Hedonic Founded Cleaning of the Estimated Property Value in the Micro Survey Panel on Household Finances by Linear Stochastic Imputation

Christoph Johannes Weißer*

April 28, 2013

Abstract

This paper develops a cleaning process for the variable of the estimated property price (hb0900) in the micro survey Panel on Household Finances. A prediction equation that takes regional differences into account is estimated while using a hedonic foundation for the explanatory variables. Furthermore for the estimation of a FGLS regression the heteroscedasticity is modeled by assuming different abilities of individuals to estimate the correct property price. The prediction equation is used to detect potential outliers, while analysing the residuals, leverage and further outlier statistics. During the editing process the relationship between explanatory and explained variables are used to find correct values. When this is not possible linear stochastic imputation is applied for an stochastic improvement, so that the distribution of the variable remains unbiased. The process used is very efficient to clean the variable of the estimated property price and to detect also excentrical observation in the explanatory variables.[1]

JEL CLASSIFICATION: C35, C53, C81, D31, D84, D1, E3, L85

KEYWORDS: Panel on Household Finances (PHF), Household Finance and Consumption Survey (HFCS), Data cleansing; Property prices; FGLS regression; Editing; Stochastic imputation

*Deutsche Bundesbank Research Centre, Georg-August-Universität Göttingen, Email: kontakt@christoph-weisser.de

[1]The estimations are done during an internship in the Research Centre of the Bundesbank. For the research a non finally edited and anonymized version of the Panel on Household Finances dataset was used, which is only available at the Research Centre of the Bundesbank and differs from the anonymized PHF data that can be requested for research projects (http://www.bundesbank.de/phf-data). Some of the variables used are not public available or only in a roughly form given data protection reasons.

Contents

List of Abbreviations

- AIC - Akaike information criterion
- asinh -inverse hyperbolic sine
- BIC - Bayesian information criterion
- c.p. - ceteris paribus
- CAPI - computer-aided personal interview
- E(X) - expected value
- exp - exponential function
- GDP - Gross domestic product
- GUM - general unrestricted model
- HFCN - Household Finance and Consumption Network
- HFCS - Household Finance and Consumption Survey
- hj - leverage point for observation j
- I - identity matrix
- INFAS - Institut für angewandte Sozialwissenschaften
- ln - natural logarithm
- $N_{1-\alpha}$ - (1-α) quantile of the standard normal distribution
- OLS - Ordinary Least Square
- PHF - Panel on Household Finances
- r - residual
- $r_i^{stand.}$ - standardized residuals
- SCF - Survey of Consumer Finances
- sinh - hyperbolic sine

- sd - standard deviation
- SSE - explained sum of squares
- SST - total sum of squares
- u - vector with residuals
- Var - variance
- VIF - Variance-influence factor
- X - covariates matrix
- X - random variable[2]
- y - vector with observation
- y_i^{imp} - imputed value
- α - confidence level
- β - OLS estimator
- ε - random drawn number from a standard normal distribution
- μ - sample mean
- Ω - error covariance matrix
- ρ - correlation
- Σ - covariance matrix
- σ - standard deviation

[2]Depending on the context it is obvious whether X is the covariates matrix or a random variable.

List of Figures

List of Tables

1 Introduction

To make good economic policy information about the economy are very important and thus economic data of high quality. In many cases the micro distribution of economic variables and not just aggregated macro data is needed. This is in particular true for financial stability analysis, where for instance the debt to income ratio is crucial to assess in how far households are able to repay their mortgages. Households with relatively high mortgages and payment obligations compared to their disposable income are more likely to default. Thus the tails of the debt and income distribution of the variables and their relationship on household level are essential for a policy maker to analyse financial stability questions. From the distributions risk measures like the Value-at-Risk (VaR) can be derived.[3] The Panel on Household Finances which is a very detailed micro data set provides the data for such analysis for households in Germany. However, the quality of any analysis and the reliability of the conclusion drawn from the data depends on the reliability of the data. Thus procedures are needed to find mistakes in the data and predict values for missing observations.[4]
In this thesis a procedure for the improvement of the property values in the micro survey Panel on Household Finances of the Deutsche Bundesbank is developed and the empirical results are shown.[5] Econometric techniques and hedonic theory will be applied to see how well a prediction equation for property values can be estimated by using the available resources of the data. It will be pointed out which of the used methods are helpful and in how far the variable of the property value can be improved. In the beginning a brief overview of the Panel of Household Finances will be given and the particular advantages of micro data will be pointed out. The estimated property price will be described with regard to the particular importance of the property value that justifies the extensive analysis in this paper used to improve the quality of this variable. In the following the cleaning process will be developed for which a prediction equation for the property values with a large explanatory power has to be estimated. As common in the literature hedonic theory is used as the theoretical foundation for the estimation of a first property price prediction equation. The prediction equation with the hedonic variables in the dataset don't deliver a sufficient explanation of the variance of the property values. Therefore the model is enhanced by further variables in order to increase the explanatory power of the model. The coefficients in the model are only reliable when

[3] See von Kalckreuth et al. (2012), p. 2-3.
[4] See ECB (2008), p. 1-4.
[5] The research is based on data work during a 6 month internship with the Research Centre of the Deutsche Bundesbank in the Household Finance and Consumption team.

the Ordinary Least Square (OLS) assumptions are fulfilled. Thus tests for heteroscedasticity and multicollinearity are applied. To deal with heteroscedasticity a transformation of the data with the inverse hyperbolic sine is used, which is a alternative transformation that has several advantages over the common transformations concerning wealth data.[6] Different approaches for the selection of variables that are not explained by hedonic theory will be used, including specification tests and statistics for model comparison. For the determination of the property values especially the location is important.[7] Assumptions about property price differences in Germany will be made and included in the model while using the resources of the data set as good as possible. Following behavioural economics it will be assumed that households have different abilities to estimate the correct property price and that their ability can be explained by their characteristics given by variables in the data. A model for heteroscedasticity based on these assumptions will be estimated and used to estimate a Feasible General Least Square (FGLS) regression that can be used in the case of heteroscedasticity to receive more efficient estimators and give households with biased estimations less weight in the regression estimation.[8] The resulting FGLS model will be used for the identification of outliers while analysing residuals and using further statistics for the identification of outliers whereby also statistics for the identification of outliers in the explanatory variables will be used. For the identified observations the property value will be compared with the covariates to see if the value is unrealistic or whether there might be an unrealistic value in one of the explanatory variables. With that it might be possible to identify the true value of the property value due to the relationship between the variables. If this is not possible the values will be replaced with a stochastic imputation procedure, which has the goal to contain the structure of the data. Changes in the distribution of the property value after the clearing procedure in particular concerning the detected outliers will be evaluated to see whether it was possible to improve the quality data, so that it can be used as a more reliable source for economic research and policy.[9]

[6] See Burbidge et al. (1988), p. 123-124.

[7] See Kiel/Zabel (2008), p. 175-178.

[8] See Greene (2002), p. 209-210.

[9] The empirical analysis is done with STATA and some parts of the code are provided in appendix 4. Also the corresponding log.file is provided with this thesis. To work with the data was only possible in the Research Centre of the Deutsche Bundesbank, given that the data is not public available at the moment and especially sensitive data can be only accessed in the Deutsche Bundesbank. Thus it was necessary take more results with me than are actually needed wherefore the do. file is larger than necessary and it will be referred to the important parts of the do.file in the text. The comments in the do.file are written in German.

2 The survey and the variable of the estimated property price

The survey Panel on Household Finances (PHF) and its goal to reflect the balance sheet of a household correctly will be briefly described.[10] Furthermore advantages of micro data and possible mistakes in the data will be pointed out. After that the importance of the variable of the estimated property value for economic policy will be described and it will be analysed in how far the PHF variable can be used for the estimation of a hedonic model.

2.1 The survey Panel on Household Finances

The Panel on Household Finances is a new panel survey carried out by the Deutsche Bundesbank. It collects micro data on household finances and wealth in Germany. The PHF wants to reflect the balance sheet of a household correctly whereby also information about income, saving and consumption as well as socioeconomic and demographic characteristics of the household are collected.[11] The PHF is designed as a full panel which means that all participating households will be contacted again in the following waves. The data worked with in this paper is from the first wave which was carried out between September 2010 and July 2011. The PHF is part of the Household Finance and Consumption Network (HFCN) which is a collaboration between the national central banks in the euro area and the European Central Bank (ECB) as well as several statistical institutes. The HFCN conducts the Household Finance and Consumption Survey (HFCS), which is a system of national wealth surveys that collects micro data in every country of the euro area.[12] The questionnaires of the different national surveys are not directly comparable, however harmonised output variables, which are called core variables are provided by all participating countries so that comparable data is received.[13] The HFCS might be comparable to the long established Survey of Consumer Finances (SCF), which is a household survey conducted by the Federal Reserve Board since 1983 every three years.[14] The micro data of the PHF is collected during face to face interviews by interviewers from the Institut für angewandte Sozialwissenschaften (INFAS). Thereby the person with the best knowledge about the finances of the household as a

[10] See von Kalckreuth et al. (2012), p. 1.
[11] The household balance sheet covered in the PHF is shown in appendix 2, figure 7.
[12] See von Kalckreuth et al. (2012), p. 1.
[13] See Deutsche Bundesbank (2012a), p. 30.
[14] See Bledsoe/Fries (2002), p. 1.

whole is interviewed. The questionnaire for the interview is programmed so that the questions are adjusted to the already given answers of the household. This is called a computer-aided personal interview (CAPI). Furthermore specific data is collected for every household member older than 16.[15] Thus very detailed micro data for the household is collected.

2.2 The advantages of micro data and mistakes in the data

Micro data has several advantages compared to aggregated macro data given that information are measured for one household simultaneously and this allows to understand structural relationships which can't be analysed on a higher level of aggregation. The representative or average household causes a substantial loss of information and limits the scope of the empirical analysis.[16] In many cases the representative household is not helpful for the understanding of the impact of monetary policy and exogenous shocks on saving and consumption. The HFCS is conducted as a panel wherefore it has the special advantage that developments on a micro level over time can be analysed. Von Kalckreuth et al. point out that central banks have two major reasons to collect micro data. The financial behaviour and condition of household have major implication for the development of the economy. Micro data is important for the understanding of individual behaviour, like saving, or what determines the ownership of house.[17] In particular the saving rate has a key role for long term economic growth according to common growth models like the Harrad-Domar or Solow growth models.[18] The distributions that can be received from micro data are especially crucial concerning the net wealth of households. For financial stability the tails of distributions are important and less the averages given that highly indebted households with low income are the ones which have problems to pay their obligations so that a credit default is more likely for them.[19] For a central bank this is important since a increase of the interest rate will also increase the payment obligations of households with mortgages with flexible interest rates and thus increase their default probability. However, the recorded information during the interview might deviate from the desired one due to various reasons. For instance the question can be wrongly interpreted and understood by the interviewed household member or the interviewer can record the answer incorrectly.[20] Possible

[15] See von Kalckreuth et al. (2012), p. 7.
[16] See Cameron/Trivedi (2005), p. 6.
[17] See von Kalckreuth et al. (2012), p. 2-3.
[18] See Blanchard/Illing (2009), p. 338-345.
[19] See von Kalckreuth et al. (2012), p. 2-3.
[20] See Bledsoe/Fries (2002), p. 1.

mistakes are pointed out in a cognitive model by Sander et al. (1992).[21] Thus to receive unbiased estimates the data has to be reviewed and the quality of the data has to be optimised. Therefore comparable to the approach of the Survey of Consumer Finances the Deutsche Bundesbank has an own team that reviews the data to correct mistakes in the dataset with editing techniques and imputation to improve the quality of the data.[22] Like in the SCF[23] in the PHF most problematic values are identified by logical and consistency checks, and only a smaller part results from outlier and influence plots due to a complex estimation procedures as it is done for the variable of the estimated property value in this thesis.[24]

2.3 The main dependent variable of the estimated property value

To analyse the net wealth of households the property wealth of the household and its corresponding financing is especially important given that the property makes a large part of household wealth. The housing crises in the USA or in Spain have shown the importance of the property markets for financial stability and the whole economy.[25] Furthermore given that property is an important part of the household wealth, property price changes have wealth effects that influence the consumption and saving behaviour of the households through different channels and thus have substantial effect on the whole economy, whereby again micro data is essentially important to quantify these effects correctly.[26] In the case that the property is not inherited or received as a gift buying a property is an incentive for the household to go into debt and thus take mortgages as the corresponding financing on the passive side of the household balance sheet.[27] Thereby the loans for property in 2007 are with 47 percent of the Gross domestic product (GDP) in the euro zone the main liability of households.[28] Therefore the property variables and their financing are very important for the net wealth of the household which distribution is essentially important for financial stability and thus monetary policy as pointed out in the last chapter. Given the importance of property and its financing it is necessary to have particular high quality data concerning them, since the advantages of micro data are of course only available when the data is reliable wherefore the property value is

[21] See Sander el al. (1992), p.818-823.
[22] See von Kalckreuth et al. (2012), p. 14-15.
[23] See Bledsoe/Fries (2002), p. 2.
[24] See von Kalckreuth et al. (2012), p. 15.
[25] See De-Bandt et al. (2010), p. vii-viii.
[26] See Carrol et al. (2010), p. 5-20.
[27] See von Kalckreuth et al. (2012), p. 18-20.
[28] See ECB (2009), p. 12.

subject to the following extensive cleaning procedure. The variable of the estimated property value in the PHF is received by the following question that asks for an estimate of the owner: "What is the current value of this property, including plots of land, if you could sell it now, how much do you think would be the price of it?".[29] Further possibilities to receive estimated values of the property are basically tax assessments and transaction prices. The estimate of the owner is due to the fact the value can be easily received by a question in a survey the most widely available source. Owner estimates can be find in many surveys like the American Housing Survey, Survey of Income and Program Participation, Panel Study of Income Dynamics and the Survey of Consumer Finances.[30] The estimated property prices will be used as the dependent variable in a hedonic model, however within the hedonic theory market prices are used.[31] Thus it is important to know how accurate the owner estimates are with regard to market prices. In this thesis the estimated property prices can't be compared with transaction data, so that only findings in the literature can be considered. Goodman and Ittner compare estimated property prices and sales prices with data from 1985 and 1987 from the American Housing Survey.[32] They find that owner estimates are upwardly bias, since the average owner overestimates the sale price by 6 percent. Above this they find a large variance of owners estimates, so that the bias and variance combined results in a 14 percent absolute error with regard to the sales price.[33] Kiel et al. (1999) find some problems in the approach of Goodman and Ittner that they try to correct. They find that the average owner over estimates his property by 5.1 percent.[34] Despite of these findings it will be assumed that the estimated property in the PHF survey is a good approximation for market prices and can be thus used within the hedonic context. In the following it will be referred to the property value instead to the estimated property value given that this is more simple since the term estimated will be used very often.

[29] The PHF questionnaire is not numbered wherefore it is not possible to cite the questions appropriately.

[30] See Goodman/Ittner (1992), p. 339-340.

[31] See Baranzini et al. (2008), p. 20.

[32] Studies before the paper of Goodman and Ittner don't compare the estimated property value to transaction prices, so that they will be not considered.

[33] See Goodman/Ittner (1992), p. 340-343.

[34] See Kiel/Zabel (1999), p. 292.

3 Hedonic property price model

The hedonic pricing theory can be used as theoretical foundation for the estimation of models for property prices. Thus an introduction to hedonic theory and its basic assumptions and implications will be given with regard to the explanation of property values. Since statistics for model comparison will be needed in the following they will be introduced. After that variables of the PHF dataset will be used to estimate a hedonic founded OLS regression model with the property price as the dependent variable. Based on this an empirical analysis follows that will assess the explanatory power of the hedonic model.

3.1 Theoretical foundation

Hedonic analysis is widely used in the context of property markets.[35] The hedonic theory is basically developed by Kelvin Lancaster (1966)[36], who developed a consumer theory concerning the demand of heterogeneous commodities with valuable attributes.[37] Lancaster's microeconomic theory was further developed by Rosen (1974)[38] which allowed it to formalize empirical models. The model developed by Rosen is accepted until now as the leading hedonic model.[39] The following theoretical outline gives only the basic idea of the hedonic theory after Rosen without giving a detailed microeconomic foundation. The following outline of Rosen's theory follows the overview given by Witte et al. (1979). To take the heterogeneity of properties into account many studies have analysed houses in hedonic terms. In the hedonic context a property is not seen as a homogeneous good but as a bundle of its components and attributes $H = (h_1, h_2, ..., h_k)$ that contribute to the housing service. The price of the components are jointly determined by the bid and offer functions of consumers and producers of the components of the house. Rosen defines a bid function for each household, which gives the maximum amount that the household is willing to pay for varying attributes of the house, which means for different bundles $H = (h_1, h_2, ..., h_k)$. The bid Θ of the household is also affected by the household income (y) and a vector α that contains characteristics that determine the taste of the household. The bid function is with that formally given as: $\Theta = \Theta(h_1, h_2, ..., h_k, y, \alpha)$. It is

[35] See Can (1992), p. 454.
[36] See Lancaster (1966), p. 132-157.
[37] See Pozo (2006), p. 3.
[38] See Rosen (1974), p. 34-55.
[39] See Pozo (2006), p. 3.

assumed that the components of the bundle are normal goods,[40] which means that the demand for the good increases when the income increases.[41] Furthermore it is assumed that the utility functions for the goods are concave. With that it can be concluded that when a component of the bundle increases the bid function Θ will increase but with a decreasing rate, so that for the first and second partial derivative the following holds: $\frac{\partial \Theta}{\partial h_i} > 0$ and $\frac{\partial^2 \Theta}{\partial^2 h_i} < 0$. The implicit bid price of the household for a component h_i can be derived from the first derivative, whereby it is assumed that the implicit bit price for the component decrease with an increased consume of h_i. The producer side is defined in a corresponding way. For a firm a offer function Φ is defined for the bundle $H = (h_1, h_2, ..., h_k)$ that the firm produces. Φ takes the minimum unit price that the firm is willing to accept for different bundles. It is assumed that the market is competitive and that the firm maximise its profit. The offer price Φ depends also on the output level of the firm (M) and parameters of the production function as well as factor prices, which are both given by the vector β. With that the offer function is given as: $\Phi = \Phi(h_1, h_2, ..., h_k, M, \beta)$. It is assumed that the profit function of the firms is convex, Φ should be constant or increase while one of the components increases. The implicit offer price can be derived from the partial derivative that increases or is constant for an increase of a component: $\frac{\partial \Phi}{\partial h_i} \geq 0$. The equilibrium of the market is defined as the tangency of the offer and bid functions for the property bundles, so that market supply equal market demand: $Q^s(H) = Q^D(H)$. Furthermore the k markets for the components also have to be in equilibrium, so that: $Q^s(h_i) = Q^D(h_i)$. Thus k markets for the k components must be in equilibrium. To solve this 2k equations have to simultaneously solved.[42] In the literature hedonic models are often estimated with an OLS regression model or a maximum likelihood estimation.[43] The OLS approach will be used with the property value as the dependent variable and the implicit prices of the components as the estimated coefficients of the explanatory variables that are the components and attributes of the house. This fits with the hedonic theory outlined above that the price of the good is given as a linear combination of the characteristics of the property.[44] In the hedonic theory market prices are used, whereas the property price in our variable is the result of a subjective estimation of the interviewed person that has the best knowledge about financial questions in the household. As pointed out in chapter 2.3 we assume that the estimated property value is a good approximation for the market value of the property.

[40] See Witte et al. (1979), p. 1151-1152.
[41] See Varian (2006), p. 96.
[42] See Witte et al. (1979), p. 1151-1152.
[43] See Baranzini et al. (2008), p. 21.
[44] See Bazyl (2009), p. 3.

3.2 Statistics for regression model comparison

In the following chapters statistics for the comparison of the estimated models will be needed. For this purpose four statistics that are commonly used for model comparison will be briefly described. The further tests and statistics used in the paper are pointed out when they are needed in the analysis. The statistics can be used to compare OLS regression models which are estimated as $y_i = X\beta + u_i$ with (y_i) as the dependent variable, X as the matrix of the covariates, β as coefficients that has to be estimated, u_i as the residual component and $\widehat{y} = X\widehat{\beta}$ as the conditional expected values.[45] The coefficient of determination (R^2) is defined as the ratio of the explained sum of squares $SSE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ to the total sum of squares $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$, whereby $\overline{y}$ is the average of the the dependent variable.[46]

$$R^2 = \frac{SSE}{SST} \tag{1}$$

Thus the R^2 is a measure for the explained variation compared to the total variation, so that it can be interpreted as the fraction of the sample variation in the dependent variable that is explained by the explanatory variables X.[47] The problem with the R^2 is that it increases always when further variables are included in the model. Given that we want to concentrate on important variables and keep the model simple, the number of parameters should also considered in a statistic for model comparison. The following statistics incorporate also the amount of parameters used. The adjusted R^2 can be derived from the R^2 by the following equation, with K as the number of regressors, including the content term and thus n-K as the degree of freedom.

$$R^2_{adjusted} = 1 - \frac{n-1}{n-K}(1-R^2) \tag{2}$$

If a further variable is included into the regression the adjusted R^2 can decline and also be negative. Thereby the adjusted R^2 rises when the contribution of a new variable is larger than the loss due to the inclusion of further parameters, which let to a decline of the degree of freedom.[48]

[45] See Heeringa et al. (2010), p. 180-182.
[46] There is also another definition of the R^2 that will be used and explained in chapter 4.
[47] See Wooldridge (2009), p. 38-40.
[48] See Greene (2002), p. 35.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) can also be used for model comparison.[49] The AIC statistic was developed by Akaike (1973) and is given by the following equation for least square estimations where the error term is normally distributed.[50] The distribution of the error term for most of the following regression is approximately normally distributed, as will be shown later. As before n is the number of observations, K is the number of parameters in the model and $\widehat{\sigma}_z^{\,2} = \frac{\Sigma \widehat{u}_i^{\,2}}{n}$ is the estimated variance of the residuals.

$$AIC_\sigma = \ln \widehat{\sigma}_z^{\,2} + \frac{K}{n} 2 \tag{3}$$

The BIC criterion was developed by Scharz (1978) and is given by the following equation:

$$BIC_\sigma = \ln \widehat{\sigma}_z^{\,2} + \frac{K}{n} \ln(n). \tag{4}$$

As it can be seen the AIC and the BIC include a penalty term, so that the statistics increase with the number of regressors. The statistics increase also with a larger variation of the residuals. Thus lower AIC and BIC are preferred.[51] The BIC criterion can be received by a multiplication of the second term of the AIC statistic with $\frac{1}{2}\ln(n)$. Thus they are very close for 8 ($\ln(8) \approx 2$) observations and differ essentially for large samples.[52] In large samples a reduction of the residual variance might be easier possible, wherefore the BIC statistic might be the better statistic for larger samples. Comparing the statistics for price estimations with Monte Carlo Simulations the BIC statistic outperforms the AIC criterion for larger samples (150 observations).[53] Given that we have more than 1500 observations the BIC criterion might be thus more reliable, whereby we will take all model comparison statistics outlined above for the model selection into account.

[49] See Akaike (1974), p. 716-723.
[50] See Anderson/Burnham (2004), p. 268.
[51] See Verbeek (2004), p. 58-59.
[52] See Schwarz (1978), p. 463.
[53] See De-Graft Acquah (2010), p. 1-5

3.3 Estimation of a hedonic model

Hedonic theory offers a developed theoretical framework, whereas the empirical results are criticized that they are underidentified and that the empirical results come from arbitrary functional form assumptions.[54] Criterions for selecting the best functional form are not pointed out by Rosen and in the further literature.[55] The functional form is usually determined by model selection criterions for the best fit.[56] Thus the following estimation of the model will be basically based on empirical interference and less on economic reasoning. Following the hedonic pricing theory, variables are chosen for the hedonic model, from which the households make use of. As common in the literature an OLS regression is used to estimate an empirical model, with (y_i) as the property values (hb0900)[57] and X as the matrix of the covariates and u_i as the residual component. The coefficients (β) in the regression model can be interpreted as implicit prices which give multiplied with the variables and aggregated the value of the property.[58]

$$y_i = X\beta + u_i \tag{5}$$

We use variables in the dataset that measures the size of the residence (SIZE) as well as the size of piece of land (LAND) which belongs directly to the property, since the household is asked to incorporate the value of this land in the price of the property so that it should be included into the model. Furthermore we argue that the age of the property might explain some utility that the household gains, whereby the relationship can't be easily described. The dataset offers only a variable that gives the time since the property acquisition to account for the age of the building. Of course this variable is a bad proxy for age but it will be included nevertheless to analyse the coefficient. The variable of the time since the property acquisition (*TIME since acq.*) is calculated by subtraction of the variable that gives the year of property acquisition (hb0700) in the questionnaire from the time of data collection, which is approximated with 2010 (hb0700_2=2010-hb0700). The following tables gives an overview of the variables. The corresponding questions and different categories for categorial variables are pointed out in the appendix 3. The description of the variables in the table are simplified in the sense that the aspects of the question are pointed out which are

[54] See Ekeland et al. (2001), p. 1.
[55] See Pozo (2006), p. 3.
[56] See Radcliffs (1984), p. 81.
[57] The variable names in brackets are the names of the variables in the STATA files.
[58] See Baranzini et al. (2008), p. 20-22.

essential for understanding the corresponding variable. The last column gives the name of the variable used in the STATA files and the questionnaire.

Table 1: Hedonic explanatory variables

Variable Name	Description	STATA Variable Name
SIZE	Size of household main residence (m^2)	hb0100
LAND	Land belonging directly to the property	dhb0151
TIME since acq.	Years since the property acquisition	hb0700_2

Source: Own table

The questionnaire of the PHF survey is designed for being programmed, since for the data collection a programme is used where the questions are dynamic fitted at the answers that the household gave before. This is called computer aided personal interview (CAPI).[59] Furthermore several categorial variables are included in the model, which take numbers according to several categories that are related to the property in the sense that they describe the property and its surrounding. Therefore they characterise the property and can be thus regarded as hedonic variables. We assume that the household make use of a property in a better shape and a better surrounding. For the model for every category a dummy variable is created, which takes the value 1 if the category is chosen and 0 if not. Thus for a categorial variable that has for instance five different categories five dummy variables are created.[60] Generally the first category, which is also the first dummy in the regression reflects the best possible rating and the last dummy for the categorial variable reflects the worst conditions. The first dummy variable is dropped automatically by STATA. The hedonic categorial variables are shown in the following table.

[59] See Deutsche Bundesbank (2012a), p. 32.

[60] The dummy variables are created by writing *xi:* before the *reg* command in STATA.

Table 2: Hedonic categorial explanatory variables

Variable Name	Description	STATA Variable Name
BUILDING TYPE	Type of dwelling	dhb0100
DWELLING RATE	Rating of the building	sc0200
DWELLING LOCATION	Rating of the location	sc0300
DWELLING OUTWARD	Rating of the outward appearance	sc0400
NEIGHBOURHOOD	Compared to the neighbourhood	sc0500
SURROUNDING	Rating of residential area	sc0600
INTERIOR	Interior conditions of the building	hr0200
FEDERAL STATE	Different countries in Germany	bland
POPULATION DENSITY	Different populations density classes	bik

Source: Own table

For property price estimations especially the location is important.[61] Therefore besides the variables *DWELLING LOCATION*, *NEIGHBOURHOOD* and *SURROUNDING* also dummy variables for every *FEDERAL STATE* and a variable for the *POPULATION DENSITY* are included. The variable *POPULATION DENSITY* is developed by the BIK Aschupurwis and Behrens Gmbh and classifies local communities into different categories (BIK-Regionsgrößenklassen) that give the population density of the functional areas into which the local communities belong. Functional areas are defined as space that belongs together due to functional aspects.[62] The *POPULATION DENSITY* thus reflects the population density in the surrounding area of the property. Given that a larger population lets to an increased demand for property, whereby the supply is especially restricted in cities due to limited space, this categorial variable should explain a part of the price differences between properties. The household might profit from living in a local community that belongs functionally to an area with a larger population given that this might offer more opportunities to choose between goods supplied and a larger and more diversified labor demand. However, also cultural or educational opportunities should be better within a larger functional space, so that the household can take also more use of this.[63] The variable has 10 different categories for the *POPULATION DENSITY* in

[61] See Kiel/Zabel (2008), p. 175-178.
[62] Functional areas are explained in the BIK (2001) paper.
[63] See BIK (2001), p. 10.

a ascending order. Thus 10 dummy variables are created, whereby the last one is dropped as the reference category.[64] The regression results can be seen in appendix 1 in the second and third column of table 15 (OLS models). The regression in the second column (hedonic) show a regression with all variables pointed out above and the regression in the third column (hed. significant) contains only the significant variables after the following tests. Only the very insignificant variables are dropped, which are the *DWELLING LOCATION*, *DWELLING OUTWARD*, *NEIGHBOURHOOD* and *INTERIOR* concerning the categorial variables according to a F-test for the simulation significance of the dummy variables. The F-test tests the null hypothesis that the dummy variables are simulation zero. Under the null hypothesis the F statistic is $F = \frac{(SST-SSE)/(K-1)}{SSE/(T-K))} \sim F_{(K-1,T-K)}$ distributed.[65] We receive p-values larger than 0.1 for the named variables, so that the null hypothesis can't be rejected for the usual confidence levels (0.01, 0.05, 0.1). The t-test tests the null hypothesis $H_0 : \beta_i = 0$ against the alternative hypothesis $H_A : \beta_i \neq 0$. The test statistic follows a t distribution $t = \frac{\widehat{\beta}}{se(\widehat{\beta})} \sim t_{(T-K)}$ under the null hypothesis.[66] According to the t-test the variable LAND as well as the TIME since acq. are not significant. The regression model with all hedonic variables has an $R^2_{adj.}$ of 0.356 and a BIC statistic of 43941. The regression model hed. significant without the variables that we dropped after the tests above has a $R^2_{adj.}$ of 0.350 and an BIC statistic of 50121. The model selection statistics for the model with and without the insignificant variables are close which underlines that their explanatory power is low. The interpretation of the coefficients is not the focus of this thesis wherefore they will be only briefly analysed. The coefficient of the variable SIZE is positive as expected. The interpretation of the coefficient is that when the size of the property increases by one m^2 c.p., the value of the property price increases by €1239. The other variables can be interpreted in the same way. The dummy variables of the *DWELLING RATE* are larger for properties in a better shape which makes also sense. This can be also seen for the variable *SURROUNDING*. The dummy variables include level effects in the model. Thus the interpretation of the *DWELLING RATE* is that if the property is for instance rated as in a very good shape the property price is c.p. €58734 smaller than the value of a property that is rated as exclusive since the dummy variable exclusive is dropped and is thus the reference category. The other dummy variables are interpreted as in the given example. Concerning the *FEDERAL STATE* the dummy variables for East Germany are negative, which fits with the expectation that there should be a price differential

[64] P is the number of the population in the area.
[65] See Hill et al. (2001), p. 173-176.
[66] See Hill et al. (2001), p. 99-100.

between the former East and West Germany. However, some of the coefficients are not significant. The *POPULATION DENSITY* dummy variables are highly significant, whereby an essential positive price effect for functional areas with a larger population density can be seen only for the last two categories.

There are several papers with hedonic regressions for property prices that receive higher R^2 and more significant parameters, whereby these papers have further hedonic variables concerning the properties of the property. Canavarro et al. estimate a hedonic model for new apartments with an R^2 of 0.81[67] and Bazyl estimates an R^2 of 0.56.[68] Verbeck estimates an R^2 of 0.68, while having highly significant coefficients.[69] Can estimates a hedonic model while using spatial statistics. The coefficients in the model are very significant and the explanatory power is with an adjusted R^2 of 0.76 very high.[70] The hedonic model with best results concerning the significance of the coefficient and an adjusted R^2 of 0.83 is estimated by Pozo.[71] A more advanced model by Dubin that includes also non hedonic variables like for instance the number of cars belonging to the household estimate an adjusted R^2 of 0.73.[72]

The explanatory power measured by the R^2 and the significance of the coefficients of this first model are not sufficient for using this regression model for data cleaning purposes. If we assume that hedonic variables in our model explain the property prices the model should explain a larger amount of the variance in the dependent variable. Thus there should be some missing hedonic variables in the model. However, to the knowledge of the authors there are no further hedonic variables in the dataset that can be used in the regression. Given that there must be missing hedonic variables in the model, because of the low explanatory power the adding of further variables will also increase the reliability of the hedonic coefficients which should be biased due to the omission of relevant variables in the model. The consequences of the omitted variable bias will be considered in detail in chapter 4.2.4.

[67] See Canavarro et al. (2010), p. 5-6.
[68] See Bazyl (2009), p. 12.
[69] See Verbeck (2004), p. 65-68.
[70] See Can (1992), p. 464.
[71] See Pozo (2006), p. 9.
[72] See Dubin (1998), p. 49-50.

4 Estimation of a prediction equation

The hedonic variables in our model are explained by economic theory. Nevertheless the variance of the property value is not explained very well so that the predicted values can't be used for an approximation of the property value. However, better estimation results are constitutive for the cleaning of the variable. Thus further variables which might explain the property price due to economic relationships apart from hedonic theory will be included into the model. For this purpose an extensive econometric analysis is required. After a description of the further variables, the model will be tested for heteroscedasticity and the data will be transformed. Furthermore it will be controlled for multicollinearity and the omitted variable bias will be considered in the model estimation strategy.

4.1 Further explanatory variables

The PHF dataset contains a large amount of variables about the wealth, income, saving, consumption and socioeconomic variables of the household which can be used to explain the property value apart from the hedonic approach. Besides the statistical approach used for the selection of variables, the variables should be also selected while using economic arguments. Statistical inference never gives certain arguments since there is always a probability of making incorrect conclusion like for instance when we reject the null hypothesis that the coefficient is zero while the null hypothesis is in fact true (type 1 error). Thus without taking economic aspects into account it is indeed possible to select regressors that are significant even though in reality no relationship with the explained variable exists. For the following variables economic arguments will be given but their inclusion in the model can't be justified in the same way with economic theory, as it was done for the hedonic founded variables. Thus we rely more on statistical arguments for the selection of variables. Therefore the probability to make incorrect choices have to be minimised by applying many econometric tests and taking the features of the data into consideration to select the important variables of the following outlined variables.[73] As common in microeconometric regression the disaggregation of the data lets to increased heterogeneity. Thus several variables have to be aggregated to one variable to increase their explanatory power but also to deal with multicollinearity what will be explained in

[73] See Verbeck (2004), p. 56.

detail in chapter 4.2.3.[74] In the following the variables used in the further models and analysis will be outlined, whereby for the aggregated variables the components are outlined in appendix 3.[75] The variables can be differentiated into variables that describe the household wealth, yearly flow variables which refer to the household income as well as socioeconomic variables and further categorial variables. The first variable is the *PURCHASING PRICE* of the property which should be obviously correlated with the current property value. The further three household wealth variables describe different components of the household wealth on the active side of the household balance sheet. The variable *MORTGAGES* belongs to the passive side of the household balance sheet and contains mortgages with the main property of the household as a collateral. The mortgages, which reflect a part of the financing of the property should be correlated with the current price. The mortgages of the other property are not included because they are correlated with the other property that is included as an explanatory variable and thus would let to a multicollinearity problem. It can be argued that the *OTHER PROPERTY* which the household owns, as well as the value of his *CARS* and his other *FINANCIAL WEALTH* are related to the value of its main property, since a wealthy household might have more expansive car, larger financial wealth and maybe also further expansive property compared to a less wealthy household. Thereby we will use all outlined wealth variables for the estimation to capture different components of the wealth, since it might be for instance possible that a household owns no expansive car but has large financial wealth. The *PURCHASING PRICE* of the property should be correlated with the named wealth variables and thus a potential cause of a multicollinearity problem.

[74] See Cameron/Trivedi (2005), p. 5-8.

[75] The STATA code for the aggregation of the variables is programmed by Dr. Tobias Schmidt from the Research Centre of the Deutsche Bundesbank and can be received upon request. The note min indicates that the lower bound is used for the aggregation of the variable in the case that the household provides an interval instead of one value.

Table 3: Further variables: Household wealth

Variable Name	Description	STATA Variable Name
PURCHASING PRICE	Value at the time of its acquisition	hb0800
OTHER PROPERTY	Aggregated value of further property	immoson_{min}
CARS	Aggregated value of cars	kfz_{min}
FINANCIAL WEALTH	Aggregated value of financial wealth	finv_{min}
MORTGAGES	Aggregated value of mortgages	hyp

Source: Own table

Table 4: Further variables: Flow variables

Variable Name	Description	STATA Variable Name
INCOME	Aggregated total income	tincome_{min}
SAVING	Aggregated total savings	spara_{min}
CONSUME	Aggregated total consume	taus_{min}

Source: Own table

The flow variables outlined in the table above are important for the further analysis, whereby only the *INCOME* will be included into the regression, while arguing that households with a higher income tend to have more expansive property. The flow variables are measured in monthly units. For the estimation of housing models usually the current income (Y) is divided into permanent income (Y^P) and transitory income (Y^T). The permanent income is then modeled as a function of human capital assets which can be modeled with variables like education and age, so that the permanent income can be explained as the fitted values of a regression of these variables on Y and the transitory income as the residual component.[76] However, given that the focus of this thesis is more on the prediction quality and we are less interested in the interpretation of the coefficients the *INCOME* will be included without to use the differentiation above. Given that the household might inherited or received the property as gift, which is the fact for about 24 percent of the household in the survey, the household might own an expansive property but have a low income.[77] The amount

[76] See Goodman (1988), p. 331-332.
[77] See von Kalckreuth et al. (2012), p. 27.

of household *MEMBERS* should be related with the value of the property, but also in particular with the *SIZE* which is included as an explanatory variable. Thus we have to control for this in the multicollinearity section. For the property value 1848 observations exists. If there are missing values in the explanatory variables of an observation, the observation is not used within the regression so that the amount of observations is reduced. This has to be also considered during the selection of explanatory variables. A lot of households in the dataset are without children, thus using the amount of *CHILDREN* in a regression would let to a significant reduction of observations used for the regression wherefore the number of household *MEMBERS* is the better variable for the regression.

Table 5: Further variables: Socioeconomic variables

Variable Name	Description	STATA Variable Name
MEMBERS	Number of household members	anzhhm
CHILDREN	Number of children	dpe1275

Source: Own table

Table 6: Further variables: Categorial variables

Variable Name	Description	STATA Variable Name
MEANS acquisition	Means of property acquisition	dhb0410
EDUCATION	Highest level of education	dpa0300
EDUCATION prof.	Highest level of professional education	dpa0400
EMPLOYED	Employed at the moment	dpa0500

Source: Own table

The *MEANS* of the property acquisition as well as the *EDUCATION* and employment status given by the variable *EMPLOYED* of the household member with the best financial knowledge, which is the one who is interviewed might be also correlated with the value of the property.[78] The results for the OLS regression with the variables named above can be seen in the table 15 (OLS models) in the first column (OLS all) in appendix 1.[79] The R^2 increases to 0.61 and the adjusted

[78] Again the education should be correlated with other explanatory variables like the income or the variables, so that we have to control for this.

[79] The histogram and quantile plots of the residuals can be find in appendix 2, figure 8-10.

R^2 to 0.59. The AIC and BIC take values about 4100 and are thus smaller than the values that we received for the hedonic regression. The coefficients differ from the hedonic model and are as well as the coefficients in the hedonic model not very reliable without further tests. However, the improvement of the model selection statistics shows that some of the selected variables should be included into the model, which will be done in the further model specification.

4.2 Model specifications

The model specifications with the further variables and the hedonic model have to be analysed with several econometric tests, given that the results of the OLS estimation are not reliable when OLS assumptions are not fulfilled. In the following we will first concentrate on heteroscedasticity. To deal with heteroscedasticity and other essential problems of the data a transformation of the variables with the inverse hyperbolic sine will be applied. For the transformed data the variables in the model will be selected with regard to different model specification approaches while taking multicollinearity and the omitted variable bias into consideration. Furthermore non linear relationships between the explanatory variables and the property value will be included in the model.

4.2.1 Heteroscedasticity

An assumption for OLS regression is that the variance of the residuals conditional on X is constant and that the residuals are uncorrelated, which means that $Var(\varepsilon_i|X) = \sigma^2$ for i=1,...,n and $Cov(\varepsilon_i, \varepsilon_j|X) = 0$ for all i $\neq$ j. The assumption of content variance is called homoscedasticity. The case when the variance of the residuals depends on the covariates so that $Var(\varepsilon_i|X) = \sigma_i^2$ for i=1,...,n is called heteroscedasticity. Survey data with household data is often heteroscedastic.[80] Given the lower level of aggregation that reduces the heterogeneity of the data, heteroscedasticity is typical for microdata. For instance heteroscedasticity can be caused by an increased variation of the property values with an increasing size of the property or income of the household.[81] Thus we have to control for heteroscedasticity.[82] Heteroscedasticity doesn't let to inconsistency or a bias of the OLS estimator, but let to a bias of the estimators of the variances of the coefficients Var($\widehat{\beta}_j$). Thus the

[80] See Greene (2002), p. 15.

[81] See Cameron/Trivedi (2005), p. 5-8.

[82] A scatterplot of the property price against the size of the property for the untransformed and in the following transformed data can be find in appendix 2, figure 11-12.

standard error of the coefficients which are directly derived from them can't be used to calculate reliable t-statistics for the interference of the significance of the coefficients.[83] Linear forms of heteroscedasticity can be detected by the Breusch Pagan test which was developed by Breusch and Pagan (1979).[84] The null hypothesis is that the residuals are homoscedastic: $H_0 = Var(\varepsilon_i|X) = \sigma^2$. Given the OLS assumption that the error term has a zero conditional expectation $E(\varepsilon_i|X = 0)$ the null hypothesis can be also written as $H_0 = E(\varepsilon_i^2|X) = \sigma^2$. To test this the squared residuals of the regression model are regressed on the explanatory variables in an auxiliary regression: $\widehat{u}^2 = X\delta + v$. The F-test for $H_0 : \delta_1 = ...\delta_n = 0$, which should be the case for homoscedastic residuals tests for the joint significance of the explanatory variables regressed on the squared residuals. If they are jointly significant the null hypothesis can be rejected, which is a strong indicator for heteroscedasticity. Under the assumption that the null hypothesis holds the F statistic is F(k,n-k-1) distributed.

$$F = \frac{R^2_{\widehat{u}^2}/k}{(1 - R^2_{\widehat{u}^2})/(n-k-1)} \tag{6}$$

However, there is also another version to test the null hypothesis above, by computing the LM statistic by the multiplication of n and $\widehat{u}^2$: $LM = n * \widehat{u}^2$. The LM statistic is asymptotically X^2_k distributed.[85] The LM version of the test can be directly calculated with the STATA command *estat hettest*. This test assumes that the error term is normally distributed which is the case for most of the used regressions as will be shown later.[86] Another test, which test also for nonlinear forms of heteroscedasticity is a test developed by White (1980).[87] As in the Breusch Pagan test the squared residual are regressed on the explanatory variables in an auxiliary regression, but to consider also nonlinear dependencies squared and interaction terms of the explanatory variables are included in the regression in the original version of the White test.[88] In an alternative version of the White test only the fitted values of the original regression and their squared terms are included in the auxiliary regression. However, the STATA command *estat imtest* computes the original version of the White test.[89] The Breusch Pagan test for the hedonic regression with the significant variables reports that X^2 is 1498.38 and thus a p-value of zero. Thus the H_0 can be rejected on the highest confidence

[83] See Wooldridge (2009), p. 264-265.
[84] See Breusch/Pagan (1979), p. 1287-1288.
[85] See Wooldridge (2009), p. 271-274.
[86] See Baum (2006), p. 145-146.
[87] See White (1980), p. 821-825.
[88] See Wooldridge (2009), p. 275-276.
[89] See Baum (2006), p. 146.

level, which indicates that the data is highly heteroscedastic. The White test reports that X^2 is 393.7 and thus a p-value of .9277. Thus for the White test the null hypothesis can't be rejected, so that it can't be rejected that the residuals are homoscedastic. For the OLS model that includes all variables X^2 is 2162.24 and thus null hypothesis can be rejected as well. However, for the White test the X^2 statistic is 1532 with a corresponding p-value of .4952. Thus the null hypothesis can't be rejected as well for this test. However, the White test might be not reliable, when some of its requirements on the data are not fulfilled. According to the Breusch Pagan test the data is highly heteroscedastic, which will be treated in the following chapter.

4.2.2 Transformation with the inverse hyperbolic sine

In the following a transformation as well as the corresponding retransformation of the data and possibilities to estimate a comparable R^2 statistic for the untransformed and transformed models will be described. A standard approach to deal with heteroscedasticity is to transform the data with the natural logarithm (ln). This approach is also commonly used in the context of hedonic regressions.[90] The data set contains no negative values, since values like debt are accounted positive. However, the problem with microdata is that many variables have the value zero for which the natural logarithm is not defined. A possibility to deal with this problem, which is often used in economic papers is a shifting of all variables by a constant c, by transforming with $\ln(x+c)$.[91] The addition of a constant leads to a bias of the original model. The deletion of observations which take the value zero is also problematic since this reduces the sample and thus gives up sample information what can be considered as statistically inefficient.[92] An alternative transformation can be find in the literature concerning wealth transformations which is the inverse hyperbolic sine.[93] Also the Box-Cox methodology is recommended by Halvorsen/ Pollakowski (1981) for hedonic price equations.[94] This methodology is the preferred functional form for hedonic price estimation equations but can be also criticized due to a number of problems that result from the transformation.[95] Furthermore it can't be used in the case that variables take on zero or negative

[90] See Diewert (2001), p. 326.
[91] See Wooldridge (2009), p. 191-192.
[92] See Burt (1971), p. 671.
[93] See Burbidge et al. (1988), p. 123-124.
[94] See Halvorsen/Pollakowski (1981), p. 38.
[95] See Cassel/Mendelsohn (1985), p. 135.

values.[96] However the inverse hyperbolic sine transformations is easier to implement and has several further advantages and thus will be used in the following.[97] The inverse hyperbolic sine is given by the following formula.

$$asinh(y) = \ln(y + \sqrt{y^2 + 1}) \tag{7}$$

The inverse hyperbolic sine is a symmetric function and defined at zero. Above that it approximates the logarithm at its right tail very well. For values larger six the inverse hyperbolic sine and ln are nearly identical. As a result for larger values the coefficient in a model in which the dependent and explanatory variables are transformed with the inverse hyperbolic sine can be interpreted as elasticities. Thus a transformation where the explanatory variables as well as the dependent variables are transformed with the inverse hyperbolic sine can be interpreted as elasticities equivalent to the interpretation of a transformation with the natural logarithm.[98] The elasticity gives the percentage change of the dependent variable for a percentage change in a explanatory variable and is therefore independent from units, so that for the change $\%\Delta x$ the interpretation of β_i is given as: $\%\Delta y = \beta_i \%\Delta x$.[99] When the data is transformed with the natural logarithm the variables that have the same relative relationship in the untransformed data have the same arithmetic differences in the transformed data as shown by the following equation.[100]

$$\ln(\frac{y_i}{\widehat{y_i}}) = \ln(y_i) - \ln(\widehat{y_i}) \tag{8}$$

This property of the transformation is essentially important for the following analysis. Given that we want to find outliers we will analyse the residuals. In the untransformed data of course expansive properties will have potentially larger residuals, whereby the relative dispersion of the value might be the same for properties which have a low value. For example for a property which has the value 1,000,000 a residuals of 100,000 means a percental change of 10 percent for a property with the value of 100,000 a residuals of 10,000 is also ten percent. In an analysis with untransformed data where we have a look at observations with a large residual only properties with large values

[96] See MacKinnon/Magee (1990), p. 315.

[97] The recommendation to use the inverse hyperbolic sine to deal with the problems pointed out above comes from Yunyi Zhu, Deutsche Bundesbank, Research Centre.

[98] See Pence (2006), p. 5-7.

[99] See Wooldridge (2009), p. 46.

[100] See Sydsaeter/Hammond (2009) p. 154-155.

would be detected and thus the analysis would be strongly biased to the households with expansive properties without considering properties with lower values. With the transformation the relative deviations are analysed. Thus the residuals for the example above should be identical so that they are both considered equally in the analysis. The inverse hyperbolic sine is linear around zero so that as a consequence of this the logarithm property described above do not work for very small values for a transformation with the inverse hyperbolic sine.[101] Given that the variables in our analysis take usually larger values this should be not problematic. Household wealth distribution are heavily skewed.[102] Most variables are normally distributed after the transformation. Also the distribution of the residuals of a regression with the transformed data is closer to a normal distribution. Some explanatory variables will be not transformed (SIZE, MEMBER, TIME since acq.) because these kinds of variables are usually not transformed.[103] Given that variable LAND has a large variation, which makes sense since it includes properties in urban as well as rural areas this variable will be transformed. After the transformation the distribution of the variable is closer to a normal distribution.[104] The regression results of the transformed data can be seen in table 16 (Transformed regression long) in appendix 1, whereby the regression in the first column (trans. all) is the regression with all transformed variables. Generally the coefficients are more significant than the coefficients of the regression with the untransformed data which might be due to the reduction of heteroscedasticity that inflates the standard error of the coefficients. As pointed out they can be interpreted as elasticities. The interpretation can be a bit complicated. For instance the coefficient of the *PURCHASING PRICE* is 0.267. This means that if the purchasing price is 100 percent larger, c.p. the current property price is 26.7 percent larger. This makes sense because the property price should have changed in absolute terms and thus 26.7 percent of the changed absolute property value might fit with the absolute value of a 100 percent change. The coefficient of *TIME since acq.* is 0.006, which means if the property is one year older, c.p. the price increases by 0.6 percent.[105] It can be interpreted as the yearly average inflation, whereby we also have to consider the depreciation of the property value for some properties and the increasing price for others. However, it is difficult to figure out the average property price inflation in Germany for the observations in the data set,

[101] See Pence (2006), p. 5.
[102] See Barceló (2006), p. 9.
[103] See Wooldridge (2009), p. 191.
[104] In the do.file the name of the transformed variables begins with asinh. The histogram for the variables LAND before and after the transformation and the histogram of the untransformed property value can be seen in appendix 2, figure 13-15. The transformed variable of the property value can be find in chapter 9.1.
[105] See Wooldridge (2009), p. 43-46

whereby the coefficient seems to make approximately sense. Applying the Breusch Pagan test for heteroscedasticity on the transformed data shows no substantial improvement since we receive X^2=58.16 and thus a p-value of zero, so that the null hypothesis is rejected. However, the White test now also indicates highly heteroscedastic residuals since X^2=1466.25 so that the p-value is zero. Thus for the transformed data we receive corresponding results for the Breusch Pagan and White test.[106] The other coefficients show the expected signs. The relevant subset from the regression with all transformed variables, also called long regression will be selected in the chapter 4.2.4 which deals with model specification.

For the analysis of the retransformation of the predicted values the literature concerning retransformation of log transformed data will be used, since to the knowledge of the author literature concerning retransformations for inverse hyperbolic sine transformed data is not available. In order to retransform the fitted values we can't just apply the exponential function or hyperbolic sine, because this would let to an under estimation bias of the retransformed data.[107] The hyperbolic sine is given by a combination of the exponential function as: $sinh(y) = \frac{1}{2}(\exp(y) - \exp(-y)$.[108] Basically, to the knowledge of the author three possibilities to calculate a correction factor are discussed in the literature for transformations with the natural logarithm. We will apply all of them to compare the results.[109] As the inverse function of the natural logarithm of course the exponential function is used and the correction factor is derived from this under the assumption that the residuals are normally distributed. With this approach the retransformed fitted values can be calculated with the exponential function multiplied with a correction factor as can be seen by the following formula:[110]

$$\widehat{y}_i = \exp(\widehat{y}_i + \frac{\widehat{\sigma}}{2}) = \exp(\widehat{y}_i)\exp(\frac{\widehat{\sigma}}{2}). \tag{9}$$

The following correction factors are computed for the final model specification of chapter 6, which will be used for the residual analysis and stochastic imputation. With the approach above we estimate a correction factor of 1.08023. This means that for an underestimation of 8 percent is corrected. This

[106] This might be due to the fact that the residuals are nearly normally distributed for the transformed data.
[107] See Wooldridge (2009), p. 211.
[108] See Bronstein et al. (2006), p. 721.
[109] The corresponding STATA code can be seen in appendix 4, page 113.
[110] See Wooldridge (2009), p. 211.

correction factor can be also calculated with the empirical distribution with the following formula.[111]

$$\sum_{i=1}^{n} \exp(r_i) \tag{10}$$

The results is 1.08234, which is very close to the results with the analytical approach. This should be the result of the fact that the empirical distribution is very close to the normal distribution. However the retransformed values can be still biased when log scaled residuals are heteroscedastic, which is the case according for our data according to the tests above.[112] Another approach is to run the following auxiliary regression, without a constant and to use the estimator of the coefficient α_0 as the correction factor.

$$E(y|X) = \alpha_0 \exp(X\beta) \tag{11}$$

Thus the estimator of α_0 is received by the following regression through the origin, whereby the equation has to be log transformed to run the regression:

$$\widehat{y} = \alpha_0 \exp(\widehat{log(y)}) \qquad ln(\widehat{y}) = \alpha_0 \widehat{log(y)} + u. \tag{12}$$

For the regression $\widehat{\alpha_0}$ is 1.09628, so that the correction factor is 1.5 percent larger than the factors estimated above.[113]

The R^2 can be only used for the comparison of models that have the same dependent variable. The dependent variable of the transformed model is measured in different units so that the R^2 calculated with formula (1) in chapter 2.4 can't be used for the comparison of the transformed and untransformed model. In the following a calculation which allows to compute comparable R^2 measures will be applied. It can be shown that the R^2 is also equivalent to the square of the correlation of the y_i and the fitted values $\widehat{y_i}$.[114] Thus the comparable R^2 for the transformed model can be computed as the squared correlation of the values that are received by retransforming the fitted values of the transformed model with the hyperbolic sine $sinh(\widehat{asinh(y_i)})$ with y_i as $\rho(y_i, sinh(\widehat{asinh(y_i)}))^2$,

[111] See Duan (1983), p. 606-608.
[112] See Manning/Mullahy (1999), p. 21.
[113] See Wooldridge (2009), p. 210-213.
[114] See Wooldridge (2009), p. 213.

so that ρ is given as:

$$\rho(y_i, sinh(\widehat{asinh(y_i)})) = \frac{Cov(y_i, sinh(\widehat{asinh(y_i)}))}{\sigma_{y_i}\sigma_{sinh(\widehat{asinh(y_i)})}}. \tag{13}$$

In this case we can directly retransform the fitted values, without using the correcting factor, because this would make no difference since the correlation is unaffected by a multiplication with a constant.[115] For the hedonic transformed regression we receive a correlation of 0.6149 and thus an R^2 of 0.378. For the final specification of the long regression we receive a correlation of 0.71 and thus an R^2 of 0.504. They R^2 of the hedonic regression is close to the hedonic regression with the untransformed data whereas the R^2 for the long regression is about 0.1 smaller than the R^2 for the untransformed long regression. The AIC and BIC statistics and the adjusted R^2 can be only used to compare differences between the transformed models.

4.2.3 Multicollinearity

Multicollinearity is defined as a high, but not perfect correlation between two or more explanatory variables. Multicollinearity lets to an inflation of the variance of β but doesn't violates the OLS assumptions. The problem of multicollinearity is thus not well defined but c.p. less multicollinearity is better wherefore we will control for it since as pointed out in chapter 4.1 relationships between several explanatory variables can be identified with economic reasoning. The variance influence factor can be used to measure the amount by which the variance of β is increased because of correlations between the predictor variables. The variance of β is given by the following formula, whereby σ^2 is the variance of the residuals and SST_j is the total sample variation of the variable x_j which is given as: $SST_j = \sum_{i=1}^{n}(x_i - \bar{x}_j)^2$. R_j^2 is defined as the R^2 of the regression of all explanatory variables in spite of the j variable on the j explanatory variable. With that the variance of β_j is given as:[116]

$$Var(\beta_j) = \frac{\sigma^2}{SST_j}\frac{1}{1-R_j^2}. \tag{14}$$

[115] See Wooldridge (2009), p. 213.
[116] See Wooldridge (2009), p. 95-98.

The last factor ($\frac{1}{1-R_j^2}$) is the Variance Influence Factor (VIF) which gives the amount by which the variance of β is inflated due to multicollinearity. The VIF can be calculated with the STATA command *estat vif*. In the case that $R_j^2 = 0$, which means that the other explanatory variables explain nothing of the variance of the j variable the VIF is 1 and thus the variance of the coefficient β_j is not inflated as a result of multicollinearity. However, when $R_j^2 \to 1$ it can be seen from the formula above that $Var(\beta_j) \to \infty$.[117] In the literature different values of the VIF are considered as problematic. Values of the VIF larger than 4 as well as 10 are seen as indicators for high multicollinearity.[118] For the regression with all transformed variables we receive very high VIF for the variables *EDUCATION* (dpa0300) and *EDUCATION prof.* (dpa0400). This makes sense because the education and professional education should be correlated. The VIF for the dummy variable *UNIVERSITY ENTRANCE* (dpa0300_6) is 384.66 and 1/VIF is 0.0026. This means that the R_j^2 is 99.97, which shows that having a university entrance is explained nearly completely by the other variables. For the variable *DWELLING RATE* (sc0200) we receive also VIF larger than 4. The VIF for the other variables is smaller than 4 and thus multicollinearity seems to be no problem in the model specification. The variable *EDUCATION* is dropped as well as *DWELLING RATE*. It can be argued that the professional education should explain the wealth of the household better since it is the relevant job qualification. Another approach to detect multicollinearity is to have a look at the correlation matrix of the explanatory variables, where a high correlation between two variables can indicate a collinearity problem.[119] No high correlations can be seen which fits with the results of the VIF. Thus the explanatory variables that were considered as problematic in chapter 4.1 don't cause multicollinearity problems according to the results of the VIF and correlation matrix.

[117] See Wooldridge (2009), p. 95-98.
[118] See O'Brien (2007), p. 679.
[119] See Belsley et al. (2004), p. 92-93.

Table 7: Correlation matrix of explanatory variables

	PROPERTY VALUE	*PURCHASING PRICE*	*PROPERTY*	*CARS*	*FINANCIAL WEALTH*	*MORTGAGES*	*TIME since acq.*	*SIZE*
PROPERTY VALUE								
PURCHASING PRICE	.62							
PROPERTY	.27	.14						
CARS	.25	.19	.15					
FINANCIAL WEALTH	.36	.22	.27	.20				
MORTGAGES	-.01	.15	-.13	-.01	-.12			
TIME since acq.	-.02	-.40	.05	-.03	.09	-.42		
SIZE	.44	.31	.19	.15	.22	-.01	-.01	
LAND	.10	-.04	.07	.03	.03	-.03	.12	.29

Source: Own table

4.2.4 Omitted variable bias and estimation strategies

In the model specification we might include irrelevant variables in our model or we might not include relevant variables. Including irrelevant variables should have coefficients with the value zero and they should be not significant. However, the irrelevant variables inflate the variance of the coefficients of the other explanatory variables and should be therefore not part of the model. The omission of relevant variables in the model is indeed more problematic since it might let to a bias of the other coefficients, this is referred to as the omitted variable bias.[120] For a model with two variables with the coefficients β_1 and β_2, where β_2 belongs in the model but is omitted the bias of $\widehat{\beta}_1$ is given as Bias($\widehat{\beta}_1$)=E($\widehat{\beta}_1$) $-\beta_1 = \beta_2\widehat{\delta}_1$, with $\widehat{\delta}_1$ as the sample covariance of x_1 and x_2. If x_2 is irrelevant for the model and $\widehat{\beta}_2$ is zero $\widehat{\beta}_1$ is not biased and in the case when x_2 is relevant but uncorrelated with x_1 and thus $\widehat{\delta}_1$ is zero there is also no bias of $\widehat{\beta}_1$.[121] To minimise the inclusion of irrelevant variables and the danger of the omitted variable bias an appropriate estimation strategy has to be applied. The omitted variable bias might occur in the hedonic model since the explanatory power is very low, when we compare the results to other hedonic models. Therefore it is likely that variables are omitted. The added further variables should reduce the omitted variable bias and let to more reliable estimations of the coefficients of the hedonic variables. Concerning the estimation of the model with all variables it can be differentiated between an estimation strategy where single variables are included into the model after another and an approach where we start with the full model. By adding single variables the change of the model evaluation statistics can be assessed for every variable, so that the contribution to the model can be assessed. Only variables that have significant explanatory power not only according to the t statistic but also according to other statistics like the R^2 adjusted, AIC and BIC will be used in the model in order to keep the model simple and to concentrate on the important variables.[122]

The method to start with the full model, which is also called general unrestricted model (GUM) is referred to as the LSE methodology. If we start with the full model the danger of an omitted variable bias is minimised, so that the coefficients are more reliable, whereby on the other hand the variance of the coefficients might be more inflated because of the inclusion of irrelevant variables. In this approach the unrestricted model is estimated first and than reduced by testing which variables are relevant. Thereby a model is searched that has the same explanatory power as the unrestricted model

[120] See Verbeck (2004), p. 55-56.
[121] See Wooldridge (2009), p. 90-91.
[122] See Verbeck (2004), p. 55-57.

but has a simpler form. For this it is assumed that the right model is a part of the unrestricted model and that due to the applied tests the true specification can be identified.[123] The LSE method identifies specifications that are close to the correct specification in practice according to Hoover/Perez.[124]
Both methods are applied for the selection of variables. The long regression can be seen in table 16 (Transformed regression model long), whereby the first column shows the regression with all variables included (trans. all) and the second column (trans. signif.) contains the regression with the coefficients considered as significant after the application of the LSE methodology. According to the t-statistics the coefficients of the variable *MEMBERS* is not significant and is thus not included into the model. *MORTGAGES* and *INCOME* are as well not significant but we leave them in the regression given that they should have explanatory power and the standard error of the coefficients might be inflated due to heteroscedasticity. The variables *EDUCATION* and *EMPLOYED* were excluded due to high multicollinearity. For the further dummy variables the F-test is applied to test their simulation significance. The null hypothesis can't be rejected for the variables *DWELLING LOCATION*, *DWELLING OUTWARD*, *NEIGHBOURHOOD* and *INTERIOR*. Therefore we drop them also. The statistics for model comparison are very close for the model with all variables and the model with the significant variables. This underlines that the dropped variable doesn't belong to the model.
The estimation of the short regression approach can be seen in appendix 1 in table 17 (Transformed regression model short). Starting with the hedonic model the variables used in the model estimated with the LSE methodology are added after another. Given that the *PURCHASING PRICE* has a large explanatory power we do not include this variable while including the variables into the model reduces the explanatory power and thus the significance of the other variables. An small improvement of the model according to the model selection statistics can be seen for the variables included step by step. For the variable *INCOME* the BIC increases minimal. When the variable of the *PURCHASING PRICE* is not included into the model all the other included variables are significant. This shows that they have explanatory power but that of course a lot of the variance of the property price is explained by the *PURCHASING PRICE*. Thus all variables identified with the LSE method will be used for the model.
The coefficients of the hedonic variables in hedonic model and the model with the further variables are different as a result of the omitted variable bias which can be seen in the table 16 (Transformed

[123] See Verbeck (2004), p. 55-57.
[124] See Hoover/Perez (1999), p. 167-191.

regression model long). In this context it can be also criticized that the authors who estimate the hedonic regression outlined in chapter 3.3 don't include non hedonic variables in their models to control for the omitted variable bias. This argument fits well in the general critic of hedonic models of Ekeland (2001) as already pointed out, since especially for potentially underidentified models it should be controlled for the omitted variable bias as done above.[125]

4.2.5 RESET test and quadratic terms

Until now only linear relationships are modelled in the regression equation. However, a linear model is a reduced and misleading form for the estimation of hedonic models.[126] With economic reasoning non linear relationships between the property value and several explanatory variables can be find. In order to model decreasing and increasing marginal effects quadratic terms can be included into the model.[127] Before we include quadratic terms in the model a test that can be used to detect general functional form misspecifications will be applied. Usually the regression specification error test (RESET) is used for this, which was developed by Ramsey (1969).[128] It can be used to test in how far the model specification is appropriate or if non linear terms and interaction terms should be included into the model. For this purpose Q exponential terms of the fitted values are included into the model.

$$y_i = X\beta + \alpha_2 \widehat{y_i}^2 + \alpha_3 \widehat{y_i}^3 ... + x_1 + ... + \alpha_Q {y_i}^Q + v_i \tag{15}$$

The F-test for the null hypothesis $H_0 : \alpha_2 = ... = \alpha_Q = 0$ is tested which means that coefficients for the powers of the fitted values are simultaneously zero. Usually the test is performed for Q=2. The command *estat ovtest* in STATA runs the RESET test for Q=4.[129] It might be possible that the RESET test is rejected not due to a functional misspecification in the sense that non linear relationships are not appropriate modelled but because of the omission of relevant variables, so that the inclusion of these variables might capture the nonlinearities which are identified by the test.[130] The null hypothesis for the RESET is rejected for the hedonic model and also for the long regression

[125] See Ekeland et al. (2001), p. 1.
[126] See Ekeland et al. (2001), p. 1.
[127] See Wooldridge (2009), p. 192-193.
[128] See Ramsey (1969), p. 361-362.
[129] See Baum (2006), p. 123.
[130] See Verbeck (2004), p. 63.

model. Because of the result of the RESET test quadratic terms for all continuous variables are included to see whether they are significant. For the quadratic terms we receive the following slope parameters, which depend on x and are thus not constant as the parameters in the linear specification: $\Delta\widehat{y} \approx (\widehat{\beta}_1 + 2\widehat{\beta}_2 X)\Delta x$. Thus for a change of $\Delta x = 1$ we have $\Delta\widehat{y} \approx \widehat{\beta}_1 + 2\widehat{\beta}_2 x$. If β_2 is negative this reflects a decreasing positive effects from x on y.[131] The regression results can be seen in table 18 (Quadratic terms). The first column (all) contains the regression with the quadratic term of *OTHER PROPERTY* and the second column the regression with all quadratic terms included (quadratic). Only the coefficients for the quadratic terms SIZE and LAND are significant and thus used in the model that can be seen in the third column (quadratic sig.). The coefficient for the quadratic term of SIZE and LAND are negative and the linear terms are still positive. For them there is also an economic justification that fits with the coefficient, since a further square meter of property size or additional land should have a diminishing impact on the price. The model comparison statistics of the model with the significant quadratic terms are very close to the statistics without the quadratic terms.

4.2.6 Out of sample prediction

To evaluate the robustness of the predictions of the estimated regression model an out of sample prediction can be used. This procedure is applied on the suggestion of Dr. habil. Ulf von Kalckreuth. Given that it is usually applied in the context of forecasting with time series data no appropriate reference for a cross section regression can be given, however with Monte Carlo Simulations it can be find that out of sample prediction can help to prevent an overfitting of the model.[132] The final prediction equation estimated in chapter 6 is used. A regression with a random drawn subsample of 1569 observations is estimated. The coefficients are very close to the coefficients of the regression estimation with the whole sample. The resulting equation is then used to predict the fitted values for the 50 observations that were not used for the estimation of the prediction equation. To see how good the prediction equation works for the subsample prediction the R^2 is calculated as the squared correlation of the predicted values with the observed values. With a correlation of $\rho = 0.738$ the R^2 is 0.5452, which shows that the prediction power is large.[133] Thus it can be seen that the prediction equation works very well for the subsample.

[131] Wooldridge (2009), p. 192-193.
[132] See Clark (2000), p. 20-21.
[133] The STATA code for these calculations can be find in appendix 4, page 113.

5 Regional models

Location is one of the most important factors for the determination of the property value.[134] Therefore is important to model regional differences as good as possible with the opportunities that the data offers to improve the prediction estimation.

5.1 Theoretical models

The estimation equation used until now models the differences of property prices due to local differences besides the categorial variables with 15 dummy variables for the *FEDERAL STATES* in Germany since one is dropped and with the *POPULATION DENSITY* dummy. Because of data protection reason it can be only pointed out which federal state belong to the former East Germany and which belong to West Germany without pointing out the name.[135] We drop the dummy variable of a large federal state in West Germany which might be regarded as average concerning several economic aspects since this is a good reference point for the other dummy variables. The current model specification assumes that there are level differences of the property prices for the different federal states and also takes demographic aspects into account but doesn't incorporate different relationships between the property price and the explanatory variables in the models concerning local differences. This will be done for the former East and West Germany given that property price differences[136], as well as differences concerning the explanatory variables exists.[137] A model will be estimated that incorporates different relationships between the variables due to the location of the property in East or West Germany as well as level effects for the location of the property in different federal states. For this purpose the following model with interaction terms for the different relationship between the variables in East and West Germany and dummy variables for the level effects in the different federal states will be estimated.

$$y = \sum_{i=1}^{k} \beta_i^{west} x_i + \beta_i^{int \cdot} (D * x_i) + \sum_{j=1}^{15} \gamma_j d_j \sum_{p=16}^{n} \gamma_p d_p + u \tag{16}$$

[134] See Kiel/Zabel (2008), p. 175-178.
[135] In the following only East and West Germany.
[136] See Milleker (2006), p. 3-5.
[137] See Schürt (2011), p. 3-20.

The interaction term is built by the multiplication of the dummy variable D with with the variables x_i for k variables. The dummy D takes the value one for properties located in East Germany and the value zero for properties located in West Germany. The coefficient for West Germany is the coefficient β_{west} in the following equation and the the coefficient for East Germany is received as $\beta_{west} + \beta_{int.}$.[138] The level effects for the federal states are given by the dummy variables d_j from $j = 1, ..., 15$. The further dummy variables in the model are included by the last term: $\sum_{p=16}^{n} \gamma_p d_p$.[139]

5.2 Empirical results

The model outlined above assumes different relationships between the dependent and explanatory variables for the East and West German subsample. In order to compare if there is a significant difference between the coefficient of these regressions the Chow test can be used.[140] The Chow test can be used to test whether the coefficients of the same model $y = X\beta + u$ applied on two different subsamples are equal. To compute the Chow test besides the model for the full sample models with the subsample of the observations in East Germany and the subsample of the observations in West Germany have to be estimated. Thereby also level effects for the German federal states within the West German and East German subsample are part of the models.

$$y = X\beta^{east} + u \qquad y_i = X\beta^{west} + u \tag{17}$$

The model for the West and East German subsample can be seen in table 19 regional models in column three (West) and four (East). Differences between the coefficients can be seen, but many coefficients for the East German subsample are not significant because less observations are used for the estimation, which increases the variance of the coefficients.[141] STATA provides no command to compute the Chow test directly, so that we will compute the test in the following, while using statistics from the regressions. The Chow test is developed by Gregory Chow (1960) [142], however the notation of the following statistics is taken from Wooldridge as well as Fisher (1970), who

[138] This can be seen by comparing the coefficients of the subsample regressions with the coefficients of the regression with the interaction terms, whereby the relationship holds only approximately. The regional models can be seen in table 19 (Regional regression models).

[139] The STATA code can be find in appendix 4, pages 114.

[140] See Chow (1960), p. 591-605.

[141] See Wooldridge (2009), p. 97.

[142] See Chow (1960), p. 591-605.

provides an easier derivation of the test.[143] The F-test tests simultaneously for $i = 1...i = k$ that $H_0 : \beta_i^{east} = \beta_i^{west}$. The test statistic is $F(k,(N_1+N_2-2k))$ distributed. For the tests $SSR_N = 253.979$ (N=1619, k= 58) as the sum of the squared residuals the model for the whole sample and the sub samples $SSR_{west} = 184.256$ (n=1422) and $SSR_{east} = 44.757$ (n=197) are used.

$$F(k,(N_1+N_2-2k)) = \frac{(SSE_N - (SSE_{west}+SSE_{east})}{(SSE_{west}+SSE_{east})}(\frac{n-2(k+1)}{k+1}) \tag{18}$$

We receive a F-statistic of 2.77 for F(58,1503) and thus a p-value smaller 0.001, so that the null hypothesis can be rejected. Thus the Chow test indicates that the coefficients for the East and West German subsample are different. Due to the inclusion of different dummy variables the k differs for the regional models, which might be problematic. Given that the Chow test is a specific F-test it is only valid for homoscedastic residuals, which is indeed problematic given the results of the Breusch Pagan and White test.[144] In spite of the problems with the test the difference of the coefficients of the regressions with the subsamples show that differences exists so that they will be included in the model.

The prices of properties in the former DDR are not the result of a market system. In the data set extremely cheap acquisition prices for properties in the former East Germany can be found. The explanatory power of these prices of the current property price is indeed in doubt and might bias the estimation. Thus an interaction term with a dummy variable is created that takes the value one for properties that are purchased after 1991 or which are located in West Germany[145] and the value zero for observations that are located in East Germany and the purchased before 1991. The interaction coefficient is not significant, this comes from the large standard error of the coefficient which is probably a result of the small amount of observations used for the estimation of the dummy for the interaction term. For the further analysis the regional specification estimated before without this specific interaction term is used. The R^2 is 0.683 and the adjusted R^2 is 0.670. The AIC is 1692.604 and BIC is 2021.368. Compared to the last model (quadratic sig.) that improved slightly due to the inclusion of non linear relationships in the model, the modeling of regional aspects also let to an improvement of the model. However, the null hypothesis is still rejected and the residuals are still heteroscedastic according to the Breusch Pagan and White test.

[143] See Fisher (1970), p. 361-366.
[144] See Wooldridge (2009), p. 243-246.
[145] The or is a non exclusive or.

6 Feasible General Least Square Regression

In the case of heteroscedasticity the OLS estimator is not the most efficient unbiased linear estimator. The feasible general least-squares estimator can be used as a more efficient estimator, which means that the variance of β is smaller.[146] Given that we have an idea what might cause the heteroscedasticity so that we can estimate a model for the heteroscedasticity a FGLS regression will be estimated. First of all the theory of Feasible General Least Square estimators (FGLS) will be described briefly and after that the FGLS model will be estimated. The theoretical outline follows basically Cameron/Trivedi (2005).

6.1 Theoretical model

First of all the theory of the General Least Square regression (GLS) estimations will be outlined and with that the theory of the FGLS regressions, which can be done by making slightly different assumptions.In contrast to the OLS assumptions for FGLS estimations it is assumed that the error variance is not constant and the errors are not independent, so that the covariance matrix of the residuals is unequal to the identity matrix multiplied with a constant residual variance for homoscedastic residuals: $\Omega \neq \sigma^2 I$. Furthermore it is assumed that Ω is nonsingular so that we can calculate $\Omega^{\frac{1}{2}}$ with $\Omega^{\frac{1}{2}}\ \Omega^{\frac{1}{2}} = \Omega$.[147] A nonsingular is a matrix for which an inverse exists.[148] Homoscedastic error terms are received by transforming the data with $\Omega^{\frac{1}{2}}$ by multiplication: $\Omega^{\frac{1}{2}} y = \Omega^{\frac{1}{2}} X\beta = \Omega^{\frac{1}{2}} u$. It can be shown that $Var[\Omega^{\frac{1}{2}} y] = E[(\Omega^{\frac{1}{2}} u)(\Omega^{\frac{1}{2}} u)^{|} X] = I$. With that, the general least square estimator of β can be derived as:

$$\widehat{\beta}_{GLS} = (X^t \Omega^{-1} X)^{-1} X^t \Omega^{-1} y \tag{19}$$

$$Var[\widehat{\beta_{GLS}}] = (X^t \Omega^{-1} X)^{-1}. \tag{20}$$

The general least square estimator can't be estimated given that Ω is not known. Ω can be estimated by specifying a constant vector γ, so that $\Omega = \Omega(\gamma)$ and by estimating a consistent estimator of γ,

[146] See Greene (2002), p. 210.
[147] See Cameron/Trivedi (2005), p. 81-82.
[148] See Anton (1998), p. 48.

so that $\widehat{\Omega} = \Omega(\widehat{\gamma})$. This can be done by estimating the conditional variance of the residuals with an auxiliary regression on the squared residuals. The conditional variance of the residuals is equivalent to the conditional expected squared residuals for an expected value of zero for the residuals. The mean of the residuals in the sample used is very close to zero, so that the following equations are equivalent.

$$Var[u|X] = \exp(z^t\gamma) \qquad E[u^2|X] = \exp(z^t\gamma) \tag{21}$$

Thereby z can include further variables that might explain the conditional variance of the residuals as well as a subset from X.[149] If we assume a direct relationship between the residuals and X, obviously z can't be a subset from X because this would bias the OLS assumption that $E[u|X] = 0$. In the empirical part also such a relationship will be tested while using variables z that are disjunct from X. The exponential function in this model seems to be used in order to have only positive predicted values. We will also assume a linear relationship and then control the predicted variables for negative values, so that also the following model for the heteroscedasticity will be tested.

$$Var[u|X] = z^t\gamma \qquad E[u^2|X] = z^t\gamma \tag{22}$$

By using $\widehat{\Omega}$=Diag[$\widehat{u_i^2}$] we receive the more efficient feasible general least squared estimator and its variance as:[150]

$$\widehat{\beta}_{FGLS} = (X^t\widehat{\Omega}^{-1}X)^{-1}X^t\widehat{\Omega}^{-1}y \qquad \widehat{Var[\beta_{FGLS}]} = (X^t\widehat{\Omega}^{-1}X)^{-1}. \tag{23}$$

[149] See Cameron/Trivedi (2005), p. 81-82.
[150] See Cameron/Trivedi (2005), p. 82-83.

6.2 Empirical results

The model for heteroscedasticity comes from the following intuition. If we assume that the mean value is a good approximation for the correct property price, property value estimations of persons which are not very good in estimating the property price should differ more from the mean value, which means that their residuals should be larger. The question is whether the over and under estimation of the predicted value, which means the heteroscedastic residuals can be modeled and whether this can be used to improve the efficiency of the coefficients. We assume that persons with several characteristics will have a lower ability to estimate the true property price and therefore tend to over and underestimate this value. As a result it should be possible to explain the variation of the residuals by these variables, so that they can be used to model the heteroscedasticity. Within a FGLS estimation such observations will receive a lower weight because they bias the conditional expected value. We assume that older persons might be less able to estimate the correct property value or persons who acquired the property a long time ago. Furthermore financial literacy should play an important role. Thus a variable that measures financial literacy is produced by counting the number of the true answers of financial literacy questions. The dataset contains the following multiple choice questions to measure the financial literacy of the interviewed person. The questions offer three answer opportunities, whereby only one of them is correct.

Table 8: Financial literacy variables

Variable Name	Description	STATA Variable Name
INTEREST RATE	Compound interest rate effect	dhnm0100
INFLATION	Inflation and purchasing power	dhnm0200
DIVERSITY	Diversification effect in a portfolio	dhnm0300

Source: Own table

We count the true and wrong answers, whereby denied questions are counted as a wrong answer.[151] To count the denied answers as wrong might be problematic, but to drop these observations would on the other hand reduce the sample size. Besides the variable for the financial literacy

[151]To use the financial literacy variables for the creation of the variables CORRECT and WRONG and analyse there relationship with the residuals is an idea of Dr. habil. Ulf von Kalckreuth, Research Centre Deutsche Bundesbank. The STATA code for the creation of the financial literacy variables can be find in appendix 4, page 114.

also the age of the interviewed person is included into the regression. The age is calculated by the subtraction of the variable of birth *YEAR OF BIRTH* given by dpe9050 from the year 2011 and with that given as the variable *ALTERPERSON*. Furthermore the variable *TIME since acq.* is included, with the assumption that when the time of the property acquisition is longer ago the household should be not as good informed about the correct market price than a household who acquired the property in the last years. The *INCOME* of the household and his *EMPLOYMENT* status are also used as proxy variables for the ability of the household to estimate the property value. We argue that higher income should as well as being employed increase the ability of the household to answer the correct property price. Above this also the *EDUCATION* dummy is included. The following table shows the variables of the model that are not already pointed out in one of the tables before. The table 20 (FGLS residual model) shows the estimation results, whereby the regression in the first

Table 9: Estimation ability variables

Variable Name	Description	STATA Variable Name
ALTERPERSON	Age of the person giving the answer	alterperson
CORRECT	Number of correct answers	correct
WRONG	Number of wrong answers	wrong

Source: Own table

column (e1sq) is an estimation for the model of equation (22) and the model in the third column (le1sq) is the regression for the model of equation (21). For the regression in the third column the natural logarithm is applied on both sides of the equation, so that the $ln(u_i^2)$ is the dependent variable. The R^2 of the models are smaller than 0.05 and thus very low. Furthermore many coefficients are not significant so that their interpretation is not reliable. Nevertheless it will be analysed in how far they fit with the expectations made. The coefficient of *ALTERPERSON* is positive for the regression in the first column, which means that c.p. with an increasing age the conditional variance increases, which fits with the assumption made that the estimations of the property price made by older persons is less precise. The *TIME since acq.* coefficient is positive as well, which fits also with the assumptions. The *EDUCATION* dummy variables for the regression in the first column are in tendency negative for a better education, which makes sense since for a better education c.p. the estimation error should be smaller. For the regression in the third column the sign of the dummy variables seems to be random. For the *EMPLOYED* dummy variables the coefficients don't fit with our assumption that being not employed might increase the tendency to over or underestimate the property price.

The coefficient for INCOME is negative and thus fits with the assumptions that a higher income is a proxy for the precision of the property price estimation. The coefficient of the variable *WRONG* is positive which means that when the household answers more financial literacy questions wrong, c.p. the conditional variance of the residuals increases. When the variable *CORRECT* is included the coefficient is the same but negative since when the household answers more question correct, c.p. the conditional variance if the residuals decreases.[152] Since the squared residuals are used as the dependent variables only relationships between the absolute over and underestimation are assessed. The regression in the fourth table shows a regression with the residuals u_i as the dependent variable. Given that non of the coefficient is significant and the R^2 is also lower this approach will be not considered further. Given that the dummy variables *EDUCATION* and *EMPLOYED* are very insignificant also a model without them is estimates with the results shown in the second column (e1sq sign.). This specification will be used for the further estimations.

In the literature a comparable approach can be find by Kain and Quingly, who replicate a study of Kish and Lansing published in 1954. They have property estimates from a to the PHF question corresponding question concerning the property value as well as an appraisal for market values. This means that they use these appraised market values as a reference value, whereas we use the conditional expected value as a reference value. They regress several socioeconomic variables on the errors, whereby they use absolute as well as percental error that are approximately comparable with the errors in the log model. They use race, education, gender dummy and the value of the property as explanatory variables. As with our regression the coefficients are not significant, but the R^2 is higher.[153] Goodman and Ittner calculate estimation errors that are received by the difference of the market price and the estimated price, which differentiates them from survey before. They use percentage error and log absolute percentage error as the dependent variable. They also find a low adjusted R^2, but have some significant coefficients, like in our regression.[154] We also regress the squared residuals on all variables used in the regression to search for significant variables. The results can be seen in table 21 (FGLS residuals model all). The first column shows the regression on the squared residuals (e1sq) and the second column the regression on the natural logarithm of the squared residuals (lesq). No further significant coefficients can be identified with this approach for which also economic arguments can be given.

[152] To include the variable *WRONG* or *CORRECT* in the regression makes no difference. However, different versions can be seen in the log-file.

[153] See Kain/Quingly (1972), p. 805.

[154] See Goodman/Ittner (1992), p. 349-353.

None of the papers read by the author include financial literacy variables. It might be also interesting to use the financial literacy variable within a regression on residuals computed with the approach of Goodman and Ittner as the difference of the estimated property price to a transaction price.

To model the heteroscedasticity seems to be indeed problematic, wherefore the question arises whether it should be used for the FGLS estimation or not. Angrist and Pischke argue that efficiency gains from GLS estimations might be modest and that poorly estimated weights can do more harm than improving the estimation.[155] However, Wooldridge argues that in the case of very heteroscedastic residuals it might be better to use a weighted regression even though the model of heteroscedasticity might the wrong.[156] Following this argument the model in the first column (e1sq) residuals will be used in the following for the estimation of a FGLS regression. Every observations receives a weight (1/w) observations with smaller disturbance variance receive a larger weight. For STATA we have to define (1/w) , for FGLS regression with *aw* as an option for the regression, where *aw* has to be the inverse of the observations conditional variance.[157] The results of the FGLS estimation can be find in table 22 (FGLS regression). The first column shows the last model OLS specification from the table regional models for the comparison (all sig.). The regression in the second column shows the FGLS regression with the weights computed according to model 22 (weight: e1sq sign.), whereas the third column contains the FGLS regression for model 21 (weight: le1sq). For both models the coefficients and their significance differ slightly from the OLS model. The model comparison statistics are nearly the same for the OLS and FGLS regression. For the FGLS regression in column one the Breusch Pagan tests null hypothesis is rejected with a p-value of zero, the RESET test has a p-value of 0.0126. For the regression in the second column the null hypothesis of both tests are also rejected. Thus concerning heteroscedasticity no difference to the OLS regression can be achieved. Nevertheless the estimated $\widehat{\beta}_{FGLS}$ which differs slightly from the $\widehat{\beta}_{OLS}$ will be used for the identification of outliers in the next chapter. Thereby the specifcation in the second column is used with the already pointed out argument that it seems to be not necessary to assume model 21 ($E[u^2|X] = \exp(z^t\gamma)$) for the heteroscedasticity.

[155] See Angrist/Pischke (2008), p. 69.
[156] See Wooldrige (2009), p. 288.
[157] See Baum (2006), p. 148-149.

7 Identification of outliers and influential points

In this chapter outliers and influential points will be identified. For this purpose in particular the residual and leverage from the prediction equation will be analysed. Furthermore also other statistics that can be used in this context will be pointed out and their advantages will be discussed.

7.1 Residual and leverage analysis

An influential point can be defined as an observation that lets to a substantial change of the OLS estimates when it is dropped from the sample. The definition of an outlier is a bit vague given that the value of one observation has to be compared with the other observations in the sample and is considered as an outlier when the observation deviates largely from the other observations of the data.[158] A distinction between the points exists because a point can be an outlier without effecting the coefficients substantially.[159] In order to detect such extreme observations in the variable of the property value and in the explanatory variables the residuals as well as the leverage statistic can be used.[160] First of all outliers in the dependent variable will be identified, since this is the main focus of the thesis. The prediction equation gives us the conditional expected value of the transformed estimated property value given the explanatory variables. We assume that the conditional expected value is a good approximation for the true average property price as already done. Large deviations from the predicted value can be thus considered as outliers. The deviation of the observations (y_i) from the conditional expected value ($\widehat{y_i}$) can be measured by the residual (r_i).

$$r_i = y_i - \widehat{y_i} \tag{24}$$

The following statistics show the mean and standard deviation (sd) and extreme values of the residuals. The residuals from the last model specification, shown in Table 22, column 2 (weight: e1sq) are used.

[158] See Wooldridge (2009), p. 325.
[159] See Stevens (1984), p. 334.
[160] See Meloun/Militky (2001), p. 174-176.

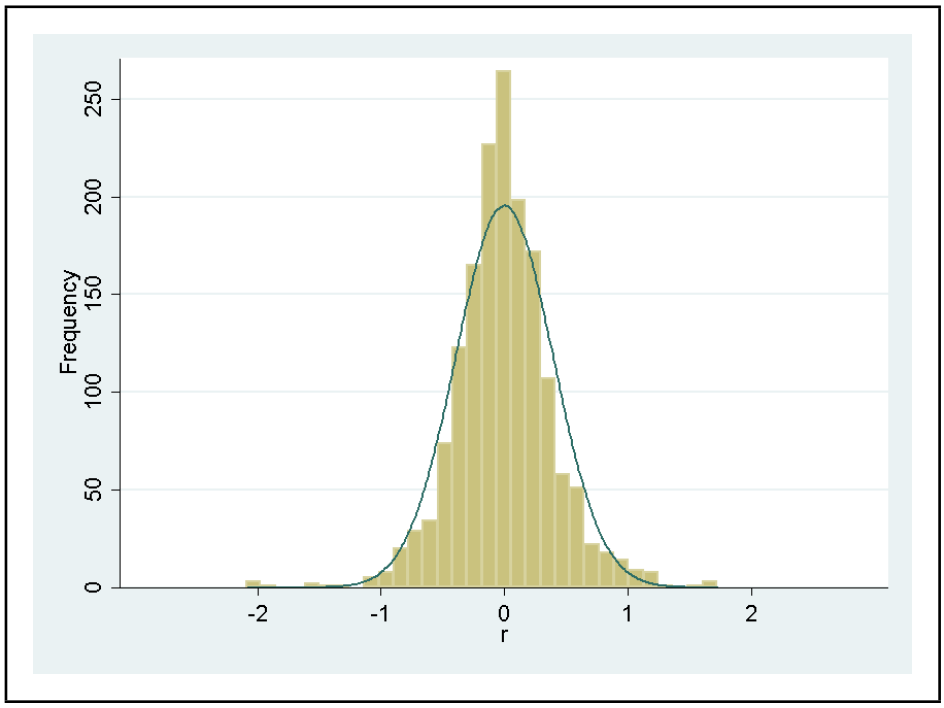
Table 10: Residuals

N	mean	sd	min	max
1619	- .000630	.393797	- 2.087974	1.726213

Source: Own graph.

The histogram of the residuals with the drawn in analytical normal distribution with the sample mean and variance of the residuals $r \sim N(-0,000630; 0,3937969)$ shows that the residuals are approximately normally distributed.[161]

Figure 1: Residuals from FGLS regression model



Source: Own graph.

For further calculations standardized residuals are more suitable for the calculations, because their interpretation is easier in the sense that the quantiles are directly known for the analytical standard normal distribution. To standardized residuals they are divided by the sample standard deviation of the residuals.[162] Due to the fact that the OLS assumption of a zero mean of the residuals

[161] This is also underlined by the quantile plots in appendix 2, figure 16-17 and fits well the OLS assumptions that the distribution of the residuals have a zero mean and that they are normally distributed.

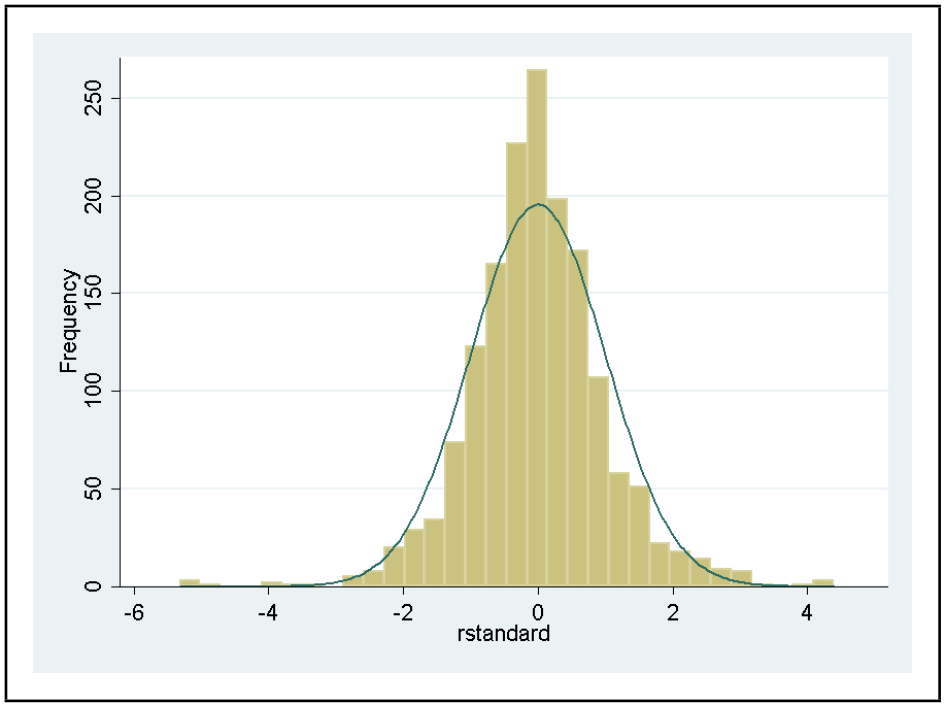
[162] See Hill et al. (2001), p. 32.

is nearly fulfilled it is not necessary to subtract the mean.

$$r_i^{stand.} = r_i/\sigma \tag{25}$$

For standardized residuals absolute smaller 1 we receive a quantile that contains $68,27\%$ of the observations, for standardized absolute smaller 2 residuals we receive the quantile that contains $95,45\%$ and for standardized residuals absolute smaller 3 we receive the $99,73\%$ quantile.[163] Another approach to receive standardized residuals would be to divide them by the conditional standard deviation. This would make the residual of observations with a large conditional standard deviation smaller.[164] Given that we want to identify exactly such observations this approach makes indeed no sense for this analysis.[165] The statistics and histogram of the standardized residuals show that they approximately standard normal distributed ($r^{stand} \sim N(0,001603; 1,002318)$).[166]

Figure 2: Standardized residuals from FGLS regression model



Source: Own graph.

[163] The values can be calculated with STATA.

[164] See Wooldridge (2009), p. 327.

[165] In the FGLS regression observations that might bias the conditional expected value where given a lower weight, since in this case it makes sense to use weights. In the case of the residuals this makes no sense.

[166] This is also underlined by the quantile plots in appendix 2.

Table 11: Standardized residuals

N	mean	sd	min	max
1619	- .001603	1.002318	- 5.314448	4.393670

Source: Own graph.

It is not obvious which quantile should be used for the identification of outliers. For an interval of +/- 2 standard deviations around the mean we receive a quantile of 95.45%. This means that 95.45% of the observations have a standardized residual absolute larger than 2 and thus the probability to have an observation with a absolute standardized residual larger 2 is 4.55%. In a first approach observation with a standardized residual larger than 2 standard deviations are considered as outliers. We receive with that 91 observations. The 91 observations are 5.62% of the 1619 residuals of the sample. This is close to the 4.55% that we receive with the analytical standard normal distribution. For these observations we have to differentiate between outliers which are obviously incorrect, given that a mistake has been made during the collection of the data for instance a zero might be added to much and observations that might be extreme but nevertheless belong to the distribution.[167] These observations should be treated as the other observations in the sample.[168] The analysis of the outliers assumes that the conditional expected value is a reliable approximation as explained above. However, the predicted value can be biased as a result of a bias in the coefficients of the equation caused by an influential observation in the explanatory variables, since the OLS regression is not robust to such influential points while minimizing the sum of the squared residuals.[169] Therefore also excentrical values in the explanatory variables have to be analysed. A statistic that can be used for this purpose is the leverage statistic, which can be used to detect observations that have large influence on the coefficients of the predictor variables. With x_j as the jth row of the regressor matrix, the leverage is given by this equation.[170]

$$h_j = x_j(X'X)^{-1}x_j' \tag{26}$$

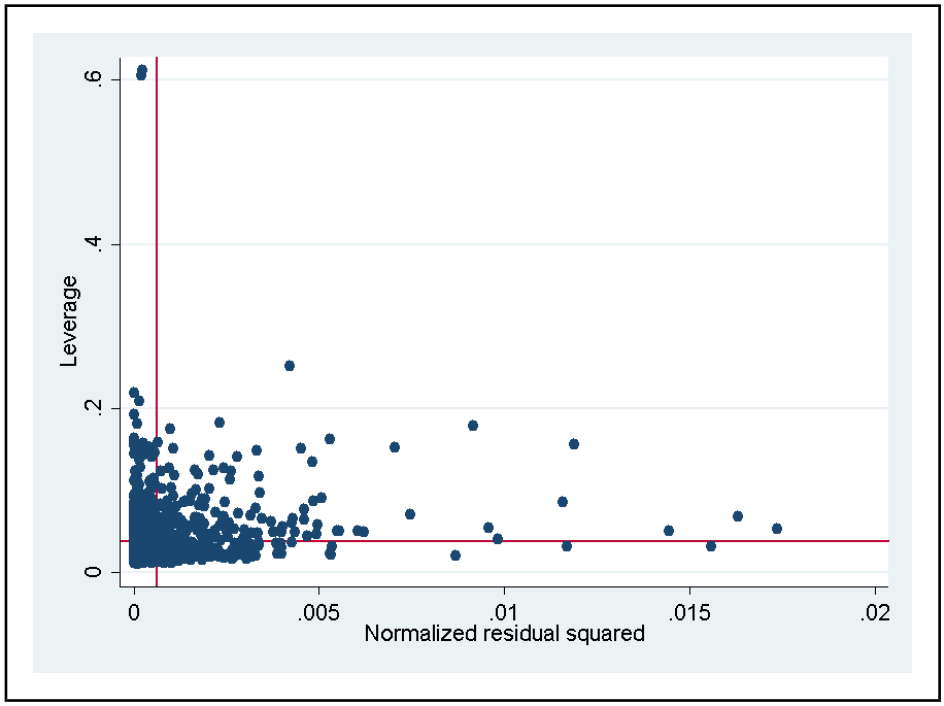
This means that the leverage statistics are diagonal elements of the hat matrix or projection matrix

[167] See Wooldridge (2009), p. 325.
[168] See Grubbs (1969), p. 1.
[169] For instance the median is robust to outliers, whereas the mean is very sensitive to outliers.
[170] See Baum (2006), p. 126.

that maps the vector of the observed values to the vector of the fitted values. For uncorrelated errors the hat matrix can be calculated with $\widehat{\beta} = (X^T X)^- 1 X^T y$, so that the fitted values are $\widehat{y} = X\widehat{\beta} = X(X^T X)^- 1 X^T y$ and thus the hat matrix is $H = X(X^T X)^- 1 X^T$, $\widehat{y} = Hy$.[171] Hoaligan and Welsch derive that the average size of a diagonal element of the hat matrix and thus the average leverage point is equal to the number of parameters in the model divided by the number of observations (k/n), which can be easily done by using the properties of eigenvalues. They suggest to analyse leverage points larger 2 (k/n), which means that observations are analysed for which: $h_i > 2\frac{k}{n}$.[172] Observation are analysed that have a large residuals as well as a large leverage or one of both. For these observations a detailed analysis is necessary.[173] Bledsoe and Fries who were involved in the editing and imputation of the SCF point out that graphical analysis is very effective and recommend its use.[174] In order to get an overview of the most influential point a scatterplot of the standardized residuals against the leverage statistic can be used. For graphical analysis it is important to use standardized residuals because they are more suggestive.[175] The red lines in the scatterplot show the average of the standardized residuals and leverage statistic as an orientation.[176] Corresponding to the experience of Bledsoe and Fries the author finds that graphical analysis is very helpful to identify extreme observations.

[171] See Hoaglin/Welsch (1978), p. 17.
[172] See Hoaglin/Welsch (1978), p. 18.
[173] See Hoaglin/Welsch (1978), p. 17.
[174] See Bledsoe/Fries (2002), p. 5.
[175] See Andrews/Pregibon (1978), p. 86.
[176] The identification numbers of the observations can be drawn in, so that the observations that belong to the point in the graph can be identified.

Figure 3: Scatterplot of standardized residuals against leverage statistic



Source: Own graph.

However, it is very important to consider that these extreme observations might occur in reality instead of being caused by a mistake. Then the outlier is part of the distribution of the variable so that changing these observations would falsify the dataset.[177]

[177] See Wooldridge (2009), p. 325.

7.2 Further outlier statistics

There are many further statistics which can be used for the identification of outliers. Because of their importance for outlier identification the DFITS statistic and the DFBETA statistic will be described briefly. The DFITS statistics combines the leverage value and the residual in one measure and gives a scaled differences between the in sample and out of sample predicted values for the jth observation. For the statistic studentized residuals $r_j = \frac{e_j}{s_{(j)}\sqrt{1-h_j}}$ are used. $s_{(j)}$ is the root mean squared error of the regression equation estimation without the jth observation.[178]

$$DFITS_j = r_j\sqrt{\frac{h_j}{1-h_j}} \tag{27}$$

As a cutoff point $|\text{DFITS}_j| > 2\sqrt{\frac{k}{N}}$ can be used to identify highly influential observations.[179]
To measure the effect of an observation on a single regressor the DFBETA statistic can be used. The DFBETA for observation jth for regressor l is defined as.

$$DFBETA_j = \frac{r_j v_j}{\sqrt{v^2(1-h_j)}} \tag{28}$$

Thereby v_j is the residuals of the partial regression with the explanatory variable x_l as the dependent variable on the other explanatory variables of X. Furthermore v^2 is defined as the sum of squares of this regression. For the regressor l the DFBETA measures the shift of the coefficient when the jth observation is included or excluded, scaled by the estimated standard error of the coefficient. As a cutoff point for influential observations $|\text{DFBETA}_j| > \frac{2}{\sqrt{N}}$ can be used.[180]

[178] The root mean squared error is defined as $MSE = \sigma(\widehat{y})$ + (Bias$(\widehat{y}, y)^2$.
[179] See Baum (2006), p. 128.
[180] See Baum (2006), p. 130.

8 Editing and imputation

The identified outliers will be analysed in this chapter. Thereby it can be differentiated between editing and imputation.[181] Data editing is defined as the procedure to correct contradictory information in the data whereas imputation deals with missing data by a substitution with an estimated value. At first the data is edited whereby if no solution for contradictory information can be find the problematic variables of an observation are set to missing and imputed.[182] The editing and imputation of the specific values can't be shown in the do.file due to the data protection regulations of the Deutsche Bundesbank.

Besides the editing and imputation also the dealing with non-response rates is crucial to receive unbiased estimates. Non-response rates take place in two ways. Households might not want to participate in the survey or it might be not possible to locate them. Second, households do not answer all questions which is referred to as item non response. Item non response occurs especially for wealth variables and is not random but depends on households characteristics. For instance wealthy households tend to have a higher item non response concerning questions about the wealth of the household. Different parametric or non-parametric methods can be used to deal with non response. This issue has to be named since it important to receive correct estimates but will not be treated further in this thesis.[183]

8.1 Editing

The HFCS dataset is provided as edited and imputed to the user.[184] This is important according to Rubin (1996), who argues that the user might not have the resources to edit and impute the data and might have less information due to data protection reasons that might be important for the editing and imputation.[185] This argument fits very well for the PHF survey and in particular for the variable of the property value given that the regional information that are very important for the editing and estimation of the prediction equation are not available for the data user. To

[181] Besides the analysis of the observations identified with the procedure above, also more general checks from the ECB are applied to identify outliers, whereby most of the observations identified with the ECB checks where already identified with the approach used in this thesis.

[182] See ECB (2008), p. 1-3.

[183] See Barceló (2006), p. 9-10.

[184] See ECB (2008), p. 4.

[185] See Rubin (1996), p. 473-474.

make visible which values are edited, for every variable a special flag variable, which is also called shadow variable is created that contains a code that specifies any changes for every observation of the variable.[186] As pointed out by Kennickell (2006) editing is essentially important to improve the quality of the data. He shows this by analysing data from the SCF. Kennickell compares the regression results of a regression with income as the dependent variable and explanatory variables that should explain the income with imputed but unedited data as well as with imputed and edited data. Kennickell finds that the coefficients of the regression with the imputed and edited data are much more significant and the R^2 is much higher compared to the results with the only imputed data.[187] This underlines that data editing is very important for the consistency of the variables in the dataset.[188] The CAPI include checks that control for inconsistencies and thus reduces the amount of consistencies checks that have to be applied.[189] Given that the data set used in this thesis is the first dataset of a panel the CAPI has to be improved according to the experiences with the data. To find logical inconsistencies and to find out which variables are reliable and which are problematic is very complicated given that economic interdependencies between many variables have to be considered.[190] According to the editing rules established by Bledsoe and Fries (2002)[191] for the SCF an observations is edited only when it can be corrected due to logical reasoning from the supporting information. This approach is also applied for the HFCS.[192] For the editing the residual and leverage statistics are considered, whereby the leverage gives a hint for influential points in the explanatory variables. The danger exists that extreme observations can't be seen in the aggregated variable when they cancel each other out. However, if the value of an aggregated variable is extreme given the other variables the components of the aggregated variable have to be analysed. Furthermore the retransformed and corrected conditional expected value are used, with which the absolute residual can be calculated easily as the difference to the untransformed property price. Describing all relationships that are taken into consideration during the editing process would be to much for the purpose of this thesis. Nevertheless the following procedure is described in detail given that it might be used for the next wave of the survey in the research centre of the Deutsche Bundesbank. As in the regression relationships of the property value with variables on the active

[186] See ECB (2008), p. 8-9.
[187] See Kennickell (2006), p. 11-16.
[188] In the next chapter the regression results of the edited and imputed data will be compared with the results of the untreated data.
[189] See ECB (2008), p. 2.
[190] Furthermore para data that contains comments from the interviewer has to be considered.
[191] See Bledsoe/Fries (2002), p. 1.
[192] See ECB (2008), p. 2.

side and on the passive side are considered, whereby variables can be analysed on a lower level of aggregation since in this case multicollinearity doesn't have to be considered as problematic, instead we want to take the economic relationships between the variables into account. Generally in the case that variables of an observations are inconsistent it is difficult to figure out which of the variables are reliable and which variable is wrong and has to be edited. For this purpose historical data from other sources have to be considered to find out which values for the economic variables are realistic at a specific point in time. The following relationship are found as most helpful due to the work with the data, whereby of course more relationships can be find due to economic reasoning.[193] The regional information that are used internally in the Deutsche Bundesbank and are not provided to the user are very helpful.[194] The value of the property has to fit with its location and size as well as other characteristics of the household. The price change, given by the difference of the property value and the purchasing price have to make sense given the time since property acquisition as well as the regional variables. An increase in the property value might be due to the fact that the household renovated the property, which might be seen by the fact that he took a loan to renovate the property. The main reason for taking a loan are given by the categorial variable *PURPOSE OF THE LOAN* (hb120$a-hb120$d). The $ symbol is a loop variable from one to three, since the household might have several mortgages. Thus three variables are created and for them further variables for categories from a to d can be created to point out different purposes in a descending importance. Deviations between the estimated net wealth given by the variable *ESTIMATE OF WEALTH* (dhi0700), where the household is asked to estimate the total net wealth of the household and the aggregated net wealth given by an aggregated variable (nvermmin) from the components of the household wealth are important since deviations might come from a mistake in the variable of the property value. On the passive side of the household balance sheet mortgages, which have the property as a collateral are essentially important. The amount of the initial mortgages with the main property as collateral is given by the variable *INITIAL AMOUNT BORROWED* (hb140$). The mortgages have to fit with the purchasing price, whereby lending limits have to be considered. Furthermore the income should be compared with the mortgages of the household. According to the European Mortgage Federation for 2009 the average mortgage to property value ratio for new mortgages is 72 percent

[193] Most of the following housing variables can be find in the PHF questionnaire in section 3: Real Assets and their financing. They are not outlined in the appendix.

[194] Besides the variable for the federal states also the so called Kreiskennziffer is provided which allows a more detailed determination of the location of the property.

and the average property price to net income ratio is about 5 to 1.[195] Thereby it might be possible that mortgage is taken on the main property but used to buy other property which are given by the variable *OTHER PROPERTY MORTGAGES* (hb280$) and it also might be possible that mortgages are refinanced. Furthermore the main property might be used as collateral for additional borrowing which is asked in the variables *ADDITIONAL BORROWING ON HMR MORTGAGE* (hb150$). For the mortgage variables the interest rates given by the variables *ADJUSTABLE INTEREST RATE* (hb180$), *EFFECTIVE INTEREST RATE*(dhb560$) and *NOMINAL INTEREST RATE* (dhb561$) are given. Furthermore the variable *INSTALLMENT PAYMENT - AMOUNT* (dhb590$) is given, whereby it can be differentiated between monthly, quarterly and yearly payments by the variable *INSTALMENT PAYMENT - TIME PERIOD* (dhb591$). The day of the raising of credit is also given by the variable *YEAR WHEN LOAN TAKEN OUT OR REFINANCED* (hb130$). With these variables the amortisation schedule of the mortgage can be calculated to control whether they are consistent. This is possible in most cases but problematic in the case of extra repayments. The average interest rate for a duration of five to ten years is in average 3.6 percent for 2011.[196] Especially the case that the outstanding amount given by *AMOUNT STILL OWED* (hb170$) is equal to the variable *INITIAL AMOUNT BORROWED* occurs several times and is interesting when the raising of credit is already some time ago, so that the variable *AMOUNT STILL OWED* should be substantially smaller than the *INITIAL AMOUNT BORROWED*. However, most often these cases make sense because the household paid only the interest rate and as a result didn't reduce the amount of the mortgage. This fits with the repayment opportunities that the household have in Germany since the households can choose between amortizing and interest only mortgages.[197] Furthermore the variable *MEANS OF PROPERTY ACQUISITION* (dhb0410) that gives the most significant mean for the property acquisition is important. Also more general checks are applied for instance that the household doesn't repays more than his income and that the aggregated saving and consume is close to the income. Besides the relationships that might be helpful during the editing process pointed out above in most cases further variables have to be considered. The amount of variables that have to be considered make it visible how important it is to have a good prediction equation to identify the important cases.

[195] See European Mortgage Federation (2012), p. 6.
[196] See Deutsche Bundesbank (2012b), p. 54.
[197] See Lindenthal/Eichholtz (2011), p. 8.

8.2 Linear stochastic imputation

The variable of the property value is imputed for the values set as missing during the editing process by using the predicted values of the estimation equation. The explanatory variables in the prediction equation are already imputed and edited. In the case that also predictor variables have to be imputed an iterative and sequential imputation process has to be applied that is based on Markov Chain and Monte Carlo Methods and will be not covered within this thesis.[198] The goal of imputation is to preserve the distribution of the data and not to replace missing data by predicted values that are the most realistic values for the missing observations.[199] A possibility of the large number of imputation procedures is the regression imputation procedure.[200] Using the predicted values for the imputation would reduce the variance of the data so that besides the variance also quantiles and correlations of the data would be biased. The stochastic approach explained in the following addresses this problems and helps to preserve the distribution of the data, so that statistics are unbiased. A stochastic component has to be added to the conditional expected value.[201] This stochastic element reflects the known variation of the prediction values.[202] Thus for this purpose a random drawn residual from the distribution of the residuals of the property value regression will be added to the predicted property value. Basically the residual can be random drawn from the analytical distribution of the residuals or from the empirical distribution of the residuals. Drawing the residuals random from the empirical distribution is of course more complicated concerning the programming but will be implement. First of all random numbers have to be generated. This can be done with pseudo random number generator of STATA, which generates random numbers of a uniform distribution. Thereby random numbers from a uniform distribution from 1 to 1533 are drawn. For these random numbers observation from the empirical distribution of the residuals are drawn, whereby the numbers of the empirical distribution are sorted from 1 to 1533. In order to receive random numbers from the analytical normal distribution a function in STATA can be used that directly draws random numbers from the analytical normal distribution given as: $u \sim N(-0,000630; 0,3937969)$.[203] An other possibility, which is applied in the case when the direct drawing of random numbers from a particular density function is not possible is to use an algorithm that generates the random variable X as pseudo

[198] See Barceló (2006), p. 19.
[199] See Barceló (2006), p. 12.
[200] See Groves et al. (2004), p. 330-331.
[201] See Barceló (2006), p. 12.
[202] See Groves et al. (2004), p. 331.
[203] The programming in STATA can be find on page 115 in appendix 4.

random uniform distributed numbers over the interval [0,1]. To receive the random drawn values from a normal distribution the normal inverse distribution function is applied on X $\varepsilon = N^{-1}(X)$. This approach can be used to receive random numbers from any cumulative distribution function which has a inverse distribution probability function.[204] The equation for the imputed values is given as the predicted value by using $\widehat{\beta}_{FGLS}$ as well as the stochastic component drawn from the distribution of the residuals, whereby the analytical distribution is given as:

$$\widehat{y}^{imp.} = X\widehat{\beta}_{FGLS} + \widehat{u} \qquad with \qquad \widehat{u} \sim N(\widehat{\mu}, \widehat{\sigma}^2). \tag{29}$$

The HFCN uses multiple stochastic imputation whereby five replications are recommended.[205] However, in this thesis only single stochastic imputation is used, whereby the prediction equation estimated above is used for the imputation of the PHF dataset for the variable of the property value. Nevertheless linear multiple stochastic imputation will be described briefly.

[204] See Jorion (2001), p. 295.
[205] See ECB (2008), p. 8.

8.3 Linear multiple stochastic imputation

The model uncertainty is not taken into account by single stochastic imputation.[206] This means that single stochastic imputation takes only the within imputation variance into consideration, without to consider the between imputation variance that results from the uncertainty about the imputation model. This can be addressed by multiple stochastic imputation.[207] For the purpose of multiple imputation m complete data sets are imputed, which differ due to the stochastic component in the imputation model. The difference between the single and multiple stochastic imputation can be seen by analysing a statistic $\widehat{\Theta}_i$ which can be estimated for every imputed dataset $i = 1, ..., m$ ($\widehat{\Theta}_i$ might be for instance the mean). The combined statistic for the m imputed dataset is then given as average as $\overline{\Theta}_m = \frac{1}{M}\sum_{i=1}^{m}\widehat{\Theta}_i$. The variance of $\overline{\Theta}_m$ has two components. The average within-imputation variance is given by the average of the within imputation variances $\widehat{\sigma}_i^2$ as: $\overline{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\widehat{\sigma}_i^2$. The between imputation variance is given as the variance between the $\widehat{\Theta}_i$ as: $\overline{\Sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\Theta}_i - \overline{\Theta}_m)^2$. The total variance T_m^2 of $\overline{\Theta}_m$ is given as a linear combination of the within-imputation variance ($\overline{\sigma}_m^2$) and between-imputation variance as $\overline{\Sigma}_m^2$.[208]

$$T_m^2 = \overline{\sigma}_m^2 + \frac{m+1}{m}\overline{\Sigma}_m^2 \qquad T_m^2 = \frac{1}{m}\sum_{i=1}^{m}\widehat{\sigma}_i^2 + \frac{m+1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\Theta}_i - \overline{\Theta}_m)^2 \tag{30}$$

Depending on how large $\overline{\Sigma}_m^2$ is the single stochastic imputation underestimates the variance of statistics. It can be seen that when m increases the efficiency of $\overline{\Theta}_m$ increases since its variance T_m^2 becomes smaller.[209] For the PHF dataset multiple imputation with five imputed datasets is used (m=5). [210] This is a generally accepted norm.[211]For multiple imputation for m larger than five the efficiency gain is very low compared to the effort.[212]

[206] See Deutsche Bundesbank (2012a), p. 38.
[207] See Barceló (2006), p. 17.
[208] See ECB (2008), p. 7-8.
[209] See Barceló (2008), p. 18.
[210] See ECB (2008), p. 8.
[211] See Deutsche Bundesbank (2012a), p. 38.
[212] See Barceló (2006), p. 18.

9 Edited and imputed data analysis

In the following the distribution of the edited and imputed data will be analysed and compared with the distribution of the original data. The estimation equation as well as the corresponding tests will be estimated with the edited and imputed data. Only the transformed data will be analysed due to the fact that the stochastic component of the imputation process and the non-linear retransformation with the hyperbolic sine or exponential function would cause further statistical problems for the retransformation of the data which are not the focus of this thesis.[213] In a second step the residuals of the regression with the edited and imputed data will be analysed and compared with the residuals of the estimation with the original data.

9.1 Comparison of the edited and imputed data with the original data

The differences of the imputed and original data can be seen by the differences of fundamental statistics and the histograms. The following table shows statistics for the original and the edited and imputed data.[214]

Table 12: Statistics of the imputed transformed estimated property values

Data	N	mean	standard deviation
Edited and Imputed	1863	13.05665	.7208222
Original	1848	13.04487	.7199863

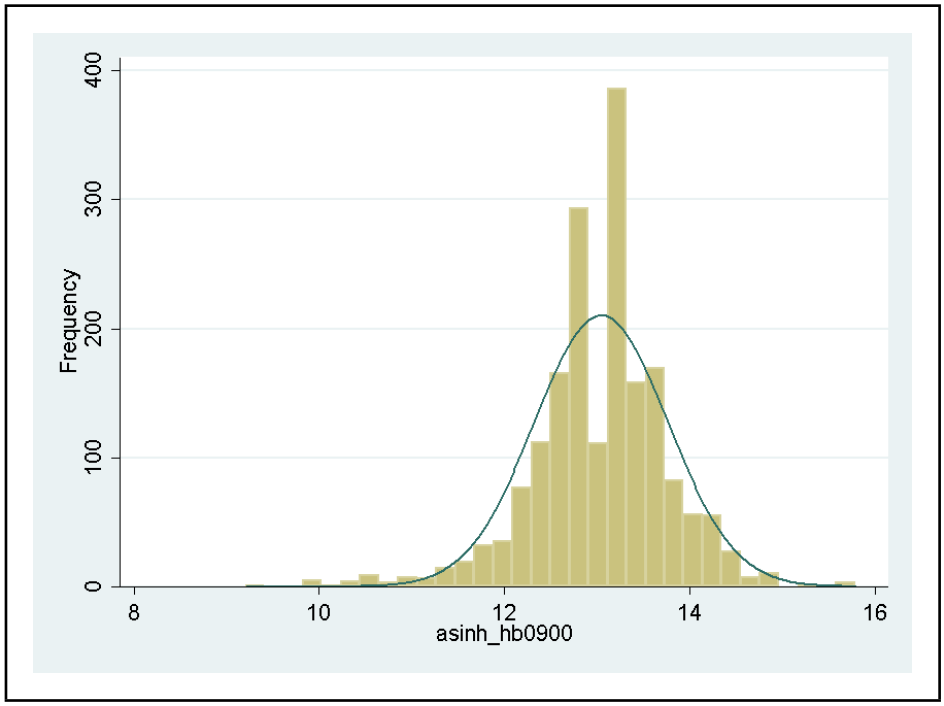
Source: Own calculation.

The mean and standard deviations are very close. This shows that the stochastic imputation helped to maintain the structure of the data and doesn't let to a reduction of the variance. The histograms of the property prices can be find on the next page. It can be seen that the distribution of the edited and imputed property value is closer to a normal distribution which should be the result of a reduction of excentrical values.

[213] See Ramirez et al. (1994), p. 289-300.

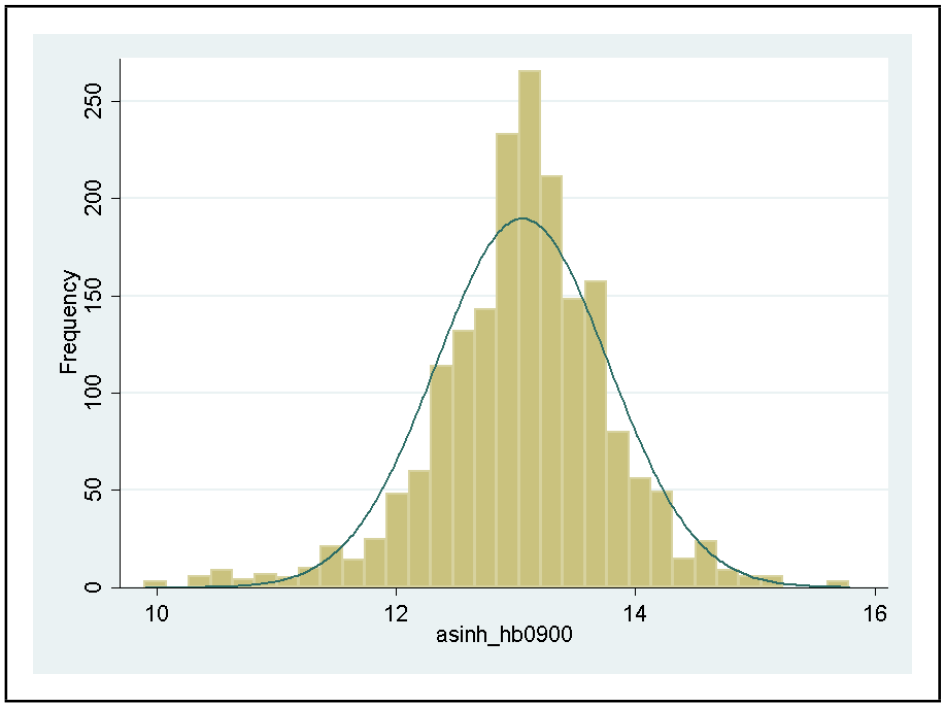
[214] Maximum and minimum of the data can't be shown due to the data security policy of the Deutsche Bundesbank.

Figure 4: Histogram of transformed variable PROPERTY PRICE



Source: Own graph.

Figure 5: Histogram of the edited and imputed variable PROPERTY PRICE



Source: Own graph.

9.2 Estimations with the edited and imputed data

The regressions with the edited and imputed data can be seen in appendix 1 in table 23 (FGLS regression imputed and edited data). The first column contains the regional OLS regression with the original data for comparison purposes (old data). The seconds column contains the FGLS regression with the edited and imputed data while using the old weights estimated for the regression with the original data (old weights). In the fourth column an auxiliary regression for the estimation of weights with the edited and imputed data is estimated. The last column contains the corresponding FGLS regression with the new weights (new weights). The coefficients of the regression differ only slightly from the coefficients of the regression with the original data. The coefficients of the regression with the old weights and weights estimated with the edited and imputed data seem to be similar. To estimate new weights seems to be the more consistent approach, so that the test results for this regression will be pointed out whereby the test results for the other FGLS regression are very similar. The R^2 increases when running the regression with the imputed and edited data and the coefficients are in tendency to be more significant. This fits with the experience of Kennickell (2006), whereby he compares the regression results of the only imputed dataset with the edited and imputed dataset and in this thesis the results of an already edited and imputed dataset are compared with a dataset where the dependent variable is substantially changed due to editing and imputation.[215] The R^2 is 0.73 and the $R^2_{adj.}$ is 0.719. The Breusch Pagan test again has a very high test statistic ($X(1) = 31.01$) and thus a p-value of 0.0000. According to this test the residuals are still highly heteroscedastic. Thus the amount of heteroscedasticity caused by the edited and imputed outliers seems be not very high. The RESET test has the following results: $F(3,1558) = 5.5$ so that the p-value is 0.0008 which is close to the results of the estimation with the untreated data. This shows that for the model estimated the results of the RESET test are robust to outliers in the data.

9.3 Comparison of the residuals

The residuals of the estimation with the original and with the edited and imputed data should be especially different in the tails of the distributions since the observations with large residuals where edited and imputed when considered as problematic. The following statistics show that the extreme values for the residuals of the estimation with the edited and imputed values are smaller. The 10%

[215] See Kennickell (2006), p. 11-16.

quantiles of the residuals differ only slightly. Whereby differences in the tails of the distribution concerning the size of the distribution can't be seen by the quantiles. Thus besides the extreme values also the histograms are compared.

Table 13: Residuals of the original and edited and imputed data

Data	N	mean	sd	min	max
Original	1619	-.0006299	.3937969	-2.087974	1.726213
Imputed and Edited	1622	-.0002338	.3577497	-1.196623	1.694757

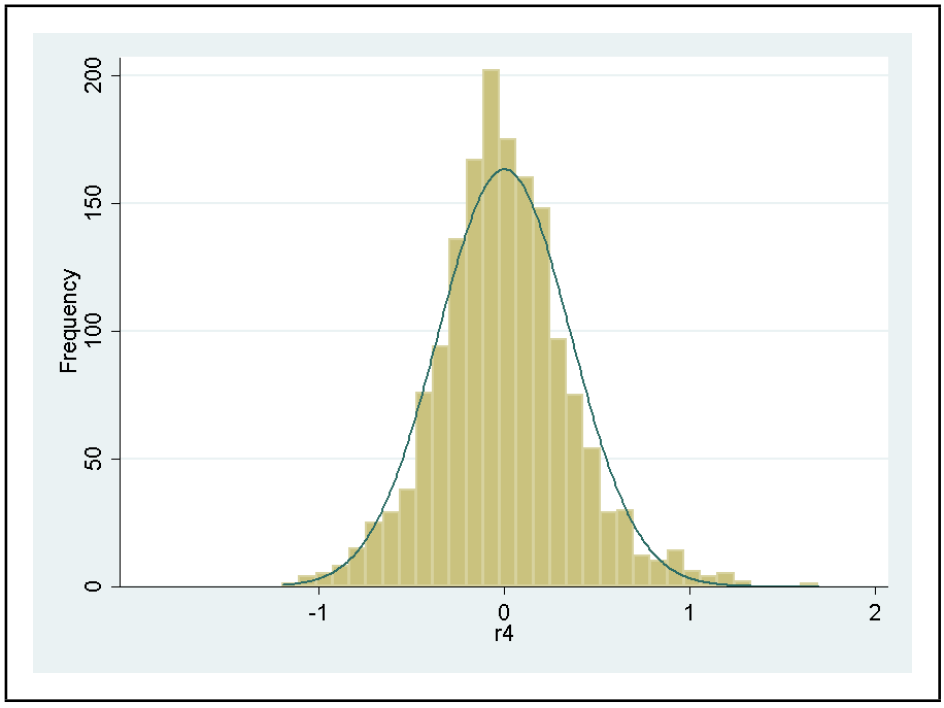
Source: Own calculation.

Table 14: Quantiles for the original and edited and imputed residuals

Data	10%	20%	30%	40%	50%	60%	70%	80%	90%	97.5%
Original	-.435	-.281	-.171	-.083	-.008	.067	.162	.267	.457	.867
Edited and Imputed	-.424	-.271	-.170	-.088	-.016	.067	.160	.260	.436	.796

Source: Own calculation.

Figure 6: Histogram of residuals with the edited and imputed data



Source: Own graph.

The histogram show that most of the very extreme and unlikely residuals are vanished for the edited and imputed data.[216] This is especially true for negative residuals, which came from a underestimation of the true property price whereby there are still some large residuals. These observations belong to the distribution and are caused by extreme property price increases for properties located in large cities. Analysing the residuals of the already edited and imputed data is in particular very helpful to identify mistakes in the editing process. Observations that were edited might have a large residual or leverage statistic for the regression with the edited and imputed data which is a hint for an editing mistake. Given the complexity of the data set mistakes in the editing process are easily possible since not all relevant variables can be considered so that the necessary reduction of the complexity can be a cause for editing mistakes. Furthermore to optimise the editing and imputation the regression estimation and outliers detection should be applied on the imputed data in an iterative process, which means that after the editing and imputation the regression is again applied to search for new outliers. This can be used for the detection of further problematic values and thus for an optimisation of the quality of the dataset.[217] Given that after the first editing and imputation no further problematic values can be identified this is not applied in this case.

[216] The histogram of the standardized residuals of the edited and imputed data can be seen in appendix 3, figure 18.
[217] See Hellerstein (2008), p. 17-18.

10 Conclusion and ideas for further research

The hedonic regression model doesn't explains the variation of the property values sufficient, so that it can't be used for outlier identification and imputation. This might be the case because further hedonic variables should be included into the model which are not in the dataset or that the hedonic regression has to be enhanced with variables which are not justified with hedonic theory. The further variables included into the model are highly significant whereby changes in the hedonic variables can be seen due to the omitted variable bias. For the purpose of imputation and outlier identification thus a hedonic based approach enhanced with further variables seems to work well. Testing the model for deviations from the OLS assumptions we find that multicollinearity is not a problem according to the standard tests. This should be especially the result of the aggregation of explanatory variables that are correlated. On the other hand the data is highly heteroscedastic as common from microdata and even after the transformation of the data and the modeling of heteroscedasticity in the FGLS regression it is not possible to deal with it. This might be due to the problems to estimate a model for heteroscedasticity for the FGLS regression based on the assumptions that the households have different abilities to estimate the true property value. The model has a very low explanatory power, however interesting is that the variable derived from the financial literary variables shows the expected sign and is significant. This approach is new in the literature to the knowledge of the author. As expected regional differences and the population density have the expected signs and are highly significant and important for the explanation of the property value. Level effects for the different federal states in Germany included by dummy variables as well as interactions terms are significant and included in the model to reflect the complex determination of property value. Analysing the residuals is a very efficient method to identify outliers in the explained variable. This should be in particular the case due to the transformation of the data with the inverse hyperbolic sine so that the residuals of properties with the same relative increase of the property price are equal wherefore no bias to properties with a large absolute value exists. Also the inverse hyperbolic sine solves the problem that the natural logarithm can't be used for data that includes zeros very well and thus fits for the micro data which includes several variables that take the value zero. Especially with the leverage statistics influential points in the explanatory variables can be detected. After the editing and stochastic imputation the outliers are not part of the distribution of the property value anymore, which can be seen at the distribution of the residuals of the models estimated with the edited and imputed data set. For the estimation of hedonic regressions further hedonic variables

should be included into the PHF dataset that provide more detailed information about the property and are commonly used for hedonic regression like the number of garages and bedrooms or spacial measure like the distance to the next bus station. This should allow to receive also hedonic models with a sufficient explanatory power like some of the hedonic models in the literature. However, for the purpose of outlier identification and stochastic imputation it can be concluded that the variables available in the dataset are sufficient to estimate a prediction equation with a high explanatory power. The procedures applied are useful to identify and replace problematic values in the property variable, so that the edited and imputed variable is a much more reliable basis for further economic research and policy decisions. Given that with the following waves of the HFCS panel data is available this will also allow to estimate hedonic panel regression, which should be especially interesting for property markets given their large variation that became very visible with the crisis.[218] Building on the research in this paper it might be interesting to compare the estimated property values with transaction data or data from other property valuation methods.[219] With that research concerning the accuracy of property price estimates fitting in the literature discussed in chapter 2.3 might be interesting. This is only one idea to use opportunities that the edited and imputed data offers for further research.

Acknowledgements

The development of the estimation procedure and outlier identification was only possible with the support of the Private Household and Finance team. I especially want thank Dr. habil. Ulf von Kalckreuth for his constant support and highly didactic explanations. Julia Le Blanc taught me how to deal with the complexity of the editing process. Dr. Tobias Schmidt helped especially with the STATA programming and Junyi Zhu, PhD with the understanding of imputation process and the transformation of the data. Martin Eisele provided the framework for working with the data and helped me to use also sensible variables in my thesis by taking the data protection regulations into account. Furthermore I would like to thank Dr. Astrid Krenz for her constant support as well as Professor Gerhard Rübel for his support to write my bachelor thesis with the Deutsche Bundesbank.

[218] See Kajuth/Schmidt (2011), p. 1.
[219] Such data might be for instance received from the BulwienGesa AG.

Appendix 1: Tables

Table 15: OLS models

Variables	OLS all	hedonic	hed. significant
PURCHASING PRICE	0.888***		
	[0.0319]		
OTHER PROPERTY	0.00865		
	[0.00966]		
CARS	-0.0288		
	[0.182]		
FINANCIAL WEALTH	-0.000771		
	[0.00260]		
MORTGAGES	-0.0206		
	[0.0478]		
INCOME	1.338**		
	[0.537]		
MEMBERS	929.8		
	[4265]		
EDUCATION			
Secondary school	36763		
	[62872]		
Higher secondary school	34334		
	[63428]		
East German standard school	45683		
	[66901]		
Entrance diploma FH	42996		
	[65099]		
University entrance diploma	39113		
	[64228]		
Other	25920		
	[82280]		
No school-leaving qualification	0		
	[0]		
EDUCATION prof.			
Apprenticeship	-53495		
	[39859]		
Vocational school/commercial college	-53203		
	[42161]		
Technical college	-40684		
	[40818]		
University of applied science	-33561		
	[41490]		
University degree/teacher training	-54010		
	[40514]		
Doctoral/postdoctoral training	-29124		
	[43781]		
Other	-5013		
	[73041]		
No training completed	-49224		

OLS models

Variables	OLS all	hedonic	hed. significant
	[43715]		
EMPLOYED			
Ordinary employed but not currently	17361		
	[29240]		
Not employed	16456		
	[10454]		
SIZE	512.8***	1289***	1239***
	[94.48]	[104.9]	[87.89]
LAND	2.315	1.704	
	[2.214]	[2.424]	
TIME since acq.	3462***	-325.4	
	[401.9]	[370.9]	
BUILDING TYPE			
Semi-detached house	-18798	-39534**	-36082**
	[13127]	[15903]	[14715]
Multiple family dwelling	22807*	18682	-21629
	[13385]	[16074]	[13504]
Farm	93076**	165567***	207768***
	[45126]	[50532]	[39569]
Building with various uses	182230***	284313***	236265***
	[32116]	[39393]	[34667]
Non-detached house	-41733***	-87462***	-82489***
	[13783]	[16820]	[15726]
Other	-66612	17567	3168
	[47269]	[54407]	[50963]
DWELLING RATE			
Very Good	13928	-39868*	-58734***
	[17733]	[21122]	[19409]
Satisfactory	19701	-66464***	-85128***
	[20945]	[24922]	[21702]
Simple	-27321	-120172***	-118177***
	[30529]	[36210]	[29148]
Very simple	-20629	-137159**	-127298***
	[52471]	[61615]	[43949]
DWELLING LOCATION			
City centre and suburbs	5255	17720	
	[21396]	[26083]	
Suburbs or city outskirts	5587	34958	
	[20684]	[25185]	
Rural area	-7652	4286	
	[20829]	[25361]	
DWELLING OUTWARD			
Small cracks in the facade	2858	4820	
	[15836]	[19029]	
Needs major renovation	91902***	78710**	
	[32909]	[38479]	

OLS models

Variables	OLS all	hedonic	hed. significant
Dilapidated	-24429	-98746	
	[113564]	[139397]	
NEIGHBOURHOOD			
same condition than others	-36573*	-33223	
	[19590]	[23209]	
building in better state	-18237	-9367	
	[22484]	[26781]	
No other building nearby	-19829	-45326	
	[39134]	[46494]	
SURROUNDING			
Good	-35673***	-61891***	-65989***
	[11952]	[14484]	[13456]
Satisfactory	-51743***	-90045***	-103230***
	[16405]	[19764]	[17706]
Adequate	-65642**	-97463***	-113658***
	[26898]	[32031]	[28430]
Unsatisfactory	-113247*	-140758**	-100941*
	[59052]	[69804]	[60739]
Poor	-243020	-202592	-368242**
	[199828]	[245432]	[145281]
INTERIOR			
Good	-17630	-22056	
	[11698]	[14218]	
Fair	-40559	-20923	
	[26819]	[32121]	
Poor	-8974	7779	
	[70495]	[80660]	
FEDERAL STATES			
West	43574**	43723*	32517
	[21954]	[26460]	[24191]
West	40086	31609	54173*
	[29019]	[35736]	[31899]
West	-53869	-120430**	-91497**
	[40707]	[50318]	[46361]
West	16382	15250	23049
	[16677]	[20419]	[18692]
West	11337	31119	17843
	[19514]	[24068]	[21696]
West	15352	8685	6241
	[23134]	[28338]	[26430]
West	32095*	79326***	70845***
	[17474]	[21272]	[19346]
West	68626***	131452***	120269***
	[17580]	[21157]	[19238]
West	-2368	-8582	15091
	[38514]	[47566]	[41713]
West	-1162	693.8	21372

OLS models

Variables	OLS all	hedonic	hed. significant
	[27984]	[34593]	[30834]
East	55237	-12940	-3548
	[41839]	[46594]	[41888]
East	29876	-2693	-11636
	[35670]	[41814]	[38633]
East	-23934	-100712**	-111600***
	[34976]	[40333]	[34859]
East	-11008	-51059	-43369
	[36032]	[40056]	[36710]
East	20111	-25264	-17266
	[39795]	[44432]	[41457]
POPULATION DENSITY			
$P < 2000$	-79999**	-183562***	-150126***
	[35493]	[41896]	[38336]
$2000 < P < 5000$	-80461**	-208400***	-195863***
	[31603]	[36765]	[32590]
$5000 < P < 20.000$	-41333**	-128189***	-118923***
	[20949]	[24895]	[22254]
$20.000 < P < 50.000$	-86039***	-187596***	-165940***
	[21559]	[25537]	[22887]
$50.000 < P < 100.000$	-77825***	-192565***	-167244***
	[22079]	[26346]	[23034]
$50.000 < P < 100.000$	-52845	-151607***	-118372**
	[48067]	[53980]	[47927]
$100.00 < P < 500.000$	-75835***	-170105***	-152345***
	[15267]	[17996]	[15778]
$100.00 < P < 500.000$	-40511**	-94978***	-77999***
	[16417]	[19830]	[17523]
$P \geq 500.000$	-37935**	-89308***	-68921***
	[15067]	[18268]	[16075]
Constant	36932	307760***	305978***
	[86823]	[48949]	[30727]
Observations	1532	1601	1834
R-squared	0.617	0.377	0.364
R-squared adj.	0.597	0.356	0.350
AIC	41094.26	43645.43	49895.77
BIC	41510.34	43941.24	50121.86
Standard errors in brackets *** $p < 0.01, ** p < 0.05, *p < 0.1$			

Source: Own calculations

Table 16: Transformed regression model long

Variables	trans. all	trans. signif.	hed. trans. all	hed. trans. signif.
PURCHASING PRICE	0.267***	0.266***		
	[0.0147]	[0.0147]		
OTHER PROPERTY	0.00413**	0.00413**		
	[0.00184]	[0.00184]		
CARS	0.00787**	0.00868**		
	[0.00366]	[0.00363]		
FINANCIAL WEALTH	0.0189***	0.0196***		
	[0.00466]	[0.00464]		
MORTGAGES	0,00281	0,0025		
	[0.00198]	[0.00192]		
INCOME	0,0129	0,0102		
	[0.00894]	[0.00875]		
Anzhhm	0,00117			
	[0.0107]			
EDUCATION				
Secondary school	0.227			
	[0.414]			
Higher secondary school	0.233			
	[0.414]			
East German standard school	0.251			
	[0.418]			
Entrance diploma FH	0.308			
	[0.416]			
University entrance diploma	0.250			
	[0.415]			
Other	0.152			
	[0.437]			
No school-leaving qualification	-0.271			
	[0.444]			
EDUCATION prof.				
Apprenticeship	-0.256**	-0.259**		
	[0.102]	[0.102]		
Vocational school/commercial college	-0.286***	-0.279***		
	[0.108]	[0.107]		
Technical college	-0.231**	-0.225**		
	[0.105]	[0.105]		
University of applied science	-0.231**	-0.203*		
	[0.107]	[0.106]		
University degree/teacher training	-0.239**	-0.227**		
	[0.104]	[0.103]		
Doctoral/postdoctoral training	-0.162	-0.149		
	[0.112]	[0.111]		
Other	-0.208	-0.179		
	[0.172]	[0.172]		
No training completed	-0.275**	-0.292***		
	[0.111]	[0.110]		
EMPLOYED				
Ordinary employed but not currently	0.141*			

Transformed regression model long

Variables	trans. all	trans. signif.	hed. trans. all	hed. trans. signif.
	[0.0738]			
Not employed	0,031			
	[0.0262]			
SIZE	0.00191***	0.00193***	0.00318***	0.00311***
	[0.000228]	[0.000226]	[0.000256]	[0.000242]
LAND	0.111***	0.110***	0.136***	0.129***
	[0.0147]	[0.0147]	[0.0171]	[0.0161]
TIME since acq.	0.00610***	0.00599***	-0.00300***	-0.00328***
	[0.00106]	[0.000982]	[0.000887]	[0.000853]
BUILDING TYPE				
Semi-detached house	-0.0624*	-0.0648*	-0.0688*	-0.0748**
	[0.0331]	[0.0330]	[0.0382]	[0.0368]
Multiple family dwelling	0,0384	0,0308	0,0188	0,00667
	[0.0343]	[0.0342]	[0.0389]	[0.0377]
Farm	0.205*	0.194*	0.154	0.140
	[0.106]	[0.106]	[0.112]	[0.107]
Building with various uses	0.392***	0.385***	0.532***	0.490***
	[0.0765]	[0.0762]	[0.0933]	[0.0853]
Non-detached house	-0.0848**	-0.0849**	-0.101**	-0.0987**
	[0.0369]	[0.0367]	[0.0425]	[0.0414]
Other	-0,0293	-0,0232	0,0383	-0,000486
	[0.115]	[0.115]	[0.129]	[0.123]
INTERIOR				
Very Good	-0.0734*	-0,0669	-0.105**	-0.124**
	[0.0433]	[0.0432]	[0.0500]	[0.0480]
Satisfactory	-0.137***	-0.133***	-0.231***	-0.261***
	[0.0489]	[0.0488]	[0.0589]	[0.0539]
Simple	-0.205***	-0.202***	-0.391***	-0.421***
	[0.0669]	[0.0669]	[0.0856]	[0.0736]
Very simple	-0.283***	-0.295***	-0.738***	-0.735***
	[0.105]	[0.105]	[0.146]	[0.113]
DWELLING LOCATION				
City centre and suburbs			-0,0598	
			[0.0618]	
Suburbs or city outskirts			-0,0508	
			[0.0597]	
Rural area			-0.120**	
			[0.0603]	
DWELLING OUTWARD				
Small cracks in the facade			0,0183	
			[0.0450]	
Needs major renovation			0,0534	
			[0.0910]	
Dilapidated			0,0717	
			[0.330]	
NEIGHBOURHOOD				

Transformed regression model long

Variables	trans. all	trans. signif.	hed. trans. all	hed. trans. signif.
same condition than others			0,00898	
			[0.0550]	
building in better state			0,00782	
			[0.0635]	
No other building nearby			-0.106	
			[0.110]	
SURROUNDING				
Good	-0.111***	-0.114***	-0.137***	-0.144***
	[0.0298]	[0.0297]	[0.0343]	[0.0333]
Satisfactory	-0.155***	-0.161***	-0.223***	-0.240***
	[0.0406]	[0.0403]	[0.0468]	[0.0446]
Adequate	-0.237***	-0.233***	-0.253***	-0.308***
	[0.0660]	[0.0660]	[0.0757]	[0.0722]
Unsatisfactory	-0.267*	-0.244*	-0.211	-0.185
	[0.146]	[0.146]	[0.165]	[0.159]
Poor	-1.036***	-1.020***	-0.723	-1.076***
	[0.312]	[0.312]	[0.579]	[0.355]
INTERIOR				
Good			-0.0907***	
			[0.0337]	
Fair			-0.157**	
			[0.0761]	
Poor			0,0721	
			[0.191]	
FEDERAL STATE				
West	0.134**	0.139**	0.108*	0,0719
	[0.0542]	[0.0542]	[0.0625]	[0.0601]
West	0.224***	0.226***	0.211**	0.213***
	[0.0700]	[0.0700]	[0.0846]	[0.0797]
West	0,0268	0,0301	-0.185	-0.151
	[0.102]	[0.102]	[0.119]	[0.115]
West	0.109***	0.109***	0.105**	0.112**
	[0.0413]	[0.0413]	[0.0483]	[0.0464]
West	0.136***	0.137***	0.168***	0.158***
	[0.0489]	[0.0489]	[0.0571]	[0.0552]
West	0.139**	0.137**	0.110	0,0918
	[0.0580]	[0.0580]	[0.0672]	[0.0651]
West	0.230***	0.229***	0.308***	0.300***
	[0.0444]	[0.0443]	[0.0511]	[0.0494]
West	0.339***	0.340***	0.411***	0.408***
	[0.0437]	[0.0434]	[0.0501]	[0.0483]
West	0.150	0.165*	0,0397	0,0812
	[0.0927]	[0.0925]	[0.113]	[0.105]
West	0,0672	0,0643	0,0737	0,0737
	[0.0696]	[0.0691]	[0.0818]	[0.0788]
East	0.280***	0.270***	-0.208*	-0.177*
	[0.101]	[0.0991]	[0.110]	[0.103]

Transformed regression model long

Variables	trans. all	trans. signif.	hed. trans. all	hed. trans. signif.
East	-0.146	-0.193**	-0.499***	-0.544***
	[0.0903]	[0.0886]	[0.0990]	[0.0957]
East	-0.646***	-0.638***	-0.980***	-1.147***
	[0.0875]	[0.0837]	[0.0955]	[0.0893]
East	-0.318***	-0.315***	-0.620***	-0.601***
	[0.0920]	[0.0886]	[0.0948]	[0.0925]
East	-0,00185	-0,00267	-0.294***	-0.298***
POPULATION DENSITY				
	[0.101]	[0.0956]	[0.105]	[0.102]
$P < 2000$	-0.207**	-0.205**	-0.451***	-0.450***
	[0.0874]	[0.0875]	[0.0992]	[0.0947]
$2000 < P < 5000$	-0.448***	-0.455***	-0.610***	-0.665***
	[0.0772]	[0.0771]	[0.0875]	[0.0822]
$5000 < P < 20.000$	-0.284***	-0.286***	-0.460***	-0.484***
	[0.0515]	[0.0513]	[0.0592]	[0.0558]
$20.000 < P < 50.000$	-0.541***	-0.543***	-0.699***	-0.719***
	[0.0540]	[0.0539]	[0.0611]	[0.0589]
$50.000 < P < 100.000$	-0.446***	-0.452***	-0.605***	-0.619***
	[0.0539]	[0.0535]	[0.0626]	[0.0587]
$50.000 < P < 100.000$	-0.486***	-0.486***	-0.621***	-0.592***
	[0.115]	[0.115]	[0.128]	[0.120]
$100.00 < P < 500.000$	-0.354***	-0.353***	-0.461***	-0.475***
	[0.0376]	[0.0374]	[0.0429]	[0.0409]
$100.00 < P < 500.000$	-0.152***	-0.151***	-0.184***	-0.177***
	[0.0405]	[0.0404]	[0.0470]	[0.0450]
$P \geq 500.000$	-0.157***	-0.161***	-0.208***	-0.210***
	[0.0369]	[0.0368]	[0.0433]	[0.0413]
Constant	8.401***	8.690***	12.281***	12.310***
	[0.485]	[0.252]	[0.159]	[0.132]
Observations	1618	1619	1601	1689
R-squared	0.675	0.671	0.565	0.569
R-squared adj.	0.661	0.659	0.550	0.558
AIC	1744.036	1742.386	2165.349	2268.29
BIC	2255.655	2049.592	2461.16	2501.861

$***p < 0.01, **p < 0.05, *p < 0.1$
Standard errors in brackets

Source: Own calculations

Table 17: Transformed regression model short

Variables	1	2	3	4	5	6	7
PURCHASING PRICE	0.268***	0.279***					
	[0.0146]	[0.0147]					
OTHER PROPERTY	0.00456**		0.00747***	0.00656***	0.00432**	0.00466**	0.00415**
	[0.00183]		[0.00198]	[0.00198]	[0.00199]	[0.00200]	[0.00202]
CARS	0.00945***			0.0167***	0.0136***	0.0137***	0.0130***
	[0.00360]			[0.00393]	[0.00393]	[0.00392]	[0.00394]
FINANCIAL WEALTH	0.0211***				0.0284***	0.0289***	0.0266***
	[0.00460]				[0.00476]	[0.00476]	[0.00492]
MORTGAGES	0,00245					0.00409*	0.00368*
	[0.00192]					[0.00209]	[0.00210]
INCOME	0,0116						0.0169*
	[0.00864]						[0.00908]
SIZE	0.00199***	0.00224***	0.00300***	0.00293***	0.00278***	0.00277***	0.00273***
	[0.000226]	[0.000226]	[0.000242]	[0.000242]	[0.000240]	[0.000240]	[0.000241]
LAND	0.108***	0.115***	0.122***	0.121***	0.121***	0.121***	0.121***
	[0.0147]	[0.0148]	[0.0162]	[0.0161]	[0.0159]	[0.0159]	[0.0159]
TIME since acq.	0.00592***	0.00596***	-0.00343***	-0.00324***	-0.00371***	-0.00295***	-0.00271***
	[0.000982]	[0.000916]	[0.000850]	[0.000847]	[0.000842]	[0.000925]	[0.000934]
BUILDING TYPE							
Semi-detached house	-0.0652**	-0.0573*	-0.0838**	-0.0832**	-0.0758**	-0.0781**	-0.0824**
	[0.0330]	[0.0334]	[0.0368]	[0.0366]	[0.0362]	[0.0362]	[0.0363]
Multiple family dwelling	0,0266	0,0293	-0,00549	-0,00293	0,00771	0,0102	0,00848
	[0.0342]	[0.0345]	[0.0377]	[0.0375]	[0.0371]	[0.0371]	[0.0371]
Farm	0.184*	0.186*	0.141	0.149	0.128	0.124	0.130
	[0.106]	[0.107]	[0.106]	[0.106]	[0.105]	[0.105]	[0.105]
Building with various uses	0.379***	0.404***	0.471***	0.484***	0.467***	0.461***	0.456***
	[0.0761]	[0.0770]	[0.0851]	[0.0847]	[0.0839]	[0.0839]	[0.0839]
Non-detached house	-0.0806**	-0.0805**	-0.106**	-0.100**	-0.0952**	-0.0987**	-0.0980**
	[0.0367]	[0.0372]	[0.0413]	[0.0411]	[0.0407]	[0.0407]	[0.0407]
Other	-0,0313	-0,0697	0,00478	-0,000632	0,0339	0,0436	0,0439
	[0.115]	[0.117]	[0.123]	[0.122]	[0.121]	[0.121]	[0.121]
DWELLING RATE							

Transformed regression model short

Variables	1	2	3	4	5	6	7
Very Good	-0.0709*	-0.0841*	-0.111**	-0.115**	-0.112**	-0.112**	-0.111**
	[0.0428]	[0.0434]	[0.0479]	[0.0476]	[0.0471]	[0.0471]	[0.0471]
Satisfactory	-0.139***	-0.171***	-0.242***	-0.240***	-0.222***	-0.221***	-0.220***
	[0.0485]	[0.0490]	[0.0539]	[0.0537]	[0.0532]	[0.0532]	[0.0531]
Simple	-0.213***	-0.254***	-0.406***	-0.394***	-0.368***	-0.365***	-0.364***
	[0.0666]	[0.0674]	[0.0734]	[0.0731]	[0.0724]	[0.0724]	[0.0723]
Very simple	-0.300***	-0.383***	-0.722***	-0.692***	-0.604***	-0.601***	-0.603***
	[0.105]	[0.106]	[0.113]	[0.112]	[0.112]	[0.112]	[0.112]
SURROUNDING							
Good	-0.119***	-0.120***	-0.145***	-0.141***	-0.141***	-0.143***	-0.141***
	[0.0296]	[0.0300]	[0.0331]	[0.0330]	[0.0326]	[0.0326]	[0.0326]
Satisfactory	-0.169***	-0.168***	-0.242***	-0.240***	-0.237***	-0.240***	-0.238***
	[0.0401]	[0.0407]	[0.0445]	[0.0442]	[0.0438]	[0.0438]	[0.0438]
Adequate	-0.242***	-0.246***	-0.305***	-0.305***	-0.309***	-0.314***	-0.308***
	[0.0660]	[0.0670]	[0.0719]	[0.0715]	[0.0708]	[0.0708]	[0.0708]
Unsatisfactory	-0.251*	-0.229	-0.199	-0.201	-0.232	-0.228	-0.220
	[0.146]	[0.148]	[0.159]	[0.158]	[0.156]	[0.156]	[0.156]
Poor	-1.020***	-1.132***	-1.036***	-1.024***	-0.960***	-0.956***	-0.953***
	[0.312]	[0.317]	[0.353]	[0.352]	[0.348]	[0.348]	[0.348]
FEDERAL STATES							
West	0.136**	0.126**	0,0729	0,0763	0,0925	0,092	0,0946
	[0.0541]	[0.0549]	[0.0599]	[0.0595]	[0.0590]	[0.0589]	[0.0589]
West	0.215***	0.214***	0.219***	0.218***	0.214***	0.214***	0.214***
	[0.0700]	[0.0710]	[0.0794]	[0.0790]	[0.0782]	[0.0781]	[0.0781]
West	0,0257	0,0184	-0.154	-0.131	-0.126	-0.130	-0.130
	[0.101]	[0.103]	[0.115]	[0.114]	[0.113]	[0.113]	[0.113]
West	0.105**	0.114***	0.107**	0.105**	0.104**	0.101**	0.102**
	[0.0413]	[0.0419]	[0.0463]	[0.0460]	[0.0456]	[0.0455]	[0.0455]
West	0.133***	0.142***	0.148***	0.146***	0.151***	0.149***	0.147***
	[0.0489]	[0.0496]	[0.0551]	[0.0548]	[0.0542]	[0.0542]	[0.0542]
West	0.133**	0.133**	0,0891	0,0891	0,0947	0,0961	0,097

Transformed regression model short

Variables	1	2	3	4	5	6	7
	[0.0579]	[0.0588]	[0.0648]	[0.0645]	[0.0638]	[0.0638]	[0.0637]
West	0.226***	0.238***	0.292***	0.289***	0.280***	0.283***	0.284***
	[0.0441]	[0.0447]	[0.0492]	[0.0490]	[0.0485]	[0.0485]	[0.0484]
West	0.332***	0.351***	0.394***	0.387***	0.383***	0.385***	0.388***
	[0.0433]	[0.0438]	[0.0483]	[0.0481]	[0.0476]	[0.0475]	[0.0475]
West	0.159*	0.135	0,0748	0,078	0.114	0.112	0.114
	[0.0921]	[0.0934]	[0.105]	[0.104]	[0.103]	[0.103]	[0.103]
West	0,0665	0,0641	0,0733	0,066	0,0737	0,0768	0,0788
	[0.0691]	[0.0702]	[0.0784]	[0.0781]	[0.0773]	[0.0772]	[0.0772]
East	0.279***	0.301***	-0.171*	-0.181*	-0.157	-0.166	-0.162
	[0.0991]	[0.100]	[0.103]	[0.102]	[0.101]	[0.101]	[0.101]
East	-0.185**	-0.225**	-0.530***	-0.524***	-0.487***	-0.488***	-0.485***
	[0.0886]	[0.0898]	[0.0954]	[0.0949]	[0.0941]	[0.0941]	[0.0940]
East	-0.622***	-0.700***	-1.125***	-1.075***	-1.040***	-1.031***	-1.025***
	[0.0836]	[0.0840]	[0.0892]	[0.0895]	[0.0887]	[0.0888]	[0.0888]
East	-0.303***	-0.333***	-0.582***	-0.552***	-0.539***	-0.546***	-0.547***
	[0.0884]	[0.0893]	[0.0923]	[0.0921]	[0.0912]	[0.0911]	[0.0911]
East	0,00772	0,0148	-0.302***	-0.293***	-0.302***	-0.297***	-0.299***
	[0.0951]	[0.0965]	[0.102]	[0.101]	[0.100]	[0.100]	[0.100]
POPULATION DENSITY							
$P < 2000$	-0.211**	-0.235***	-0.445***	-0.433***	-0.416***	-0.421***	-0.415***
	[0.0875]	[0.0886]	[0.0943]	[0.0939]	[0.0930]	[0.0929]	[0.0929]
$2000 < P < 5000$	-0.457***	-0.498***	-0.646***	-0.649***	-0.589***	-0.595***	-0.584***
	[0.0770]	[0.0775]	[0.0820]	[0.0816]	[0.0814]	[0.0813]	[0.0815]
$5000 < P < 20.000$	-0.299***	-0.333***	-0.470***	-0.471***	-0.435***	-0.436***	-0.436***
	[0.0511]	[0.0515]	[0.0557]	[0.0554]	[0.0552]	[0.0552]	[0.0551]
$20.000 < P < 50.000$	-0.559***	-0.590***	-0.708***	-0.691***	-0.677***	-0.679***	-0.677***
	[0.0536]	[0.0542]	[0.0587]	[0.0585]	[0.0580]	[0.0579]	[0.0579]
$50.000 < P < 100.000$	-0.469***	-0.479***	-0.604***	-0.613***	-0.600***	-0.608***	-0.601***
	[0.0532]	[0.0536]	[0.0587]	[0.0584]	[0.0578]	[0.0579]	[0.0580]
$50.000 < P < 100.000$	-0.502***	-0.580***	-0.560***	-0.563***	-0.514***	-0.503***	-0.502***
	[0.115]	[0.116]	[0.120]	[0.119]	[0.118]	[0.118]	[0.118]

Transformed regression model short

Variables	1	2	3	4	5	6	7
$100.00 < P < 500.000$	-0.356***	-0.373***	-0.462***	-0.469***	-0.449***	-0.451***	-0.449***
	[0.0371]	[0.0374]	[0.0409]	[0.0407]	[0.0404]	[0.0404]	[0.0404]
$100.00 < P < 500.000$	-0.151***	-0.149***	-0.179***	-0.180***	-0.174***	-0.175***	-0.171***
	[0.0404]	[0.0410]	[0.0448]	[0.0446]	[0.0442]	[0.0441]	[0.0441]
$P \geq 500.000$	-0.161***	-0.167***	-0.209***	-0.209***	-0.197***	-0.200***	-0.200***
	[0.0366]	[0.0371]	[0.0411]	[0.0409]	[0.0405]	[0.0405]	[0.0405]
Constant	8.412***	8.650***	12.33***	12.20***	11.93***	11.90***	11.78***
	[0.233]	[0.226]	[0.132]	[0.135]	[0.141]	[0.142]	[0.155]
Observations	1619	1619	1689	1689	1689	1689	1689
R-squared	0.668	0.656	0.573	0.578	0.586	0.587	0.588
R-squared adj.	0.658	0.647	0.562	0.566	0.575	0.576	0.577
AIC	1740.989	1786.993	2255.655	2239.181	2204.94	2203.022	2201.467
BIC	2005.078	2024.134	2494.658	2483.616	2454.807	2458.321	2462.198

$***p < 0.01, **p < 0.05, *p < 0.1$
Standard errors in brackets

Source: Own calculations

Table 18: Quadratic terms

Variables	all	quadratic	quadratic sig.
PURCHASING PRICE	0.268***	0.256***	0.258***
	[0.0146]	[0.0147]	[0.0146]
OTHER PROPERTY	0.00456**	-0.0127	0.00428**
	[0.00183]	[0.0134]	[0.00182]
CARS	0.00945***	-0.000281	0.00790**
	[0.00360]	[0.0133]	[0.00360]
FINANCIAL WEALTH	0.0211***	0.0223*	0.0198***
	[0.00460]	[0.0133]	[0.00460]
MORTGAGES	0,00245	-0.0516**	0.00268
	[0.00192]	[0.0216]	[0.00190]
INCOME	0,0116	-0.0106	0.00821
	[0.00864]	[0.0248]	[0.00868]
EDUCATION			
Apprenticeship		-0.240**	-0.243**
		[0.101]	[0.101]
Vocational school/commercial college		-0.269**	-0.266**
		[0.106]	[0.107]
Technical college		-0.202*	-0.202*
		[0.104]	[0.104]
University of applied science		-0.180*	-0.187*
		[0.105]	[0.105]
University degree/teacher training		-0.216**	-0.219**
		[0.102]	[0.102]
Doctoral/postdoctoral training		-0.127	-0.131
		[0.110]	[0.110]
Other		-0.144	-0.160
		[0.170]	[0.170]
No training completed		-0.272**	-0.270**
		[0.109]	[0.109]
SIZE	0.00199***	0.00601***	0.00602***
	[0.000226]	[0.000767]	[0.000765]
LAND	0.108***	0.146*	0.154*
	[0.0147]	[0.0811]	[0.0812]
TIME since acq.	0.00592***	0.000751	0.00587***
	[0.000982]	[0.00266]	[0.000975]
BUILDING TYPE			
Semi-detached house	-0.0652**	-0.0570*	-0.0579*
	[0.0330]	[0.0327]	[0.0328]
Multiple family dwelling	0,0266	0.0635*	0.0615*
	[0.0342]	[0.0344]	[0.0344]
Farm	0.184*	0.232**	0.220**
	[0.106]	[0.109]	[0.109]
Building with various uses	0.379***	0.378***	0.395***
	[0.0761]	[0.0756]	[0.0755]
Non-detached house	-0.0806**	-0.0718**	-0.0737**
	[0.0367]	[0.0365]	[0.0365]
Other	-0,0313	0.0515	0.0511
	[0.115]	[0.116]	[0.116]

Quadratic terms

Variables	all	quadratic	quadratic sig.
DWELLING RATE			
Very Good	-0.0709*	-0.0620	-0.0791*
	[0.0428]	[0.0432]	[0.0429]
Satisfactory	-0.139***	-0.115**	-0.136***
	[0.0485]	[0.0489]	[0.0484]
Simple	-0.213***	-0.190***	-0.203***
	[0.0666]	[0.0665]	[0.0663]
Very simple	-0.300***	-0.285***	-0.287***
	[0.105]	[0.104]	[0.104]
SURROUNDING			
Good	-0.119***	-0.106***	-0.109***
	[0.0296]	[0.0294]	[0.0294]
Satisfactory	-0.169***	-0.145***	-0.149***
	[0.0401]	[0.0399]	[0.0399]
Adequate	-0.242***	-0.223***	-0.234***
	[0.0660]	[0.0652]	[0.0654]
Unsatisfactory	-0.251*	-0.248*	-0.258*
	[0.146]	[0.144]	[0.144]
Poor	-1.020***	-0.869***	-0.932***
	[0.312]	[0.311]	[0.310]
FEDERAL STATES			
West	0.136**	0.148***	0.153***
	[0.0541]	[0.0537]	[0.0538]
West	0.215***	0.220***	0.229***
	[0.0700]	[0.0694]	[0.0693]
West	0,0257	0.0300	0.0300
	[0.101]	[0.101]	[0.101]
West	0.105**	0.112***	0.115***
	[0.0413]	[0.0408]	[0.0409]
West	0.133***	0.142***	0.138***
	[0.0489]	[0.0483]	[0.0485]
West	0.133**	0.131**	0.132**
	[0.0579]	[0.0573]	[0.0574]
West	0.226***	0.225***	0.229***
	[0.0441]	[0.0438]	[0.0439]
West	0.332***	0.335***	0.343***
	[0.0433]	[0.0429]	[0.0430]
West	0.159*	0.166*	0.162*
	[0.0921]	[0.0916]	[0.0916]
West	0,0665	0.0706	0.0727
	[0.0691]	[0.0688]	[0.0685]
East	0.279***	0.278***	0.285***
	[0.0991]	[0.0980]	[0.0982]
East	-0.185**	-0.136	-0.168*
	[0.0886]	[0.0881]	[0.0879]
East	-0.622***	-0.608***	-0.618***
	[0.0836]	[0.0829]	[0.0829]
East	-0.303***	-0.293***	-0.292***

Quadratic terms

Variables	all	quadratic	quadratic sig.
	[0.0884]	[0.0877]	[0.0879]
East	0,00772	0.0142	0.0120
	[0.0951]	[0.0946]	[0.0947]
POPULATION DENSITY			
$P < 2000$	-0.211**	-0.203**	-0.221**
	[0.0875]	[0.0868]	[0.0867]
$2000 < P < 5000$	-0.457***	-0.430***	-0.446***
	[0.0770]	[0.0763]	[0.0764]
$5000 < P < 20.000$	-0.299***	-0.287***	-0.291***
	[0.0511]	[0.0507]	[0.0508]
$20.000 < P < 50.000$	-0.559***	-0.527***	-0.530***
	[0.0536]	[0.0536]	[0.0535]
$50.000 < P < 100.000$	-0.469***	-0.441***	-0.450***
	[0.0532]	[0.0530]	[0.0530]
$50.000 < P < 100.000$	-0.502***	-0.454***	-0.468***
	[0.115]	[0.114]	[0.114]
$100.00 < P < 500.000$	-0.356***	-0.344***	-0.352***
	[0.0371]	[0.0372]	[0.0371]
$100.00 < P < 500.000$	-0.151***	-0.152***	-0.149***
	[0.0404]	[0.0401]	[0.0401]
$P \geq 500.000$	-0.161***	-0.165***	-0.166***
	[0.0366]	[0.0365]	[0.0365]
OTHER PROPERTY squared	8.412***	0.00133	
	[0.233]	[0.00104]	
CARS squared		0.000783	
		[0.00117]	
FINANCIAL WEALTH squared		-0.000235	
		[0.000835]	
MORTGAGES squared		0.00437**	
		[0.00174]	
INCOME squared		0.00148	
		[0.00193]	
TIME since acq. squared		0.000116**	
		[5.19e-05]	
SIZE squared		-1.02e-05***	-9.99e-06***
		[1.80e-06]	[1.79e-06]
LAND squared		-0.00309	-0.00363
		[0.00581]	[0.00582]
Constant		8.427***	8.301***
		[0.364]	[0.358]
Observations	1619	1619	1619
R-squared	0.668	0.681	0.678
R-squared adj.	0.658	0.668	0.666
AIC		1708.453	1713.631
BIC		2058.774	2031.616

$***p < 0.01, **p < 0.05, *p < 0.1$
Standard errors in brackets

Quadratic terms

Variables	all	quadratic	quadratic sig.

Source: Own calculations

Table 19: Regional regression models

Variables	all	all sig.	West	East
PURCHASING PRICE	0.277***	0.272***	0.295***	0.170***
	[0.0168]	[0.0169]	[0.0160]	[0.0439]
PURCHASING PRICE int.	-0.0878***	-0.0509**		
	[0.0270]	[0.0255]		
OTHER PROPERTY	0.00399**	0.00416**	0.00413**	0.00820
	[0.00186]	[0.00181]	[0.00172]	[0.0106]
OTHER PROPERTY int.	0.00549			
	[0.00676]			
CARS	0.000711	0.00754**	0.00163	0.0281**
	[0.00385]	[0.00358]	[0.00357]	[0.0140]
CARS int.	0.0404***			
	[0.00977]			
FINANCIAL WEALTH	0.0149***	0.0122**	0.0163***	0.0344*
	[0.00483]	[0.00484]	[0.00448]	[0.0194]
FINANCIAL WEALTH int.	0.0405***	0.0602***		
	[0.0132]	[0.0126]		
MORTGAGES	0.00246	0.00244	0.00253	0.00454
	[0.00199]	[0.00189]	[0.00185]	[0.00832]
INCOME	0.00224	0.0114	0.00438	0.0586
	[0.00902]	[0.00865]	[0.00838]	[0.0400]
INCOME int.	0.0879***			
	[0.0284]			
EDUCATION prof.				
Apprenticeship	-0.223**	-0.235**	-0.192*	-0.473
	[0.0994]	[0.100]	[0.101]	[0.353]
Vocational school/commercial college	-0.251**	-0.253**	-0.218**	-0.472
	[0.105]	[0.106]	[0.106]	[0.389]
Technical college	-0.185*	-0.192*	-0.175*	-0.337
	[0.102]	[0.103]	[0.103]	[0.372]
University of applied science	-0.168	-0.180*	-0.165	-0.338
	[0.104]	[0.104]	[0.105]	[0.370]
University degree/teacher training	-0.205**	-0.212**	-0.211**	-0.290
	[0.101]	[0.102]	[0.102]	[0.362]
Doctoral/postdoctoral training	-0.121	-0.125	-0.108	-0.283
	[0.109]	[0.109]	[0.110]	[0.389]
Other	-0.188	-0.160	-0.184	
	[0.168]	[0.169]	[0.160]	
No training completed	-0.250**	-0.253**	-0.216**	-0.743*
	[0.107]	[0.108]	[0.108]	[0.401]
SIZE	0.00615***	0.00593***	0.00614***	0.00385
	[0.000756]	[0.000760]	[0.000734]	[0.00457]

Regional regression models

Variables	all	all sig.	West	East
SIZE squared	-1.02e-05***	-9.73e-06***	-1.04e-05***	-5.30e-06
	[1.77e-06]	[1.78e-06]	[1.70e-06]	[1.26e-05]
LAND	0.148*	0.155*	0.111	0.505
	[0.0800]	[0.0807]	[0.0757]	[0.727]
LAND squared	-0.00328	-0.00376	-0.00116	-0.0240
	[0.00573]	[0.00578]	[0.00545]	[0.0494]
TIME since acq.	0.00612***	0.00608***	0.00685***	0.00282
	[0.000968]	[0.000975]	[0.000951]	[0.00435]
BUILDING TYPE				
Semi-detached house	-0.0672**	-0.0625*	-0.0417	-0.230
	[0.0323]	[0.0325]	[0.0317]	[0.145]
Multiple family dwelling	0.0648*	0.0617*	0.0428	0.262*
	[0.0339]	[0.0342]	[0.0335]	[0.157]
Farm	0.202*	0.196*	0.184*	0.494
	[0.108]	[0.109]	[0.106]	[0.498]
Building with various uses	0.377***	0.387***	0.330***	0.793*
	[0.0744]	[0.0751]	[0.0714]	[0.422]
Non-detached house	-0.0747**	-0.0735**	-0.0722**	-0.0620
	[0.0359]	[0.0362]	[0.0351]	[0.169]
Other	0.0957	0.0772	0.0773	-0.346
	[0.114]	[0.115]	[0.114]	[0.471]
DWELLING RATE				
Very Good	-0.0813*	-0.0724*	-0.0902**	-0.0641
	[0.0423]	[0.0426]	[0.0417]	[0.174]
Satisfactory	-0.142***	-0.131***	-0.157***	-0.111
	[0.0477]	[0.0481]	[0.0467]	[0.211]
Simple	-0.192***	-0.191***	-0.195***	-0.455
	[0.0653]	[0.0659]	[0.0641]	[0.295]
Very simple	-0.185*	-0.236**	-0.0818	-0.624*
	[0.104]	[0.104]	[0.108]	[0.351]
SURROUNDING				
Good	-0.105***	-0.105***	-0.0821***	-0.150
	[0.0290]	[0.0292]	[0.0292]	[0.115]
Satisfactory	-0.148***	-0.148***	-0.123***	-0.159
	[0.0394]	[0.0397]	[0.0391]	[0.177]
Adequate	-0.228***	-0.224***	-0.195***	-0.443*
	[0.0644]	[0.0650]	[0.0665]	[0.238]
Unsatisfactory	-0.300**	-0.262*	-0.224	-0.795
	[0.142]	[0.144]	[0.140]	[0.620]
Poor	-1.057***	-0.985***	-1.102***	
	[0.306]	[0.309]	[0.288]	
FEDERAL STATES				
West	0.152***	0.152***	0.151***	
	[0.0530]	[0.0535]	[0.0491]	
West	0.228***	0.229***	0.234***	
	[0.0682]	[0.0688]	[0.0632]	
West	0.0315	0.0358	0.0512	

Regional regression models

Variables	all	all sig.	West	East
	[0.0992]	[0.100]	[0.0919]	
West	0.117***	0.115***	0.119***	
	[0.0402]	[0.0406]	[0.0373]	
West	0.137***	0.135***	0.135***	
	[0.0477]	[0.0481]	[0.0442]	
West	0.133**	0.131**	0.134**	
	[0.0565]	[0.0571]	[0.0525]	
West	0.223***	0.226***	0.206***	
	[0.0432]	[0.0436]	[0.0402]	
West	0.344***	0.343***	0.337***	
	[0.0423]	[0.0427]	[0.0393]	
West	0.164*	0.159*	0.168**	
	[0.0901]	[0.0910]	[0.0835]	
West	-0.445	0.0516		-0.283
	[0.388]	[0.342]		[0.364]
East	-0.301	0.252		0.286
	[0.370]	[0.326]		[0.239]
East	-0.587	-0.112		0.422**
	[0.360]	[0.316]		[0.206]
East	-0.965***	-0.591*		-0.319
	[0.355]	[0.310]		[0.237]
East	-0.725**	-0.287		
	[0.367]	[0.320]		
East	-0.521	-0.0101		0.0775
	[0.369]	[0.321]		[0.292]
POPULATION DENSITY				
$P < 2000$	-0.230***	-0.225***	-0.307***	0.194
	[0.0856]	[0.0862]	[0.0827]	[0.526]
$2000 < P < 5000$	-0.449***	-0.450***	-0.400***	-1.098***
	[0.0753]	[0.0760]	[0.0751]	[0.398]
$5000 < P < 20.000$	-0.295***	-0.297***	-0.281***	
	[0.0503]	[0.0508]	[0.0470]	
$20.000 < P < 50.000$	-0.507***	-0.515***	-0.368***	-1.443***
	[0.0527]	[0.0532]	[0.0538]	[0.299]
$50.000 < P < 100.000$	-0.452***	-0.456***	-0.447***	-1.016***
	[0.0524]	[0.0529]	[0.0515]	[0.330]
$50.000 < P < 100.000$	-0.482***	-0.485***	-0.474***	
	[0.112]	[0.113]	[0.104]	
$100.00 < P < 500.000$	-0.354***	-0.356***	-0.340***	-1.110***
	[0.0369]	[0.0372]	[0.0349]	[0.348]
$100.00 < P < 500.000$	-0.159***	-0.150***	-0.162***	-0.609*
	[0.0396]	[0.0400]	[0.0379]	[0.327]
$P \geq 500.000$	-0.165***	-0.166***	-0.152***	-0.816**
	[0.0360]	[0.0363]	[0.0339]	[0.333]
Constant	8.214***	8.166***	8.056***	7.971***
	[0.367]	[0.370]	[0.349]	[2.691]
Observations	1619	1619	1422	197
R-squared	0.689	0.683	0.625	0.791

Regional regression models

Variables	all	all sig.	West	East
R-squared adj.	0.677	0.670	0.611	0.730
AIC	1665.177	1692.604	1235.615	357.1189
BIC	2015.499	2021.368	1514.386	504.8631

$***p < 0.01, **p < 0.05, *p < 0.1$
Standard errors in brackets

Source: Own calculations

Table 20: FGLS residual model

Variables	e1sq	e1sq sign.	le1sq	e1
ALTERPERSON	0,00035	-.0000783	0,00318	-0,000389
	[0.000977]	[0.0007949]	[0.00679]	[0.00117]
TIME since acq.	0.00149*	.0014704	0.0172***	-0,000153
	[0.000815]	[0.000798]	[0.00566]	[0.000973]
EDUCATION prof.				
Apprenticeship	-0,011		0.255	0,00139
	[0.0837]		[0.582]	[0.0999]
Vocational school/commercial college	0,0068		0.599	-0,0022
	[0.0886]		[0.616]	[0.106]
Technical college	-0,00375		0.278	0,00154
	[0.0864]		[0.600]	[0.103]
University of applied science	-0,0273		-0.178	0,00296
	[0.0880]		[0.612]	[0.105]
University degree/teacher training	0,00135		0.247	-0,000692
	[0.0856]		[0.595]	[0.102]
Doctoral/postdoctoral training	-0,0382		0.265	0,000331
	[0.0920]		[0.640]	[0.110]
Other	-0,0929		-0.368	-0,0117
	[0.140]		[0.971]	[0.167]
No training completed	0,0542		0.793	0,0016
	[0.0900]		[0.625]	[0.107]
EMPLOYED				
Ordinary employed but not currently	0,0311		-0.185	0.115
	[0.0591]		[0.410]	[0.0705]
Not employed	-0,0159		-0.253	0,0352
	[0.0230]		[0.159]	[0.0274]
INCOME	-0,00859	-0.0093569	-0.114**	0,00198
	[0.00665]	[0.0062844]	[0.0462]	[0.00793]
CORRECT		-0.0178067	0,0105	0,0125
		[0.012449]	[0.0882]	[0.0152]
WRONG	0,0149			
	[0.0127]			
Constant	0.186*	0.2586747	-3.152***	-0,0435
	[0.102]	0.077973	[0.743]	[0.128]
Observations	1618	1618	1618	1618
R-squared	0.011	0.0077	0.024	0.003
R-squared adj.	0,00282	0.0053	0,0153	-0,00578

$***p < 0.01, **p < 0.05, *p < 0.1$
Standard errors in brackets

Source: Own calculations

Table 21: FGLS residuals model all

Variables	e1sq	lesq
ALTERPERSON	0,000543	0,00522
	[0.000983]	[0.00695]
EMPLOYED		
ordinary employed but not currently	0,0214	-0.257
	[0.0569]	[0.402]
not employed	-0,00756	-0.184
	[0.0224]	[0.158]
CORRECT	-0,000854	0,091
	[0.0128]	[0.0908]
PURCHASING PRICE	-0.0650***	-0.360***
	[0.0133]	[0.0939]
PURCHASING PRICE int.	0,0107	0.291**
	[0.0200]	[0.142]
OTHER PROPERTY	0,00174	0,0165
	[0.00142]	[0.0101]
CARS	-0,00215	-0,0235
	[0.00283]	[0.0200]
FINANCIAL WEALTH	-0,000287	-0,0337
	[0.00381]	[0.0269]
FINANCIAL WEALTH int.	0,00107	-0,0793
	[0.00989]	[0.0699]
MORTGAGES	0,000404	0,0143
	[0.00154]	[0.0109]
INCOME	-0,00397	-0.0892*
	[0.00692]	[0.0489]
EDUCATION prof.		
Apprenticeship	0,00788	0.240
	[0.0811]	[0.573]
Vocational school/commercial college	0,0247	0.529
	[0.0857]	[0.605]
Technical college	0,0339	0.325
	[0.0835]	[0.590]
University of applied science	0,0114	-0.110
	[0.0850]	[0.601]
University degree/teacher training	0,0369	0.367
	[0.0827]	[0.584]
Doctoral/postdoctoral training	-0,00619	0.295
	[0.0892]	[0.630]
Other	-0,0466	-0.124
	[0.136]	[0.960]
No training completed	0,0761	0.691
	[0.0878]	[0.621]
SIZE	-0,000233	-0,00167
	[0.000598]	[0.00422]
SIZE squared	1.60e-06	1.02e-05
	[1.40e-06]	[9.90e-06]
LAND	0,0504	0.891**
	[0.0634]	[0.448]

FGLS residuals model all

Variables	e1sq	lesq
LAND squared	-0,00204	-0.0548*
	[0.00454]	[0.0321]
TIME since acq.	-0,000581	0,00823
	[0.000903]	[0.00638]
BUILDING TYPE		
Semi-detached house	-0,0332	-0.268
	[0.0256]	[0.181]
Multiple family dwelling	0.0581**	0.359*
	[0.0269]	[0.190]
Farm	0,0721	0.490
	[0.0857]	[0.606]
Building with various uses	0.131**	0.776*
	[0.0591]	[0.418]
Non-detached house	-0.0557*	-0.294
	[0.0285]	[0.201]
Other	0.298***	1.911***
	[0.0905]	[0.639]
DWELLING RATE		
Very Good	-0,00165	-0,0105
	[0.0335]	[0.237]
Satisfactory	-0,0133	-0.260
	[0.0378]	[0.267]
Simple	0,000529	-0.105
	[0.0518]	[0.366]
Very simple	0,0425	-1.167**
	[0.0821]	[0.581]
SURROUNDING		
Good	-0,00315	-0,0433
	[0.0230]	[0.163]
Satisfactory	0,00384	0.402*
	[0.0313]	[0.221]
Adequate	-0,0368	-0.164
	[0.0510]	[0.361]
Unsatisfactory	0.129	1.265
	[0.113]	[0.799]
Poor	-0.250	1.015
	[0.243]	[1.715]
FEDERAL STATES		
West	0,00106	-0.211
	[0.0420]	[0.297]
West	-0,00326	-0.570
	[0.0541]	[0.383]
West	-0,00283	-0.365
	[0.0788]	[0.557]
West	0,00393	0.294
	[0.0319]	[0.225]
West	0,0365	0.435

FGLS residuals model all

Variables	e1sq	lesq
	[0.0379]	[0.268]
West	0,0514	0.816**
	[0.0448]	[0.317]
West	0.0657*	0.870***
	[0.0344]	[0.243]
West	0.0895***	0.598**
	[0.0335]	[0.237]
West	-0,0105	-0.106
	[0.0715]	[0.506]
West	-0.143	-2.358
	[0.268]	[1.897]
East	0,0115	-1.372
	[0.256]	[1.808]
East	-0,00609	-1.497
	[0.248]	[1.756]
East	0.209	-0.895
	[0.244]	[1.721]
East	0.280	-1.517
	[0.251]	[1.776]
East	-0,00488	-0.643
	[0.252]	[1.780]
POPULATION DENSITY		
$P < 2000$	-0.126*	-0.400
	[0.0679]	[0.480]
$2000 < P < 5000$	-0.111*	-1.154***
	[0.0599]	[0.423]
$5000 < P < 20.000$	-0,0217	-0,0673
	[0.0401]	[0.283]
$20.000 < P < 50.000$	-0.136***	-0.592**
	[0.0418]	[0.296]
$50.000 < P < 100.000$	-0.0753*	-0.981***
	[0.0417]	[0.295]
$50.000 < P < 100.000$	-0,0635	-0.119
	[0.0896]	[0.634]
$100.00 < P < 500.000$	-0.0809***	-0.298
	[0.0292]	[0.207]
$100.00 < P < 500.000$	-0.0630**	-0.518**
	[0.0314]	[0.222]
$P \geq 500.000$	-0.0706**	-0.401**
	[0.0286]	[0.202]
Constant	0.735**	-2.240
	[0.295]	[2.082]
Observations	1618	1618
R-squared	0.134	0.115
R-squared adj.	0,0978	0,0781
Standard errors in brackets $***p < 0.01, **p < 0.05, *p < 0.1$		

FGLS residuals model all

Variables	e1sq	lesq

Source: Own calculations

Table 22: FGLS regression

Variables	all sig. (from regional models)	weight: e1sq sign.	weight: le1sq
PURCHASING PRICE	0.272***	0.297***	0.305***
	[0.0169]	[0.0172]	[0.0172]
PURCHASING PRICE int.	-0.0509**	-0.0681***	-0.0890***
	[0.0255]	[0.0261]	[0.0261]
OTHER PROPERTY	0.00416**	0.00395**	0.00387**
	[0.00181]	[0.00177]	[0.00174]
CARS	0.00754**	0.00675*	0.00676*
	[0.00358]	[0.00353]	[0.00357]
FINANCIAL WEALTH	0.0122**	0.0129***	0.0140***
	[0.00484]	[0.00479]	[0.00483]
FINANCIAL WEALTH int.	0.0602***	0.0570***	0.0587***
	[0.0126]	[0.0127]	[0.0129]
MORTGAGES	0,00244	0,00247	0,00192
	[0.00189]	[0.00182]	[0.00179]
INCOME	0,0114	0,0139	0,0154
	[0.00865]	[0.00963]	[0.0105]
EDUCATION			
Apprenticeship	-0.235**	-0.204**	-0.204**
	[0.100]	[0.0982]	[0.0852]
Vocational school/commercial college	-0.253**	-0.216**	-0.206**
	[0.106]	[0.104]	[0.0945]
Technical college	-0.192*	-0.161	-0.154*
	[0.103]	[0.101]	[0.0888]
University of applied science	-0.180*	-0.155	-0.156*
	[0.104]	[0.102]	[0.0880]
University degree/teacher training	-0.212**	-0.188*	-0.185**
	[0.102]	[0.0997]	[0.0869]
Doctoral/postdoctoral training	-0.125	-0,0928	-0,0862
	[0.109]	[0.107]	[0.0945]
Other	-0.160	-0.149	-0.169
	[0.169]	[0.175]	[0.140]
No training completed	-0.253**	-0.233**	-0.218**
	[0.108]	[0.107]	[0.101]
SIZE	0.00593***	0.00577***	0.00575***
	[0.000760]	[0.000731]	[0.000710]
SIZE squared	-9.73e-06***	-9.25e-06***	-9.33e-06***
	[1.78e-06]	[1.69e-06]	[1.63e-06]
LAND	0.155*	0.134*	0.110
	[0.0807]	[0.0782]	[0.0794]
LAND squared	-0,00376	-0,00244	-0,000359
	[0.00578]	[0.00562]	[0.00574]

FGLS regression

Variables	all sig. (from regional models)	weight: e1sq sign.	weight: le1sq
TIME since acq.	0.00608***	0.00648***	0.00652***
	[0.000975]	[0.000982]	[0.000990]
BUILDING TYPE			
Semi-detached house	-0.0625*	-0.0587*	-0.0547*
	[0.0325]	[0.0314]	[0.0308]
Multiple family dwelling	0.0617*	0.0587*	0.0623*
	[0.0342]	[0.0337]	[0.0335]
Farm	0.196*	0.139	0.149
	[0.109]	[0.111]	[0.116]
Building with various uses	0.387***	0.382***	0.390***
	[0.0751]	[0.0730]	[0.0759]
Non-detached house	-0.0735**	-0.0706**	-0.0744**
	[0.0362]	[0.0354]	[0.0348]
Other	0,0772	0,0899	0.102
	[0.115]	[0.117]	[0.118]
DWELLING RATE			
Very Good	-0.0724*	-0,0668	-0,0621
	[0.0426]	[0.0407]	[0.0395]
Satisfactory	-0.131***	-0.127***	-0.131***
	[0.0481]	[0.0463]	[0.0453]
Simple	-0.191***	-0.184***	-0.195***
	[0.0659]	[0.0653]	[0.0655]
Very simple	-0.236**	-0.195*	-0.180*
	[0.104]	[0.105]	[0.105]
SURROUNDING			
Good	-0.105***	-0.104***	-0.112***
	[0.0292]	[0.0283]	[0.0273]
Satisfactory	-0.148***	-0.141***	-0.130***
	[0.0397]	[0.0388]	[0.0380]
Adequate	-0.224***	-0.225***	-0.220***
	[0.0650]	[0.0648]	[0.0646]
Unsatisfactory	-0.262*	-0.225	-0.270*
	[0.144]	[0.142]	[0.144]
Poor	-0.985***	-0.975***	-0.966***
	[0.309]	[0.320]	[0.308]
FEDERAL STATES			
West	0.152***	0.163***	0.151***
	[0.0535]	[0.0528]	[0.0517]
West	0.229***	0.239***	0.236***
	[0.0688]	[0.0664]	[0.0660]
West	0,0358	0,0394	0,0676
	[0.100]	[0.0976]	[0.0925]
West	0.115***	0.118***	0.121***
	[0.0406]	[0.0397]	[0.0386]
West	0.135***	0.142***	0.139***
	[0.0481]	[0.0467]	[0.0460]

FGLS regression

Variables	all sig. (from regional models)	weight: e1sq sign.	weight: le1sq
West	0.131**	0.130**	0.136**
	[0.0571]	[0.0562]	[0.0551]
West	0.226***	0.232***	0.225***
	[0.0436]	[0.0431]	[0.0419]
West	0.343***	0.345***	0.336***
	[0.0427]	[0.0419]	[0.0410]
West	0.159*	0.160*	0.177**
	[0.0910]	[0.0901]	[0.0901]
West	0,0516	0.326	0.581*
	[0.342]	[0.348]	[0.348]
East	0.252	0.488	0.743**
	[0.326]	[0.333]	[0.332]
East	-0.112	0.155	0.410
	[0.316]	[0.324]	[0.326]
East	-0.591*	-0.275	-0,0316
	[0.310]	[0.318]	[0.318]
East	-0.287	-0,0219	0.206
	[0.320]	[0.327]	[0.326]
East	-0,0101	0.241	0.497
	[0.321]	[0.327]	[0.328]
POPULATION DENSITY			
$P < 2000$	-0.225***	-0.238***	-0.236***
	[0.0862]	[0.0833]	[0.0826]
$2000 < P < 5000$	-0.450***	-0.438***	-0.410***
	[0.0760]	[0.0748]	[0.0751]
$5000 < P < 20.000$	-0.297***	-0.280***	-0.279***
	[0.0508]	[0.0503]	[0.0497]
$20.000 < P < 50.000$	-0.515***	-0.489***	-0.482***
	[0.0532]	[0.0521]	[0.0518]
$50.000 < P < 100.000$	-0.456***	-0.432***	-0.417***
	[0.0529]	[0.0525]	[0.0520]
$50.000 < P < 100.000$	-0.485***	-0.455***	-0.447***
	[0.113]	[0.108]	[0.108]
$100.00 < P < 500.000$	-0.356***	-0.334***	-0.329***
	[0.0372]	[0.0365]	[0.0358]
$100.00 < P < 500.000$	-0.150***	-0.132***	-0.127***
	[0.0400]	[0.0389]	[0.0378]
$P \geq 500.000$	-0.166***	-0.158***	-0.144***
	[0.0363]	[0.0351]	[0.0345]
Constant	8.166***	7.858***	7.800***
	[0.370]	[0.364]	[0.361]
Observations	1619	1619	1618
R-squared	0.683	0.687	0.690
R-squared adj.	0.670	0.675	0.678
AIC	1692.604	1629.291	1572.845
BIC	2021.368	1958.054	1901.571

Standard errors in brackets

FGLS regression

Variables	all sig. (from regional models)	weight: e1sq sign.	weight: le1sq
$***p < 0.01, **p < 0.05, *p < 0.1$			

Source: Own calculations

Table 23: FGLS regression with imputed and edited data

VARIABLES	old data (no weights)	old weights	no weights	r3sq	new weights
PURCHASING PRICE	0.272***	0.342***	0.321***		0.342***
	[0.0169]	[0.0165]	[0.0162]		[0.0165]
PURCHASING PRICE int.	-0.0509**	-0.0968***	-0.0837***		-0.0968***
	[0.0255]	[0.0241]	[0.0236]		[0.0241]
OTHER PROPERTY	0.00416**	0.00504***	0.00543***		0.00504***
	[0.00181]	[0.00161]	[0.00164]		[0.00161]
CARS	0.00754**	0.00538*	0.00590*		0.00538*
	[0.00358]	[0.00320]	[0.00324]		[0.00320]
FINANCIAL WEALTH	0.0122**	0.0117***	0.0116***		0.0117***
	[0.00484]	[0.00434]	[0.00438]		[0.00434]
FINANCIAL WEALTH int.	0.0602***	0.0560***	0.0598***		0.0560***
	[0.0126]	[0.0115]	[0.0114]		[0.0115]
MORTGAGES	0,00244	0.00276*	0.00271		0.00276*
	[0.00189]	[0.00165]	[0.00171]		[0.00165]
INCOME	0,0114	0.00757	0.00454	-0.00307	0.00757
	[0.00865]	[0.00875]	[0.00785]	[0.00443]	[0.00875]
EDUCATION					
Apprenticeship	-0.235**	-0.179**	-0.208**	-0.0269	-0.179**
	[0.100]	[0.0893]	[0.0909]	[0.0557]	[0.0893]
Vocational school/commercial college	-0.253**	-0.205**	-0.242**	-0.000800	-0.205**
	[0.106]	[0.0945]	[0.0961]	[0.0590]	[0.0945]
Technical college	-0.192*	-0.117	-0.148	-0.0219	-0.117
	[0.103]	[0.0920]	[0.0936]	[0.0575]	[0.0920]
University of applied science	-0.180*	-0.134	-0.160*	-0.0446	-0.134
	[0.104]	[0.0930]	[0.0948]	[0.0586]	[0.0930]
University degree/teacher training	-0.212**	-0.160*	-0.187**	-0.0290	-0.160*
	[0.102]	[0.0906]	[0.0924]	[0.0570]	[0.0906]
Doctoral/postdoctoral training	-0.125	-0.0885	-0.123	-0.0475	-0.0885
	[0.109]	[0.0968]	[0.0993]	[0.0613]	[0.0968]
Other	-0.160	-0.113	-0.130	-0.0705	-0.113

	[0.169]	[0.159]	[0.154]	[0.0930]	[0.159]
No training completed	-0.253**	-0.215**	-0.237**	0.0499	-0.215**
	[0.108]	[0.0970]	[0.0981]	[0.0598]	[0.0970]
SIZE	0.00593***	0.00517***	0.00522***		0.00517***
	[0.000760]	[0.000662]	[0.000687]		[0.000662]
SIZE squared	-9.73e-06***	-7.58e-06***	-7.69e-06***		-7.58e-06***
	[1.78e-06]	[1.54e-06]	[1.61e-06]		[1.54e-06]
LAND	0.155*	0.0950	0.114		0.0950
	[0.0807]	[0.0711]	[0.0733]		[0.0711]
LAND squared	-0,00376	3.92e-05	-0.00117		3.92e-05
	[0.00578]	[0.00510]	[0.00524]		[0.00510]
TIME since acq.	0.00608***	0.00706***	0.00666***	0.00194***	0.00706***
	[0.000975]	[0.000893]	[0.000886]	[0.000540]	[0.000893]
BUILDING TYPE					
Semi-detached house	-0.0625*	-0.0624**	-0.0644**		-0.0624**
	[0.0325]	[0.0285]	[0.0295]		[0.0285]
Multiple family dwelling	0.0617*	0.0437	0.0482		0.0437
	[0.0342]	[0.0304]	[0.0309]		[0.0304]
Farm	0.196*	0.100	0.147		0.100
	[0.109]	[0.101]	[0.0990]		[0.101]
Building with various uses	0.387***	0.362***	0.362***		0.362***
	[0.0751]	[0.0664]	[0.0682]		[0.0664]
Non-detached house	-0.0735**	-0.0774**	-0.0790**		-0.0774**
	[0.0362]	[0.0321]	[0.0329]		[0.0321]
Other	0,0772	0.226**	0.223**		0.226**
	[0.115]	[0.106]	[0.105]		[0.106]
DWELLING RATE					
Very Good	-0.0724*	-0.0803**	-0.0840**		-0.0803**
	[0.0426]	[0.0370]	[0.0387]		[0.0370]
Satisfactory	-0.131***	-0.132***	-0.132***		-0.132***
	[0.0481]	[0.0421]	[0.0437]		[0.0421]
Simple	-0.191***	-0.149**	-0.153**		-0.149**
	[0.0659]	[0.0590]	[0.0597]		[0.0590]
Very simple	-0.236**	-0.239**	-0.241**		-0.239**
	[0.104]	[0.0944]	[0.0940]		[0.0944]

SURROUNDING				
Good	-0.105***	-0.0842***	-0.0859***	-0.0842***
	[0.0292]	[0.0257]	[0.0266]	[0.0257]
Satisfactory	-0.148***	-0.130***	-0.137***	-0.130***
	[0.0397]	[0.0352]	[0.0361]	[0.0352]
Adequate	-0.224***	-0.168***	-0.171***	-0.168***
	[0.0650]	[0.0589]	[0.0590]	[0.0589]
Unsatisfactory	-0.262*	-0.280**	-0.293**	-0.280**
	[0.144]	[0.125]	[0.125]	[0.125]
Poor	-0.985***	-0.892***	-0.931***	-0.892***
	[0.309]	[0.290]	[0.280]	[0.290]
FEDERAL STATES				
West	0.152***	0.148***	0.141***	0.148***
	[0.0535]	[0.0479]	[0.0486]	[0.0479]
West	0.229***	0.281***	0.273***	0.281***
	[0.0688]	[0.0603]	[0.0625]	[0.0603]
West	0,0358	-0.0404	-0.0481	-0.0404
	[0.100]	[0.0883]	[0.0906]	[0.0883]
West	0.115***	0.123***	0.120***	0.123***
	[0.0406]	[0.0361]	[0.0369]	[0.0361]
West	0.135***	0.141***	0.135***	0.141***
	[0.0481]	[0.0424]	[0.0437]	[0.0424]
West	0.131**	0.134***	0.138***	0.134***
	[0.0571]	[0.0511]	[0.0518]	[0.0511]
West	0.226***	0.213***	0.205***	0.213***
	[0.0436]	[0.0392]	[0.0396]	[0.0392]
West	0.343***	0.355***	0.352***	0.355***
	[0.0427]	[0.0382]	[0.0388]	[0.0382]
West	0.159*	0.155*	0.157*	0.155*
	[0.0910]	[0.0818]	[0.0826]	[0.0818]
West	0,0516	0.709**	0.483	0.709**
	[0.342]	[0.320]	[0.315]	[0.320]
East	0.252	0.908***	0.718**	0.908***
	[0.326]	[0.306]	[0.301]	[0.306]
East	-0.112	0.520*	0.301	0.520*

	[0.316]	[0.298]	[0.291]		[0.298]
East	-0.591*	0.178	-0.0805		0.178
	[0.310]	[0.293]	[0.287]		[0.293]
East	-0.287	0.509*	0.284		0.509*
	[0.320]	[0.301]	[0.295]		[0.301]
East	-0,0101	0.595**	0.380		0.595**
	[0.321]	[0.301]	[0.296]		[0.301]
POPULATION DENSITY					
$P < 2000$	-0.225***	-0.253***	-0.237***		-0.253***
	[0.0862]	[0.0757]	[0.0784]		[0.0757]
$2000 < P < 5000$	-0.450***	-0.449***	-0.461***		-0.449***
	[0.0760]	[0.0674]	[0.0681]		[0.0674]
$5000 < P < 20.000$	-0.297***	-0.282***	-0.299***		-0.282***
	[0.0508]	[0.0455]	[0.0459]		[0.0455]
$20.000 < P < 50.000$	-0.515***	-0.499***	-0.522***		-0.499***
	[0.0532]	[0.0474]	[0.0483]		[0.0474]
$50.000 < P < 100.000$	-0.456***	-0.385***	-0.396***		-0.385***
	[0.0529]	[0.0476]	[0.0480]		[0.0476]
$50.000 < P < 100.000$	-0.485***	-0.444***	-0.471***		-0.444***
	[0.113]	[0.0986]	[0.103]		[0.0986]
$100.00 < P < 500.000$	-0.356***	-0.308***	-0.327***		-0.308***
	[0.0372]	[0.0332]	[0.0338]		[0.0332]
$100.00 < P < 500.000$	-0.150***	-0.138***	-0.152***		-0.138***
	[0.0400]	[0.0353]	[0.0363]		[0.0353]
$P \geq 500.000$	-0.166***	-0.152***	-0.157***		-0.152***
	[0.0363]	[0.0319]	[0.0330]		[0.0319]
AGE				-4.59e-05	
				[0.000647]	
EMPLOYED					
Ordinary employed but not currently				-0.0126	
				[0.0393]	
Not employed				0.00586	
				[0.0153]	
CORRECT				0.00240	
				[0.00841]	

Constant	8.166*** [0.370]	7.520*** [0.336]	7.785*** [0.341]	0.132* [0.0711]	7.520*** [0.336]
Observations	1619	1622	1622	1621	1622
R-squared	0.683	0.730	0.725	0.028	0.730
R-squared adj.	0.670	0.719	0.715	0.0196	0.719
$***p < 0.01, **p < 0.05, *p < 0.1$ Standard errors in brackets					

Source: Own calculations

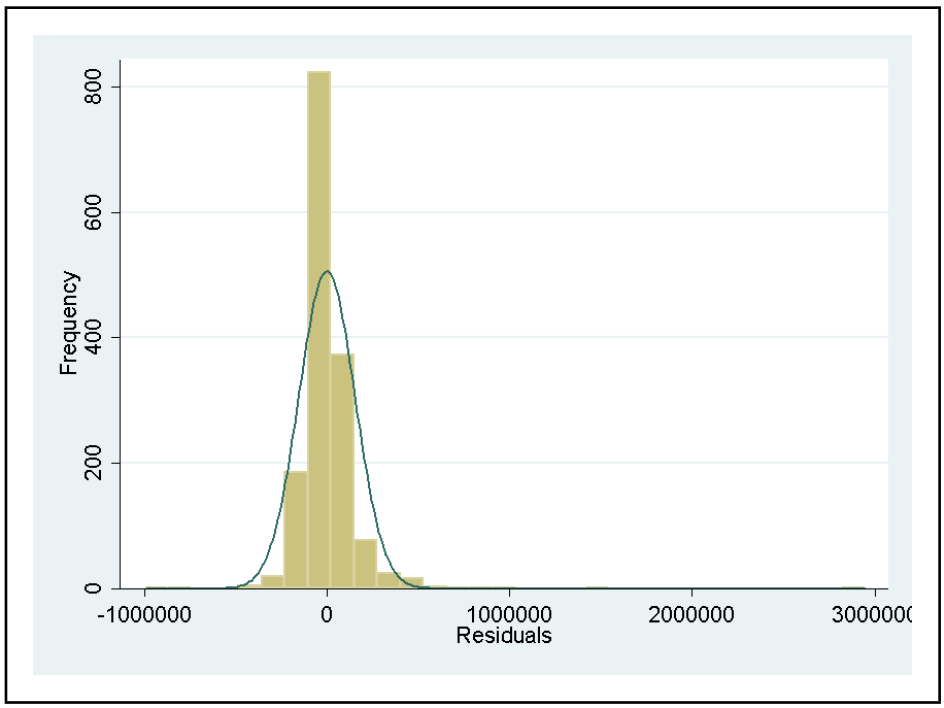
Appendix 2: Graphs

Figure 7: Household balance sheet

Assets	**Liabilities**
Non-financial assets – Owner-occupied housing – Other ownership of homes and property – Established businesses (net value) – Vehicles, collections, jewellery etc	*Liabilities* – Mortgages – Consumer loans (including credit card debt, current account credit, unpaid invoices, student loan debt) – Loans for business activity
Financial assets – Savings and current accounts, savings under building loan contracts – Mutual fund shares/units, debt securities, shares, derivatives and certificates – Balances from private pension and life insurance policies – Long-term equity investment – Assets under management	*Net wealth*
Total assets	Total liabilities

Source: Von Kalckreuth et al. (2012), p. 4.

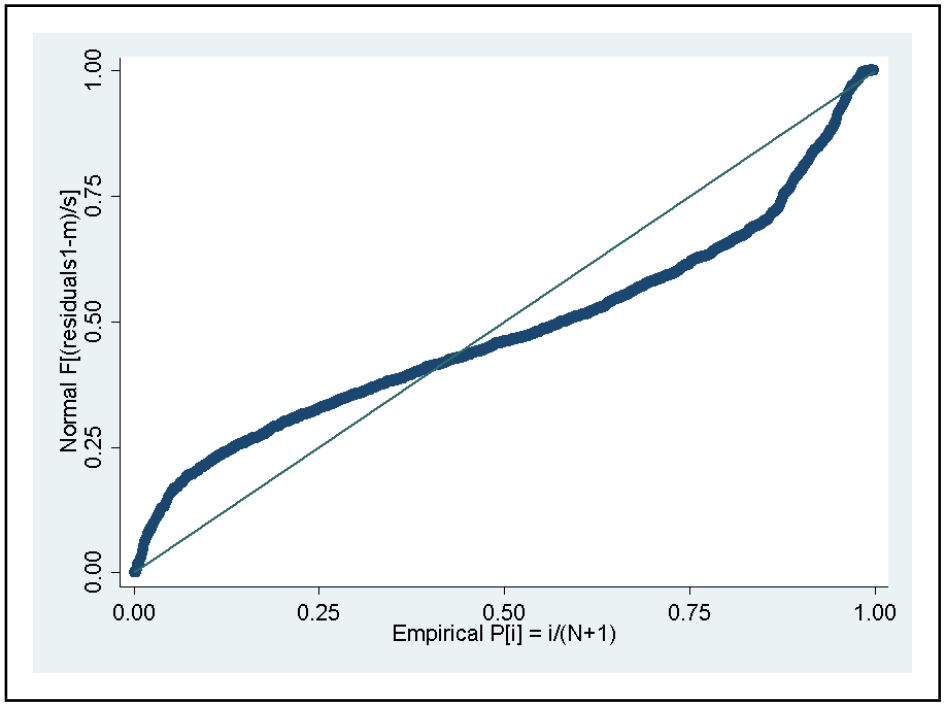
Figure 8: Residuals of regression model OLS all



Source: Own graph.

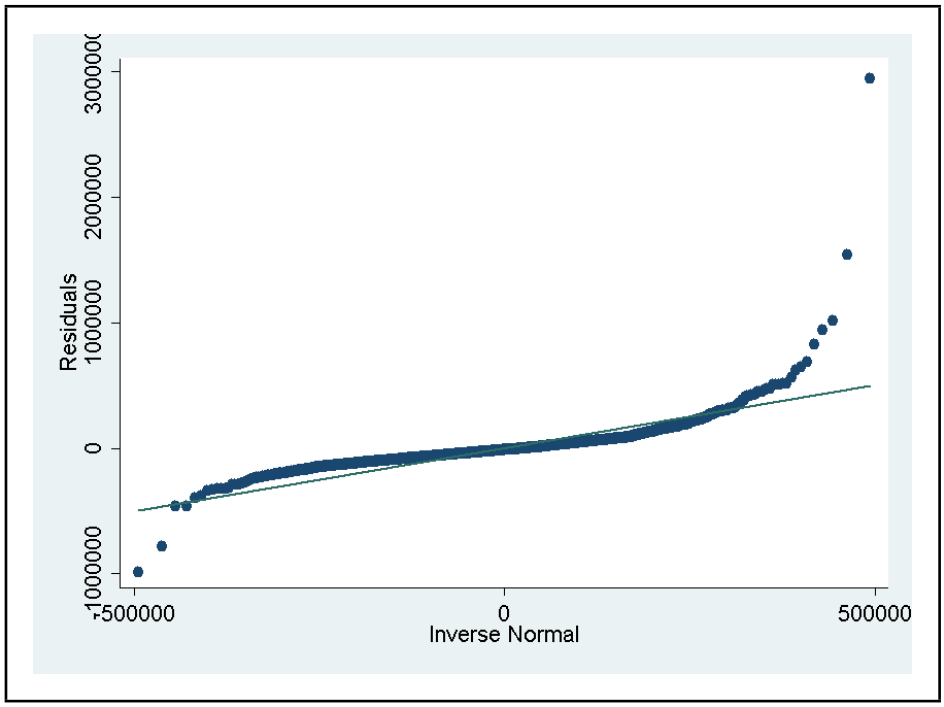
The pnorm command graphs a standardized normal probability plot. The qnorm command plots the quantiles of a variable against the quantiles of a normal distribution. This information can be received by using the help function of STATA. The graphs are named after the STATA commands.

Figure 9: Pnorm for residuals of regression model OLS all



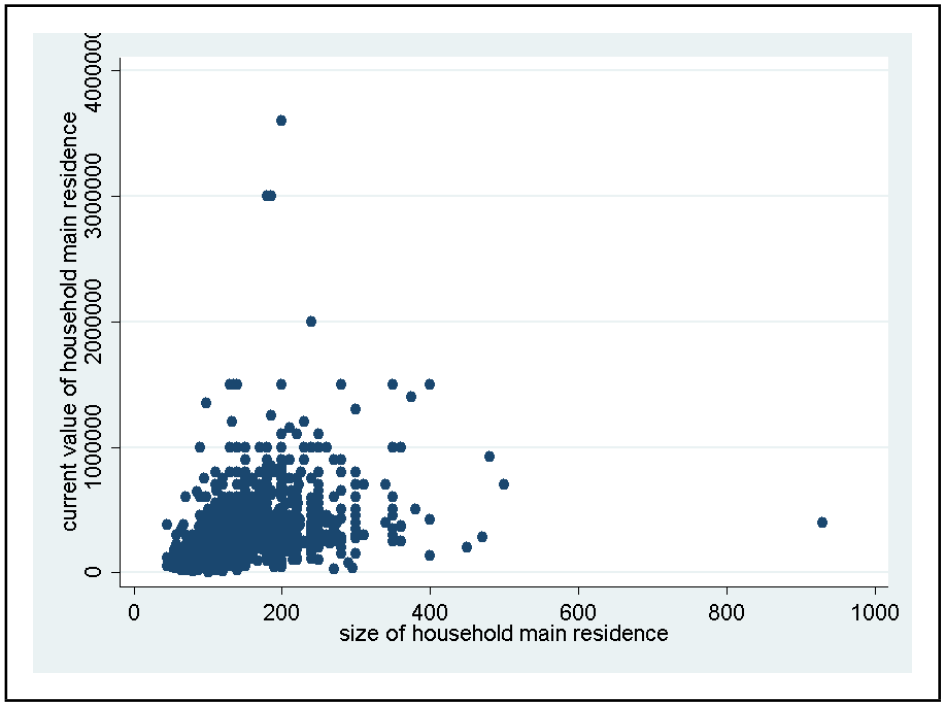
Source: Own graph.

Figure 10: Qnorm for residuals of regression model OLS all



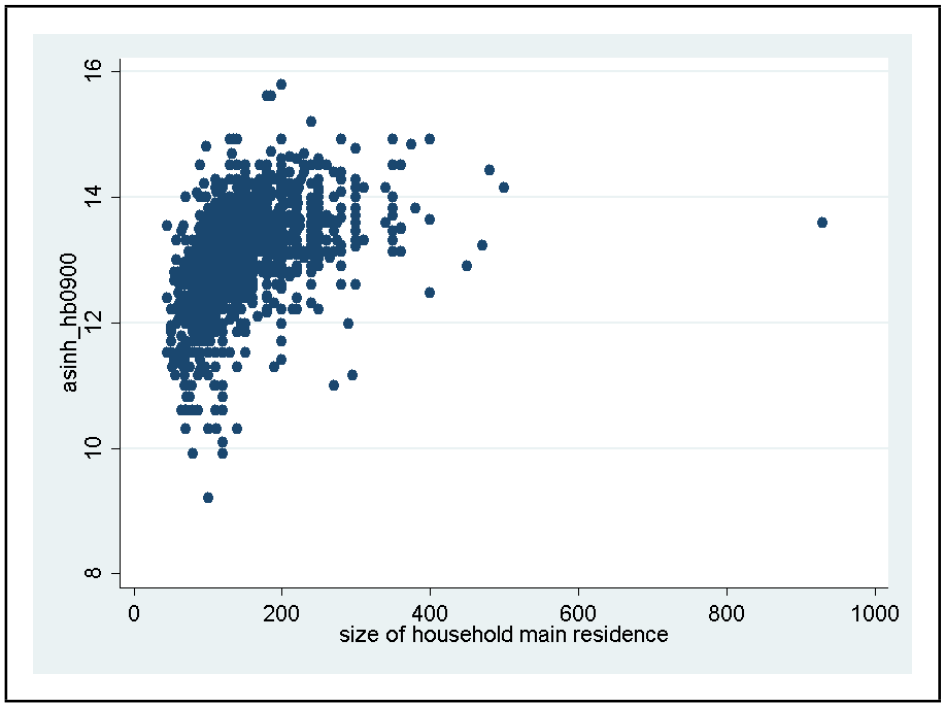
Source: Own graph.

Figure 11: Scatterplot property value against size of property



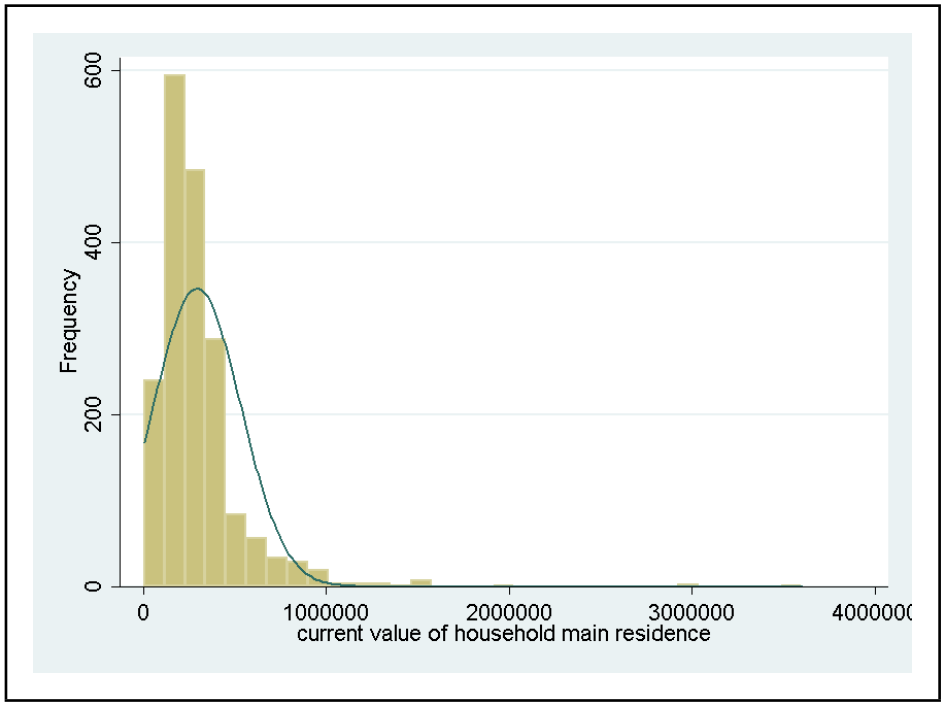
Source: Own graph.

Figure 12: Scatterplot transformed property value against transformed size of property



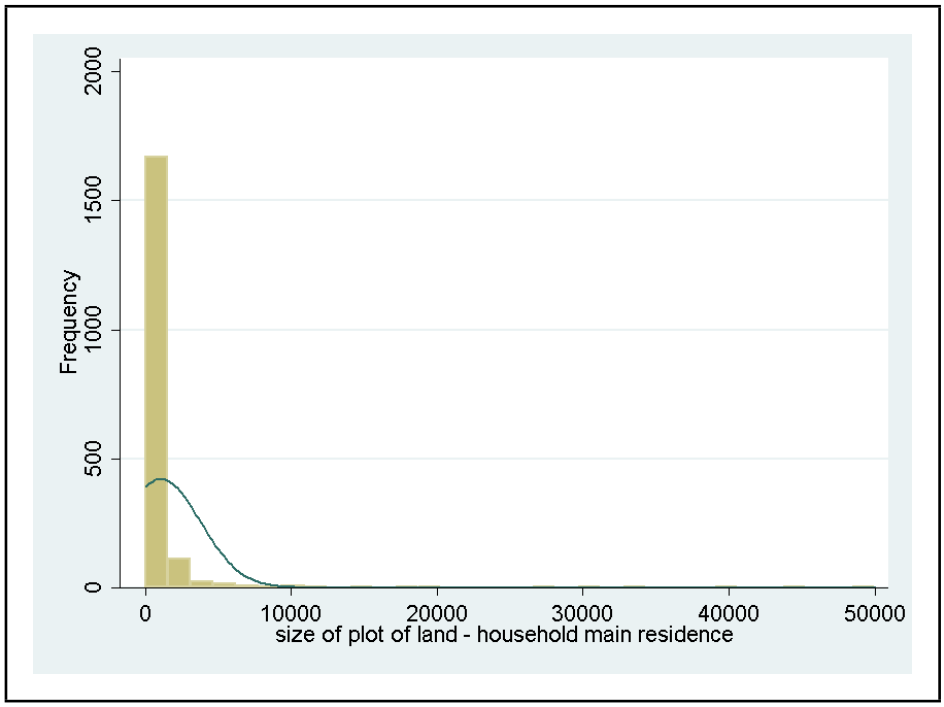
Source: Own graph.

Figure 13: Histogram of the variable PROPERTY PRICE



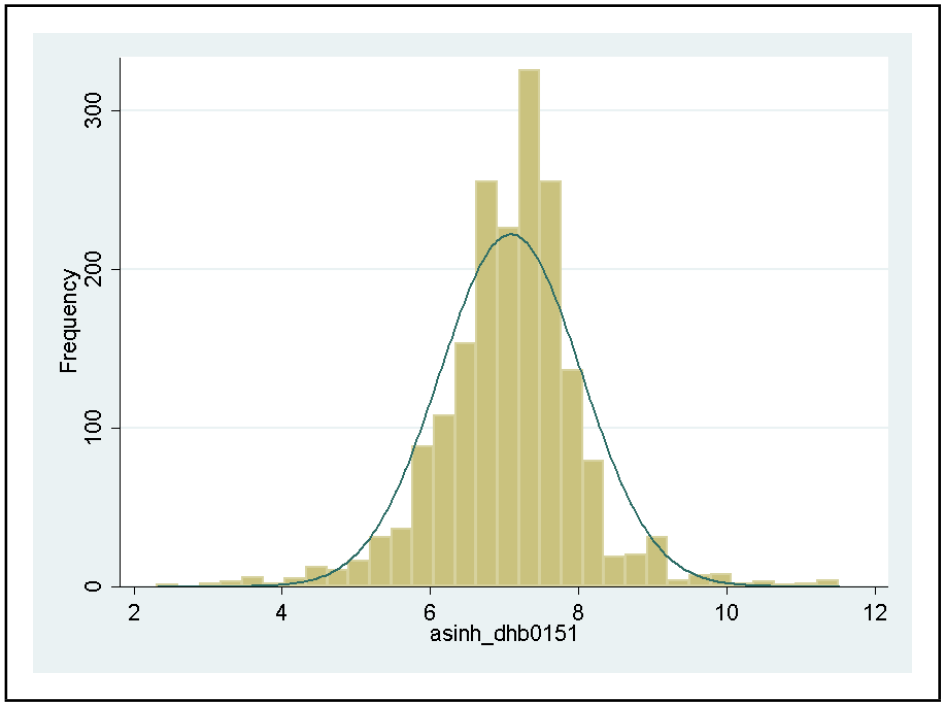
Source: Own graph.

Figure 14: Histogram of the variable LAND



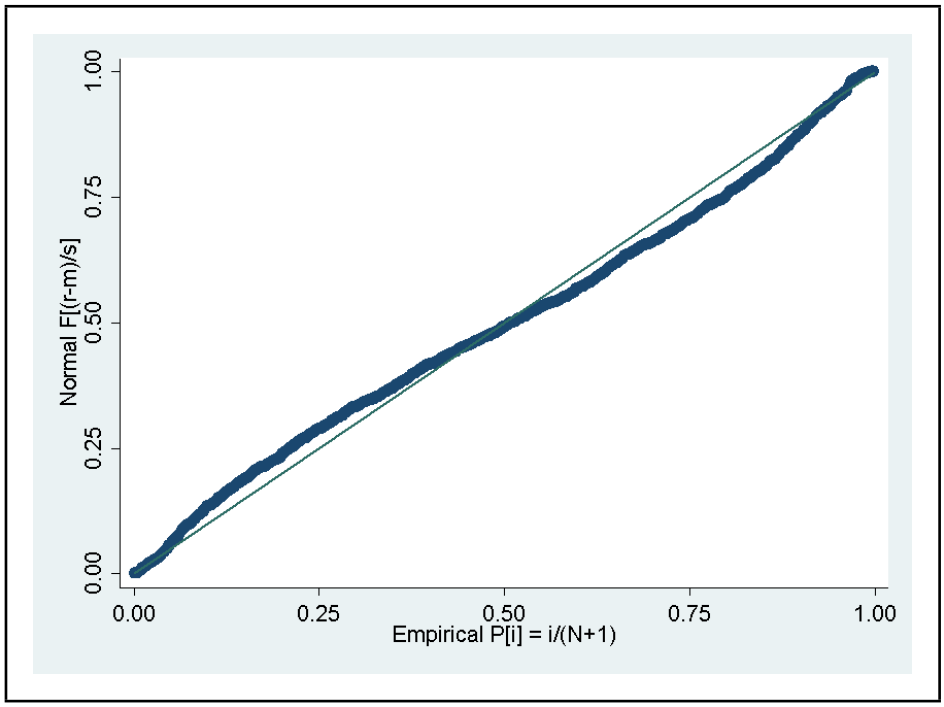
Source: Own graph.

Figure 15: Histogram of the transformed variable LAND



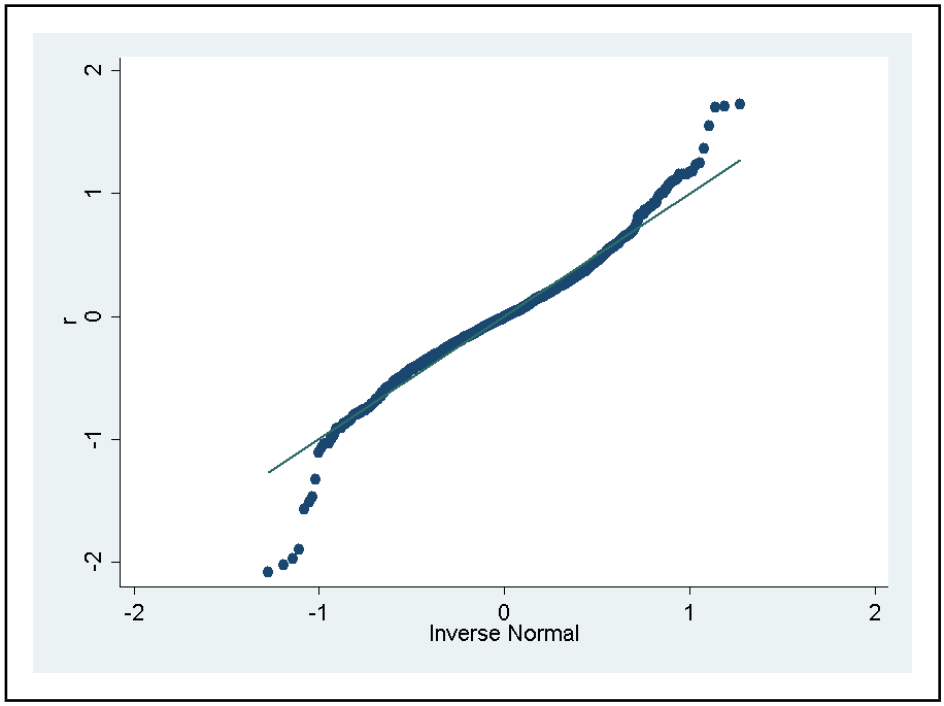
Source: Own graph.

Figure 16: Pnorm for residuals of regression model with edited and imputed data



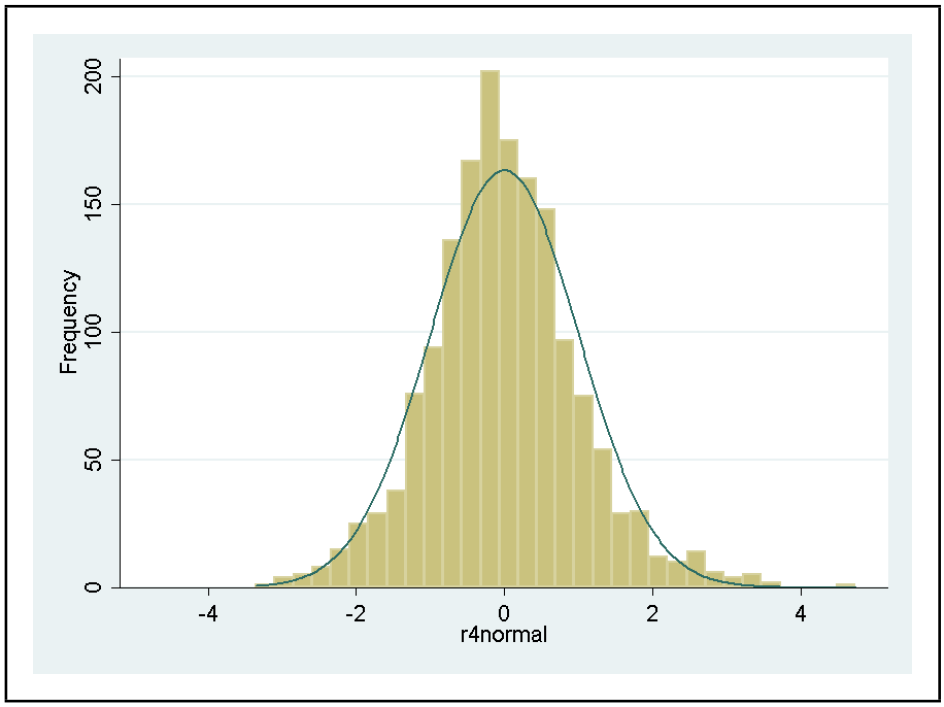
Source: Own graph.

Figure 17: Qnorm for residuals of regression model with edited and imputed data



Source: Own graph.

Figure 18: Standardized residuals for regression model with edited and imputed data



Source: Own graph.

Appendix 3: Variables

The PHF questionnaire is not numbered so that it can't be referred to the pages where the question for the variables are pointed out. However, the questions for all variables used and categories for the categorial variables are outlined in the following. Thereby only the important part of the questions are pointed out and adjusted so that they can be easily understand since the PHF questionnaire is designed for programming the CAPI and therefore very complex.

Table 1: Hedonic explanatory variables

Size - household main residence (hb0100)

What is the size of the property in square metres? Please only include the living space. Any additional space will be recorded later.

Size of plot of land - household main residence (dhb0151)

What is the size of the piece of land belonging directly to this property in square metres?

Year of property acquisition - household main residence (hb0700)

In what year did (you / your household / the household) acquire the property / the undeveloped plot of land?

Table 2: Hedonic categorial explanatory variables

Building type - household main residence (dhb0100)

In what type of building (do you / does your household / does the household) live?

1 - Detached single-family house

2 - Semi-detached house

3 - Multiple family dwelling or communal housing (e.g. apartment building)

6 - Non-detached house

4 - Farm

5 - Building with various uses (e.g. multiple family building with office, medical pratice or shop)

9 - Other

For category 5 and 9 a text is collected with a specification described.

Dwelling rate (sc0200)

Please rate the building

1 - Exclusive

2 - Very good

3 - Satisfactory

4 - Simple

5 - Very simple

Dwelling - location (sc0300)

Describe the location of the building

1 - City centre

2 - Located between the city centre and the suburbs

3 - Suburbs or city outskirts

4 - Rural area

Dwelling - outward appearance (sc0400)

Describe the condition of the building

1 - Clean and well-maintained

2 - Some small cracks in the facade and some crumbling paintwork

3 - Needs major renovation

4 - Dilapidated

Comparison with other dwellings in the neighbourhood (sc0500)

Describe the condition of the building compared with the neighbourhood

1 - The building is in poorer state than the surrounding buildings

2 - The surrounding buildings and the building itself are in the same condition

3 - The building is in a better state than the surrounding buildings

4 - No other building nearby

Rating of surrounding building (sc0600)

Assessment of residential area

1 - Very good

2 - Good

3 - Satisfactory

4 - Adequate

5 - Unsatisfactory

6 - Poor

Interior conditions (hr0200)

Describe the conditions in the interior of the dwelling

1 - Excellent to very good. Ceiling has no cracks, paintwork on the walls in very good to fairly good condition.

2 - Good. Needs repainting and some minor refinishing work.

3 - Fair. Needs some major interior work. (Holes and/or cracks need patching, broken windows etc).

4 - Poor. Some walls and ceilings need replacement.

Bland Categories from 1 to 16 for every federal state in Germany. It can be only differentiated between the former West and East German federal states due to data protection reasons.

Population density (bik)

P gives the amount of the population in the functional area.

$P < 2000$

$2000 < P < 5000$

$5000 < P < 20.000$

$20.000 < P < 50.000$

$50.000 < P < 100.000$

$50.000 < P < 100.000$

$100.00 < P < 500.000$

$100.00 < P < 500.000$

$P > 500.000$

Table 3: Further variables: Household wealth

Property value at the time of its acquisition (HB0800)

How much was the property, including the plots of land, worth at the time (you / someone in your household / someone in the household) acquired it?

Aggregated variables

The following variables are aggregated variables with other variables in the questionnaire. There components are outlined in German since the corresponding STATA code and the names of the variables are in written in German.

Value of the property besides the main property (immoson$_{min}$)

- derzeitiger Wert der Immobilie (neben dem Hauptwohnsitz) (hb280$)
- derzeitiger Wert der Immobilie (neben dem Hauptwohnsitz) (hb2900)

Value of the cars

- Wert aller PKW im Eigentum HH (dhb0810)
- Wert sonstige Fahrzeuge (hb4600)

Financial wealth ($finv_{min}$)

- Betriebsvermögen (bv)
- Konten und Bausparverträge (kon)
- Wertpapiere + Summe der sonstigen Vermögenswerte (wp)
- Geldschulden gegenüber dem HH (gs)

The components used for the aggregation of the variable Financial wealth are also received by an aggregation of other variables which are listed in the following:

Betriebsvermögen (bv)

- Wert des UN (Anteil) (hd070$)
- Wert aller anderen UN +3 (hd0900)
- Gesamtwert aller UN (dhd3100)
- Wert der Anteile (stille Beteiligung) (hd1010)

Konten und Bausparverträge (kon)

- Guthaben auf Sparkonten (hd1210)

- Guthaben auf Bausparverträgen (dhd0610)
- Wert sonstiger Vermögenswerten auf treuhänderisch verwalteten Konten (hd1620)

wp Wertpapiere + Summe der sonstigen Vermögenswerte

- Wertpapierdepot - geschätzter Marktwert (ausser Riester/Rürup) (dhd0750)
- Summe der sonstigen Vermögenswerte (hd1920)

gs Geldschulden gegenüber dem HH

- Höhe des dem HH geschuldeten Betrags (hd1710)

hyp Hypothekenschuld

- Outstanding amount of mortgages

Table 4: Further variables: Flow variables

The aggregation of income, savings and consume of the household is very complex since the values for every household member have to the aggregated. Thus the components of these variables will not be pointed out in the following but can be received upon request.

Table 5: Further variables: Socioeconomic variables

Number of children(dpe 1275)

In total, how many children (do you / does [Name]) have?

Table 6: Further variables: Categorial variables

Household main residence - Most significant means of property acquisition (dhb0410)

Of the possibilities stated, which was the most important when it came to the size of the property (including plots of land)?

1 - Purchased

2 - Constructed (yourself)

3 - Inherited

4 - Received as a gift

Highest level of education completed (dpa0300)

What is the highest level of education (you have / [Name] has) completed?

1 - Still at school

2 - Completed lower secondary school

3 - Completed higher secondary school

4 - Completed East German standard school up to 10th grade

5 - University of applied sciences entrance diploma / completed technical school

6 - General or subject-specific university entrance diploma / senior school-leaving certificate (from a grammar school) / East German secondary school up to 12th grade (also with apprenticeship)

7 - Other

8 - No school-leaving qualification

Highest level of professional education completed (dpa0400)

Have you / Has [Name]) completed a training qualification or course of study?

1 - Currently in training or studying

2 - Yes, vocational training completed (apprenticeship)

3 - Yes, vocational training completed (vocational school or commercial college)

4 - Yes, training at a technical or commercial college, school for master craftsmen or engineers or university of cooperative education completed

5 - Yes, university of applied sciences degree

6 - Yes, university degree obtained / teacher training completed

7 - Yes, doctorate / postdoctoral qualification obtained

8 - Other

9 - No, no training completed

Employed (dpa0500)

Are you / Is [NAME]) currently employed?

1 - Yes, employed (full-time, part-time, apprenticeship, low-paid part-time job or irregular employment)

2 - Yes, ordinarily employed but not currently (maternity leave / long-term sick leave / other period of leave)

3 - No, not employed (in training, unemployed, retired, homemaker)

Table 8: Financial literacy variables

Financial literacy- compound interest effect (dhnm0100)

Let us assume that you have a balance of 100 on your savings account. This balance bears interest at a rate of 2% per year and you leave it for 5 years on this account. How high do you think your balance will be after 5 years?

1 - More than 102

2 - Exactly 102

3 - Less than 102

Financial literacy - inflation (dhnm0200)

Let us assume that your savings account bears interest at a rate of 1% per year and the rate of inflation is 2% per year. Do you think that in one year's time the balance on your savings account will buy the same as, more than or less than today?

1 - More

2 - The same

3 - Less than today

Financial literacy - diversification (dhnm0300)

Do you agree with the following statement: "Investing in shares of one company is less risky than investing in a fund containing shares of similar companies"?

1 - Agree

2 - Disagree

Table 9: Estimation ability variables

Year of birth (dpe9050)

Before we start with the questions about employment, could you please tell me in which year (you were / [Name] was) born?

Appendix 4: Code

```
The code shows some of the programming solutions that might be interesting,
the other parts of the code can be received upon request.

* Retransformation of the conditional expected value

* Estimation of the correction factor under the assumption of a normal distribution

di exp((e(rss)/e(df_r))/2)

* Empirical estimation of the correction factor

predict re, residuals
gen eres = exp(re)
sum eres

* Further opportunities to estimate the correction
factor with an auxiliary regression

predict yhat3,xb
gen sinhyhat3 = sinh(yhat3)
reg hb0900 sinhyhat3, noconstant

* Estimation of an R^2 that can be compared with the R^2
of the regression with the non transformed data (smearing estimate)

corr hb0900 sinhyhat3  if e(sample)==1
return list
di r(rho)^2

* Out of sample prediction

xi:reg asinh_hb0900 asinh_hb0800 ost_asinh_hb0800
asinh_immoson_min asinh_kfz_min asinh_finv_min
ost_asinh_finv_min asinh_hyp asinh_tincome_min i.dpa0400
hb0100 hb0100sq asinh_dhb0151 asinh_dhb0151sq hb0700_2
i.dhb0100 i.sc0200 i.sc0600 bland_1 bland_2 bland_4-bland_16
bik_1 - bik_9

gen asinh_hb0900p = asinh_hb0900 if e(sample)==1
sort asinh_hb0900p

*over [a,b], use a+int((b-a+1)*runiform()).
set seed 1
gen n_p=1+int((e(N))*runiform()) if e(sample)==1
sort n_p
gen asinh_pgroup1= 1 if _n<=50
replace asinh_pgroup1= 0 if _n>50 & n_p!=.
count if asinh_pgroup1== 1
count if asinh_pgroup1== 0
*br asinh_hb0900 n_p asinh_p asinh_pgroup1
* Estimation of weights.
predict pe1, residuals
generate pe1sq= pe1^2
xi: regress pe1sq alterperson hb0700_2 asinh_tincome_min correct
if asinh_pgroup1 == 0
predict pzd, xb
generate pw=zd

* Regression with subsample.

xi:reg asinh_hb0900 asinh_hb0800 ost_asinh_hb0800
asinh_immoson_min asinh_kfz_min asinh_finv_min
ost_asinh_finv_min asinh_hyp asinh_tincome_min
i.dpa0400 hb0100 hb0100sq asinh_dhb0151 asinh_dhb0151sq
hb0700_2 i.dhb0100 i.sc0200 i.sc0600
bland_1 bland_2 bland_4-bland_16
```

```
bik_1 - bik_9 [aweight=1/pw] if asinh_pgroup1 == 0

* Out of sample prediction.
predict fitted, xb
corr asinh_hb0900 fitted if asinh_pgroup1 ==1
return list
di r(rho)^2

* Variables for regional model specifications

* East, West Dummy
gen ost=.
replace ost=1 if (bland == 13 | bland == 12 |
bland == 11 | bland == 15 | bland == 14 | bland == 16)
replace ost=0 if ost==.
count if ost==1
gen west=1-ost

*Interaction terms

gen ost_asinh_hb0800 = ost*asinh_hb0800
gen ost_asinh_finv_min = ost*asinh_finv_min
gen ost_asinh_hyp = ost*asinh_hyp
gen ost_asinh_tincome_min = ost*asinh_tincome_min
gen ost_asinh_kfz_min = ost*asinh_kfz_min
gen ost_asinh_immoson_min = ost*asinh_immoson_min

* East, West dummy for observations in East Germany before 1991.

gen dumhb0800w =1 if (west==1 | hb0700>1991)
replace dumhb0800w =0 if (ost==1 & hb0700<1991)
gen dumhb0800o = 1-dumhb0800w
gen westhb0800 = dumhb0800w * asinh_hb0800
gen osthb0800 = dumhb0800o * asinh_hb0800

* Variable for the estimation of economic education

*i.dhnm0100 financial lit. interest rate
*i.dhnm0200 financial lit. inflation
*i.dhnm0300 financial lit. diversification

* Variables that count the number of correct and wrong answers
of the financial literacy questions

* Number of correct answers
gen correct =0
replace correct =1 if dhnm0100==1
replace correct = correct + 1 if dhnm0200==3
replace correct = correct +1 if dhnm0300==2


* Number of wrong answers (Missing values are counted as wrong)

gen wrong =0
replace wrong =1 if (dhnm0100==2 | dhnm0100==3 | dhnm0100==.)
replace wrong = wrong + 1 if (dhnm0200==1 | dhnm0200==2 | dhnm0200==.)
replace wrong = wrong + 1 if (dhnm0300==1 | dhnm0300==.)
```

```
* Stochastic imputation

* Drawing randomly from the analytical normal distribution of the residuals

foreach counter of numlist 1 2 {
set seed `counter'
gen x`counter' = rnormal(0,(sqrt( e(rss))/ e(df_r)))
}
gen erwartungswert_plusx=yhat+x1

* Drawing randomly from the empirical normal distribution of the residuals

sort r
set seed 1
gen n_pick=1+int((1532)*runiform())
gen r_picked=r[n_pick]
gen yimput= y+r_picked
*gen sinh_yimput= sinh(yimput)
```

References

Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transaction on Automatic Control, Vol. 19, No. 6, pages 716-723.

Anderson, D.; Burnham, K. P. (2004), "Multimodel Inference. Understanding AIC and BIC in Model Selection", Sociological Methods & Research, Vol. 33, No. 2, pages 261-304.

Andrews, D. F.; Pregibon, D. (1978), "Finding the Outliers that Matters", Journal of the Royal Statistical Society, Series B (Methodological), Vol. 40, No. 1, pages 85-93.

Angrist, J. D.; Pischke J.-S. (2008), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press, Princeton.

Anton, H. (1998), Lineare Algebra. Einführung, Grundlagen, Übungen. Spektrum, Akad. Verlag, Heidelberg, Berlin.

Baum, C. F. (2006), An Introduction to Modern Econometrics Using Stata, Stata Press, College Station.

Baranzini, A.; Ramirez, J.; Schaerer, C.; Tahlman, P. (Editors) (2008), Hedonic Methods in Housing Markets: Pricing Enviromental Amenities and Segregation, Springer Verlag, Berlin, Heidelberg.

Barceló, C. (2006), "Imputation of the 2002 Wave of the Spanish Survey of Household Finances (EFF)", Banco de Espana, Documentos Ocasionales No. 0603.

Bazyl, M. (2009), "Hedonic price model for Warsaw housing market", Department of Applied Econometrics Working Papers, Warsaw School of Economics, Working paper No. 8-09.

Belsley, D. A.; Kuh, E.; Welsch, R. E. (2004), Regression Diagnostics Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, Hoboken.

BIK Aschpurwis + Behrens GmbH (2001), BIK Regionen: Ballungsräume, Stadtregionen, Mittel- / Unterzentrengebiete, Methodenbeschreibung zur Aktualisierung 2000, Hamburg.

Blanchard, O.; Illing, G. (2009), Makroökonomie, 5. Ed., Pearson Studium, München.

Bledsoe, R.; Fries, G. (2002), "Editing the 2001 Survey of Consumer Finances", Federal Reserve Board. Working Paper September 2002.

Breusch, T. S.; Pagan, A. R. (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation", Econometrica, Vol. 47, No. 5, pages 1287-1294.

Bronstein, I. N.; Semendjajew, K. A.; Musiol, G.; Mhlig, H. (2006), Taschenbuch der Mathematik, 6. Ed., Verlag Harri Deutsch, Frankfurt.

Burbidge, J. B.; Magee, L.; Robb, A. L. (1988), "Alternative Transformations to Handle Extreme Values of the Dependent Variable", Journal of the American Statistical Association, Vol. 83, No. 401, pages 123-127.

Burt, O. R. (1971), "Effects of Misspecification of Log-Linear Functions When Sample Values Are Zero or Negative: Comment", American Journal of Agricultural Economics, Vol. 53, No. 4, pages 671-673.

Cameron, A. C.; Trivedi, P. K. (2005), Microeconometrics: Methods and Application, Cambridge University Press, Cambridge.

Can, A. (1992), "Specification and estimation of hedonic housing price models", Regional Science and Urban Economics, Vol. 22, pages 453-474.

Canavarro, M. C.; Caridad, J. M.; Ceular, N. (2010), "Hedonic Methodologies in the Real Estate Valuation", in: Mathematical Methods in Engineering. International Symposium, Coimbra, 21 - 24 October. Coimbra 10 pages.

Carroll, C. D.; Otsuka, M.; Slacalek, J. (2010), "How Large Are Housing Financial Wealth Effects? A New Approach", ECB Working Paper Series No 1283.

Cassel, E.; Mendelsohn, R. (1985), "The Choice of Functional Forms for Hedonic Price Equations: Comment", Journal of Urban Economics, Vol. 18, No. 2, pages 135-142.

Chow, B. C. (1960), "Tests of Equality between Sets of Coefficients in Two Linear Regressions", Econometrica, Vol. 28, No. 3, pages 591-605.

Clark, T. E. (2000), "Can out-of-sample forecast comparisons help prevent overfitting?", Research Division Federal Reserve Bank of Kansas City, Research Working

Paper 00-05.

De-Bandt, O. de; Knetsch, T.; Penalosa, J.; Zollino, F. (Editors) (2010), Housing Markets in Europe: A Macroeconomic Perspective, Springer-Verlag, Berlin, Heidelberg.

De-Graft Acquah, H. (2010), "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship", Journal of Development and Agricultural Economics, Vol. 2, No. 1, pages 1-6.

Deutsche Bundesbank, Question Programme "Households and their finances", *http://www.bundesbank.de/Redaktion/EN/Downloads/Bundesbank/Research_Centre/phf_codebook_en.pdf?__blob=publicationFile.*

Deutsche Bundesbank (2012a), "The PHF: a survey of household wealth and finances in Germany", Monthly Report January 2012, pages 29-45.

Deutsche Bundesbank (2012b), "Die Preise für Wohnimmobilien in Deutschland", Monatsbericht Februar 2012, 64. Jahrgang, Nr. 2, pages 54-58.

Diewert, W. E. (2001), "Hedonic Regressions: A Consumer Theory Approach", Discussion Paper 01-12, Department of Economics, University of British Columbia, Vancouver.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method", Journal of the American Statistical Association, Vol. 78, No. 383, pages 605-610.

Dubin, R. A. (1998), "Predicting House Prices Using Multiple Listings Data", Journal of Real Estate Finance and Economics, Vol. 17, pages 35-59.

Ekeland, I.; Heckman, J. J.; Nesheim, L. (2002), "Identifying hedonic models", cemmap working papers CWP 06/02.

European Central Bank (2008), "Imputation and data editing", Household Finance and Consumption Network.

European Central Bank (2009), "Housing Finance in the Euro Area", Occasional Paper Series No 101.

European Mortgage Federation (2012), Factsheet 2012 Germany.

Fisher, F. M. (1970), "Tests of Equality between Sets of Coefficients in Two Linear Regressions: An Expository Note", Econometrica, Vol. 38, No. 2, pages 361-366.

Goodman, J. L.; Ittner, J. B. (1992), "The Accuracy of Home Owners' Estimates of House Value", Journal of Housing Economics, Vol. 2, No. 4, pages 339-357.

Goodman, A.C. (1988) "An econometric model of housing price, permanent income, tenure choice, and housing demand", Journal of Urban Economics, Vol. 23, No. 3, pages 327-353.

Greene, W. H. (2002), Econometric Analysis, Fifth Edition, Prentice Hall International, New York.

Groves, R. M.; Fowler Jr. F. J.; Couper, M. P.; Lepkowski, J. M.; Singer, E.; Tourangeau, R. (2004), Survey Methodology, John Wiley and Sons, Hoboken.

Grubbs, F. E. (1969), "Procedures for Detecting Outlying Observations in Samples", Technometrics, Vol. 11, No. 1, pages 1-21.

Halvorsen, R.; Pollakowski, H. O. (1981), "Choice of Functional Form for Hedonic Price Equations", Journal of Urban Economics, Vol. 10, pages 37-49.

Heeringa, S. G.; West, B. T.; Berglund, P. A. (2010), Applied Survey Data Analysis, Chapman and Hall, Boca Raton, London, New York.

Hellerstein, J. M. (2008), Quantitative Data Cleaning for Large Databases, Report for UNECE, *http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.*

Hill, R. C.; Griffith, W. E.; Judge, G. G. (2001), Undergraduate Econometrics. 2. Ed., John Wiley & Sons, Hoboken.

Hoaglin, D. C.; Welsch, R. E. (1978) "The Hat Matrix in Regression and ANOVA", The American Statistician, Vol. 32, No. 1, pages 17-22.

Hoover, K. D.; Perez, S. J. (1999), "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search", Econometrics Journal, Vol. 2, 167-191.

Jorion, P. (2001), Value at risk: the new benchmark for managing financial risk, 2. Ed., McGraw-Hill, New York.

Kain, J. F.; Quigley, J. M. (1972), "Note on Owner's Estimate of Housing Value", Journal of the American Statistical Association, Vol. 67, Issue 340, pages 803-806.

Kajuth, F.; Schmidt, T. (2011), "Seasonality in house prices", Deutsche Bundesbank, Discussion Paper Series 1: Economic Studies No 08/2011.

Kennickell, A. B. (2006), "How Do We Know if We Aren't Looking? An Investigation Of Data Quality in the 2004 SCF", Federal Reserve Board, Working Paper, September 2006.

Kiel, K. A.; Zabel, J. E. (1999), "The Accuracy of Owner-Provided House Values: The 1978-1991 American Housing Survey", Real Estate Economics, Vol. 27, No. 2, pages 263-298.

Kiel, K. A.; Zabel, J. E. (2008), "Location, location, location: the 3L approach to house price determination", Journal of Housing Economics, Vol. 17, No. 2, pages 175-190.

Lancaster, K. J. (1966), "A New Approach to Consumer Theory", The Journal of Political Economy, Vol. 74, No. 2, pages 132-157.

Lindenthal, T.; Eichholtz, P. (2011), "Prolonged Crisis: Housing in Germany and Berlin", in: Bardok, A.; Edelstein, R., Kroll, C. (Editors) Global Housing: Markets, Crises, Institutions, and Policies. John Wiley & Sons, New Jersey.

MacKinnon, J. G.; Magee, L. (1990), "Transforming the Dependent Variable in Regression Models", International Economic Review, Vol. 31, No. 2, pages 315-339.

Manning, W. G.; Mullahy, J. (1999), "Estimating Log Models: To Transform Or Not To Transform", NBER Technical Working Paper 246.

Meloun, M.; Militky, J. (2001), "Detection of single influential points in OLS regression model building", Analytica Chimica Acta, Vol. 439, pages 169-191.

Milleker, D. F. (2006), German residential property: signs of a pick-up in prices, Allianz Dresdner Economic Research Working Paper No: 65, 2006.

O'Brien, R. M. (2007), "A Caution Regarding Rules of Thumb for Variance Inflation Factors", Quality & Quantity International Journal of Methodology, Vol. 41, pages 673-690.

Pence, K., M., (2006), "The Role of Wealth Transformation: An Application to Estimating the Effect of Tax Incentives on Saving", Contributions to Economic Analysis & Policy, Vol. 5, pages 1-24.

Pozo, A. G. (2006), "Housing Market in Malaga: An Application of the Hedonic Methodology", European Regional Science Association ERSA 2006, paper 101.

Radcliffs, G. (1984), "A Theoretical Basis for Hedonic Regression. A Research Primer", Areuea Journal, Vol. 12, No. 1, pages 72-85.

Ramirez, O. A.; Moss, C. B.; Boggess, W. G. (1994), "Estimation and use of the inverse hyperbolic sine transformation to model non-normal correlated random variables", Journal of Applied Statistics, Vol. 21, No. 4, pages 289-304.

Ramsey, J. B. (1969), "Test for Specification Errors in Classical Linear Least-Squares Regression Analysis", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 31, No. 2, pages 350-371.

Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", The Journal of Political Economy, Vol. 82, No. 1, pages 34-55.

Rubin, D. B. (1996), "Multiple Imputation After 18+ Years", Journal of the American Statistical Association, Vol. 91, No. 434, pages 473-489.

Sander, J. E.; Conrad, F. G.; Mullin, P. A.; Herrmann D. J. (1992), "Cognitive modelling of the survey interview", Proceedings of the Annual Meetings of the American Statistical Association, Section on Survey Research Methods, pages 818-823.

Schwarz, G. (1978), "Estimating the Dimension of a Model", The Annals of Statistics, Vol. 6, No. 2, pages 461-464.

Stevens, J. P. (1984) , "Outliers and Influential Data Points in Regression Analysis", Psychological Bulletin, Vol. 95, No. 2, pages 334-344.

Sydsaeter, K.; Hammond, P. (2009), Mathematik für Wirtschaftswissenschaftler, 3. Ed., Pearson Studium, München.

Schürt, A. (2012), Housing and Property Markets in Germany 2011 at a Glance, German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR), BBSR-Analysen Kompakt 02/2012.

Varian, H. R. (2006), Intermediate Microeconomics. A Modern Approach, 7. Ed., International Student Edition, Norton.

Verbeek, M. (2004), A Guide to Modern Econometrics, 2. Ed., John Wiley & Sons Ltd, Chichester.

Von Kalckreuth, U.; Eisele, M.; Le Blanc, J.; Schmidt, T.; Zhu, J. (2012), "The PHF: A comprehensive panel survey on household finances and wealth in Germany", Deutsche Bundesbank Discussion Paper No 13/2012.

Welsch, R. E. (1986), "Influential Observations, High Leverage Points, and Outliers in Linear Regression: Comment", Statistical Science, Vol. 1, No. 3, pages 403-405.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", Econometrica, Vol. 48, No. 4, pages 817-838.

Witte, A. D.; Sumka, H. J.; Erekson, H. (1979), "An Estimate of a Structural Hedonic Price Model of the Housing Market: An application of Rosen's Theory to Implicit Markets", Econometrica, Vol. 47, No. 5, pages 1151-1173.

Wooldridge, J. M. (2009), Introductory Econometrics: A Modern Approach, 4. Ed., South-Western.