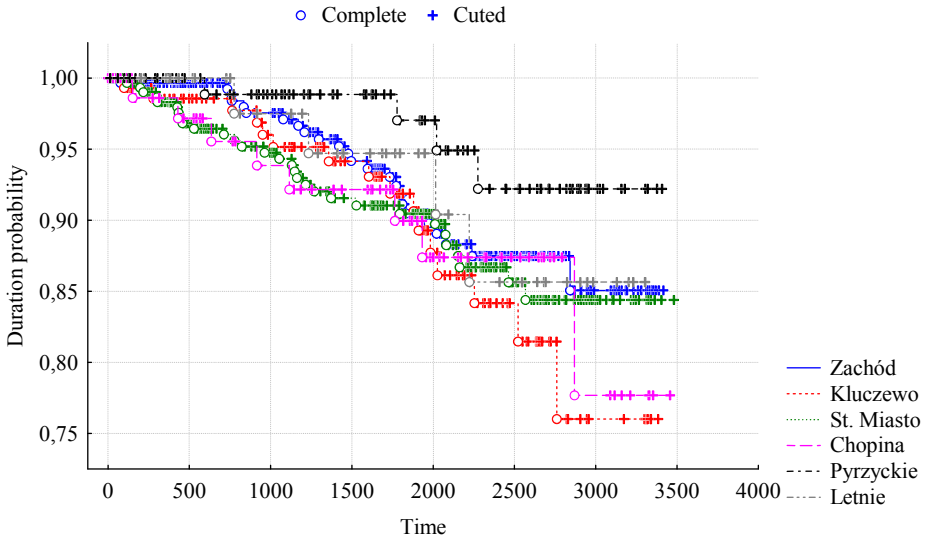

Fig. 2. Kaplan-Meier estimate of the duration function for the local cooperative district
Source: own study on the basic of data from SM in Stargard Szczeciński.

The histogram of the total of points (the points calculated for the Mantel test[9]) gave the opportunity to select those housing estate that differed significantly (see Figure 3). Two of them scored polar number of points: Osiedle Pyrzyckie (3574) and Stare Miasto (-3 980). Stare Miasto is conveniently located but its housing resources are old, not functional and small, contrary to apartments available in Osiedle Pyrzyckie for which the highest probability was calculated. It is an attractively located area with new buildings. The study also covered Kluczewo estate which was built as a part of a project to revitalize a former Soviet military area and where the highest likelihood of selling an apartment occurred after eight years. The estate is home mostly to young couples who started moving there in 1992-1998.

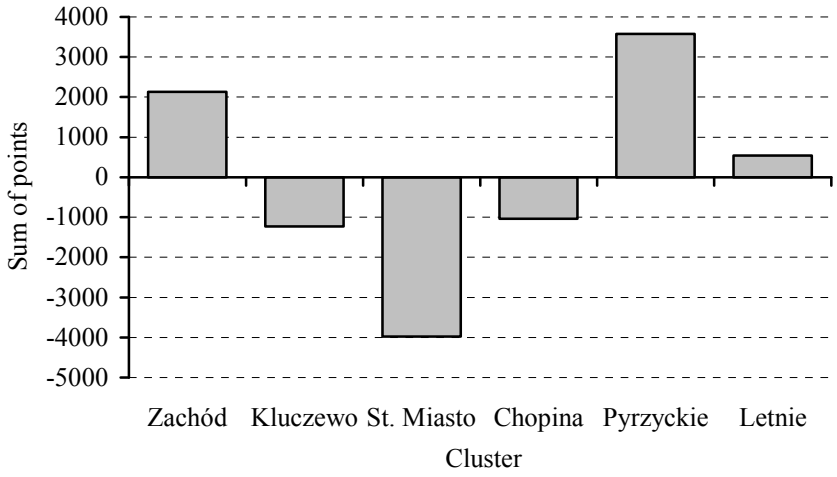

Fig. 3. Histogram sum of point for the local cooperative district
Source: own study on the basic of data from SM in Stargard Szczeciński.

The introduction of additional attributes to the analysis is a good reason to apply the Cox proportional hazards model. The variables that could have had effect on probability of selling the apartment include such commonly used attributes as: the floor it is on, its size (in m^2), number of rooms, the height of a building (a low building – no more than 5 floors, a tall one – over 5 floors) and the type of ownership (a cooperative right to the premise and an sole ownership title).

In order to decide on the relevance of the analyzed variables the author applied the single-factor Cox model with a variable grouping the location in a given estate (see Table 2). If the chi-square value is significant ($p<0.05$), we should abandon the zero hypothesis (a zero value of regression rates in population) and decide that independent variables are significantly related to duration. This condition has not been met, which means that there are no independent variables significantly related to duration. This conclusion is confirmed by the value of test *t* which evaluates statistical significance of the estimated coefficients and which is a quotient of a coefficient value divided by standard error (see Table 2).

Table 2. Listing of the One-Factor Cox' Model with Clustering Value (location on the local cooperative district)

Independent Value	Beta	Standard Error	t- Value	Models Exponent	Wald Statistics	p
Building level	-0.6466	0.3303	-1.9575	0.5238	3.8320	0.0503
Number of rooms	-0.1172	0.1399	-0.8379	0.8894	0.7021	0.4021
Space [m^2]	-0.0060	0.0097	-0.6133	0.9941	0.3761	0.5397
Building Height	0.0293	0.2978	0.0985	1.0298	0.0097	0.9215
Occupancy Form	0.9478	0.6667	1.4216	2.5799	2.0210	0.1551

Source: own study on the basic of data from SM in Stargard Szczeciński.

Values of *t* significantly higher than 2 (when is $p<0.05$) indicate that the analyzed variable affects the time when an apartment is owned by the same individual owner. But in case of each of the five independent variables the author obtained the values of statistics *t* that made her reject the hypothesis about significant influence of the analyzed features on the probability that the apartment will be sold in a specific time.

However in case of each of the five independent variables such values of the statistics *t* were obtained that the hypothesis about significant influence of the examined features on the location-based probability that the apartment will be sold in specific time must be abandoned.

The absence of the above mentioned dependence means that it is not justifiable to construct the multi-factor Cox proportional hazards model for the chosen independent variables.

Conclusions

The conducted considerations confirm the high probability of further '*duration of the apartment ownership right'* of the same owner, thus the „*continuity of inhabitance'* of households in the resources of the Housing Cooperative in Stargard Szczeciński. The observed phenomenon is the effect of low incomes earned by local households who do not decide to improve their living conditions although the existing ones are far below their actual needs (not suited to the family structure or to the age of its members). Apart from the economic factors, local residents represent different psychological and social attitudes towards mobility.

Additionally, the probability that the present ownership will last depends on the fact where exactly in the housing estate the apartment is located, which is associated with the age, technical condition and the structure of housing resources. But, contrary to a popular opinion, none of the examined apartment features affects the probability of the apartment to be sold. It means that neither its size nor the floor it is on are such significant sale incentives as it is suggested by surveys examining housing preferences.

Notes

[1] See Blossfeld, Hamerle, Mayer (1989), Frątczak, Gach-Ciepiela, Babiker (2005).

[2] Frątczak (2005).

[3] This formula see in Frątczak (2005).

[4] See Bowers, Gerber, Hivkman, Jones, Nesbitt (1986), Gerber (1990). The software of this distributions is presented in STATISTICA 8.0.

[5] Hosmer, Lemeshow (1999).

[6] Stanisz (2007).

[7] See Frączak, Gach-Ciepiela, Babiker (2005).

[8] Stanisz (2007).

[9] Stanisz (2007).

References

Bowers, N.L., Gerber, H.U., Hivkman, J.C., Jones, D.A. & Nesbitt, C.J. (1986). *Actuarial Mathematics*. The Society of Actuaries.

Blossfeld, H.P., Hamerle, A. & Mayer, K.U. (1989). *Event History Analysis, Statistical Theory and Application in the Social Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publ.

Gerber, H.U. (1990). *Life Insurance Mathematics.* Berlin Heidelberg: Springer-Verlag.

Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*. John Wiley and Sons Inc.

Frątczak, E., Gach-Ciepiela, U. & Babiker, H. (2005). *Analiza historii zdarzeń. Elementy teorii, wybrane przykłady zastosowań*. Warszawa: SGH.

Stanisz, A. (2007). *Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny*, T.3, *Analizy wielowymiarowe*. Kraków: Wydawnictwo StatSoft Polska Sp. z o.o.