Application of Hedonic Methods in Modelling Real Estate Prices in Poland

Chapter · January 2015

DOI: 10.1007/978-3-662-44983-7_44

1 author:

Anna Król
Wroclaw University of Economics
25 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Application of Hedonic Methods in Modelling Real Estate Prices in Poland

Anna Król

Wrocław University of Economics anna.krol@ue.wroc.pl

Abstract. This paper concentrates on empirical and methodological issues of the application of econometric methods to modelling real estate market. The presented hedonic analysis of apartments' prices in Wrocław is based on the dataset consisting of over 10 thousands offers from the secondary real estate market. The models estimated as the result of the research allow for pricing the appartments, as well as its characteristics.

The foundations of hedonic methods are formed by the so-called hedonic hypothesis which states that heterogeneous commodities are characterized by a set of attributes relevant both from the point of view of the customer and the producer. As a consequence, the price of a commodity is determined as an aggregate of values estimated for each significant characteristic of this commodity. The hedonic model allows to price the commodity as well as to identify and estimate the prices of respective attributes, including the prices which are not directly observable on the market. The latter is particularly useful for the real estate market as it enables pricing location-related, neighbourhood-related and structure-related characteristics of housing whose values cannot be obtained otherwise.

1 Introduction

This paper presents the results of hedonic analysis of apartments' prices in Wrocław, with special attention paid to the problems of data gathering and model specification. The primary research objective was to investigate the relationship between the price of an apartment and its significant characteristics, and to create an efficient tool for pricing the apartments on Wrocław real estate market.

Hedonic methods were invented in the early twentieth century (first attempts at hedonic analyses were made by Haas (1922), Wallace (1926), Waugh (1929) and Court (1939)), and revisited at the end of the century when the rapid progression of technology started (cf. Boskin et al. (1996), Berndt et al. (1995)). The sudden development of production engineering and the emergance of completely new, advanced technologies caused acceleration in changes of the quality of goods. This shortening of products' life cycles have led to the situation in which the goods present on the market were not comparable with the goods whose prices were observed in the past. As a result, standard methods of price changes measurement have yielded biased results, usually an overestimation of price growth rates.

The problem of quantification of the so called ,,true" price change affects two classes of heterogeneous goods:

- goods which undergo very rapid technological development (e.g. consumer electronics, household appliances, cars, IT and ICT devices),
- goods which are strictly heterogeneous. For heterogeneous goods it is highly unlikely to encounter two identical specimens whose prices could be compared without the risk of quality-related biases (e.g. apartments, houses, land parcels).

For the first class of goods the difficulties in comparing prices from periods t and $t+1$ arise either from a significant change in characteristics of a commodity (e.g. when a given computer is endowed with the hard drive of a considerably larger capacity), or from the final withdrawal of a certain good from the market (e.g. withdrawal of obsolete, dangerous, or otherwise needless good whose price has been measured in the period t), or from development of completely new technology (e.g. replacement of CD drives by DVD drives). In the class of strictly heterogeneous goods the price of a good observed in the period t may only be compared with the price in period $t+1$ of a ,,similar" good. Because of the above mentioned fact the problem of quality difference is immanent feature of this measurement process.

Other application areas of hedonic methods include (cf. Dziechciarz (2004), Triplett (1986)) facilitating the development of a pricing strategy, pricing the commodities, as well as estimating the prices of respective attributes, including the prices which are not directly observable on the market. The latter is particularly useful for the real estate market (cf. Sheppard (1999), Can (1992)), as it enables pricing location-related, neighbourhood-related and structure-related characteristics of housing whose values cannot be obtained otherwise (e.g. the distance from city centre to the apartment, the quality of air in the vicinity of the housing, the age of the building).

2 Hedonic models

2.1 Basic concepts

The foundations of hedonic methods are formed by the so-called hedonic hypothesis which states that heterogeneous commodities are characterized by a set of relatively homogeneous attributes (characteristics) relevant both from the point of view of the customer and the producer (cf. Brachinger (2002), Triplett (2006)). For a given good described by m attributes this set may be formally written as a vector $\boldsymbol{X}$:

$$\boldsymbol{X}^T = [x_1 \;\; x_2 \;\; \ldots \;\; x_m] \,. \tag{1}$$

Moreover it is assumed that there exists a relationship between the price of the good and its significant characteristics which may be described by a certain function f, called the hedonic function. As a consequence, the price of a commodity is determined as an aggregate of values estimated for each significant characteristic of this commodity. A commonly used method for obtaining the hedonic function is the application of regression model described in the following general notation:

$$P = f\left(\boldsymbol{X}; \boldsymbol{\beta}; \epsilon\right), \tag{2}$$

where P is the commodity price, $\boldsymbol{\beta}$ a vector of parameters, and ϵ the error term.

The estimate of the vector of parameters $\boldsymbol{\beta}$, obtained by estimation of a correctly specified hedonic regression model using a data set, allows to calculate the theoretical price of a given good with a specified set of significant characteristics. This property of hedonic models is crucial for their application in revealing the ,,true" price change because of a change in characteristics. In general, implementation of hedonic methods to make adjustments for quality changes consists in incorporating results obtained from hedonic regressions into classic Laspeyres, Paasche and Fischer price index formulas.

2.2 Selected problems in hedonic modelling

Hedonic regression models face all the classical problems of econometric modelling of cross-sectional data, such as heteroskedasticity of the error term, collinearity of the independent variables or spatial autocorrelation (cf. Greene (2011), Wooldridge (2010)). In this section two crucial issues for hedonic modelling will be briefly addressed: model specification and data collection.

Specification problems

The specification problems in hedonic modeling comprise the following three aspects:

1. specifying correctly the class of good,
2. correctly determining the set of characteristics,
3. getting the right functional form.

In the theory of hedonic model the class of good signifies all the variants of the given commodity which may be correctly described by a common hedonic function. In practice it is not always easy to correctly define the class of the commodity. For example, it is very probable that the group of commodities specified as ,,apartments in Wrocław" in its majority is comprised by flats of size 50-100m^2 and price 150-500 thousands of PLN. However, it is almost certain that in that group much bigger (e.g. 300m^2), and much more luxurious (e.g. 1 500 thousands of PLN) units occur as well. Such heterogeneity of goods might by difficult to capture by the same hedonic function, causing the indispensability of a compromise between the accuracy of the estimated hedonic model and the generality of obtained results.

The number and type of characteristics which comprise the vector $\boldsymbol{X}$ depend on the nature of a given good and its technical properties. The determination of significant attributes may be based on technological information concerning production processes, marketing data on the needs and preferences of the consumers, as well as the use of statistical information obtained from the data set. Frequently, however, the choice of independent variables is limited by the availability of data. Therefore, the possibility biases caused by omitted variables must be taken into account.

The theory of hedonic methods gives little suggestions as to how the relationship between the price of a good and its characteristics should be specified. Many studies present either the a priori assumption of the functional form of hedonic regression, or the approach of using the functional form which fits the data best (for the survey of hedonic empirical research on real estate market see e.g. Herath (2010)). The commonly used functional forms of hedonic regressions are linear ($P = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \epsilon$), exponential (log-lin) ($\ln P = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \epsilon$), double-log (log-log) ($\ln P = \beta_0 + \sum_{j=1}^{m} \beta_j \ln X_j + \epsilon$) and logarithmic (lin-log) ($P = \beta_0 + \sum_{j=1}^{m} \beta_j \ln X_j + \epsilon$). A convinient metod for choosing functional form of hedonic regression is Box-Cox transformation (cf. Box et al. (1964)) of the dependent variable (a similar transformation may be applied for the independent variables of the model leading to further extensions of the set of the potential functional forms):

$$B\left(P_i, \lambda\right) = \begin{cases} \frac{P_i^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln P_i & \text{for } \lambda = 0 \end{cases} \tag{3}$$

The Box-Cox method consists in comparing the whole spectrum of potential functional forms of the model for various values of parameter λ and choosing the one which yields the highest value of likelihood function. This procedure allows to consider a large family of functional forms, including linear (for $\lambda = 1$) and exponential (for $\lambda = 0$) approaches.

Data requirements

The discussed research problem is not new, however it has not been sufficiently fathomed, especially from the Polish perspective. The main reasons and possible explanations of such unsatisfactory level of research in the hedonic method

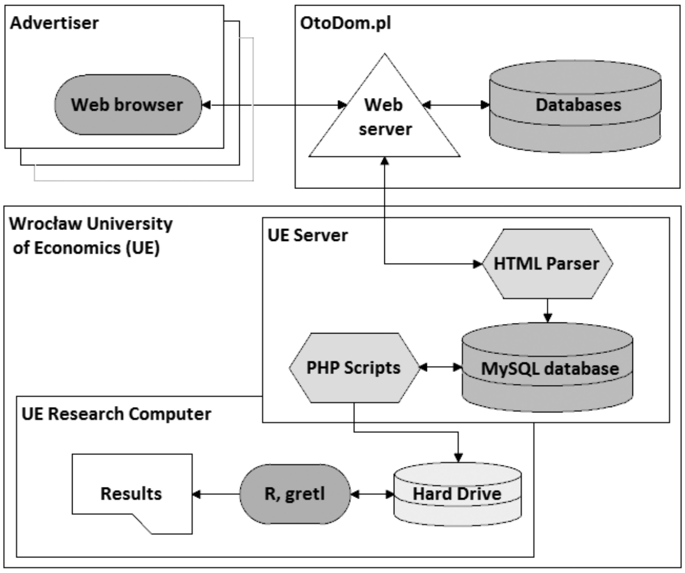

Fig. 1. Infrastructure for data gathering and analysis

area are on the one hand the complexity and difficulty of the related issues, and on the other shortage of suitable databases. The demands on the data for the purposes of hedonic analysis are numerous. The data should consist of large number of observations on prices of real estates and sufficiently extensive set of characteristics. Since such datasets are difficult to obtain, for this research special infrastructure designed for data collection was developed. The schematic diagram of the infrastructure is presented in the Figure 1.

In essence, the collected data came from Internet site `www.OtoDom.pl`, which is one of the biggest Polish advertising site for both individual and institutional real estate sellers. Each advertiser for the price of small fee can place sale offer, which is then stored in `www.OtoDom.pl` databases and visible for the potential buyers on the website. Since the direct access to the databases is impossible, special tool (called HTML Parser) was designed. The tool itself is a set of mutually related PHP scripts integrated with SQL database, and is using PHP framework *Simple HTML DOM Parser* written by S.C. Chen (`http://simplehtmldom.sourceforge.net/`). In order to gather the data the user has to determine the specifics of the chosen commodity. In case of real estates the user defines the type of the real estate (e.g. houses, apartments, parcels), the location (e.g. region, city, district) and the type of the market (secondary or primary). Basing on the information provided by the user the programme from all the offers published on `www.OtoDom.pl` website chooses only the suitable advertisments. Subsequently necessary for the reasearch data is collected and stored in the SQL database. Afterwards, another PHP script transfers the data in the desired by the user format to the research computer, ready to be processed and analysed. The main disadventage of such data source is the fact that the collected prices are offer prices, and not transaction prices.

3 Description of the data set

Using the above mentioned infrastructure for data collection extensive dataset on apartments from the secondary real estate market in Wrocław has been collected. The database consists of 13 920 offers of apartments for sale advertised in the year 2011, including 3 340 offers from Fabryczna district (D1), 1 525 from Psie Pole district (D2), 2 349 from Śródmieście district (D3), 1 768 from Stare Miasto district (D4) and 4 938 from Krzyki district (D5). The information gathered included such variables as price, size, address (from which, using geolocation, distance to the city centre was calculated), number of rooms, building age, number of floors in the building as well as the floor on which the apartment was located. Descriptions and basic summary statistics for those variables are given in the Table 1.

Table 1. Variables description and summary statistics (NA - the percentage of unavailable observations)

Variable name	Description	Mean	Min	Max	NA
PRICE	Apartment price [PLN]	366 620	110 000	676 000	0.00%
SIZE	Apartment size [m^2]	56.32	16.00	105.20	0.00%
DISTANCE	Distance to city centre [km]	3.87	0.2	13.23	21.12%
NOROOM	No. of rooms	2,51	0	5	0.00%
YEAR	Year of the building	1990.5	1867	2012	61.38%
FLOOR	Apartment floor	2.74	0	14	0.00%
NOFLOOR	No. of floors in building	4.50	0	14	0.00%

The second part of the information collected concerned additional features of the apartments, such as: balcony, garage, terrace, separate kitchen, garden, basement, whether the apartment is located on the ground floor, whether there is a lift in the building etc. Such characteristics may be incorporated into the model in the form od dummy variables, taking the value of 1 if an apartment posseses given feature, and 0 otherwise. Table 2 presents only some of the collected dummy variables, namely those which were the most common in the dataset.

The described characteristics may be divided into two groups: structure-related (such as size, number of rooms, building age, number of floors in the building, additional features of the apartment), and location-related (distance to city centre, city district).

Table 2. Dummy variables description and percentages of their occurrence in dataset

Variable name	Description	Percentage
GRFLOOR	Dummy for ground floor	14.50%
BALCONY	Dummy for balcony	39.40%
GARAGE	Dummy for garage	7.15%
BASEMENT	Dummy for basement	19.19%
GARDEN	Dummy for garden	0.91%
TERRACE	Dummy for terrace	1.83%
LIFT	Dummy for lift	5.77%
TLEVELS	Dummy for two levels	0.92%
SKITCH	Dummy for separate kitchen	23.38%

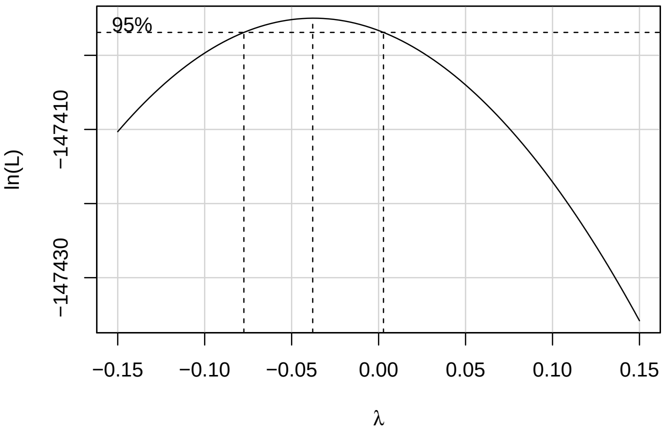

Fig. 2. The values of log-likelihood function for various λ

4 Research steps and results

In order to determine the correct functional relationship f for the model 2, the Box-Cox transformation of the dependent variable was performed. The results are presented in the Figure 2, where the values of log-likelihood function for various values of parameter λ are compared.

The highest value of log-likelihood function equal to -147 395 was found for $\lambda = -0.03788$. However, the 95% confidence interval included $\lambda = 0$ as well, which indicates the correctness of the exponential (log-lin) function (to be more specific, a mixed approach was applied, as one of the independent variables - *SIZE* - was transformed using the logarithm function as well).

The model $\ln P = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \epsilon$ was estimated using Ordinary Least Squares method, and since the presence of heteroskedasticity was detected (White's test statistics $LM = 796.92$) re-estimated using White's Weighted Least Squares method. The procedure involves OLS estimation of the model of interest, followed by an auxiliary regression to generate an estimate of the error variance, then finally weighted least squares, using as weight the reciprocal of the estimated variance (cf. White (1980)). Afterwards, using the results of of t-test, the insignificant variables were rejected, and the best models were chosen with the help of AIC and SC information criteria, when necessary. For both models the Variance Inflation Factors for each variable did not exceed 3, indicating lack of serious collinearity problems. The highest correlation coefficients were observed for the pairs (*DISTANCE*, *D1_FAB*) about 0.41, and (*DISTANCE*, *D4_SM*) about -0.4. The results of OLS estimation (model (1) W_OLS) and WLS estimation (model (2) W_WLS) are presented in the table 3.

Both models were estimated using 9 565 observations (some parts of the initial data were discarded as outliers, and some were missing). The differences in parameters estimates are in most cases small, however in view of the presence of heteroskedasticity the results of model (2) will be further analysed. All the variables are statistically significant, and the signs of the parameters are consistent with the expectations. The prices of the apartments in Wrocław are influenced by their size: 1% increase in the size of the apartment result in a 0.72% increase in the apartment's price, ceteris paribus (moreover the parameter for the variable *NOROOM* is positive and significant). Another factor is the location of the apartment within the building. The more floors the building has and the higher the apartment is located (except the ground floor), the lower the price. Furthermore, the apartments located on the ground floor are significantly cheaper. The presence of additional features significantly increases the apartment price. The price of a flat ednowed with garage is about 8.6% higher ($(exp\,(\beta) - 1) \cdot 100\%$); a balcony and terrace increase the price by 3.3% and 10% respectively. The last factor is the location of the apartment in the city - the apartments closer to the city center are more expensive than similar apartments in the city outskirts.

One of the research hypothesis was that the central districs of Wrocław (namely D4 and D3) along with the southern Krzyki district (D5) are more expensive locations than the remaining two districts. The obtained results support this hypothesis. The location of Wrocław's districts from the most to the least expensive is presented in the Figure 3. The apartments in the city centre (Stare miasto (D4)) are the most costly and the succession of the other districts is as follows: (D5) Krzyki district (similar in size and other characteristics apartment located in Krzyki is cheaper by almost 11%), (D3) Śródmieście district (cheaper by over 14%), (D1) Fabryczna district (cheaper by about 15.5%) and (D2) Psie Pole district (cheaper by approximately 17%).

The fit of the model is relatively high with adjusted R^2 equal to about 78%. In order to additionaly evaluate the model performance, 1% of the initial

Table 3. Estimation results for models of Wrocław apartments' prices (t-ratios in parenthesis)

ln PRICE	(1) W_OLS	(2) W_WLS
n	9565	9565
Adj. R^2	0.7543	0.7820
constant	9.971***	10.02***
	(447.8)	(408.8)
ln SIZE	0.7392***	0.7243***
	(122.8)	(101.7)
DISTANCE	−0.004947***	−0.006605***
	(-6.315)	(-9.444)
NOROOM	0.007440***	0.01267***
	(4.311)	(4.986)
FLOOR	−0.002482***	−0.001170*
	(-2.979)	(-1.819)
GRFLOOR	−0.01300***	−0.01363***
	(-2.694)	(-2.98)
NOFLOOR	−0.006149***	−0.007044***
	(-9.515)	(-13.63)
BALCONY	0.03519***	0.03228***
	(11.38)	(11.63)
GARAGE	0.08477***	0.08279***
	(14.87)	(14.35)
TERRACE	0.1019***	0.09540***
	(9.423)	(10.07)
TLEVELS	0.04822***	0.05756***
	(3.292)	(3.911)
SKITCH	−0.008729**	−0.009316***
	(-2.553)	(-3.091)
D1_FAB	−0.1694***	−0.1607***
	(-26.66)	(-23.01)
D2_PP	−0.1871***	−0.1791***
	(-24.49)	(-22.41)
D3_SROD	−0.1525***	−0.1509***
	(-26.24)	(-20.74)
D5_KRZ	−0.1158***	−0.1178***
	(-21.43)	(-18.07)

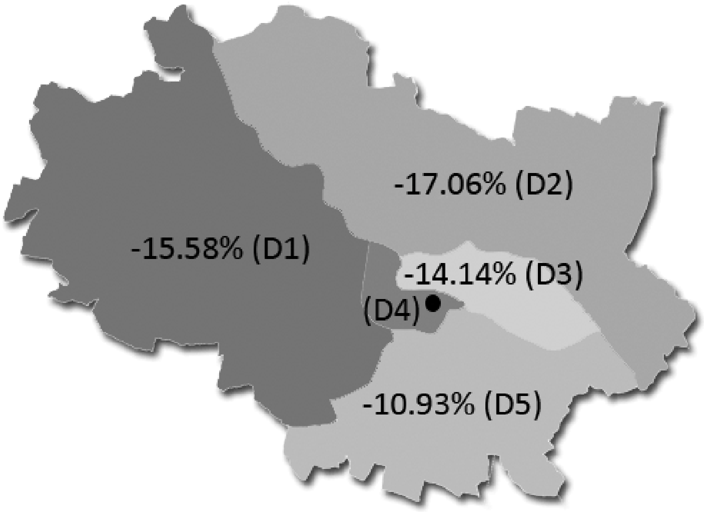

Fig. 3. Differences in apartments' prices in Wrocław districts: Fabryczna (D1), Psie pole (D2), Śródmieście (D3) and Krzyki (D5) in comparison to Stare miasto (D4).

dataset was not used in the estimation process and was intended for the out-of-sample testing. Only 3% of the obtained forecasts were outside of the 95% confidence intervals, and the following values of error measures were reported RMSE = 51 632, MAE = 37 500, MAPE = 9.90%.

5 Final remarks

The study attempted the estimation of a hedonic model for the secondary apartment market in Wrocław. The obtained results provide the means for pricing apartments on the basis of their most significant characteristics. Moreover, hedonic prices for particular characteristics were estimated, including the prices which are not directly observable on the market (such as distance to the city centre, location in a given district).

The estimated model is relatively well-behaved and yields acceptable forecasts, however additional research could provide improvements in its predictive power. Further research directions should include:

- estimating the models with more location and neighbourhood-related variables (e.g. distance to green areas, communication facilities),
- missing data handling (imputation mechanisms),
- investigating in greater detail the problem of outliers and influential observations (e.g. DFFITS (cf. Belsley et al. (1980)), robust methods),

- researching usefulness of obtained results in creating quality-adjusted price indices.

References

BELSLEY, D.A., KUH, E. and WELSH, R.E. (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

BERNDT, E.R., GRILICHES, Z. and RAPPAPORT, N.J. (1995): Econometric Estimates of Prices Indexes for Personal Computers in the 1990s. *Journal of Econometrics, 68(1), 243–268.*

BOSKIN, M.J., DULBERGER, E.R. and GRILICHES, Z. (1996): *Toward a More Accurate Measure of the Cost of Living: Final Report.* Diane Publishing Co., Washington.

BOX, G.E.P. and COX, D.R. (1964): An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252.*

BRACHINGER, H.W. (2002): *Statistical Theory of Hedonic Price Indices.* DQE Working Papers, 1, Department of Quantitative Economics, University of Freiburg/Fribourg, Switzerland.

CAN, A. (1992): Specification and Estimation of Hedonic Housing Price Models. *Regional Science and Urban Economics, 22 (3), 453–474.*

COURT, A. (1939): Hedonic Price Indexes with Automotive Examples. *The Dynamics of Automobile Demand, General Motors Corporation, New York, 99–117.*

DZIECHCIARZ, J. (2004): Regresja Hedoniczna. Próba Wskazania Obszarów Stosowalności. In: A. Zelias (Ed.): *Przestrzenno-czasowe Modelowanie i Prognozowanie Zjawisk Gospodarczych.* Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 163–175.

GREENE, W.H. (2011): *Econometric Analysis, 7th edition.* Prentice Hall, New Jersey.

HAAS, G.C. (1922): *Sale Prices as a Basis for Farm Land Appraisal.* Technical Bulletin 9, University of Minnesota, Agricultural Experiment Station.

HERATH, S. and MAIER, G. (2010): *The Hedonic Price Method in Real Estate and Housing Market Research. A Review of the Literature.* SRE - Discussion Papers, 2010/03, WU Vienna University of Economics and Business, Vienna.

SHEPPARD, S. (1999): Hedonic Analysis of Housing Markets. In: P.C. Cheshire and E.S. Mills (Eds.): *Handbook of Regional and Urban Economics.* Elsevier, 3, 1595–1635.

TRIPLETT, J. (2006): *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes.* OECD Directorate for Science, Technology and Industry, OECD Publishing, Paris.

TRIPLETT, J. (1986): The Economic Interpretation of Hedonic Methods. *Survey of Current Business, 36(1), 36–40.*

WALLACE, H.A. (1926): Comparative Farmland Values in Iowa. *The Journal of Land and Public Utility Economics, 2 (4), 385–392.*

WAUGH, F.V. (1929): *Quality as a Determinant of Vegetable Prices.* Columbia University Press, New York.

WhHITE, H. (1980): A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica, 48 (4), 817–838.*

WOOLDRIDGE, J.M. (2010): *Econometric Analysis of Cross Section and Panel Data, 2nd edition.* The MIT Press, Cambridge.