UNIVERSITY OF KWAZULU-NATAL

Listing Price Estimation of Flats in KwaZulu-Natal Coastal Sub-markets: A Novel Econometric Model

by

Dane Bax

214580191

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Business Administration

Graduate School of Business and Leadership

College of Law and Management Studies

Supervisor: Dr Mihalis Chasomeris

2016

Supervisors Permission to Submit Dissertation for Examination

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

Name: Dane Bax	Student No: 214580191	
Title: Listing Price Estimation of Flats in KwaZulu-Natal Coastal Sub-markets: A Novel Econometric Model		
Qualification: Masters in Business Administration	School: Graduate School of Business & Leadership	
	Yes	No
To the best of my knowledge, the thesis/dissertation is primarily the student's own work and the student has acknowledged all reference sources	X	
The English language is of a suitable standard for examination without going for professional editing.	X	
Turnitin Report %		
Comment if % is over 10%: The 8% of the 13% of the similarity index on the Turnitin Report is attributable to the students, Dane Bax, research methodology assignment which was submitted in the previous semester and formed his research proposal for this dissertation. Many changes were made, however, some were kept which are populating in the Turnitin Report.		
I agree to the submission of this thesis/dissertation for examination	X	
Supervisors Name: Dr Mihalis Chasomeris		
Supervisors Signature:		
Date:		
Co- Supervisors Name: N/A		
Co- Supervisors Signature: N/A		
Date:		

Declaration

I, **Dane Bax** declare that

1. The research reported in this dissertation, except where otherwise indicated, is my original research.
2. This dissertation has not been submitted for any degree or examination at any other university.
3. This dissertation does not contain other persons "data, pictures, graphs or other information, unless specifically acknowledged as being sources from other persons.
4. This dissertation does not contain other persons "writing, unless specifically acknowledged as being sources from other researchers. Where other written sources have been quote then:
    a. Their words have been re-written but the general information attributed to them has been referenced;
    b. Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
5. This dissertation does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference section.

Signature:.....

Date: ….....

Acknowledgements

I wish to express my sincere appreciation and gratitude to the following individuals, without whose assistance, this study would not have been possible:

- My family who have supported me through this journey.
- My supervisor Dr Mihalis Chasomeris who has provided invaluable guidance.
- Private Property (Pty) Ltd for allowing me to use their data, especially Mr Grant Elliot for the motivation he provided.

Abstract

The aim of this study was to derive a hedonic price function for flats within KwaZulu-Natal coastal sub-markets, filling a gap in residential hedonic price studies in South Africa. This study set out to develop a model to estimate listing prices of flats located in sub-markets along the KwaZulu-Natal coast. Identifying the appropriate distribution of listing prices and a set of statistically significant structural and locational attributes was paramount in achieving the research objectives. This was accomplished through a set of research hypotheses that were formulated and tested through rigorous statistical techniques. A generalised linear model based on the gamma distribution and log-link function was developed as a novel alternative to derive a hedonic price function for a segment of the residential property market in KwaZulu-Natal. The generalised linear model based on the gamma distribution and log-link function proved to be a more effective model for this research problem than the traditional ordinary least squares modelling approach. Based on the findings, a software application was developed to disseminate the results of the generalised linear model for potential commercial use by real estate businesses, bridging the gap between academia and business.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE
Introduction

1.0 Introduction

Residential property is perceived as a fundamental barometer of individual and collective wealth where its cumulative value is closely tracked by government statistical bureaus, banks and other economic establishments. Individual households, financial institutions and policy makers closely monitor residential property price trends in order to gauge real house price growth, financial stability as well as monitor the activity and condition of the of the credit market (de Haan and Erwin, 2011).

The residential property market is a driver of economic growth with many countries using property prices as a proxy for economic stability. Consumer spending is a significant component of gross domestic profit in most economies and can be affected by changes in residential property prices as a result of household debt levels and overvalued asset prices (Hill, 2011). Globally, housing wealth has been identified as having an important role in business cycles, further emphasising the economic significance thereof (Girouard and Blondal 2001).

Residential property is an important segment of the property market in South Africa, the large portfolio of residential property contributes significantly towards the wealth of the country where it is capitalized on the household balance sheet in the set of national accounts (South African Reserve Bank, 2015). Residential property transactions are typically infrequent and relate to a highly differentiated set of items rendering effective measurement techniques complex and difficult (Hill, 2011). Although several techniques exist to estimate residential property prices, no standard or accepted set of guidelines currently exist in South Africa and further research is required to provide insight into this problem.

This chapter presents an overview of the research project and introduces the background to the study which contextualises the research and subsequently details the problem statement. The specific research objectives and hypotheses are articulated, providing insight into how the research problem was explored. Finally, an outline is provided detailing the structure of the study.

1.1 Background to the Study and Problem Statement

The residential property market is a significant source of wealth for households, investors and countries, however, estimating and modelling residential property prices is a particularly difficult endeavour due to the heterogeneity of residential properties. De Haan and Erwin (2011) propound that residential property markets are typically characterised as heterogeneous, where property sales are infrequent and the listing (asking) price is normally negotiable, making measurement techniques difficult.

A residential property market consists of willing buyers and sellers that are each trying to maximise their objective of obtaining value in their respective transactional position, however, due to the heterogeneity of residential properties, pricing a property at where the market will clear is often extremely difficult due to the mechanics of supply and demand (Day, 2003). A plethora of international research exists employing hedonic models to estimate residential property prices and determine the significant structural and locational attributes thereof. However, in South Africa research conducted in this field has been limited to only a few studies. Examples of these studies are discussed in depth in Chapter Two.

In South Africa, currently there is no freely available online tool or software application that facilitates the estimation of residential property prices based on structural and locational attributes. To derive practical value from the hedonic pricing model developed in this study, a software application was built where a property market participant can submit selected structural and locational attributes as inputs to the application and the arithmetic mean listing price is calculated as the output by calling the hedonic model. This study bridges the gap between academia and business by providing a software application that can be hosted online by Private Property (Pty) Ltd using the empirically tested model developed through rigorous academic research.

The problem underpinning this research was to construct a hedonic pricing model to estimate listing prices for flats within KwaZulu-Natal coastal sub-markets based on statistically significant structural and locational attributes.

1.2 Research Objectives

The overall objective of this study was to develop a hedonic pricing model for flats located in KwaZulu-Natal coastal sub-markets.

The following sub-research objectives are as follows:

1. To determine an appropriate hedonic price model for flats located in KwaZulu-Natal coastal sub-markets based on the distribution of listing prices.
2. To develop a model to estimate listing prices of flats that are located in sub-markets along the KwaZulu-Natal coast based on structural and locational attributes.
3. To build a software application that facilitates the estimation of listings prices for flats in KwaZulu-Natal coastal sub-markets given a set of structural and locational attributes.

1.3 Research Hypotheses

In order to achieve the overall research objective and sub-research objectives, the following research hypotheses were formulated:

H0a: The distribution of listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast follow a gamma distribution.
H1a: The distribution of listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast do not follow a gamma distribution.

H0b: Floor Area is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
H1b: Floor Area is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H0c: Number of bedrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H1c: Number of bedrooms is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0d*:** Number of bathrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H1d*:** Number of bathrooms a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H0e: The suburb, a dummy locational variable, is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H1e: The suburb, a dummy locational variable, is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

1.4 Significance of the Study

Online property portals and realtor's are ideally positioned to leverage a plethora of property data by developing property valuation solutions and sharing these services downstream with households and investors. The ubiquity of heuristics in determining residential property listing prices facilitates the need for greater transparency in the property industry. This study provides residential property market participants with a contextual framework and model to analyse, assess and determine listing prices for residential property and the significant structural and locational attributes thereof.

Through scientific rigour, empirical evidence can proliferate transparency and assist buyers and sellers in determining and valuing the significant structural and locational property attributes to establish listing prices. This study develops a hedonic price function using two techniques, namely ordinary least squares based on the log-normal distribution and a generalized linear model based on the gamma distribution and log-link function. Deriving a hedonic price function for residential property using a generalized linear model based on the gamma distribution and log link function has never been attempted before in South African

research and finding global research employing this methodology has proven extremely difficult where no existing studies have been identified. This study presents a novel alternative to the ordinary least squares method of deriving a hedonic price function for residential property through the use of a generalized linear model based on the gamma distribution and log-link function where the arithmetic mean is computed as the expected value and not the geometric mean.

Another important feature of this study was the use of property listing prices and not historic sales prices which is geared towards making greater use of larger amounts of available online property data that is aggregated by property portal industry participants, namely Private Property (Pty) Ltd. The application of the study presents a framework that has the potential to add value to the property industry including investors and buyers and sellers of residential property by presenting a scientific and empirically tested method for evaluating whether a property of interest is aligned to market expectations. Hedonic pricing models can potentially be utilised by residential property market participants to determine the statistically significant attributes of a property of interest when entering into a transaction and assign implicit prices to to each significant attribute. Furthermore, the expected prices for residential properties can easily be determined with a correctly specified hedonic model which will position residential property market participants to make more informed decisions about purchase and sale transactions.

A software application was developed to disseminate the findings of the this study to facilitate the practical application of the proposed generalized linear model for real estate businesses, bridging the gap between academia and business.

This study contributes to the existing body of knowledge by enriching current literature on hedonic pricing in the South African context. This was achieved through rigorous statistical analysis where the expected values for flats located in KwaZulu-Natal coastal sub-markets were estimated and the functional relationships between listing prices and the associated attributes examined. While Dodds (2011) and Du Preez *et al.* (2014) focused on property segments within the Gauteng province and Els and Von Fintel (2010) focused on property segments within the Western Cape province, it appears that no research on the hedonic price function of residential property has been conducted on any segment of the Kwazulu-Natal province. This study aimed to contribute to the limited research done on residential property

pricing by creating a framework using a novel econometric model that can be applied to the plethora of residential property data which is democratised via the internet and online real estate participants. This study augments the existing body of knowledge of residential property market participants with a scientific model to estimate residential property listing prices.

1.5 Scope and Delimitations of the Study

The scope of the study relates solely to the domain of residential property, a subset of the entire property market which excludes commercial, industrial or retail property markets. Moreover, the study focuses on residential properties that are marketed online as the data was obtained from an online property portal, however, the same data is contained in traditional print mediums such as property magazines.

The sample that was used pertains to flats listed for sale within three coastal sub-markets of KwaZulu-Natal. The study comprised solely of the property type flats. The primary reasons for using flats as the property type for the analysis was due to the lack of research done on flats in South Africa and concomitantly the amount of data that was available.

1.6 Structure of the Study

This study will comprise of five chapters which are introduced below in Table 1.1.

Table 1.1: Chapter Outline

Chapter	Description
Chapter One: Overview of the Study	This chapter introduces the study by articulating the background and problem statement. This chapter also sets out the research objectives and the research hypotheses.
Chapter Two: The Literature Review	This chapter presents a comprehensive and rigorous literature review of international and local studies relevant to the research objectives and research hypotheses.
Chapter Three: The Research Methodology	This chapter presents and defines the approach of data analysis in order to answer the research hypotheses and research objectives. This chapter also defines key concepts such as the sample and the statistical techniques used to perform the rigorous analysis.
Chapter Four: The Presentation and Discussion of Results	This chapter presents the results of the study and unpacks the statistical models and tests employed using the statistical software R. This chapter explains the results in the context of the literature presented in the Chapter Two.
Chapter Five: Conclusions, Limitations and Recommendations	This chapter discusses and presents the conclusions of the study as well as the research limitations and recommendations of the study and further research.

1.7 Conclusion

This chapter has presented an overview of the research project and introduced the background to the study, contextualising the research and detailing the problem statement. The research objectives and hypotheses were articulated providing insight into how the research problem will be explored. An outline of the study was provided detailing the structure of the research.

The next chapter reviews various sources of literature in order to gain a deeper understanding of the economic importance of residential property prices and unpack the concepts and complexities of deriving a residential property hedonic price function.

CHAPTER TWO

The Review of the Literature:
The Economic Relevance of Residential Property Markets and Hedonic Valuation Theory

2.1 Introduction

Interest in residential property prices is ubiquitous, a myriad of economic participants base decision making on available property price data. Therefore, deriving an accurate model to assess and value residential property prices is paramount in facilitating transparent transactions and creating accurate growth indicators. This literature review introduces the economic importance of residential property markets and presents an empirical analysis of econometric hedonic pricing in global and South African residential property markets. An analysis of existing research underpins the important considerations and complexities involved in constructing a residential property hedonic price function. The discourse reviews ordinary least squares and the log-normal distribution as a candidate model to derive a hedonic price function. A generalised linear model based on the gamma distribution is reviewed for its potential application in developing a novel approach to derive a residential property hedonic price function. Finally, an evaluation on bootstrapping is presented with special focus on its use in model validation.

2.2 The Economic Importance of the Residential Property Market

A country's inventory of residential property is a portion of its national wealth, therefore residential property price trends influence decision making and inform economic policy formulation. Residential property is reported as a significant share of wealth in most countries national accounts where house price trends are used to monitor economic activity and financial stability, to conduct economic policy and effect decision making (Hill, 2011). In South Africa, residential property is capitalised on the household balance sheet in the national accounts (South African Reserve Bank, 2015).

2.2.1 A Macro-economic Indicator of Economic Growth

Shifts in residential property prices have an effect on various areas of the economy such as macro economic stability, consumption spending, wealth accumulation, and labour mobility (Siaz, 2012).

Empirical evidence suggests that periods of economic expansion and upward trajectories of business cycles are often associated with burgeoning house prices. Girouard and Blondal (2001) identified housing wealth in OECD countries as having an important role in business cycles where house price developments have a positive effect on private consumption. Goodhart and Hoffman (2008) using vector auto-regression, estimated by fixed effects ordinary least squares, found a significant relationship between house prices and the macro-economy where they provide evidence of multiple relationships between house prices, broad money and credit. The relationship between housing bubbles and bank crises exhibit an identifiable pattern, booms in real house prices and subsequent busts or corrections have been linked to the start of bank crises. In a study of advanced economies from 1970-2001 Bordo and Jeanne (2002) found a greater probability of a financial crisis occurring either at the peak of a boom or shortly after a bust in real estate markets. Steep declines in residential property prices have the potential to negatively affect credit ratings and the debt to equity ratio, adversely affecting the stability of the financial sector (de Haan and Erwin, 2011).

Should house prices increase, households view their net wealth or net asset value as increasing which could trigger greater consumption spending and affect aggregate demand. An increase in consumption spending and household borrowing is generally experienced due to the positive spillover that increased house prices have on household balance sheets, known as a wealth effect (Girouard and Blondal, 2001). This manifests if households are influenced to spend based on their net asset value, where houses are regarded as wealth. Girouard and Blondal (2001) concluded that through wealth effects or by ameliorating households' liquidity constraints, private consumption can be bolstered which will have an effect on aggregate demand. They found that movements in house prices can affect private consumption if households access to credit is contingent on housing wealth or equity, however, the propensity for households to adjust current spending based on movements in house prices is largely dependent on the functioning of the financial system. Indeed, Friedman (1957) asserted that consumption expenditure should be viewed as a function of income and wealth

which differs to the Keynesian consumption function where consumption is viewed as mainly a function of current income. Keynes propounded that aggregate demand is a function of consumption spending, investment spending, government spending and net exports where he regarded consumption spending as being contingent on income and the marginal propensity to consume (Friedman, 1997). It is discernible that consumer spending is a significant component of aggregate demand which measures gross domestic profit and can be affected by changes in residential property prices as a result of household debt levels and overvalued asset prices.

Higher house prices have the propensity to stimulate and bolster construction activity which can lead to increased employment and higher wages for an array of workers in the housing market. Girouard and Blondal (2001) assert that private residential construction can be influenced by changes in property prices where it becomes less or more profitable to build new houses. Higher incomes are often experienced for construction workers, real estate agents and professionals in legal and financial property roles from increased activity in housing construction (de Haan and Erwin, 2011).

2.2.2 Asset Prices and Price Stability

In many countries, central banks follow an inflation targeting mandate or regime which effectively pegs the lower and upper bounds for inflation using monetary policy in order to achieve price stability, an important macro-economic objective (Duttagupta and Fernandez, 2004).

In South Africa, inflation targeting is used to remove price volatility and assist in creating a stable financial environment though monetary policy. The South African Reserve Bank (SARB) initiates monetary policy to maintain its inflation target, although in the long run monetary policy is not able to contribute directly to economic growth and employment creation, it does fulfil an essential requirement for the attainment of these objectives, namely price stability in the interest of sustainable economic development and growth (Kumo, 2015). South Africa's continuous 3% to 6% headline consumer price index inflation target is mandated by the Monetary Policy Committee, a special unit within SARB that is responsible for monetary policy decisions (South African Reserve Bank, 2015). Although asset prices may not always be targeted directly, asset prices are monitored closely for decision making.

Rises in asset prices such as residential property could reflect developments in the financial sector and real economy which may result in upward inflationary pressure. This usually coincides with a rise in credit and debt, therefore, taking cognisance of the movements in the prices of fixed property and the effects thereof on final demand and inflation is paramount (van der Merwe, 2004). Expectations of higher inflation will manifest in higher interest rates affecting interest bearing loans such as in the case of mortgage bonds. (South African Reserve Bank, 2007). Jorda, Schularick and Taylor (2015) assert that a trade off exists in using interest rates to curb asset price booms where large increases in interest rates to slow rising house prices could cost the economy elsewhere in terms of unemployment and inflation and that a policy of increased interest rates hikes is contingent on how responsive housing demand is to increased interest rates.

2.2.4 Household Decision Making

Purchasing residential property is a significant decision for people seeking a place of residence and plays a large role for investors seeking to augment investment portfolios. Residential property is an important barometer of wealth and a paramount factor in the cost of living and is therefore a key indicator to consider in terms of its economic significance (Els and Von Fintel, 2010).

The act of purchasing a residential property is perceived as one of the largest financial transactions an individual or household will undergo, therefore changes in property prices are likely to affect when, where and what type of property an individual or household will purchase (Els and Von Fintel, 2010). Purchasing residential property provides a means of shelter as well as a capital investment where capital gains may accrue in the long term (de Haan and Erwin, 2011). The opportunity cost associated with the outlay of funds for a property or leverage undertaken in terms of a mortgage is significant and can affect home purchasers' decisions (de Haan and Erwin, 2011). Many individuals have an indirect stake in property though pension funds and mutual funds (Lee, 2006). Significant declines in residential property prices can have an adverse impact on the financial position of households by reducing the value of their collateral (Girouard and Blondal, 2001).

Online property businesses provide valuable services to households and investors where a plethora of property data is aggregated and disseminated to facilitate purchasing decisions.

South Africa is home to Private Property (Pty) Ltd, an online property portal which aggregates property listings throughout South Africa from estate agents, private sellers, property developers and banks (Private Property, 2015). Zillow a USA based property portal aggregates property listings throughout the USA and provides an online user interface where a purchaser or seller of property can determine an estimate of the property they wish to buy or sell using a proprietary algorithm (Corcoran and Liu, 2014). The use of such statistical techniques to model residential property data makes it possible to enhance household decision making with regards to purchasing or selling a home. The ability to leverage technology that is combined with advanced econometric and statistical techniques is paramount in providing South African households and investors with the ability to make better decisions when purchasing or selling residential property.

2.3 Hedonic Pricing Theory

Informed decisions are essential in order to ascertain fair and unbiased market prices of residential properties. However, determining market property prices is multivariate and extremely difficult due the heterogeneity of residential properties. Prices of residential property are difficult to measure due to the heterogeneous nature thereof, where it can be observed that dwellings are not identical even by the sole virtue of occupying different locations (Hill, 2011). Hedonic price modelling is pervasive in economic literature, it has been employed to model property prices where the price of the property is valued according to its set of structural and locational attributes (Shulz, 2003). Hedonic pricing is a mathematical technique used in economics which aims to measure the price of a good through its utility bearing attributes, where a vector of attributes determines the price of the good (Rosen, 1974). Econometrics is a field of economics that applies rigorous mathematical statistics including statistical inference to empirically determine and measure relationships of economic theory postulated by economic thought (Greene, 1993). Based on this definition one can consider the derivation of a hedonic price function for a residential property market to belong to the domain of econometrics.

2.3.1 Residential Property and the Hedonic Pricing Function

Residential property is a single class of good or commodity in the eyes of individuals, households and investors, however, it is differentiated or heterogeneous nature Hill (2011).

Heterogeneous products consist of a range of products that differ in a set of attributes, however, this set of attributes is considered closely related by consumers and is therefore defined as a single product (Day, 2003).

A residential property is a collection of attributes that each hold certain utility and value which can be characterised as structural, like size and the number of bedrooms, relate to how accessible the property is to amenities like schools and may include location specific attributes, such as being in a specific geographic area or suburb. Typically hedonic pricing techniques model property prices as a function of a set of inherent structural attributes, neighbourhood or location characteristics and accessibility to amenities (Lyons, 2015). Market forces regulate heterogeneous product prices and these prices are contingent on the individual products' set of attributes. Hedonic methods express that residential properties can be decomposed by the constituent attributes thereof and although no market for the individual attributes exists, each attributes marginal contribution to the property's price can be determined implicitly by supply and demand forces in the property market (de Haan and Erwin, 2011). Competitive market conditions for homogeneous products depend on supply and demand mechanics to reach an equilibrium price, whereby the market clears and the needs of consumers and firms are reconciled (Day, 2003). However, heterogeneous product markets, like the residential property market, comprise of many differentiated properties commanding a myriad of different prices. Market forces are responsible for the different prices of residential properties which is contingent on each individual property's set of attributes. Generally the market will settle on a set of prices for the various combinations of residential properties that will clear the market through the reconciliation of supply and demand (Day, 2003). Rosen (1979) propounds that economic agents can ascertain hedonic prices from the observed prices of heterogeneous products, where the hedonic prices equate to the implicit prices of the attributes of the heterogeneous products. In hedonic theory purchasers of properties are equivalent to consumers and the sellers are equivalent to producers.

The hedonic pricing model describes each property by a vector of Z quantifiable and inseparable attributes which determines its price:

$$Zj = (Zj1, Zj2, Zj3, \ldots Zjk) \tag{2.1}$$

Source: Day, 2003, p.3.

Therefore, a consumer chooses to purchase a vector of attributes when choosing to purchase a particular property (Day, 2003). Rosen (1974) defines hedonic pricing as the functional relationship between the price of a heterogeneous product and the associated attributes:

$$Pj = P(Zj) = P(Zj1, Zj2, Zj3, \ldots, Zjk) \tag{2.2}$$

Source: Goodman, 1978, p.472.

where: Pj is the price of the product. Simply stated $Pj = f(Zj)$ where the price of a property is a function of a set of a smaller number of attributes (Goodman, 1978).

This results in the complete hedonic pricing function where the price of property is a function the inherent attributes. Notably an increase in price is experienced by more positive attributes and a decrease in price is experienced by more negative attributes, *ceteris paribus* (Els and Von Fintel, 2010).

Hedonic prices can be measured with the use of regression techniques that aim to establish the relationship between a set of property attributes and property prices. Using regression analysis it is possible to calculate the implicit price for attribute i of property j by taking the partial derivative, represented as:

$$Pi(Zj) = \frac{\partial P}{\partial Zi} \qquad (i = 1 \text{ to } Z) \tag{2.3}$$

Source: Day, 2003, p.5.

This function describes the additional amount to be paid to obtain a marginally higher level of attribute Zi, *ceteris paribus* (Day, 2003).

Multiple regression is widely used in econometrics where many independent variables are introduced to explain a dependent variable (Greene, 1993). Hedonic prices can be measured

with the use of multiple regression, a statistical technique which aims to establish the relationship between a set of property attributes and property prices by regressing a set of property attributes on price. Multiple regression assumes an implicit relationship between the dependent variable and several independent variables (Wooldridge, 2010). This makes multiple regression an effective technique for modelling the relationship between listing price and property attributes. Multiple regression can be used to measure the correlation between each attribute or independent variable and the listing price given a collection of property data with the aim of predicting property prices (Monson, 2009).

It is important that measurement techniques allow for samples to adjust for compositional change over time so that quality change is not interpreted as pure price change. de Haan and Erwin (2011) present several important attributes to consider when constructing a residential property price index that adjusts for quality change over time, ensuring that changes in property prices reflect pure price changes and not merely changes in the composition of samples at different points in time. Table 2.1 depicts the important attributes necessary to adjust for quality change.

Table 2.1: Important Attributes in Determining Prices of Residential Properties

Attribute	**Description**
Size of the property	Measured in squared meters or squared feet
Area of the land and structure	Measured in squared meters or squared feet
Location of the property	A geographic variable
Age of the property	How old the property is
Type of property	Examples: detached, apartment, house etc.
Materials used in the construction of the property	Examples: primarily wood, brick, concrete etc.
Other	Number of bedrooms, bathrooms, garages, swimming pool, distance to amenities etc.

Source: de Haan and Erwin, 2011, p.25.

The omission of important attributes in hedonic price analysis has the propensity to bias estimates of the implicit prices measured, however, many models are subject to data availability. Model misspecification may arise in a hedonic analysis due to data availability constraints and subjective judgements by the researcher where important variables are not included in the analysis (Jiang, Phillips and Yu, 2015). An important consideration in developing a model is the principle of parsimony, where the aim is to choose a parsimonious or simpler model that explains the data well and is more generalisable. Parsimonious models typically feature fewer independent variables that explain the the most variability in the dependent variable (Tan, 2011). Simplicity through parsimony of parameter selection is a desired feature of any model as complexity is reduced and should the model be correctly specified, the result will be better predictions (McCullagh and Nelder, 1989).

2.3.2 A Review of Residential Hedonic Models

Typically international and local residential property hedonic price studies use ordinary least squares to derive the hedonic pricing function. Given a vector of a dependent variable and a matrix of independent variables, ordinary least squares makes it possible to express the dependent variable as a linear combination of the independent variables (Greene, 1993).

Day (2003) modelled house prices in Glasgow using hedonic pricing and ordinary least squares where a set of structural, accessibility, neighbourhood and environmental attributes were regressed on the selling price of properties sold. The natural logarithm of sales price was regressed on a linear combination of independent variables to derive the hedonic pricing function. Day (2003) also applied the natural logarithm transform to the floor area variable in his study. Day (2003) found all the structural attributes in the model to be statistically significant where he observed that larger properties with bigger gardens had higher prices. Day (2003) found that the inclusion of spatial data was an extremely important consideration in the estimation of the hedonic price function. A widely accepted tenet is that location is a significant determinant of a property's price (Ozyurt, 2014).

Bourassa, Cantoni and Hoesli (2010) derived a hedonic price function for the Auckland housing market using several ordinary least squares models, applying the natural logarithm to the dependent variable in each model. They took cognisance of the fact that property prices are closely related to adjacent properties and effectively modelled the spatial dependence

thereof. Broadly speaking, spatial autocorrelation can be defined as the dependence of observations across geographic locations which has the propensity to render the standard errors of ordinary least squares models inefficient and biased (Liao, Wang, 2012). Bourassa, Cantoni and Hoesli (2010) found that property price predictions were more accurate when sub-market dummy locational variables were used in contrast to using traditional statistical methods alone, however, incorporating both methods yielded the best results. Notably they argue that the use of sub-market dummy locational variables in ordinary least squares is a far simpler technique than trying to model the structure of the errors using complicated statistical methods and the benefit was evident in their results. Adding a dummy spatial variable to the combination of independent variables can remove the misspecification of the model which can be seen in the ordinary least squares regression diagnostics, making the interpretation of the results straightforward (Thayn and Simanis, 2013). In order to test for the presence of spatial autocorrelation, Borcard and Legendre (2012) found that the Mantel test was a reliable method which they used on univariate and multivariate data in an ecological study that investigated the relationship between grain and spatial autocorrelation using various statistical tests. Despite recent criticism of the Mantel test Diniz-Filho *et al.* (2013) found that the Mantel test was a powerful technique to analyse the amount of spatial variation in multivariate data where the results were congruent with *a priori* knowledge.

Els and Von Fintel (2010) conducted a time series hedonic analysis of the Stellenbosch, Somerset West, Strand and Gordon's Bay housing markets from 2004 to 2007 where they employed ordinary least squares and quantile regression techniques. Two models were derived using ordinary least squares, the first was a standard approach not including location or neighbourhood effects and the second incorporating dummy variables for the area, thereby introducing neighbourhood effects. In both ordinary least squares hedonic models, the natural logarithm of sales price was used as the dependent variable due to being the standard in literature and improving the R-squared diagnostic which measured the proportion of the variation in the sale prices explained by the variation in the set of attributes. By taking the natural logarithm of the sale price variable, all the coefficients were interpreted as percentage effects. The results showed that by capturing neighbourhood effects through the inclusion of area dummy variables, bias was reduced and the presence of spatial autocorrelation was mitigated, thus improving the overall fit of the model and increasing the R-squared diagnostic. Furthermore, Els and Von Fintel (2010) found that the ordinary least squares model without locational dummy variables was misspecified by not satisfying the functional

form requirement, whereas by incorporating locational dummy variables the functional form of the model was satisfied. Els and Von Fintel, (2010) tested this assumption using the Ramsey RESET test. Functional form of ordinary least squares is paramount and can be violated by the omission of explanatory effects or when the model does not account for important non-linearities (Tserkezos, 2009). The study of Els and Von Fintel (2010) included many structural attributes and interestingly the results revealed that the number of bedrooms was not a statistically significant variable, however, the size of the residence and the number of bathrooms were. Moreover, the number of bedrooms coefficient in the ordinary least squares model without locational effects had a negative sign whilst the same coefficient in the ordinary least squares model that included locational effects through dummy variables had a positive coefficient. The presence of the sign change could have been attributed to adding the locational dummy variables. Kennedy (2005) asserts that an omitted explanatory variable in a hedonic regression model can change the sign of one or more existing coefficients already specified in the model and to consider adding an independent variable to correct the misspecification. Els and Von Fintel (2010) were concerned over presence of heteroscedasticity in their study using ordinary least squares and therefore endeavoured to develop a non-parametric quantile regression model that is typically more robust to heteroscedasticity.

Heteroscedasticity is a common problem in cross sectional econometric studies and is endemic to spatial studies (Anselin, 2013). Using linear transformations such as taking the natural logarithm of the dependent variable often reduces the effects of heteroscedasticity and mitigates the presence thereof by changing the variance of the error term or residuals (Malpezzi, 2003). Heteroscedasticity violates one of the fundamental assumptions of ordinary least squares namely, that there is constant variance of the residuals (Stohldreier, 2012). Formally stated, the error term must be independently and identically distributed (Rawlings, Pantula and Dickey, 1998). The presence of heteroscedasticity may render the ordinary least squares coefficient estimates inefficient where standard errors and p-values may be biased or incorrect making hypothesis testing or deriving confidence intervals problematic. However, heteroscedasticity does not affect the consistency nor impair the unbiasedness of the actual ordinary least squares coefficient estimates (Gujarati, 2005). This makes the presence of heteroscedasticity a serious problem for inference in econometric models. Long and Ervin (2000) conducted a study where they explored heteroscedasticity in small samples and assert that if there is *a priori* reason to suspect the presence of

heteroscedasticity in the model, heteroscedasticity consistent covariance matrix (HCCM) techniques should be utilised. White (1980) derived an asymptotically or large sample form of HCCM which he introduced to the field of econometrics. White (1980) asserts that linear transformations may assist in eliminating the effects of heteroscedasticity. However, should the model fail the available tests for heteroscedasticity, the use of a covariance matrix estimator will be consistent in the presence of heteroscedasticity. Asymptotic HCCM techniques can mitigate type one errors of hypothesis tests for regression coefficients for large sample models (Cai, 2008). HCCM facilitates the mitigation of heteroscedasticity by producing a consistent estimator of the covariance matrix of the regression coefficients when the form of heteroscedasticity is unknown (Long, Ervin, 2000). A prominent assertion of Long and Ervin (2000) is that the decision to correct for heteroscedasticity should not be based on explicit tests or screening the results of the model, but rather this should be a routine procedure.

Dodds (2011) conducted an analysis of residential properties that were sold in the Westrand area in the Gauteng province where he aimed to predict property prices using an ordinary least squares hedonic pricing model based on statistically significant structural variables and location. Whilst the choice of structural attributes was contingent on the data, there was a total of eleven structural variables and one dummy location variable. An important observation made by Dodds (2011) was that the number of bedrooms and number of bathrooms had the highest positive correlations with the dependent variable, sale price. However, the output of the hedonic model revealed a negative coefficient for the number of bedrooms. This may have been due to an important omitted variable or multicollinearity as the model specified twelve independent variables in total. Multicollinearity is ubiquitous in ordinary least squares regression models and it's presence can be identified when independent variables are highly correlated with each other. The variance inflation factor is a useful method for detecting the magnitude of multicollinearity (Chen, 2010). Hedonic models are often subject to the presence of multicollinearity and the effects thereof can result in measurement errors and negative coefficients where based on *a priori* belief, a positive sign should be present (Triplett, 2005). Tan (2011) chose to develop a parsimonious hedonic model for the Malaysian housing market using six independent variables as a function of the dependent variable. Tan (2011) found that the model was less likely to suffer from multicollinearity issues by regressing fewer independent variables on the dependent variable through the use of stepwise regression. Interestingly Dodds (2011) found that the inclusion of

an area dummy locational variable in the initial model of the entire region reduced the R-squared model diagnostic. However, when a segment of selected suburbs were modelled in an isolated area model the R-squared diagnostic increased significantly. Dodds (2011) found that heteroscedasticity was present in the hedonic model and by applying a natural logarithm to the dependent variable or sale price made the error term less heteroscedastic.

2.4 Log-normal or Gamma Distribution and Appropriate Model Selection

Selecting an appropriate model for many econometric data is often compounded by non-normal sample distributions where transformations are required to derive a suitable model. Ordinary least squares is a prevalent technique in health econometric modelling where the dependent variable is often positively skewed and therefore transformed using the natural logarithm to obtain an error structure that is approximately normal (Manning, Basu and Mullahy, 2002). A similar assertion is made by de Haan and Erwin (2011) for house prices where the semi-log ordinary least squared model is propounded. The log-normal distribution of a random continuous variable is normally distributed through its logarithm and where the variability of residuals increase for larger values of the dependent variable, the logarithmic transform may prove effective (Rawlings, Pantula and Dickey, 1998). The log-normal distribution is a distribution that permits only positive variables (Harvey, Gavin, Scruggs, 2016). An assumption of transforming the dependent variable using the natural logarithm is that the transformed dependent variable will follow a log-normal distribution (Fu and Moncher, 2004). In ordinary least squares the error structure is assumed to be normally distributed, implying that the dependent variable is also normally distributed (Rawlings, Pantula and Dickey, 1998). A possible caveat with transforming the dependent variable is the interpretation and back transformation thereof as the dependent variable is no longer reported on the original scale (Olivier, Johnson and Marshall, 2008). The natural logarithm transformation results in predicted values on the logarithmic scale and back transformation is necessary where the geometric mean and not the arithmetic is obtained on the original scale (Olivier, Johnson and Marshall, 2008). The back transformed mean is different to the arithmetic mean of the original data, the back transformed estimate should be interpreted as the median if the log-transformation made the data normally distributed as the log transformation is monotonic (Musset, 2006). However, if the distribution was not made normal by the log-transformation then geometric mean estimates are obtained through back transforming or exponentiating to the original scale (Musset, 2006).

An alternative approach to using ordinary least squares where the dependent variable does not follow a normal or log-normal distribution and where the arithmetic mean is computed as the expected value is to use a generalised linear model. Residential property prices are a class of continuous non-negative variables where the variance is not constant thus facilitating the use of generalised linear models. House prices are generally positively skewed which reflects the heterogeneous nature of residential property (de Haan and Erwin, 2011). McCullagh and Nelder (1989) assert that normality and constant variance are no required for all generalised linear models, however, an understanding of how the variance depends on the mean is necessary. Exponential distributions such as the gamma distribution can be used to model a positive continuous dependent variable where the conditional variance of the dependent variable increases with the mean and the coefficient of variation is constant (McCullagh and Nelder 1989). Generalised linear models include distributions that are useful in the analysis of continuous measurements that have non-normal error distributions (McCullagh and Nelder 1989). The exponential gamma distribution can be used to model a non-negative, positively skewed continuous dependent variable where the variance is proportional to the square of the mean unlike the Gaussian or normal distribution where the variance is constant (Jones 2010). Fu and Moncher (2004) propound that the log-normal and gamma distributions are both widely used to model non-negative data that is positively skewed. Bromideh and Valizadeh (2013) assert that similarities exist between log-normal and gamma exponential distributions in terms of fit on moderate data sizes and both can prove effective in analysing non-negative positively skewed data. In a study of household expenditure Battese and Ponyhandy (1981) found that the assumption of normality and hypothesis of equal variance for the expenditure observations was rejected and that the gamma distribution was better for the model than the log-normal distribution in dealing with the heteroscedasticity. Moran, Solomon, Aaron and Martin (2007) in a study of patients' intensive care units costs, found that cost models employing ordinary least squares using a logarithmic transform could be augmented with the use of correctly specified generalised linear models which more effectively modelled the error structure.

The gamma distribution, a class of the exponential distribution family where non-negative positively skewed data can be modelled effectively, provides a possible solution to modelling non-negative, non-normal distributions such as residential property prices. Based on the literature presented it is apparent that ordinary least squares using the natural logarithm transformation on the dependent variable is a common method to derive hedonic price

functions for residential properties. Little or no research has been conducted on constructing a residential property hedonic price function using a generalised linear model based on the gamma distribution although this approach has been used to model actuarial data effectively. Fu and Moncher (2004) conducted an analysis on insurance claims, a non-negative continuous and positively skewed variable, using a generalised linear model model where they found that the gamma distribution resulted in better predictive accuracy and efficiency than the log-normal distribution. Furthermore, they suggest that examining the residual plots is a good measure to gauge the distribution assumptions. Checking the residuals for correct model specification is extremely important when examining the results of a generalised linear model as in the case of ordinary least squares. The error structure of a model is an important consideration in modelling data. Testing the error structure through diagnostic plots will provide guidance to how well the model fits the data (Murphy, Brockman and Lee, 2000). Carruthers, *et al* (2008) assert that inflated residual deviance or over-dispersion may arise in the form of model misspecification and can be identified where the residual deviance is greater than the residual degrees of freedom when using generalised linear models. Over-dispersion may manifest if one or more important variables are not accounted for, or if an incorrect error distribution is specified due to an inappropriate link function which can increase the chances of a type one error occurring (Carruthers *et al.,* 2008). In the case of generalised linear models, the parameter estimates are obtained by maximizing the log likelihood of the parameters for the observed data (McCullagh and Nelder, 1989).

An appealing feature of a generalised linear model using a log linear function for strictly non-negative and positively skewed data is that the predictions are kept in the natural units of measurement, making the estimation of the model more attractive than estimations through log-normal ordinary least squares.

2.5 Model Validation using Bootstrapping

A flexible and general approach to statistical inference is bootstrapping where the sample is treated as the population and repeated samples are drawn from it. Bootstrapping builds a sampling distribution of a statistic by re-sampling from the data and is considered a general approach to statistical inference (Fox, 2002). The flexibility is derived where asymptotic results cannot be relied upon or the assumptions made about the population are incorrect. Specifically the non-parametric bootstrap facilitates a practical estimate of the sampling

distribution of a statistic without knowing or deriving the explicit sampling distribution (Fox and Weisberg, 2010). Drawing a large amount of repeated samples accounts for the variance in the estimates of parameters which augments accuracy significantly. The bootstrap can be used as general tool for assessing statistical accuracy (Hastie, Tibshirani and Friedman, 2005). Testing and validation of the model objectives can be accomplished through several techniques including bootstrapping and randomisation (Carruthers *et al.,* 2008).

Gandy and Kvaloy (2013) applied non-parametric bootstrapping to circumvent estimation errors of parameters in control charts where it was found that non-parametric bootstrapping was robust against model specification errors. Moreover, Gandy and Kvaloy (2013) found that the procedure was particularly relevant when the "in-control" distribution was not known. Validating the results of regression results is an important part of the analysis where the rigour of the model can be tested. The validity of regression models are paramount when making inferences and model validation is an important step in the analysis (Oredein, Olatayo and Loyinmi, 2011). Oredein, Olatayo and Loyinmi (2011) conducted a study on regression model validation where they examined data splitting techniques and bootstrapping. The data splitting techniques involved separating the data, one part of the data used for modelling and another part for testing the model. They found that bootstrapping was a better model validation method for regression models, producing a more stable and higher R-squared model diagnostic. Wilcox (2008) used bootstrapping as a strategy to determine the correctness of hypothesis tests of coefficients in a multiple regression analysis which was found to be highly effective.

2.6 Conclusion

Residential property is a paramount determinant and indicator of economic policy formulation and the stability of an economy where many participants share an interest in tracking and understanding the movements thereof.

Hedonic price models are a pervasive and useful technique that decompose residential properties into constituent utility bearing attributes and derive the implicit prices thereof. Based on a review of international and local literature, the construction of residential property hedonic price models are often characterised by complex caveats such as spatial autocorrelation, heteroscedasticity and the optimal choice of a plethora of independent

variables. A deep understanding of these effects is required to develop an accurate reflection of residential property prices. Whilst ordinary least squares has been used to model the hedonic pricing function of residential property, novel techniques such as the use of generalised linear models may be of more value in constructing such hedonic price functions. Understanding the distribution of residential property prices is paramount in selecting an appropriate model and actuarial research provides evidence that the gamma and log-normal distribution are effective for non-negative data that is positively skewed. Model validation is extremely important and bootstrapping can provide an indispensable tool to determine the effectiveness of the results.

The following chapter presents the statistical techniques used to derive the hedonic price function for flats within KwaZulu-Natal coastal sub-markets and describe the rigorous tests required to meet the specific assumptions.

CHAPTER THREE

The Research Methodology

3.1 Introduction

The literature presented in the previous chapter outlined the economic importance of measuring residential property prices and provided an empirical evaluation of the techniques and complexities experienced in the global and local context. It is evident that more research is needed to supplement the limited number of studies conducted on the South African residential property market.

This chapter presents a thorough analysis of the research process adopted in the study, outlining all the statistical procedures involved in the research methodology.

3.2 Aim, Objectives and Hypotheses of the Study

3.2.1 Aim

The aim of this study was to investigate existing empirical research in order to develop a hedonic pricing model for flats located in KwaZulu-Natal coastal sub-markets.

3.2.2 Objectives

The research objectives of this study were as follows:

1. To determine an appropriate hedonic price model for flats located in KwaZulu-Natal coastal sub-markets based on the distribution of listing prices.
2. To develop a model to estimate listing prices of flats that are located in sub-markets along the KwaZulu-Natal coast based on structural and locational attributes.
3. To build a software application that facilitates the estimation of listings prices for flats in KwaZulu-Natal coastal sub-markets given a set of structural and locational attributes.

3.2.3 Hypotheses

In order to achieve the research objectives, the following research hypotheses were formulated:

H0a: The distribution of listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast follow a gamma distribution.
H1a: The distribution of listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast do not follow a gamma distribution.

H0b: Floor Area is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
H1b: Floor Area is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H0c: Number of bedrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
H1c: Number of bedrooms is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0d*:** Number of bathrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
***H1d*:** Number of bathrooms a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

H0e: The suburb, a dummy locational variable, is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H1e**:* The suburb, a dummy locational variable, is a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

3.3 Location of the Study

The location of the study involved three coastal sub-markets within KwaZulu-Natal, a subset of the Ethekweni municipality. The location of the study was chosen due the lack of research done on residential hedonic property pricing in KwaZulu-Natal. Dodds (2011) conducted a study to investigate the use of hedonic price analysis to determine what structural characteristics determine the open market value of houses in the Gauteng province. Dodds (2011) primarily focused on ordinary least squares regression in testing the hypotheses he propounded. Another prominent study was that of Els and Von Fintel (2010) who applied several hedonic pricing techniques including ordinary least squares regression and quantile regression to determine property market growth rates for houses in the Western Cape province. Du Preez *et al.* (2014) conducted a study where hedonic pricing was used to model house values and proximity to a landfill site using quantile regression. It appears that no research on residential property hedonic pricing has been done in KwaZulu-Natal. Based on data availability the three coastal sub-markets of this study were chosen.

The Ethekweni municipality is located on the east coast of South Africa and spans approximately 2 297 squared kilometres (Statistics South Africa, 2011). The key statistics of Ethekweni are presented in Table 3.1.

Table 3.1: Key Statistics of Ethekweni

Total Population	Number of Households	Average Household Size	Formal Dwelling	Housing Owned / Paying Off
3 442 361	956 713	3.4	79.00%	54.50%

Source: Adapted from Statistics South Africa 2011.

The three coastal sub-markets of the study were Ballito, Umhlanga, and Durban Central illustrated as red, blue and green respectively, presented in Figure 3.1. The different coloured points on the map indicate the sample observations of the study.

Figure 3.1: Sub-markets of Study

Source: Author generated map using R Package ggmap created by Kahle and Wickham (2013) and Google Maps (2016).

3.4 Type of Study

Paradigms are a set of universally accepted thinking habits of researchers (Kuhn, 2012). These paradigms guide the processes, the research design and the methods that will be adopted in the research. This study can be classified as quantitative research as the statistical nature of the study aims to incorporate multiple variables to examine correlations with the aim of making estimates and inferences. Specifically, how a set of independent variables explain the variation in a dependent variable. Quantitative research implores experiments and models that include multiple variables to examine correlation and causation analysis (Creswell, 2009). Quantitative analysis involves data collection with the aim of using statistical procedures and testing hypotheses in order to support or refute claims. Quantitative research is also the called the scientific method which embodies the post-positivist paradigm and seeks to determine cause and effect relationships, reducing ideas into a discrete set variables for research questions and hypotheses (Creswell, 2009). Therefore, the scientific method adopted by post-positivists begins with a theory, followed by data collection to determine whether the statistical analysis of the data supports or refutes the theory.

3.6 Research Design

The dependent and independent variables used in the study as well as the class and description are outlined below in Table 3.2.

Table 3.2: Outline of Variables in the Study

Name of Variable	**Type of Variable**	**Class of Variable**	**Description of Variable**
Price	Dependent	Continuous	The price of a flat in ZAR
Size	Independent	Continuous	The floor area of a flat in square meters
NumBaths (Number of Bathrooms)	Independent	Count/discrete	The number of bathrooms in a flat
NumBeds (Number of Bedrooms)	Independent	Count/discrete	The number of bedrooms in a flat
Suburb	Independent	Categorical/factor	The suburb a flat resides in

3.7 Sample

The sample data for the research was provided by Private Property (Pty) Ltd by the Chief Technology Officer, Mr Grant Elliot. The sample data was a snapshot of all the flat listings for sale in Ballito, Durban Central and Umhlanga as at 2016-02-23. The data is secondary data as it was assimilated by another party and not the researcher through primary data research initiatives. Secondary data is data that has already been collected and been made available (Kothari, 2004). The sample provided comprised of 5916 observations before cleaning the data for missing fields, duplicates and incorrect data. Upon checking the data for accuracy, integrity, completeness and uniqueness it became apparent that some data cleaning procedures would be need to be implemented. The data was received in the form a comma separated value file. Missing values or fields were removed resulting in the removal of an entire row or observation. Furthermore, rows with incorrect geographic co-ordinates were identified and removed by plotting the data on a map. Lastly duplicates were removed by applying a heuristic that identified duplicate rows on the basis of having the same residential street address. This resulted in a final sample of 1314 observations for the study.

3.8 Ethical Considerations

To obtain ethical clearance for this study the researcher first obtained a gatekeepers from the Chief Technology Officer, Mr Grant Elliot, granting permission for the use of Private Property (Pty) Ltd's data. Once the gatekeepers letter was received the research proposal was submitted to the University of KwaZulu-Natal Ethical Committee for approval of the study. Both the gatekeepers letter form Private Property (Pty) Ltd and the approval letter from the Ethical Committee can be found in the appendix of this research report.

3.9 Statistical Modelling Framework

All statistical modelling and analysis was performed in the open source statistical programme R and the integrated development environment Rstudio. Furthermore, the development of the software application was also built in Rstudio. R is a statistical programming language and free software to use under the GNU General Public License (Peng, 2015). R has its roots in the the old S language. It was created by Ross Ihaka and Robert Gentleman in the Department of Statistics at the University of Auckland and first made available to the public

in 1993 (Peng, 2015). The researcher has endeavoured to learn the R programming language in order to facilitate this study where all the analysis and modelling presented is from code written by the researcher using R and the packages provided.

3.9.1 Identifying and Testing the Distribution of the Dependent Variable

Probability distributions serve as models for the mechanisms that create observed data (Greene, 1993). Choosing the best estimator depends on the statistical properties of the sample distribution, efficiency, unbiasedness and consistency. (Greene, 1993). Based on this premise, identifying the correct distribution of the flat listing prices in the sample data will be a fundamental feature of this study.

3.9.1.1 Log-Normal Distribution

A random variable x has a log-normal distribution if its probability density function is given by:

$$f\,x^{(x)} = \frac{1}{\sigma\, x \sqrt{2\pi}} \exp\left[\frac{-(lnx-\mu)^2}{2\sigma^2}\right], 0 < x < \infty \tag{3.1}$$

Source: Kerns 2011, p.161.

where the two parameters of interest are the mean μ and the standard deviation σ which are reported on the log scale. Evident from the log-normal probability density function is that the log-normal distribution extends for positive continuous variables greater than zero. In order to determine if a distribution is normally distributed statistical tests can be utilised. The Jarque-Bera Test for normality is an asymptotic test that uses higher order moments where under the null hypothesis, the data is normally distributed (Gujarati, 2005; DeBenedictis and Giles, 1998). This test was utilised to determine whether the dependent variable was log-normally distributed.

3.9.1.2 Gamma Distribution

A random variable x has a gamma distribution if its probability density function is given by:

$$f\,x^{(x)}=\frac{\lambda^{(\alpha)}}{\Gamma(\alpha)}x^{\alpha-1}e^{\lambda x},\ x>0 \tag{3.2}$$

Source: Kerns 2011, p.158.

where the two parameters of interest are the shape α and the rate λ. Evident from the gamma probability density function is that gamma distribution extends for positive continuous variables greater than zero.

Villaseñor and Gonzalez-Estrada (2014) devised a new goodness-of-fit test for the gamma distribution based on the ratio of two variance estimators, namely the ratio of the sample variance and the moment estimator. A Monte Carlo simulation provided evidence of the efficiency of the goodness-of-fit test. The Villasenor and Gonzalez-Estrada test was applied in this study to determine whether a gamma distribution was appropriate for the dependent variable.

3.9.2 Correlation Analysis

The Pearson coefficient of correlation measures the strength and existence of a relationship between two variables and is defined as the covariance of the variables divided by the standard deviations of the variables (Keller, 2012). The Pearson coefficient of correlation is presented as:

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}} \tag{3.3}$$

Source: Egghe and Leydesdorff, 2009, p.6

The Pearson coefficient of correlation was used to measure the relationship between the set of independent variables and the dependent variable. Independent variables that are strongly correlated with the dependent variable might be good predictors in the model.

3.9.3 Model Selection and Assumptions

Two models were developed in this study namely, an ordinary least squares model based on a log-normal distribution and generalised linear model based on the gamma distribution and log link function. The ordinary least squares model was developed in order to make relevant comparisons to the existing literature presented in Chapter Two and the generalised linear model was developed as an alternative, novel model in which the research hypotheses were tested.

3.9.3.1 Ordinary Least Squares

Ordinary least squares is a popular parametric method employed to estimate the parameters of a multiple regression model (Wooldridge, 2010). The ordinary least squares method aims to express the dependent variable as a linear function of the independent variables while minimizing the sum of squared deviations of the dependent variable from the estimates of the true mean responses produced by the model (Rawlings, Pantula and Dickey, 1998). The ordinary least squares can be expressed as a linear function:

$$\hat{y}i = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + ... \hat{b}\, n_x + \varepsilon \tag{3.4}$$

Source: Rawlings, Pantula and Dickey, 1998, p.2.

$\hat{b}0$ is the intercept and each $\hat{b_1}, \hat{b_2}, ... \hat{b}n$ are the respective slope coefficients or rates of change in $\hat{y}i$ per unit change in x (Keller, 2012). Each of the slope coefficients are partial derivatives of $\hat{y}i$ with respect to the applicable variable which they multiply.

Ordinary least squares assumes the random error to have a zero mean, common variance and to be independent and identically distributed (Rawlings, Pantula and Dickey, 1998). This assumption applies to the stochastic part of the model namely, the random error, however, this implicitly means that these assumptions apply to the dependent variable and can be expressed mathematically as:

$$\varepsilon_i \sim NID(0, \delta^2) \tag{3.5}$$

Source: Rawlings, Pantula and Dickey, 1998, p3.

After estimating the ordinary least squares model coefficients, the significance of the independent variables need to be assessed with regards to predicting the dependent variable. P-values establish whether the coefficients obtained in the model are statistically significant and not arrived at by pure chance. P-values are a consistency measure of whether the results obtained in a trial are attributable to pure chance (Thisted, 2010). The p-value of a test is the probability of observing a test statistic at least as extreme as the one computed given that the null hypothesis is true. P-values smaller than the level of significance are evidence against the null hypothesis (Demortier, 2007). The level of significance used in this study for all two tailed hypothesis tests is 0.05.

The coefficient of determination, also known and the R-squared diagnostic, shows the amount of variation in the dependent variable that is explained by the variation in the independent variables. Generally the higher the R-squared the better the model fits the data (Keller, 2012). The R-squared can be represented as:

$$R^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y}_i)^2}{\Sigma(Y_i - \bar{Y}_i)^2} = \frac{EES}{TSS} \tag{3.6}$$

Source: Gujarati, 2005, p.176.

where ESS is the explained sum of squares and TSS is the total sum of squares. This metric was used to assess the ordinary least squares explanatory power.

Different types of tests exist to determine whether the assumptions of a model have been met namely, formal or informal tests. Formal tests involve statistical hypothesis tests and informal tests involve examining plots to determine model fit (Carruthers, *et al.*, 2008). Both informal and formal tests were used in this research, where informal tests were supplemented with formal tests.

Residual analysis is paramount to assess how the model fits the data (Muchabaiwa, 2013). In ordinary least squares the assumptions of homoscedasticity, or equal variance of the response residuals, is satisfied if a random scattering of the residual points around the zero horizontal line in a residual plot against the fitted values is present (Rawlings, Pantula and Dickey, 1998). This is illustrated in Figure 3.2 below.

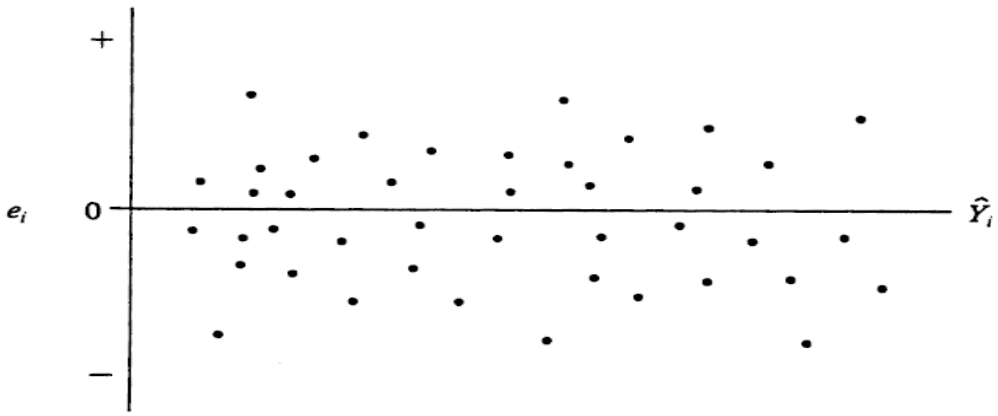


Figure 3.2: Residual Plot Against the Fitted Values

Source: Rawlings, Pantula and Dickey, 1998, p.347.

The first step to test the assumption of homoscedasticity in an ordinary least squares model is to perform a visual inspection of the the residual plots, if evidence of heteroscedasticity is present the Breusch-Pagan test can be performed to detect any linear form of heteroscedasticity. The Breusch-Pagan test, tests the null hypothesis that the response residual variances are equal or homoscedastic (Williams, 2012). However, the Breusch-Pagan test is sensitive to the normality of residuals and can lead to a type one error (the incorrect rejection of the true null hypothesis) (Coenders and Saez, 2000). The scale location plot can also be used for detecting monotone spread or heteroscedasticity (Hinkins, Mulrow and Scheuren, 2009). The Normal Q-Q plot is an informal visual plot that can be used to assess the normality of response residuals (Hinkins, Mulrow and Scheuren, 2009). Cooks distance, an informal visual plot, is a way of observing influential observations in an ordinary least squares model and can be useful in detecting outliers (Williams, 2012).

The Ramsey's Regression Specification Error Test (RESET) can be employed to test the functional form of an ordinary least squares model where the unknown non-zero random error is approximated, showing the degree of the model misspecification by using a function of the conditional mean of the model (DeBenedictis and Giles, 1998).

Model selection between a set of candidate ordinary least squares models can be achieved by a host of tests. The Akaike Information Criterion (AIC) has been proposed to compare two or more models (Greene, 1993) which is based on the asymptotic maximum likelihood estimator where the model with the lowest AIC is preferred as it penalizes for over fitting (Muchabaiwa, 2013). The AIC is given by:

$$-2\mathrm{L}(\beta)+2(p) \tag{3.7}$$

Source: Muchabaiwa, 2013, p.30, 2013

where L is the log-likelihood of the model and p is the number of parameters in the model plus 1 (Muchabaiwa, 2013).

The Wald test can be used to assess the significance of individual coefficients in the model (Muchabaiwa, 2013). The Wald test is given by:

$$W=\frac{\hat{\beta}_i}{SE(\hat{\beta})'} \tag{3.8}$$

Source: Muchabaiwa, 2013, p.31, 2013

where $\hat{\beta}_i$ is the estimated coefficient of the respective independent variable and SE is the respective standard error (Muchabaiwa, 2013).

3.9.3.2 Generalised Linear Model using a Gamma Distribution and Log Link Function

Generalised linear models use the iterative weighted least squares algorithm to obtain maximum likelihood estimates of the parameters for observations that belong to an exponential distribution family where the systematic effects can me made linear through a

link function which is a type of transformation (Nelder and Wedderburn, 1972). Generalised linear models comprise of three components namely, a random or stochastic component, a systematic component and a link function (Nelder and Wedderburn, 1972). The notation of the generalised linear model is expressed as:

$$\eta_i = g_i(\mu_i) = x_i^T \beta \tag{3.9}$$

Source: Lindsey, 1997, p.18.

where the link function $g(.)$ relates the conditional mean to the covariates or systematic component denoted by $x_i^T \beta$ (Jones, 2010) and η_i is the linear predictor (McCullagh and Nelder 1989). Generalized linear models provide a consistent way of linking together the systematic elements in a model with the stochastic elements (Nelder and Wedderburn, 1972).

The difference between the generalised linear model using a log link function and ordinary least squares using the natural logarithm transformation on the dependent variable is presented below:

$$\ln(E(y|x)) = x\beta \tag{3.10a}$$
$$E(\ln(y)|x) = x\beta \tag{3.10b}$$

Source: Glick, 2008, p.5

3.10a is the generalised linear model with the log link function and 3.10b is the ordinary least squares model with a natural logarithm transformation on the dependent variable. Clearly illustrated is the circumvention of the issue of back transformation from the log scale to the original scale with the use of the generalised linear model and the log link function as the dependent variable is kept in its original unit using the generalised linear model (Glick, 2008).

Literature presents disparate views abut the assumptions of parametric generalised linear models (Carruthers *et al.*, 2008). Carruthers *et al* (2008) assert that similar to ordinary least squares the assumption of homogeneity of residuals is a key assumption with the use of

generalised linear models, except those used for binomial and beta-binomal distributions. A primary reason for using generalised linear models over ordinary least squares is to correctly account for the error structure and through the appropriate link function, the deviance residuals should be homogeneous (Murphy, Michael and Brockman, 2000). However, Glick (2008) states that generalised linear models do not require the normality or homoscedasticity assumptions to be met. McCullagh and Nelder (1989) present the view that in order to assess the goodness of fit of a generalised linear model, an understanding of how the variance depends on the mean is required. McCullagh and Nelder (1989) assert that the standardized deviance residuals plotted against the fitted values provides a measure to test whether the assumption of constant coefficient of variation is met which is required under a gamma distribution. They suggest that if the variance function is indeed quadratic, the plot should resemble a normal theory residual plot. Nelder and Wedderburn (1972) state that through the appropriate link function, linearity of the systematic component and the desired error distribution can be achieved. Generalised linear models require an extended definition of residuals for distributions other than normal, therefore deviance residuals are preferred for model diagnostics (McCullagh and Nelder, 1989).

Specification tests have been developed for linear models including ordinary least squares such as the Ramseys RESET test, however, this test is not appropriate in its original form for generalised linear models (Sapra, 2005). When the observed variance from the data is larger than the variance accounted for in a generalised linear model, over-dispersion is present (Muller, 2012). Over-dispersion may arise due to a lack of independence in the data or due to misspecification of the model (Muller, 2012). Testing for over-dispersion is applicable to generalised linear models and can be detected when the residual deviance is greater than the residual degrees of freedom or formally stated, if the ratio of deviance and degrees of freedom is greater than 1 (Carruthers, *et al.,* 2008).

3.10 Validation and Reliability of Results

3.10.1 Heteroscedasticity

In the presence of heteroscedasticity White (1980) has formulated a covariance estimator which is consistent, providing correct asymptotic results for linear hypothesis tests for linear

models. This has been named the heteroscedasticity consistent covariance matrix, denoted HC0 (Long and Ervin, 2000). The HC0 estimator is presented as:

$$HC0=(X'X)^{-1}X'\hat{\Phi}X(X'X)^{-1}=(X'X)^{1}X'diag[\varepsilon_i^2]X(X'X)^{-1} \quad (3.11)$$

Source: Long and Ervin, 2000, p.5.

Where $\Sigma=(X'X)^{-1}X'\hat{\Phi}X(X'X)^{-1}$ is the covariance matrix for the regression coefficients and $\hat{\Phi}=diag[\varepsilon_i^2]$ is a diagonal matrix with the residuals on the diagonal and all off diagonal entries equal to zero. The ith squared residual is placed into the ith diagonal of $\hat{\Phi}$ (Dawson, Redden and Beasley, 2015). This technique was employed to mitigate the presence of heteroscedasticity based on the results of informal and formal tests.

3.10.2 Spatial Autocorrelation

The presence of spatial autocorrelation in the residuals of a statistical model has the propensity to increase type one errors for parameter estimates, falsely rejecting the null hypothesis of no effect (Dormann *et al*., 2007). Constructing a spatial autocorrelation function can be achieved with a Mantel test which produces a standardized Mantel statistic similar to the Pearson correlation coefficient, however, distance formulas other than Euclidean can be used (Borcard and Legendre, 2012). The Mantel test is formulated as:

$$Z_m=\sum_{i=1}^{n}\sum_{i=1}^{n}g_{ij}\times d_{ij} \quad (3.12)$$

Source: Diniz-Filho *et al.*, 2013, p.476.

Where g_{ij} and d_{ij} are the respective variable and geographic distances between the distributions i and j and where Z_m is the sum of products of distances which is compared to a null distribution (Diniz-Filho *et al.*, 2013). This technique was used to formally test for the presence of spatial autocorrelation.

3.10.3 Multi-collinearity

Multi-collinearity has the potential to produce parameter estimates of the incorrect sign and magnitude by increasing parameter variance (O'brien, 2007). The variance inflation factor is a suitable measure for detecting the effects of multi-collinearity where a value greater than 10 is indicative of multi-collinearity (Kennedy, 1985). By testing for the presence of multi-collinearity, correct model specification and results can be obtained which was a primary initiative of all the modelling done in this study. Principal components regression and ridge regression are models for coping with multi-collinearity, however, the variance inflation factor provides a popular method for the measurement thereof (Williams, Grajales and Kurkiewics, 2013). The variance inflation factor is calculated as:

$$VIF = \frac{1}{r_{23}^2} \tag{3.13}$$

Source: Gujarati, 2005, p.328

where as r_{23}^2 tends towards 1, the VIF approaches infinity. This means that the variance of an estimator increases as the extent of collinearity increases and a score of 1 indicates no multi-collinearity between $X2$ and $X3$ (Gujarati, 2005).

3.10.4 Bootstrapping

Bootstrapping is used to ascertain a description of the sampling distribution of an estimator using the sample data (Greene, 1993). Bootstrapping accounts for variance in the parameters estimated by drawing a large amount of repeated samples. This technique was used as general tool for assessing statistical accuracy in this study. The notation for bootstrapping is:

$$\bar{T}^* = \hat{E}^*(T^*) = \frac{\sum_{b=1}^{R} T_b^*}{R} \tag{3.14}$$

Source: Fox and Weisberg, 2010, p.2

where $\bar{T}^{\star}$ is the estimator or averaged bootstrapped estimate derived by $\frac{\sum_{b=1}^{R} T_b^{\star}}{R}$ where R is the number of bootstraps applied (Fox and Weisberg, 2010).

3.11 Comparison of Ordinary Least Squares and Generalised Linear Model Fitted Values

To determine the accuracy of each model, the Root Squared Mean Error (RMSE) of the fitted values from each model is compared against the the observed values. The RMSE is a statistical measure of the dispersement that compares the closeness of model outcomes to observed outcomes where a lower RSME indicates greater accuracy (Vastrad, 2013).

In order to illustrate how closely the fitted values of the ordinary least squares model and generalised linear model approximate the observed values in the data, a comparison is made by random sampling where a subset of the data is randomly selected and a comparison of the different model results is made against the observed prices in the data.

3.12 Confidence Intervals

The true estimate or population estimate is likely to differ from the estimate obtained in a model due to sampling fluctuations. However, in repeated sampling of a model, the mean obtained is likely to more closely approximate the true mean of the population of interest (Gujarati, 2005). Confidence intervals typically create an interval around a point estimate using a level of significance which effectively means that the probability of the true population parameter falls within the interval given a certain level of significance (Gujarati, 2005).

In simple regression confidence intervals are presented as:

$$\hat{y} \pm t_{\alpha}/2, n-2^{se} \sqrt{1+\frac{1}{n}+\frac{(x_g-\bar{x})^2}{(n-1)s^2}} \tag{3.15}$$

Source: Keller, 2012, p.642

This equation can be extended for multiple regression where there are multiple coefficients and standard errors. To obtain a 95% confidence interval with a 5% margin of error for a population of 1 000 000, a sample size of 384 observations is required as illustrated in Table 3.3 below (The Research Advisor's, 2006). Without knowing the exact population size for this study a large estimate was taken to err on the side of caution. The sample size in this study was 1314 observations therefore the threshold is more than adequate.

Table 3.3: Required Sample Size

	Confidence = 95% Margin of Error			
Population Size	5%	3.5%	2.5%	1%
50000	381	772	1491	8056
75000	382	776	1506	8514
100,000	383	778	1513	8762
250,000	384	782	1527	9248
500,000	384	783	1532	9423
1,000,000	**384**	783	1534	9512
2,500,000	384	784	1536	9567
10, 000,000	384	784	1536	9594
100,000,000	384	784	1537	9603
300,000,000	384	784	1537	9603

Source: Adapted from The Research Advisor's, 2006.

Many populations, and samples drawn from such populations, are not normally distributed, however, regardless of the shape of underlying distribution, the Central Limit Theorem states that the sampling distribution of a random variable with a well defined expected value and finite standard deviation approaches normal asymptomatically (Kerns, 2011). Confidence intervals were produced for the software application in order to provide statistical inferences for the point estimates derived from the final model.

3.13 Conclusion

This chapter presented the research process and introduced the sample along with the statistical procedures involved in the research methodology. Understanding the distribution of a sample is paramount in building an effective model that describes the data well. When building parametric statistical models to explain data, various assumptions need to be met in order to satisfy the correctness of the model. Informal methods and formal tests provide ways to determine the validity of these assumptions. Incorporating techniques to adjust for model misspecification is important and where available, these should be leveraged such as in the case of Whites HCCM. Validating a model though bootstrapping provides a robust measure of accuracy. The subsequent chapter presents the detailed results of the research.

CHAPTER FOUR

Presentation and Discussion of Results

4.1 Introduction

The type of research and the data used in this study was presented in the previous chapter along with the data cleaning processes employed. Various statistical techniques were also introduced which are implemented in this chapter through a series of models and hypothesis tests using the programming language R. A discussion of key findings of the results ensues which contextualises the relevant literature discussed in Chapter Two.

4.2 Descriptive Statistics

The descriptive statistics of the data are shown in Table 4.1 where the spread and measures of central tendency are presented

Table 4.1 Tabular Descriptive Statistics of the Data

Price	Size	NumBeds	NumBaths	Suburb
Min. : 120000	Min. : 24.0	Min. :0.50	Min. :1.000	Min. : 1.00
1st Qu.: 795000	1st Qu.: 69.0	1st Qu.:2.00	1st Qu.:1.000	1st Qu.: 4.00
Median : 1695000	Median :112.0	Median :2.00	Median :2.000	Median :14.00
Mean : 2409668	Mean :139.9	Mean :2.35	Mean :1.822	Mean :15.55
3rd Qu.: 3180000	3rd Qu.:169.0	3rd Qu.:3.00	3rd Qu.:2.000	3rd Qu.:25.00
Max. :30000000	Max. :749.0	Max. :6.00	Max. :6.000	Max. :36.00

From Table 4.1 it is discernible that the lowest price flat in the sample is R 120 000 and highest price flat is R 30 000 000. The mean flat price in the sample is R 2 409 668 and the median is R 169 5000, clearly indicating the listing price distribution for flats is not normally distributed. The suburb column is a categorical data type therefore the only meaningful row is the maximum row which indicates that there are a total of 36 suburbs within the three sub-markets in the sample.

Previously it was stated that the sample of flat prices is not normally distributed, this can be clearly illustrated with the use of a histogram. Figure 4.1 shows the histogram of flat prices

from the sample with an overlay of three measures of central tendency namely, the arithmetic mean, geometric mean and median.

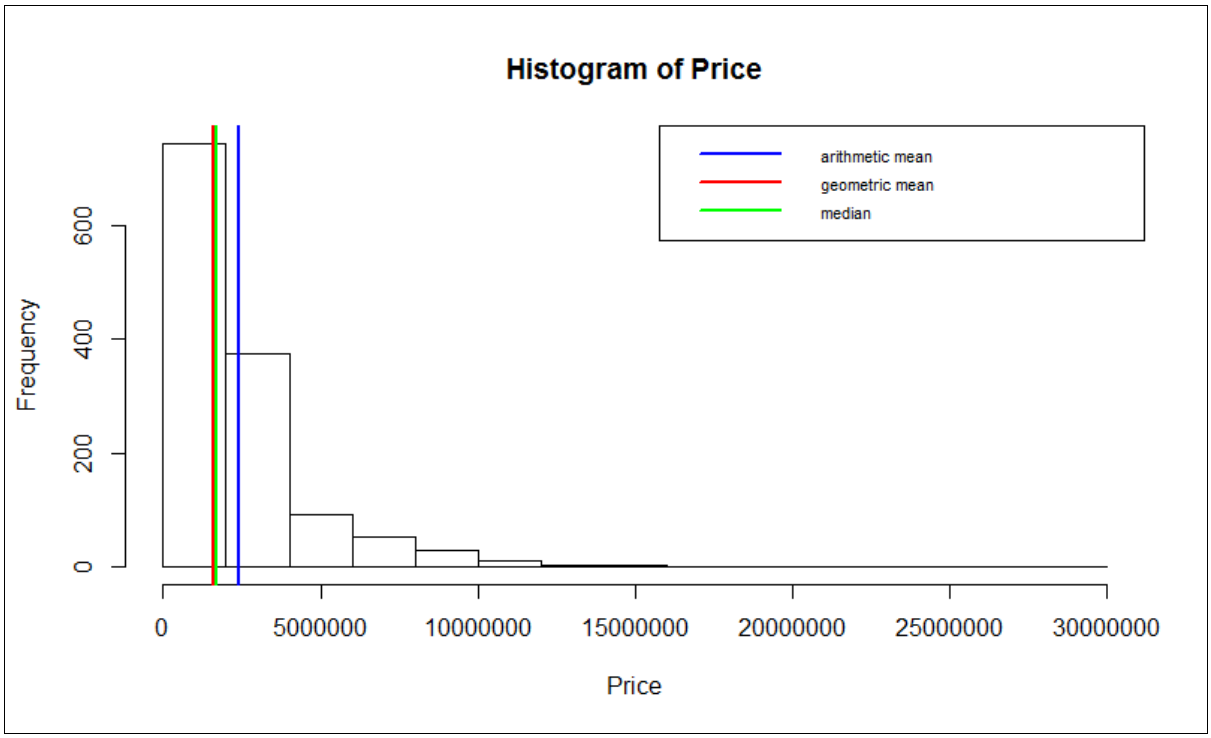


Figure 4.1: Histogram of Flat Prices

4.3 Determining the Appropriate Distribution of Flat Listing Prices

Determining the appropriate distribution of listing prices for flats in the sample data will facilitate the correct model selection.

4.3.1 Log-Normal Distribution

To determine whether the listing prices for flats in the sample follow a log-normal distribution the natural logarithm is applied. Figure 4.2 presents the histogram for the flat prices on the log scale where the kernel estimator, a non-parametric technique to estimate the probability density function, is computed and displayed along with the arithmetic mean and median.

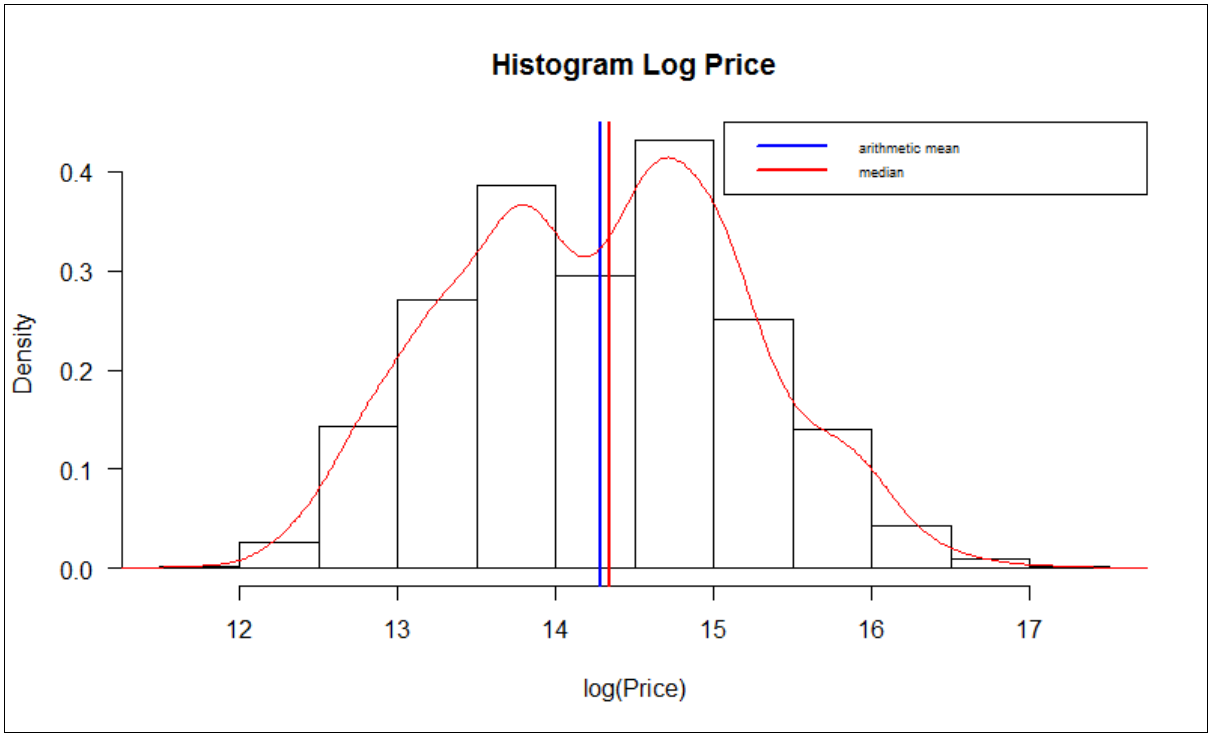


Figure 4.2: Histogram of Flat Prices with the Natural Logarithm Applied

In Figure 4.2 the flat prices appear almost log-normal, however, histogram is bi-modal and the arithmetic mean and median are not exactly equal.

To formally test whether the flat prices follow a log-normal distribution the Jarque-Bera test of normality is computed. Table 4.2 presents the results of the Jarque-Bera test for normality

Table 4.2: Jarque-Bera Test of Normality Results

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 17.7776
  P VALUE:
    Asymptotic p Value: 0.0001379

Description:
 Sat Apr 23 12:01:08 2016 by user: Dane
```

The asymptomatic p-value of the Jarque-Bera Test for normality is 0.0001379 which indicates the null hypothesis of normality is rejected. Therefore, according to this test, the flat prices do not follow a normal distribution even after the log-transform is applied. The assumption that the dependent variable is normally distributed is an important assumption of ordinary least squares (Rawlings, Pantula and Dickey, 1998).

Similar to the study on expenditure by Battese and Ponyhandy (1981), the results of this research finds that the assumption of normality is invalid. This leads to the rejection of the log-normal hypothesis and hence the need to investigate an alternative distribution, namely the gamma distribution.

4.3.2 Gamma Distribution

The gamma distribution is characterised as non-normal and is suitable for non-negative continuous data. The listing prices for flats in the sample data meets this description with Figure 4.3 illustrating the kernel density estimator and cumulative density function respectively.

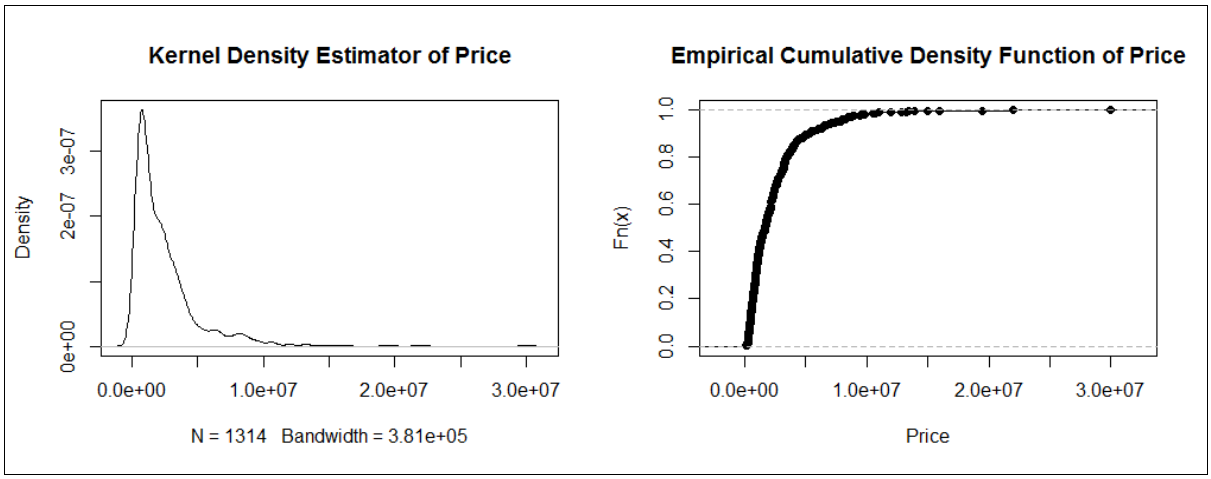


Figure 4.3: Kernel Density Estimator and Empirical Density Function of Flat Prices

The moment generating function facilitates the accurate identification of a distribution and computes the respective moments (Kerns, 2011). In order to ascertain whether the distribution of the listing prices is indeed gamma the shape and rate parameters are

determined using the matching moments method in the fitdistrplus package in R . Table 4.3 shows the computed moments, namely the shape and rate parameters for the listing price distribution of flats.

Table 4.3: Shape and Rate Parameters for Flat Price Distribution

```
Fitting of the distribution ' gamma ' by matching moments
Parameters :
             estimate
shape 0.9094943994262
rate  0.0000003774355
Loglikelihood:  -20646.2   AIC:  41296.4   BIC:  41306.76
```

To test the first hypothesis of this study, that the flat prices follow a gamma distribution, the Villasenor and Gonzalez-Estrada (2014) test is applied using the shape and rate parameters obtained from the matching moments function. Table 4.4 shows the results of the first hypothesis of this study.

Table 4.4: Villasenor and Gonzalez-Estrada Test for a Gamma Distribution

```
        Test of fit for the Gamma distribution

data:  gamtest
V = -1.2267, p-value = 0.3857
```

The p-value is 0.3857 indicating that there is sufficient evidence not to reject the null hypothesis that the flat prices follow a gamma distribution.

Notably, no previous studies discussed in the literature review sought to determine the underlying distribution of property prices through informal or formal tests but rather assumed a log-normal distribution where Day (2003); Els and Von Fintel (2010); Bourassa, Cantoni and Hoesli (2010); Dodds (2011) all applied the natural logarithm to the dependent variable in order to derive the hedonic price function using ordinary least squares. This research makes a similar assumption. Flat prices are asymptotically log-normally distributed, in order to produce a comparative log-normal ordinary least squares model. However, there is strong evidence that the distribution of flat listing prices in this study follows a gamma distribution.

Therefore, the second model, a generalised linear model, will model flat prices according to a gamma distribution.

4.4 Correlation Analysis

4.4.1 Correlation Matrix

The correlation between the numeric covariates (independent variables) in the sample can be presented in a tabular or graphical format using a correlation matrix. Table 4.5 tabulates the correlation matrix illustrating the magnitude of correlation between the numeric covariates.

Table 4.5: Tabular Correlation Matrix

	Price	Size	NumBeds	NumBaths
Price	1.0000000	0.7548592	0.6136805	0.7497797
Size	0.7548592	1.0000000	0.7528000	0.8041192
NumBeds	0.6136805	0.7528000	1.0000000	0.7777339
NumBaths	0.7497797	0.8041192	0.7777339	1.0000000

The correlation matrix indicates that the number of bedrooms and bathrooms is moderately to highly correlated with the dependent variable price which is similar to the results of Dodds (2011), however, the size of the property is the highest correlated independent variable with price.

Figure 4.4 presents the pairwise correlation matrix illustrating the magnitude of correlation between the variables graphically. The visual positive correlation between price and the size, number of bedrooms and number of bathrooms variables are evident through this plot.

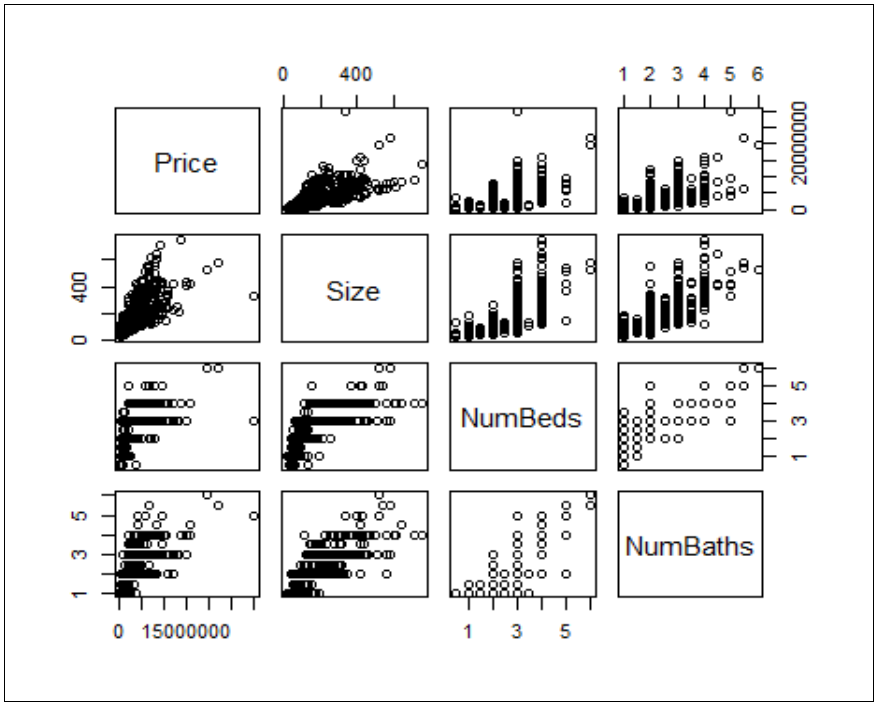


Figure 4.4: Pairwise Correlation Matrix Plot

4.4.2 Variance Inflation Factor

The Variance Inflation Factor (VIF) was computed using the usdm package in R in order to test for the presence of multi-collinearity in the independent variables which will adversely affect all the subsequent results for both the ordinary least squares model and the generalised linear model. The VIF results are presented in Table 4.6.

Table 4.6: Variance Inflation Factor Results

```
No variable from the 4 input variables has collinearity problem.

The linear correlation coefficients ranges between:
min correlation ( Suburb ~ NumBeds ):  0.1762471
max correlation ( NumBaths ~ Size ):  0.8041192

---------- VIFs of the remained variables --------
  Variables      VIF
1      Size 3.226289
2   NumBeds 2.885970
3  NumBaths 3.588363
4    Suburb 1.093962
```

Evident from the VIF results is that there was no multi-collinearity between the set of independent variables used in this study.

No multi-collinearity tests were presented in the previous studies therefore a comparison cannot be made. However, Tan (2011) used stepwise regression to mitigate the effects of multi-collinearity producing a parsimonious model.

4.5 Ordinary Least Squares Model Based on the Log-normal Distribution

4.5.1 Ordinary Least Squares Model without A Dummy Location Variable

The first approach to derive the hedonic price function for flats in the sub-markets of the study was to develop an ordinary least squares model without taking the dummy locational variable, suburb, into account. The natural logarithm was applied to the price variable. This model is named OLS_1A , Table 4.7 presents the results thereof.

Table 4.7: OLS_1A Model Output

```
Call:
lm(formula = log(Price) ~ Size + NumBaths + NumBeds, data = DATA)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74147 -0.33193 -0.00257  0.31905  1.72621

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 12.5178157  0.0394521 317.291 < 0.0000000000000002 ***
Size         0.0018179  0.0002366   7.683   0.0000000000000303 ***
NumBaths     0.4446203  0.0303460  14.652 < 0.0000000000000002 ***
NumBeds      0.2967736  0.0239544  12.389 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4894 on 1310 degrees of freedom
Multiple R-squared:  0.717,     Adjusted R-squared:  0.7163
F-statistic:  1106 on 3 and 1310 DF,  p-value: < 0.00000000000000022
```

All the coefficients are statistically significant with the variation in the set of independent variables explaining 71.63% of the variation in price as indicated by the adjusted R-squared.

Interpreting the coefficients of the OLS_1A model:

- A one unit increase in squared meters or size of a flat increases the price by approximately 0.18%.
- A one unit increase in the number of bathrooms of a flat increases the price by approximately 44.45%.
- A one unit increase in the number of bedrooms of a flat increases the price by approximately 29.68%.

To informally determine whether OLS_1A violates any of the parametric axioms, a plot of the residuals versus fitted values and normal Q-Q plot is and examined in Figure 4.5

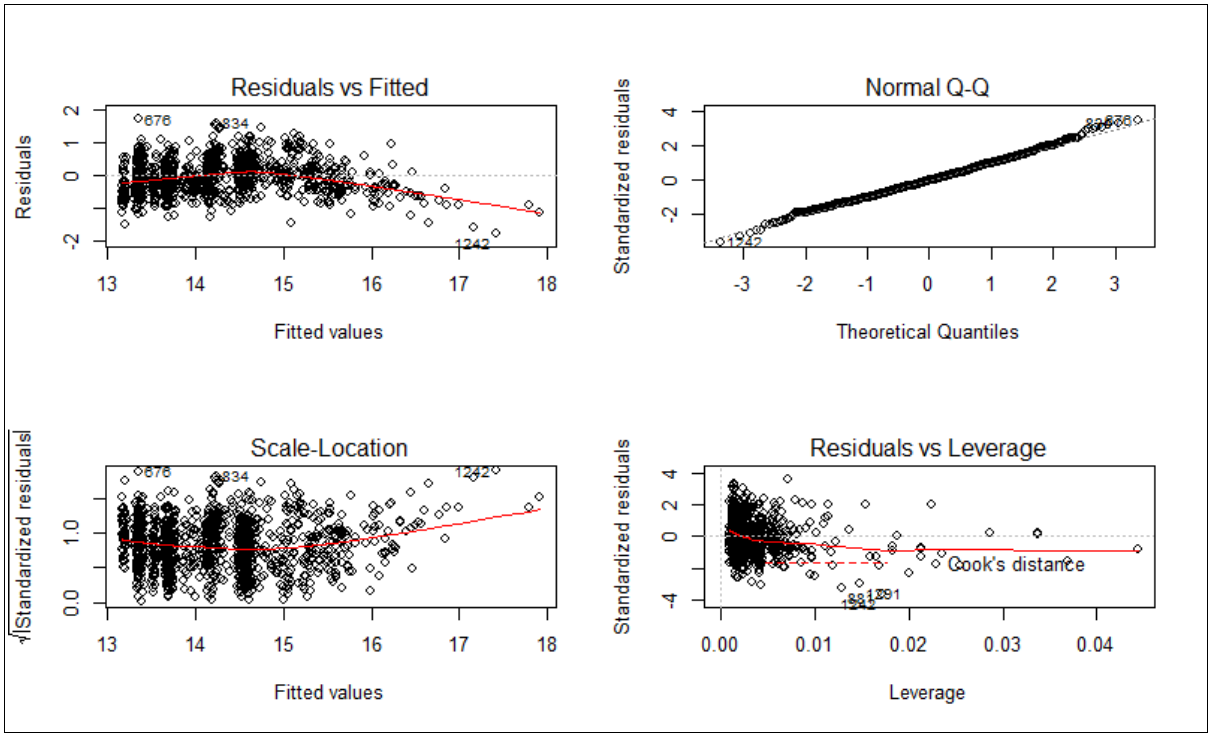


Figure 4.5: OLS_1A Diagnostic Plots

Evident from the residuals versus fitted values plot in Figure 4.4 is that the OLS_1A model suffers from heteroscedasticity, however the Q-Q plot indicates that the residuals are approximately normally distributed.

The Breusch-Pagan test is performed to formally determine if the OLS_1A model suffers from heteroscedasticity, Table 4.8 presents the results thereof.

Table 4.8: OLS_1A Breusch-Pagan Test

```
        studentized Breusch-Pagan test

data:  OLS_1A
BP = 70.363, df = 3, p-value = 0.000000000000003568
```

The p-value obtained from the Breucsh-Pagan test indicates that the null hypothesis of homoscedasticity is rejected.

A second attempt of the first ordinary least squares model is presented in Table 4.9 where the natural logarithm is applied to the size variable to mitigate the presence of heteroscedasticity. This model is named OLS_1B and presented in Table 4.9.

Table 4.9: OLS_1B Model Output

```
Call:
lm(formula = log(Price) ~ log(Size) + NumBaths + NumBeds, data = DATA)

Residuals:
     Min       1Q   Median       3Q      Max
-1.51838 -0.31192 -0.02271  0.31938  1.44728

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  9.98065    0.15001  66.534 < 0.0000000000000002 ***
log(Size)    0.72364    0.04371  16.556 < 0.0000000000000002 ***
NumBaths     0.35275    0.02725  12.947 < 0.0000000000000002 ***
NumBeds      0.09868    0.02620   3.766             0.000173 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.455 on 1310 degrees of freedom
Multiple R-squared:  0.7554,    Adjusted R-squared:  0.7548
F-statistic:  1349 on 3 and 1310 DF,  p-value: < 0.00000000000000022
```

All the coefficients are still statistically significant and interesting the adjusted R-squared has increased to 75.48% by transforming the size variable using the natural logarithm. Day

(2003) also applied the natural logarithm to the property size variable in his study where he reportedly obtained satisfactory results.

Interpreting the coefficients of the OLS_1B model:

- A 1% increase in squared meters or size of a flat increases the price by approximately 0.72%.
- A one unit increase in the number of bathrooms of a flat increases the price by approximately 35.28%.
- A one unit increase in the number of bedrooms of a flat increases the price by approximately 9.87%.

Figure 4.6 depicts the degree of heteroscedasticity has been reduced by log-transforming the size variable. The Q-Q plot indicates the residuals are approximately normally distributed and the log-transform of size has improved the normality. The residual versus leverage plot or cooks distance indicates that several observations are influential and upon investigation do appear to be outliers.

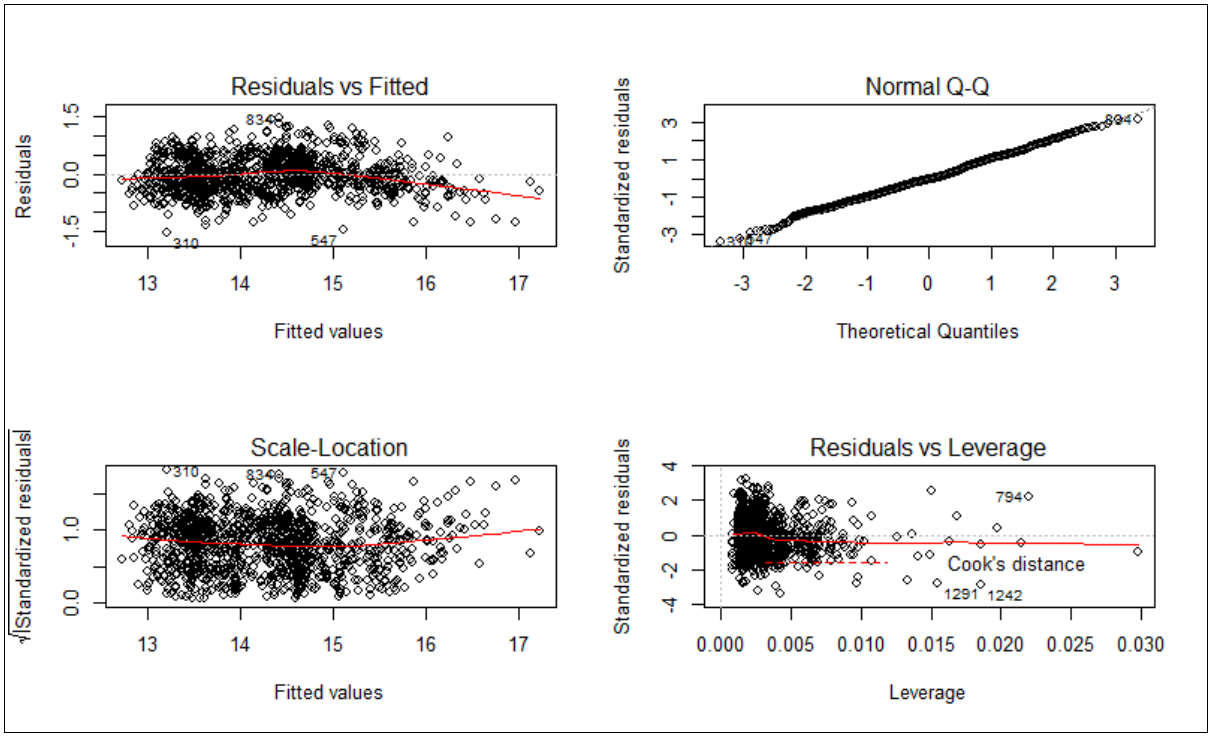


Figure 4.6: OLS_1B Diagnostic Plots

The Breusch-Pagan test is performed to formally determine if OLS_1B model suffers from heteroscedasticity, Table 4.10 presents the results thereof.

Table 4.10: OLS_1B Breusch-Pagan Test

```
        studentized Breusch-Pagan test

data:  OLS_1B
BP = 47.751, df = 3, p-value = 0.0000000002406
```

The results of the Breusch-Pagan test indicate that the null hypothesis of homoscedasticity is rejected, however, there is evidence that taking the natural of the variable size reduced the effect of heteroscedasticity as illustrated by the residual plots. Dodds (2011) reported heteroscedatsic errors in the ordinary least squares model of his study which may have been reduced if the natural logarithm was applied to the size variable in his study as experienced in this study and others.

Next the Ramsey RESET test is computed to determine whether the functional form of the OLS_1B model is misspecified. The results of the Ramsey RESET test are presented in Table 4.11.

Table 4.11: OLS_1B Ramsey RESET Test

```
        RESET test

data:  OLS_1B
RESET = 47.577, df1 = 2, df2 = 1308, p-value < 0.00000000000000022
```

The p-value is very low indicating that OLS_1B model is not correctly specified, formally the null hypothesis of correct model specification is rejected.

One method to determine whether this model, OLS_1B, exhibits spatial autocorrelation is to use an informal test which examines a plot of the residuals on the latitude and longitude co

ordinates. Should clear clustering appear, it indicates that the residuals are spatially correlated. Figure 4.7 presents a spatial residual plot of the OLS_1B model.

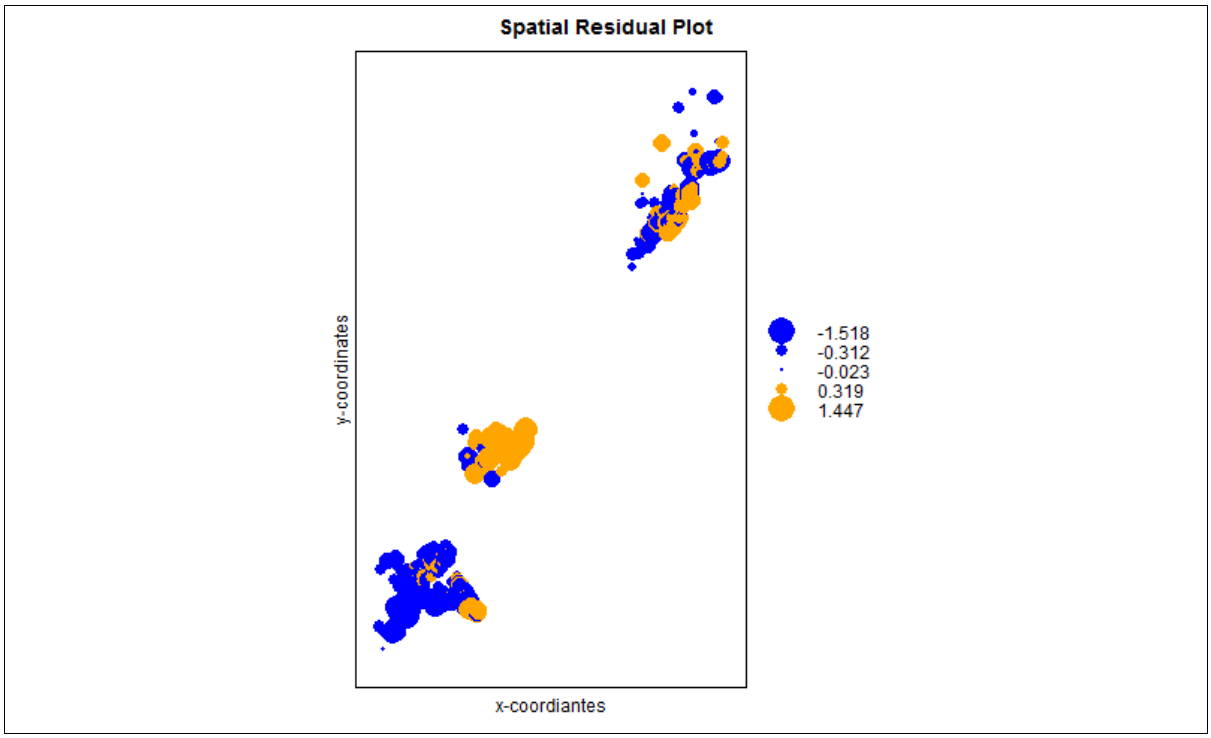


Figure 4.7: OLS_1B Spatial Residual Plot

Figure 4.6 clearly illustrates clustering of the residuals in certain areas on the plot which may mean spatial autocorrelation is a problem with this model.

In order to formally test for the presence of spatial autocorrelation the Mantel test is computed. First the distances between the flats are computed in kilometres and saved as a distance matrix based on the spherical shape of the earth using an algorithm from the fossil package in R. The distance matrix is then passed to the Mantel test function along with a matrix of the residuals from the OLS_1B model using the vegan package in R. The Mantel test then tests the null hypothesis that there is no relationship between the residuals and the distance matrix, providing a test of spatial autocorrelation. The results of the Mantel test are presented in Table 4.12.

Table 4.12: OLS_1B Mantel Test for Spatial Autocorrelation

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = earth.df, ydis = OLS_1B_Residuals)

Mantel statistic r: 0.02495
      Significance: 0.001

Upper quantiles of permutations (null model):
    90%     95%   97.5%     99%
0.00858 0.01137 0.01361 0.01551
Permutation: free
Number of permutations: 999
```

The results in Table 4.12 show that the correlation between the distance matrix and the residual matrix is 0.02495 and the significance which is the p-value in this case is 0.001, indicating the rejection of the null hypothesis of no spatial autocorrelation.

4.5.2 Ordinary Least Squares Model Including Dummy Location Variable

In order to determine if the suburb locational dummy variable is statistically significant and reduces the presence of spatial autocorrelation a second ordinary least squares model is built named OLS_2. Again the the natural logarithm was applied to the price variable and the natural logarithm was also applied to the size variable as it proved effective in reducing heteroscedasticity in the OLS_1B model. Table 4.13 presents the results of the OLS_2 model.

Table 4.13: OLS_2 Model Output

```
Call:
lm(formula = log(Price) ~ log(Size) + NumBaths + NumBeds + Suburb,
    data = DATA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91176 -0.17105 -0.01155  0.16153  0.87789

Coefficients:
                                Estimate Std. Error t value             Pr(>|t|)
(Intercept)                      9.51046    0.14890  63.871 < 0.0000000000000002 ***
log(Size)                        0.71981    0.03033  23.734 < 0.0000000000000002 ***
NumBaths                         0.18781    0.01834  10.243 < 0.0000000000000002 ***
NumBeds                          0.04699    0.01787   2.630              0.008650 **
SuburbBallito                    1.13878    0.10468  10.879 < 0.0000000000000002 ***
SuburbBeachfront                 0.49355    0.10487   4.706 0.000002800389198523 ***
SuburbBerea                      0.72421    0.10645   6.803 0.000000000015712835 ***
SuburbBrettenwood Coastal Estate 1.19867    0.22600   5.304 0.000000133476677558 ***
SuburbCarrington Heights         0.49174    0.17515   2.807              0.005069 **
SuburbCongella                  -0.46610    0.17488  -2.665              0.007792 **
SuburbDunkirk Estate             0.83645    0.14587   5.734 0.000000012234340617 ***
SuburbDurban CBD                 0.21714    0.10936   1.986              0.047296 *
SuburbEsplanade                  0.27120    0.10927   2.482              0.013199 *
SuburbEssenwood                  0.83324    0.13599   6.127 0.000000001189299623 ***
SuburbGlenwood                   0.53592    0.10659   5.028 0.000000566751052227 ***
SuburbLa Lucia                   1.32678    0.11638  11.400 < 0.0000000000000002 ***
SuburbMorningside                0.70483    0.10663   6.610 0.000000000056383555 ***
SuburbMt Edgecombe               1.04686    0.14385   7.277 0.000000000000593666 ***
SuburbMusgrave                   0.70765    0.11481   6.163 0.000000000953056153 ***
SuburbNew Town Centre Gateway    1.28501    0.11429  11.243 < 0.0000000000000002 ***
SuburbOverport                   0.40251    0.13588   2.962              0.003112 **
SuburbPalm Lakes Estate          0.78186    0.15511   5.041 0.000000530572864061 ***
SuburbPoint Waterfront           1.13809    0.10619  10.717 < 0.0000000000000002 ***
SuburbPrestondale                0.96769    0.22572   4.287 0.000019462844606262 ***
SuburbSalt Rock                  0.93442    0.13714   6.813 0.000000000014658214 ***
SuburbSeaward Estates            0.57799    0.17642   3.276              0.001081 **
SuburbShakas Rock                1.19012    0.10758  11.062 < 0.0000000000000002 ***
SuburbSheffield Beach            0.89532    0.10850   8.252 0.000000000000000385 ***
SuburbSherwood                   0.54774    0.22670   2.416              0.015827 *
SuburbSimbithi Ballito           0.89184    0.11078   8.051 0.000000000000001869 ***
SuburbSunningdale                0.78182    0.22713   3.442              0.000596 ***
SuburbSydenham                   0.39087    0.16460   2.375              0.017713 *
SuburbTinley Manor and surrounds 0.87810    0.22667   3.874              0.000113 ***
SuburbUmbilo                     0.29843    0.11910   2.506              0.012343 *
SuburbUmhlanga Ridge             1.38294    0.10986  12.588 < 0.0000000000000002 ***
SuburbUmhlanga Rocks             1.67989    0.10637  15.793 < 0.0000000000000002 ***
SuburbWestridge                  0.58612    0.14303   4.098 0.000044305254240661 ***
SuburbWindermere                 0.77772    0.16326   4.764 0.000002117206446743 ***
SuburbZimbali                    1.10995    0.12034   9.224 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2854 on 1275 degrees of freedom
Multiple R-squared:  0.9063,    Adjusted R-squared:  0.9035
F-statistic: 324.6 on 38 and 1275 DF,  p-value: < 0.00000000000000022
```

The OLS_2 model output indicates that all the coefficients are statistically significant including all the suburbs or levels of the dummy locational variable. Furthermore, the adjusted R squared has increased to 90.35% with the inclusion of the suburb dummy variable which is consistent with the findings of Bourassa, Cantoni and Hoesli (2010).

Interpreting the coefficients of the OLS_2 model:

- A 1% in increase in squared meters or size of flats increases the price by approximately 0.72%.
- A one unit increase in the number of bathrooms of a flat increases the price by approximately 18.78%.
- A one unit increase in the number of bedrooms of a flat increases the price by approximately 4.7%.
- Each suburb coefficient is the percentage difference between the reference suburb.

Interesting the sign for all the coefficients is as expected based on *a priori* expectations unlike what was experienced in the study of Dodds (2011). The number of bedrooms is statistically significant which was not the case in the study of Els and Von Fintel, (2010). The OLS_2 model results show that there are 35 suburb coefficients, yet there are a total of 36 in the sample data. This is because one of the suburbs is withheld from the model output which all the other suburbs (levels) are compared to. The suburb that has been withheld is Albert Park, a suburb in Durban Central. Interpreting the suburb coefficients presented in the results of model the will always be in comparison with the Albert Park suburb. For example flats located in the suburb of Berea are 205% (100*exp(0.72421)-1) more expensive than flats located in Albert Park.

Figure 4.8 presents the diagnostic plots of the OLS_2 model. It is evident that the presence of heteroscedasticity has decreased and the residuals appear homoscedastic. However, the residuals seem to be less normally distributed than before where a deviation from normality is pronounced at the top and lower quantiles.

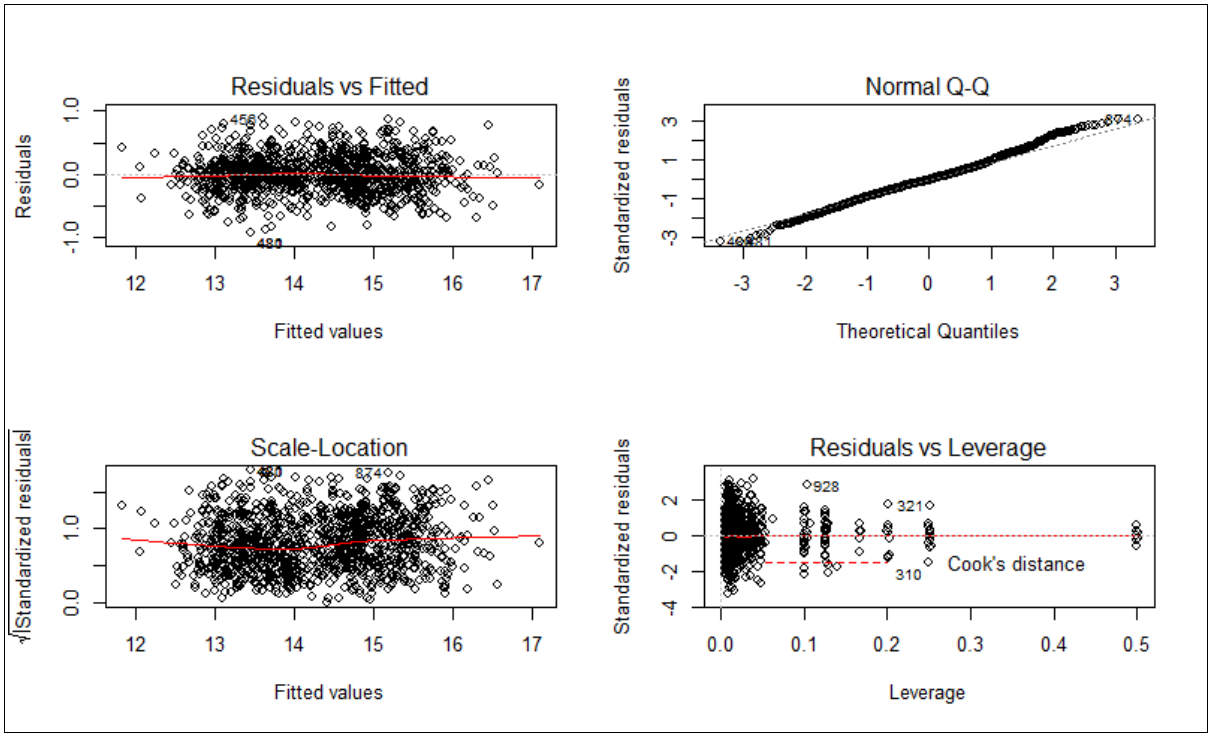


Figure 4.8: OLS_2 Diagnostic Plots

The Breusch-Pagan test is performed to formally determine if the OLS_2 model suffers from heteroscedasticity, Table 4.14 presents the results thereof.

Table 4.14: OLS_2 Breusch-Pagan Test

```
        studentized Breusch-Pagan test

data:  OLS_2
BP = 254.41, df = 38, p-value < 0.00000000000000022
```

The results of the Breusch-Pagan test indicate that the null hypothesis of homoscedasticity is rejected with the p-value lower than the OLS_1B model, however, the Breusch-Pagan test is sensitive to normality of residuals which may have had an effect on this result. Evident from the Q-Q plot in Figure 4.7 is a deviation from normality which is manifesting in this test.

In order to mitigate any effects of heteroscedasticity the Whites (1980) heteroscedasticity consistent covariance matrix can be employed. In the presence of heteroscedasticity Whites

(1980) asymptotic covariance matrix estimator can provide consistent results. Using Whites (1980) HCCM will provide insight into whether the coefficients in the OLS_2 model are indeed significant. The HCCM is run in R using the lmtest package. The results are presented in Table 4.15.

Table 4.15: OLS_2 HCCM

t test of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.510461	0.137205	69.3159	< 0.00000000000000022	***
log(Size)	0.719810	0.032820	21.9322	< 0.00000000000000022	***
NumBaths	0.187806	0.020563	9.1332	< 0.00000000000000022	***
NumBeds	0.046985	0.019442	2.4166	0.0158044	*
SuburbBallito	1.138776	0.071867	15.8456	< 0.00000000000000022	***
SuburbBeachfront	0.493548	0.079732	6.1900	0.00000000080936689	***
SuburbBerea	0.724207	0.073487	9.8549	< 0.00000000000000022	***
SuburbBrettenwood Coastal Estate	1.198673	0.091494	13.1011	< 0.00000000000000022	***
SuburbCarrington Heights	0.491738	0.072039	6.8260	0.00000000001346438	***
SuburbCongella	-0.466099	0.209391	-2.2260	0.0261904	*
SuburbDunkirk Estate	0.836447	0.107055	7.8132	0.00000000000001158	***
SuburbDurban CBD	0.217138	0.075118	2.8906	0.0039098	**
SuburbEsplanade	0.271196	0.070357	3.8546	0.0001217	***
SuburbEssenwood	0.833236	0.083615	9.9651	< 0.00000000000000022	***
SuburbGlenwood	0.535915	0.071203	7.5266	0.00000000000009807	***
SuburbLa Lucia	1.326779	0.087400	15.1806	< 0.00000000000000022	***
SuburbMorningside	0.704834	0.071325	9.8820	< 0.00000000000000022	***
SuburbMt Edgecombe	1.046860	0.166287	6.2955	0.00000000042065062	***
SuburbMusgrave	0.707652	0.082760	8.5506	< 0.00000000000000022	***
SuburbNew Town Centre Gateway	1.285007	0.075014	17.1303	< 0.00000000000000022	***
SuburbOverport	0.402508	0.105426	3.8179	0.0001410	***
SuburbPalm Lakes Estate	0.781860	0.092004	8.4981	< 0.00000000000000022	***
SuburbPoint Waterfront	1.138087	0.072989	15.5926	< 0.00000000000000022	***
SuburbPrestondale	0.967693	0.079673	12.1458	< 0.00000000000000022	***
SuburbSalt Rock	0.934416	0.148839	6.2781	0.00000000046908061	***
SuburbSeaward Estates	0.577986	0.110293	5.2404	0.00000018728460519	***
SuburbShakas Rock	1.190117	0.074036	16.0749	< 0.00000000000000022	***
SuburbSheffield Beach	0.895321	0.072133	12.4120	< 0.00000000000000022	***
SuburbSherwood	0.547739	0.170407	3.2143	0.0013404	**
SuburbSimbithi Ballito	0.891843	0.075067	11.8806	< 0.00000000000000022	***
SuburbSunningdale	0.781820	0.081056	9.6454	< 0.00000000000000022	***
SuburbSydenham	0.390872	0.172831	2.2616	0.0238905	*
SuburbTinley Manor and surrounds	0.878103	0.069342	12.6634	< 0.00000000000000022	***
SuburbUmbilo	0.298428	0.091592	3.2582	0.0011507	**
SuburbUmhlanga Ridge	1.382945	0.075027	18.4326	< 0.00000000000000022	***
SuburbUmhlanga Rocks	1.679892	0.077390	21.7069	< 0.00000000000000022	***
SuburbWestridge	0.586120	0.091690	6.3924	0.00000000022851560	***
SuburbWindermere	0.777721	0.107694	7.2216	0.00000000000088081	***
SuburbZimbali	1.109949	0.094234	11.7787	< 0.00000000000000022	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The HCCM results show that all the coefficients have remained unchanged which is as expected. Furthermore, all the coefficients are still statistically significant.

As asserted by White (1980) and Long and Ervin (2000), if a model fails tests for homoscedasticity the HCCM produces consistent results.

The Ramsey RESET test is computed to determine whether the functional from of OLS_2 is misspecified. The results of the Ramsey RESET test are presented in Table 4.16.

Table 4.16: OLS_2 Ramsey RESET Test

```
        RESET test

data:  OLS_2
RESET = 2.4324, df1 = 2, df2 = 1273, p-value = 0.08823
```

The p-value is above 0.05 indicating that the OLS_2 model is correctly specified. The null hypothesis of correct model specification is not rejected.

The addition of the suburb dummy locational variable facilitated the correct model specification which is congruent to the results of the Els and Von Fintel, (2010) .

The spatial residual plot is presented in Figure 4.9 where it is evident that the inclusion of the dummy locational variable has reduced the clustering of the residual thereby mitigating the effects of spatial autocorrelation.

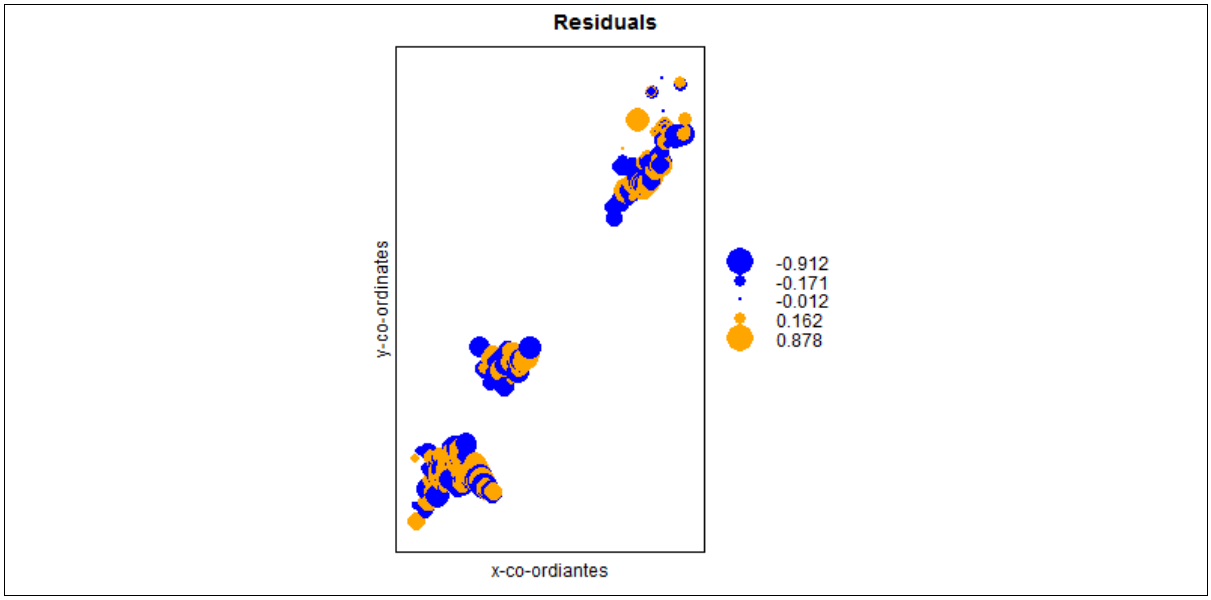


Figure: 4.9 OLS_2 Spatial Residual Plot

The results of the Mantel test are presented in Table 4.17. The correlation between the distance matrix and the residual matrix is -0.02711 and the significance (p-value) of 1 means there is sufficient evidence not to reject the null hypothesis of no spatial autocorrelation.

Table 4.17: OLS_2 Mantel Test for Spatial Autocorrelation

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = earth.df, ydis = OLS_2_Residuals)

Mantel statistic r: -0.02711
      Significance: 1

Upper quantiles of permutations (null model):
    90%     95%   97.5%     99%
0.00876 0.01165 0.01408 0.01706
Permutation: free
Number of permutations: 999
```

The use of the Mantel test to detect the presence of spatial autocorrelation in this study is consistent with the assertion of Borcard and Legendre (2012) and Diniz-Filho *et al.*, 2013, where the test adequately detects spatial autocorrelation augmented by the use of the spatial residual plots.

The results of adding the suburb locational dummy variable to mitigate the effects of spatial autocorrelation is consistent with the findings of Bourassa, Cantoni and Hoesli (2010) and Els and Von Fintel (2010).

4.5.3 Bootstrapping the Final Ordinary Least Squares Model

To test the reliability of the results, the OLS_2 model is bootstrapped. The bootstrapping procedure involved taking 5000 samples with each sample using sampling with replacement and deriving 5000 models. The bootstrapped results are derived by taking the average of the 5000 models. The bootstrapped model significantly accounts for variance in the estimates of parameters and can be used as a general tool for assessing statistical accuracy. The results are presented in Table 4.18.

Table 4.18: Bootstrapped Ordinary Least Squares_2 Model

	R	original	bootBias	bootSE	bootMed	bootSkew	bootKurtosis
(Intercept)	2300	9.510461	0.001022187	0.134381	9.511676	-0.029157	0.016553
log(Size)	2300	0.719810	-0.000797611	0.032236	0.718867	-0.027924	-0.100647
NumBaths	2300	0.187806	0.002075178	0.019640	0.190185	0.024119	-0.147927
NumBeds	2300	0.046985	-0.000303655	0.018457	0.046218	0.040518	-0.041282
SuburbBallito	2300	1.138776	-0.000292360	0.067808	1.137272	0.110034	0.383124
SuburbBeachfront	2300	0.493548	0.001792537	0.078135	0.494489	0.161465	0.124432
SuburbBerea	2300	0.724207	-0.000275424	0.068883	0.724310	0.098115	0.151797
SuburbBrettenwood Coastal Estate	2300	1.198673	-0.000231440	0.072081	1.195768	0.177484	0.350932
SuburbCarrington Heights	2300	0.491738	0.001389994	0.067290	0.490978	0.111160	0.351808
SuburbCongella	2300	-0.466099	0.003122337	0.183399	-0.466139	0.134997	-0.050683
SuburbDunkirk Estate	2300	0.836447	0.000755326	0.105349	0.842673	-0.435062	1.106471
SuburbDurban CBD	2300	0.217138	0.000750046	0.072683	0.217235	0.116195	0.081472
SuburbEsplanade	2300	0.271196	0.001512223	0.067391	0.273271	0.075430	0.349617
SuburbEssenwood	2300	0.833236	0.001952745	0.079455	0.833489	0.147106	0.067840
SuburbGlenwood	2300	0.535915	0.001407562	0.067605	0.536511	0.202500	0.499573
SuburbLa Lucia	2300	1.326779	-0.003523863	0.083160	1.323338	0.111143	0.212480
SuburbMorningside	2300	0.704834	0.001389169	0.067184	0.703898	0.126043	0.207659
SuburbMt Edgecombe	2300	1.046860	-0.003345031	0.153593	1.041710	-0.069121	0.173681
SuburbMusgrave	2300	0.707652	0.000551596	0.079293	0.708182	0.140142	0.327871
SuburbNew Town Centre Gateway	2300	1.285007	-0.000139914	0.072812	1.283082	0.253393	0.346992
SuburbOverport	2300	0.402508	0.000517652	0.101167	0.399703	0.221030	0.076674
SuburbPalm Lakes Estate	2300	0.781860	-0.002764051	0.087890	0.780780	-0.050024	0.344150
SuburbPoint Waterfront	2300	1.138087	-0.001458903	0.069440	1.136613	0.071900	0.181674
SuburbPrestondale	2300	0.967693	0.000680755	0.067148	0.968069	0.106285	0.259756
SuburbSalt Rock	2300	0.934416	0.001930000	0.140418	0.935572	0.055919	0.298848
SuburbSeaward Estates	2300	0.577986	-0.002718614	0.098201	0.573284	0.118467	-0.130314
SuburbShakas Rock	2300	1.190117	-0.001015857	0.070531	1.189000	0.130417	0.258875
SuburbSheffield Beach	2300	0.895321	0.000084518	0.068755	0.893732	0.202258	0.335839
SuburbSherwood	2300	0.547739	-0.000557509	0.105363	0.544562	0.077010	-0.418682
SuburbSimbithi Ballito	2300	0.891843	-0.001209389	0.071888	0.889225	0.148495	0.309770
SuburbSunningdale	2300	0.781820	-0.000843489	0.071945	0.779937	0.124720	0.285690
SuburbSydenham	2300	0.390872	0.001388950	0.156097	0.387529	0.293617	0.503515
SuburbTinley Manor and surrounds	2300	0.878103	-0.000373065	0.065941	0.875120	0.191872	0.338588
SuburbUmbilo	2300	0.298428	0.000184620	0.087401	0.298024	0.025672	0.149108
SuburbUmhlanga Ridge	2300	1.382945	-0.001253693	0.072237	1.380605	0.129806	0.193353
SuburbUmhlanga Rocks	2300	1.679892	-0.000814253	0.073456	1.677501	0.084196	0.258542
SuburbWestridge	2300	0.586120	0.002670067	0.087198	0.590583	-0.016832	0.249445
SuburbWindermere	2300	0.777721	0.001869987	0.099148	0.783074	-0.324631	0.594153
SuburbZimbali	2300	1.109949	-0.001687758	0.090067	1.106463	0.025086	0.287835

The results indicate that bootstrapped coefficients are similar to the OLS_2 model coefficients which is evident by comparing the column headed “original” with the column headed “bootMed”. These results coincide with the views of Carruthers, *et al* (2008) and Hastie, Tibshirani and Friedman (2005) where the bootstrap methodology can be used as a general tool for assessing statistical accuracy.

4.5.4 AIC and Wald Tests of the OLS_1B and OLS_2 Models

4.5.4.1 AIC

The AIC scores for the respective ordinary least squares models are presented in Table 4.19.

Table 4.19: Ordinary Least Squares AIC Scores

	df	AIC
OLS_1B	5	1665.470
OLS_2	40	474.251

The AIC indicates that the OLS_2 model which includes the dummy variable is preferred as it is lower.

4.5.4.2 Wald Test

The results from the Walt test are presented in Table 4.20.

Table 4.20: Ordinary Least Squares Walt Test

```
Wald test

Model 1: log(Price) ~ log(Size) + NumBaths + NumBeds
Model 2: log(Price) ~ log(Size) + NumBaths + NumBeds + Suburb
  Res.Df Df      F                Pr(>F)
1   1310
2   1275 35 58.696 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Wald test indicates that the dummy location variable is statistically significant and contributes to the OLS_2 model.

Evidently transforming the size variable using the natural logarithm and including the dummy locational variable improved the overall fit of the model and removed the spatial autocorrelation that was present. However, the price variable follows a gamma distribution therefore a new model is built, namely, a generalised linear model using a gamma distribution and log- link function.

4.6 Generalised Linear Model Based on the Gamma Distribution and Log-link Function

The second approach to derive the hedonic price function for flats in the three sub-markets of the study was to develop a generalised linear model based on the gamma distribution and log-link function.

4.6.1 Generalised Linear Model Output and Hypotheses Results

Upon developing the generalised linear model based on the gamma distribution and log-link function, the natural logarithm was applied to the size variable due to its effectiveness in the OLS_2 model and the dummy locational variable, suburb, was included for this reason too. Table 4.21 presents the results of the generalised linear model.

Table 4.21: Generalised Linear Model Output

```
Call:
glm(formula = Price ~ log(Size) + NumBaths + NumBeds + Suburb,
    family = Gamma(link = "log"), data = DATA)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.87193  -0.20102  -0.04121   0.13275   0.98862

Coefficients:
                                Estimate Std. Error t value             Pr(>|t|)
(Intercept)                      9.46220    0.15252  62.038 < 0.0000000000000002 ***
log(Size)                        0.73296    0.03107  23.593 < 0.0000000000000002 ***
NumBaths                         0.20298    0.01878  10.808 < 0.0000000000000002 ***
NumBeds                          0.03851    0.01830   2.104             0.035548 *
SuburbBallito                    1.16161    0.10722  10.833 < 0.0000000000000002 ***
SuburbBeachfront                 0.58267    0.10742   5.424  0.00000006967844448 ***
SuburbBerea                      0.74698    0.10904   6.850  0.00000000001143516 ***
SuburbBrettenwood Coastal Estate 1.17038    0.23149   5.056  0.00000049122724858 ***
SuburbCarrington Heights         0.48521    0.17941   2.704             0.006933 **
SuburbCongella                  -0.43305    0.17914  -2.417             0.015772 *
SuburbDunkirk Estate             0.80940    0.14942   5.417  0.00000007245484070 ***
SuburbDurban CBD                 0.23417    0.11202   2.090             0.036776 *
SuburbEsplanade                  0.27414    0.11193   2.449             0.014449 *
SuburbEssenwood                  0.83692    0.13930   6.008  0.00000000244357992 ***
SuburbGlenwood                   0.54902    0.10918   5.029  0.00000056482204129 ***
SuburbLa Lucia                   1.33553    0.11921  11.203 < 0.0000000000000002 ***
SuburbMorningside                0.71846    0.10923   6.578  0.00000000006955161 ***
SuburbMt Edgecombe               1.07865    0.14735   7.320  0.00000000000043708 ***
SuburbMusgrave                   0.71563    0.11761   6.085  0.00000000153903826 ***
SuburbNew Town Centre Gateway    1.28471    0.11707  10.973 < 0.0000000000000002 ***
SuburbOverport                   0.41942    0.13919   3.013             0.002635 **
SuburbPalm Lakes Estate          0.76619    0.15888   4.822  0.00000158922126327 ***
SuburbPoint Waterfront           1.14713    0.10878  10.546 < 0.0000000000000002 ***
SuburbPrestondale                0.95626    0.23121   4.136  0.00003767351592657 ***
SuburbSalt Rock                  0.98652    0.14048   7.022  0.00000000000353592 ***
SuburbSeaward Estates            0.55325    0.18071   3.061             0.002249 **
SuburbShakas Rock                1.19157    0.11020  10.813 < 0.0000000000000002 ***
SuburbSheffield Beach            0.89679    0.11114   8.069  0.00000000000000162 ***
SuburbSherwood                   0.53323    0.23222   2.296             0.021822 *
SuburbSimbithi Ballito           0.86659    0.11347   7.637  0.00000000000004336 ***
SuburbSunningdale                0.75334    0.23265   3.238             0.001235 **
SuburbSydenham                   0.43097    0.16861   2.556             0.010700 *
SuburbTinley Manor and surrounds 0.85732    0.23219   3.692             0.000232 ***
SuburbUmbilo                     0.32250    0.12199   2.644             0.008305 **
SuburbUmhlanga Ridge             1.38917    0.11254  12.344 < 0.0000000000000002 ***
SuburbUmhlanga Rocks             1.71270    0.10895  15.719 < 0.0000000000000002 ***
SuburbWestridge                  0.58978    0.14650   4.026  0.00006016838359940 ***
SuburbWindermere                 0.77823    0.16723   4.654  0.00000359969880730 ***
SuburbZimbali                    1.09251    0.12327   8.863 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.08546552)

    Null deviance: 1091.47  on 1313  degrees of freedom
Residual deviance:  104.44  on 1275  degrees of freedom
AIC: 38026

Number of Fisher Scoring iterations: 5
```

Table 4.21 shows that all the coefficients are statistically significant including all the levels of the dummy locational variable or suburbs. Moreover, the residual deviance is less than the

residual degrees of freedom indicating the model is not misspecified and that the saturated model which includes all the coefficients is better than the null model which includes only the intercept.

Interpreting the coefficients of the generalised linear model:

- A 1% increase in squared meters or size of a flat increases the price by approximately 0.733%.
- A one unit increase in the number of bathrooms of a flat increases the price by approximately 20.3%.
- A one unit increase in the number of bedrooms of a flat increases the price by approximately 3.85%.
- Each suburb coefficient is the percentage difference between the reference suburb.

The generalised linear model results also show that there are 35 suburb coefficients. The same suburb has been withheld as in the case of the OLS_2 model.

The results are not dissimilar to the results obtained from the OLS_2 model which included the suburb locational dummy variable. This is consistent with the assertion of Bromideh and Valizadeh (2013) where they postulate that similarities exist between log-normal and gamma exponential distributions in terms of fit on moderate data sizes.

The first plot in Figure 4.10 is the jackknife deviance residuals against the fitted values, indicating homogeneous variance and no curvilinear pattern of the deviance residuals. The second plot is a Q-Q plot of the standardized deviance residuals indicating normality thereof. The bottom two plots relate to cooks distance, indicating there are observations with high leverage or influence, however, upon inspection of these data points no clear outliers are evident indicating that these point may have high degrees of leverage or influence on the model.

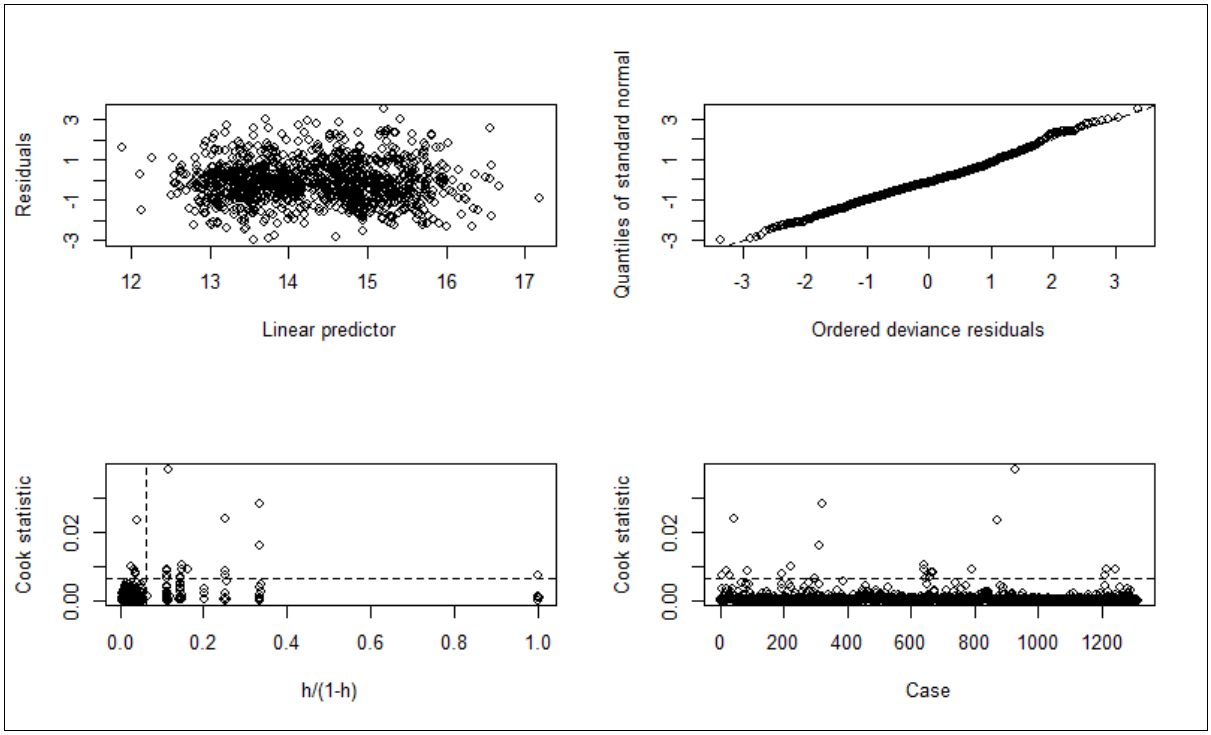


Figure 4.10: Generalised Linear Model Residual Plot

McCallullagh and Nelder (1989) assert that the standardized deviance residuals plotted against the fitted values should resemble a normal theory residual plot. This is evident Figure 10 where the standardized residuals appear normal with no curvilinear pattern. Therefore the based on the informal tests in Figure 10, the model appears correctly specified in terms of the parametric assumptions.

This model rejects the following hypotheses of the study:

> ***H0b****:* Floor area is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
>
> ***H0c****:* Number of bedrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0d*:** Number of bathrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0e**:* A dummy locational variable is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

The spatial residual plot for the generalised linear model is presented in Figure 4.11 where it is evident that clustering of the residuals is not pronounced.

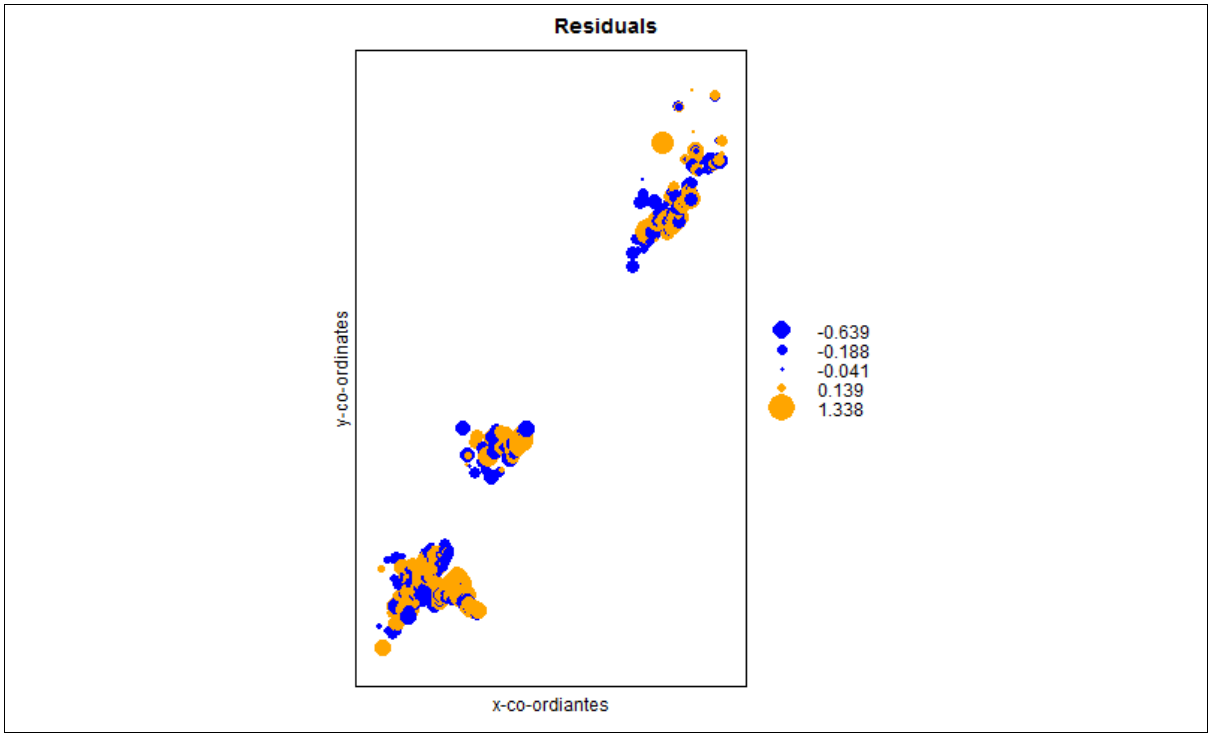


Figure 4.11: Generalised Linear Model Spatial Residual Plot

The results of the Mantel test for the generalised linear model are presented in Table 4.23. The correlation between the distance matrix and the residual matrix is -0.02411 and the significance (p-value) is 1, providing sufficient evidence not to reject the null hypothesis of no spatial autocorrelation.

Table 4.23: Generalised Linear model Mantel Test for Spatial Autocorrelation

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = earth.df, ydis = GLM_Residuals)

Mantel statistic r: -0.02441
      Significance: 1

Upper quantiles of permutations (null model):
    90%     95%   97.5%     99%
0.00936 0.01184 0.01432 0.01656
Permutation: free
Number of permutations: 999
```

These results indicate that there is no presence of spatial autocorrelation which would affect the standard errors and p-values of the model.

The bootstrapping technique used for the generalised linear model is similar to the ordinary least squares where 5000 samples were used with each of the samples using sampling with replacement and the average of the 5000 models computed was taken. The results are presented in Table 4.24.

4.6.2 Bootstrapping the Generalised Linear Model

The results indicate that bootstrapped coefficients are extremely similar to the generalised linear model coefficients which is evident by comparing the column headed “original” with the column headed “bootMed” in Table 4.24. Through simulating many random sampling distributions where increased variance is accounted for to test for reliability against the generalized linear models parametric assumptions, the results are consistent and represent an accurate reflection of the data.

Table 4.24: Bootstrapped generalised linear Model

	R	original	bootBias	bootSE	bootMed	bootSkew	bootKurtosis
(Intercept)	2194	9.462201	-0.008291175	0.139005	9.449233	0.0052818	0.056513
log(Size)	2194	0.732960	0.001169352	0.033635	0.734793	-0.0500038	0.031915
NumBaths	2194	0.202984	0.000409610	0.021085	0.203209	0.0350099	-0.170459
NumBeds	2194	0.038513	-0.000670899	0.019742	0.038008	0.0231656	-0.086310
SuburbBallito	2194	1.161614	0.003347619	0.068875	1.164911	0.1122976	0.543795
SuburbBeachfront	2194	0.582668	0.002622434	0.076079	0.584300	0.1251940	0.416777
SuburbBerea	2194	0.746979	0.003865832	0.069661	0.749166	0.1680494	0.400954
SuburbBrettenwood Coastal Estate	2194	1.170375	0.002221222	0.071806	1.172913	0.0014494	0.171672
SuburbCarrington Heights	2194	0.485213	0.003705877	0.067818	0.490015	0.0679586	0.604251
SuburbCongella	2194	-0.433046	-0.013349784	0.190698	-0.437814	-0.1058730	-0.146001
SuburbDunkirk Estate	2194	0.809397	-0.000937352	0.099751	0.813618	-0.4495154	1.141019
SuburbDurban CBD	2194	0.234167	0.002723281	0.072944	0.237637	0.0048971	0.438575
SuburbEsplanade	2194	0.274145	0.002687679	0.067512	0.276546	0.0848500	0.595012
SuburbEssenwood	2194	0.836923	0.001812316	0.080513	0.835307	0.2792094	0.593904
SuburbGlenwood	2194	0.549024	0.003786160	0.068967	0.553202	0.0449517	0.710317
SuburbLa Lucia	2194	1.335534	0.003588477	0.088316	1.341386	-0.0160357	0.095600
SuburbMorningside	2194	0.718460	0.003579625	0.068544	0.721921	0.1681599	0.487340
SuburbMt Edgecombe	2194	1.078650	0.001625004	0.147574	1.085874	-0.4527827	0.833344
SuburbMusgrave	2194	0.715626	0.000504024	0.081211	0.717503	0.0498025	0.150944
SuburbNew Town Centre Gateway	2194	1.284705	0.003353466	0.072259	1.287007	0.0984200	0.305774
SuburbOverport	2194	0.419416	0.001347456	0.101720	0.423324	-0.0787322	0.322682
SuburbPalm Lakes Estate	2194	0.766187	0.001666709	0.087186	0.769822	-0.1710463	0.521268
SuburbPoint Waterfront	2194	1.147131	0.003243269	0.069162	1.152322	-0.0049976	0.646309
SuburbPrestondale	2194	0.956261	0.003711508	0.068127	0.957237	0.1088961	0.509912
SuburbSalt Rock	2194	0.986517	-0.004324199	0.152297	0.987189	-0.0282274	0.074519
SuburbSeaward Estates	2194	0.553246	-0.000642886	0.099687	0.554347	-0.0453956	0.048786
SuburbShakas Rock	2194	1.191571	0.002634724	0.070973	1.192938	0.1090303	0.445877
SuburbSheffield Beach	2194	0.896794	0.003826300	0.069548	0.901282	0.1155085	0.471040
SuburbSherwood	2194	0.533233	0.001232588	0.105212	0.533948	0.0455417	-0.539622
SuburbSimbithi Ballito	2194	0.866593	0.002369934	0.072558	0.869840	0.0447431	0.364322
SuburbSunningdale	2194	0.753337	0.004092389	0.073277	0.758305	0.0191157	0.380495
SuburbSydenham	2194	0.430975	-0.003935970	0.162656	0.430606	0.0401224	0.028909
SuburbTinley Manor and surrounds	2194	0.857316	0.003783832	0.066366	0.861646	0.1004432	0.535554
SuburbUmbilo	2194	0.322495	0.001902980	0.083675	0.324751	0.0367692	0.099604
SuburbUmhlanga Ridge	2194	1.389172	0.002586008	0.071693	1.392348	-0.0371497	0.461434
SuburbUmhlanga Rocks	2194	1.712698	0.002458290	0.074615	1.715394	-0.0223394	0.312240
SuburbWestridge	2194	0.589776	0.002838880	0.086824	0.594515	-0.0940166	0.279275
SuburbWindermere	2194	0.778229	-0.001156938	0.097210	0.782636	-0.3147873	0.529745
SuburbZimbali	2194	1.092508	0.000032188	0.099351	1.091839	0.1042661	0.092194

4.7 Comparison and Testing of the Models

In order to assess and compare the accuracy of the ordinary least squares and generalised linear models concurrently, the RMSE is computed by determining the dispersement of the fitted values for the ordinary least squares model and generalised linear model against the observed values. The RMSE for each model is presented in table Table 4.25.

Table 4.25: Root Mean Squared Error Model Comparison

RMSE_OLS_2	RMSE_GLM_Mod
1097749	1089502

The RMSE is computed on the original scale therefore the anti-log of the OLS_2 fitted values were taken, in order to ensure a direct model comparison. Interestingly the RMSE is lower

for the generalised linear model, indicating the dispersion of the models fitted values are lower, suggesting the generalised linear model is a better fit for the data.

To illustrate the accuracy of the two models, a random sample of 15 rows is generated where each models fitted price values are compared to the observed price. Interestingly both models produce similar fitted value estimates which is evident in Table 4.26.

Table 4.26: Ordinary Least Squares and Generalised Linear Models Comparison

	Observed Price	OLS Fitted Price	GLM Fitted Price
42	725000	650000	660000
97	1050000	1020000	1060000
126	845000	740000	760000
129	1035000	1070000	1100000
181	6500000	6330000	6840000
257	750000	750000	770000
350	450000	420000	430000
479	350000	300000	310000
632	2950000	3000000	3110000
683	980000	970000	980000
747	1950000	2010000	2110000
924	5950000	5810000	6000000
905	2700000	2500000	2620000
1145	3200000	3210000	3340000
1314	950000	870000	880000

The first column in Table 4.26 is the row selected from the 1314 listed flats in the sample dataset via random sampling. All estimates have been rounded. Notably the generalised linear model consistently produces higher values than the ordinary least squares model because the ordinary least squares produces geometric mean estimates whilst the generalised linear model produces arithmetic mean estimates (Morgan et al., 2006; Baum, 2013). Although the results are similar between the two models, the generalised linear model performs slightly better in this randomly selected sample producing 8 closer estimates to the the observed prices.

4.8 Software Application

The final objective of this study was to build a software application that facilitates the estimation of listing prices for flats in KwaZulu-Natal coastal sub-markets given a set of structural and locational attributes. A software application was built using the shiny package in R to render the results of the generalised linear model through a simple user interface. The

generalised linear model was selected as the final model as all the hypotheses and parametric axioms were met using this technique. The lack of model violations as examined by the diagnostic plots and the similarity of the bootstrapped standard errors warranted the use of the generalised linear model for the software application.

The software application calls the generalised linear model once a user has selected a set of attributes. The generalised linear model then calculates the expected price or arithmetic mean listing price and the confidence interval, rendering it on the user interface. Three examples of the software application are presented, one for a suburb within each sub-market using the same input parameters to gauge price differences between flats in different suburbs within different sub-markets. Note all figures are rounded to the nearest thousand.

The first example presented in Figure 4.12 details the average price for a flat in Morningside within the Durban Central sub-market that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms. The confidence interval is also computed which provides the statistical inference of the point estimate of a flat given the selected attributes.

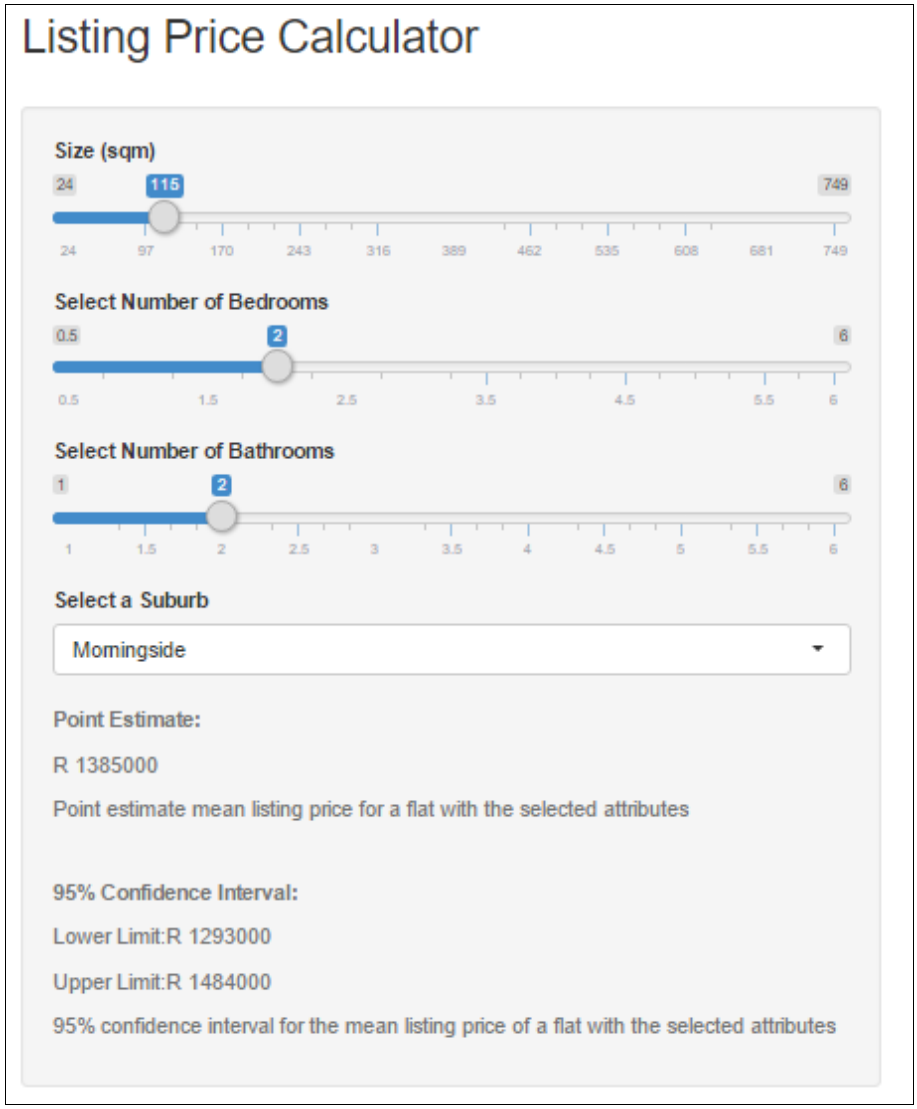


Figure 4.12: Software Application Example 1

The point estimate for a flat in Morning that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms is R1 385 000. The lower and upper limits of the 95% confidence interval are R1 293 000 and R1 484 000 respectively, which means that the true average listing price for a flat with the user selected attributed falls within this range.

The second example presented in Figure 4.13 details the average price for a flat in Umhlanga Rocks within the Umhlanga sub-market that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms. The confidence interval is also computed.

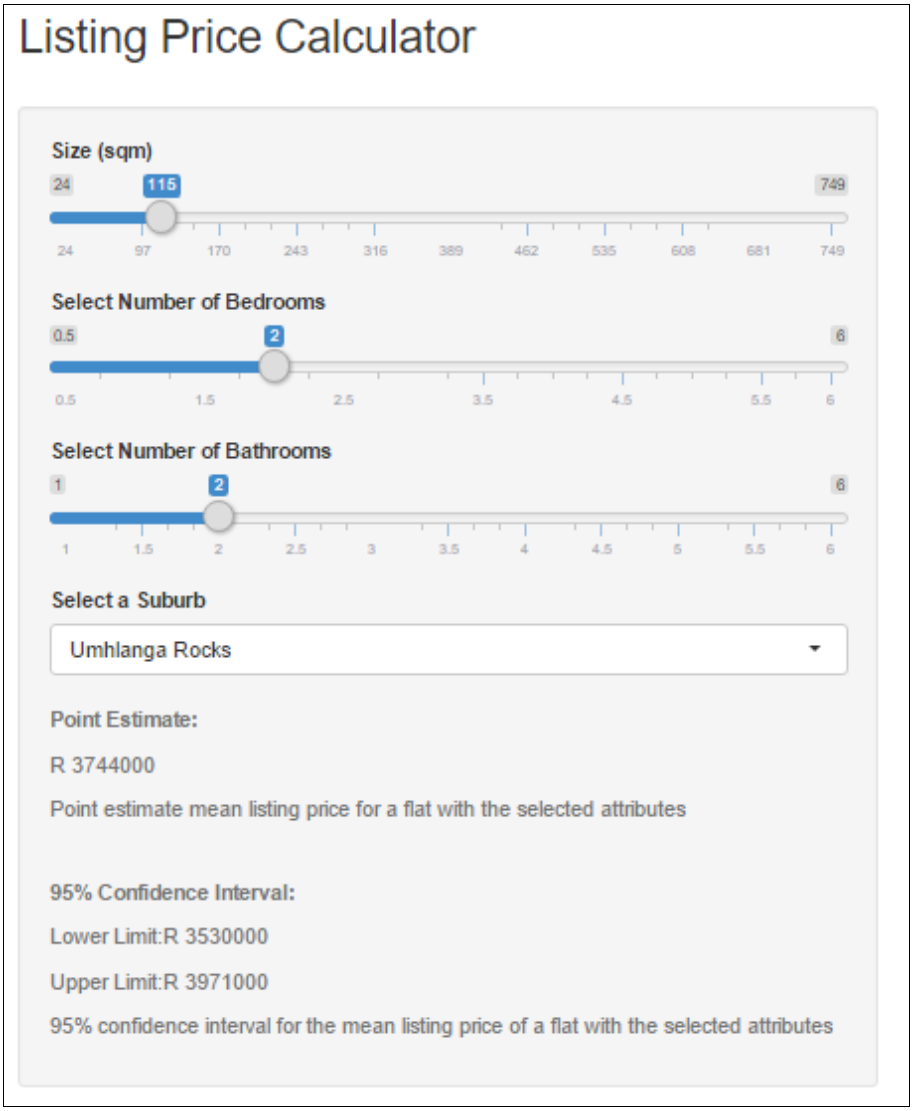


Figure 4.13: Software Application Example 2

The point estimate for a flat in Umhlanga Rocks that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms is R3 744 000. The lower and upper limits of the 95% confidence interval are R3 530 000 and R3 971 000 respectively, which means that the true average listing price for a flat with the user selected attributed falls within this range.

The third example details presented in Figure 4.14 details the average price for a flat in Simbithi Ballito within the Ballito sub-market that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms. The confidence interval is also computed.

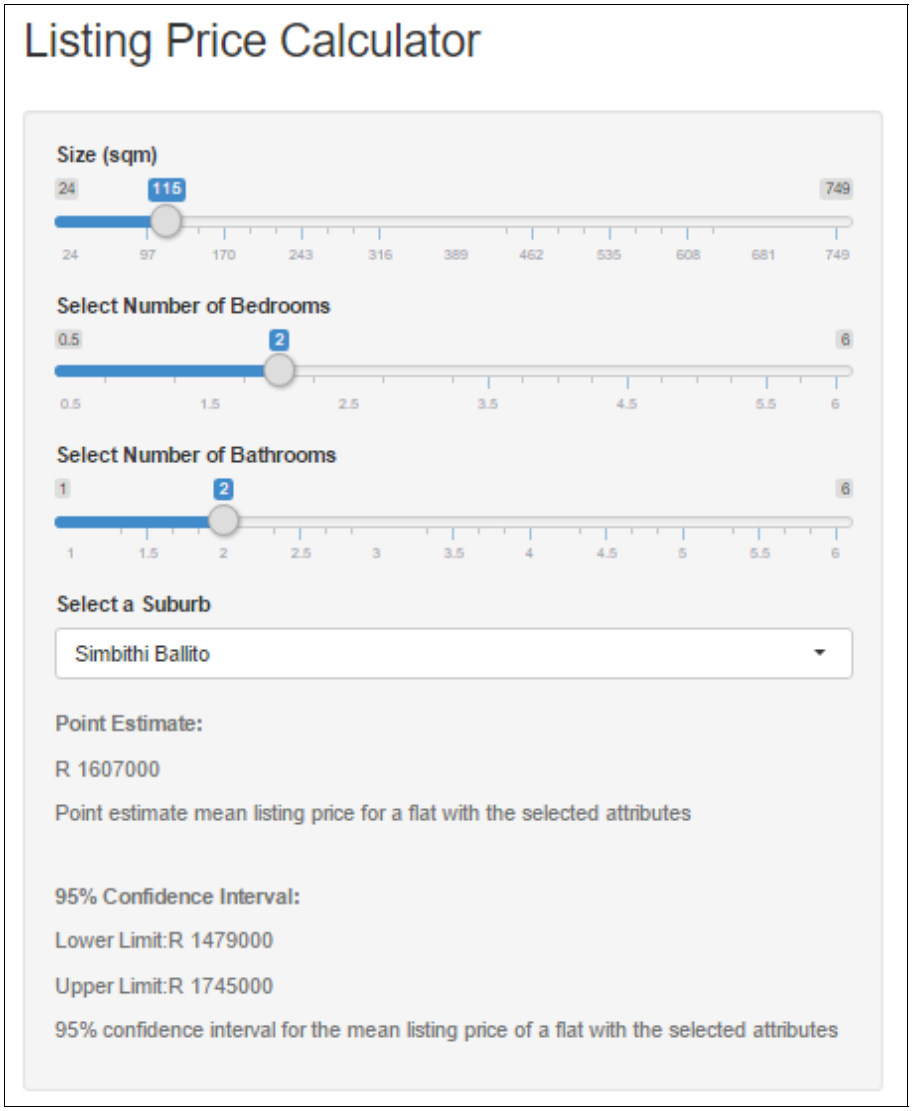


Figure 4.14: Software Application Example 3

The point estimate for a flat in Simbithi Ballito that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms is R1 607 00. The lower and upper limits of the 95% confidence interval are R1 479 000 and R1 745 000 respectively, which means that the true average listing price for a flat with the user selected attributed falls within this range.

The input parameters are dynamic so the user can select the size of the flat, the number of bedrooms, the number of bathrooms and the suburb which will all be within the range of the independent variables used to build the model. This software application provides an easy and accurate tool to understand the pricing dynamics of flats within the three sub-markets of the study. This software tool has the propensity to add value to households wishing to sell

their flats where they can obtain an understanding of market pricing dynamics and obtain listing price estimates. Furthermore, credit providers such as banks and mortgage originators can use this tool to assess how a flat is priced relative to the market in order to determine the fair market value of the asset. Real estate agencies could use this software application as a tool to value new flats and determine listing prices that are congruent to the general market consensus.

4.9 Conclusion

This chapter presented two separate hedonic models with the aim of deriving a hedonic price function for flats in KwaZulu-Natal coastal sub-markets and performed a rigorous set of statistical tests.

The findings of this study were discussed within the context of previous studies introduced in Chapter Two and the methodology introduced in Chapter Three. Various models and statistical tests were presented which showed how each research objective had been met.

The null hypothesis that the distribution of listing prices followed a gamma distribution was not rejected resulting in the development of a novel generalized linear model based on a gamma distribution and log link function. This model was correctly specified where no over-dispersion was detected and all the covariates were statistically significant. Adding a dummy locational variable resulted in the correct error structure thereby removing the presence of spatial autocorrelation which was evident from the formal and informal tests that were applied. Bootstrapping provided a measure of reliability about the models accuracy where the coefficients and standard errors were similar to the original model. A software application was successfully developed to disseminate the results of the model by producing point estimates and confidence intervals.

Significant conclusions can be drawn from the research findings of this study which are further presented in the following chapter. The following chapter outlines the limitations of this study together with recommendations for this research and for further research.

CHAPTER FIVE
Conclusions, Limitations and Recommendations

5.1 Introduction

Residential property is an important segment of the property market in South Africa where transactions are typically infrequent and relate to a highly differentiated set of commodities rendering effective measurement techniques complex. The intention of this research was to construct a hedonic pricing model to estimate listing prices for flats within KwaZulu-Natal coastal sub-markets based on statistically significant structural and locational attributes. The research objectives and subsequent research hypotheses were formulated in order to derive a hedonic price function for flats within KwaZulu-Natal coastal sub-markets and develop a software application to disseminate the results to households and investors. The analysis of the data and discussion of the results was presented in Chapter Four. This chapter presents the key findings of the study as well as the limitations and recommendations for practical use of the framework presented and potential future research initiatives.

This study sought to produce an econometric framework to augment the existing body of knowledge on the South African residential property market by developing a novel model to estimate residential property listing prices which is discussed further below.

5.2 Key Findings

Many key findings were identified in this study and are subsequently discussed in the context of the research objectives.

5.2.1 Objective One: Determining the Distribution of Flat Listing Prices

The first research objective of this study was to determine an appropriate hedonic price model for flats located in KwaZulu-Natal coastal sub-markets based on the distribution of listing prices. In previous hedonic residential property price studies a fundamental assumption about the distribution of property prices were made, namely that the observations were log-normally distributed where Day (2003); Els and Von Fintel (2010); Bourassa, Cantoni and Hoesli (2010); Dodds (2011) all applied natural logarithm transformations to ordinary least

models to derive their respective hedonic price functions without explicitly testing their log-normal distribution assumptions. However, Els and Von Fintel (2010) settled on quantile regression for their final model.

This research finds that although the distribution for listing prices of flats seemed approximately log-normal by presentation of a histogram, the Jarque-Bera test of normality revealed that there was sufficient evidence to reject the null hypothesis of normality. This means that the assumption of an asymptotic log-normal distribution for flat prices in this study was not a suitable parametric assumption and an alternative distribution was investigated to derive an appropriate hedonic price model.

The listing prices for flats in the sample data appeared to be gamma distributed which formulated the first hypothesis of this study. Upon testing this hypothesis there was strong evidence not to reject the null hypothesis of the Villasenor and Gonzalez-Estrada test, that the flat listing prices followed a gamma distribution, indicating that a model based on the gamma distribution was more suitable for the data.

5.2.2 Objective Two: Developing an Appropriate Model

The second objective of this study was to develop a model to estimate listing prices of flats located in sub-markets along the KwaZulu-Natal coast based on structural and locational attributes. This involved the construction and testing of an ordinary least squares model and a generalised linear model.

5.2.2.1 The Ordinary Least Squares Model

In order to make relevant comparisons to the literature presented in Chapter Two, an assumption that the distribution of flat prices were asymptotically log-normally distributed was made in order to develop an ordinary least squares model. Several ordinary least squares models were developed in order to find the optimal one. Taking the natural logarithm of the size variable reduced the presence of heteroscedasticity. Furthermore, the addition of the suburb variable ensured the correct specification of the model which was shown using the Ramsey RESET test. The introduction of the suburb variable also improved the R-squared diagnostic of model fit from 75.48% to 90.35% concomitantly removing the presence of

spatial autocorrelation as illustrated by the spatial residuals plots and the Mantel test. All of the variables in the study, that is, size, number of bedrooms, number of bathrooms and suburb were statistically significant. The expected values of the ordinary least squares model were reported as the geometric mean estimates and not the arithmetic mean estimates. Whites (1980) heteroscedasticity consistent covariance matrix provides a method to account for the presence of heteroscedasticity where the standard errors are biased resulting in incorrect p-values and type one errors of which none were detected in the final ordinary least squares model. However, based on the parametric assumption of the distribution of listing prices a novel alternative model was constructed, namely a generalized linear model based on a gamma distribution and log-link function.

5.2.2.2 The Generalized Linear Model

A novel approach to deriving a hedonic price function for residential property was the highlight of this study where a generalized linear model based on the gamma distribution and log-link function was developed to meet the research objectives. To the best knowledge of the researcher, this has not been attempted before in South Africa and no relevant international studies were identified that use such a novel technique. The generalised linear model based on the gamma distribution and log link function yielded better results than the ordinary least squared model. This is evident by the identification and selection of the correct distribution for the parametric model, namely gamma, and a lower root mean squared error. An important difference was that the arithmetic mean was computed as the expected value and not the geometric mean which is a fundamental reason for using a generalised linear model over the ordinary least squares model. The research hypotheses of this study involved testing whether a set of structural and locational independent variables were statistically significant for estimating listing prices of flats in the sub-markets. The following null hypotheses of the research were all rejected:

> ***H0b****:* Floor Area is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.
>
> ***H0c****:* Number of bedrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0d*:** Number of bathrooms is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

***H0e**:* The suburb, a dummy locational variable, is not a statistically significant independent variable to measure listing prices for flats that are located in sub-markets along the KwaZulu-Natal coast.

These findings indicates that this set of independent variables, namely floor area, number of bedrooms, number of bathrooms and suburb were all significant determinants of listing prices for flats within the KwaZulu-Natal coastal sub-markets of this study. Similar to the ordinary least squares model, the results of the generalised linear model show that the inclusion of the suburb variable removed the presence of spatial autocorrelation. The generalised linear model showed no sign of over dispersion where the residual deviance was less than the residual degrees of freedom.

5.2.2.3 Reliability of Results

Non-parametric bootstrapping provided a good measure of model validation for both models by introducing variance through re-sampling with replacement where the coefficients were all similar to the respective original models, indicating that the original model results were accurate when introducing variability through random sampling distributions.

The root mean squared error (RMSE) for each model was computed where the generalised linear model achieved the lower value indicating that the dispersion of the models fitted values versus the observed values was lower and that the generalised linear model is a better fit for the data. This was further illustrated by generating a random sample of 15 sample observations where each models fitted price values was compared to the observed price. Interestingly both models produced similar fitted value estimates, however, the generalised linear model produced slightly better estimates.

5.2.3 Objective Three: Developing a Software Application

The third and final objective of the study was to build a software application that facilitates the estimation of listings prices for flats in KwaZulu-Natal coastal sub-markets given a set of structural and locational attributes. This application provides a simple front end user interface to determine the average listing price of a flat given a set of structural and locational attributes. The third objective of this study was successfully accomplished, facilitating the potential to disseminate valuable information to households and investors wanting to purchase or sell a flat in one of the sub-markets of the study. This study sought to bridge the gap between academia and business by providing a software application that can be hosted online by Private Property (Pty) Ltd using an empirically tested model developed through rigorous academic research.

The software application calls the generalised linear model once a user has selected a set of attributes. The generalised linear model then calculates the expected price or arithmetic mean listing price and the confidence interval, rendering it on the user interface. The example presented in Figure 5.1 details the average price for a flat in Musgrave within the Durban Central sub-market that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms. The confidence interval is also computed which provides the statistical inference of the point estimate of a flat given the selected attributes. Note all figures are rounded to the nearest thousand.

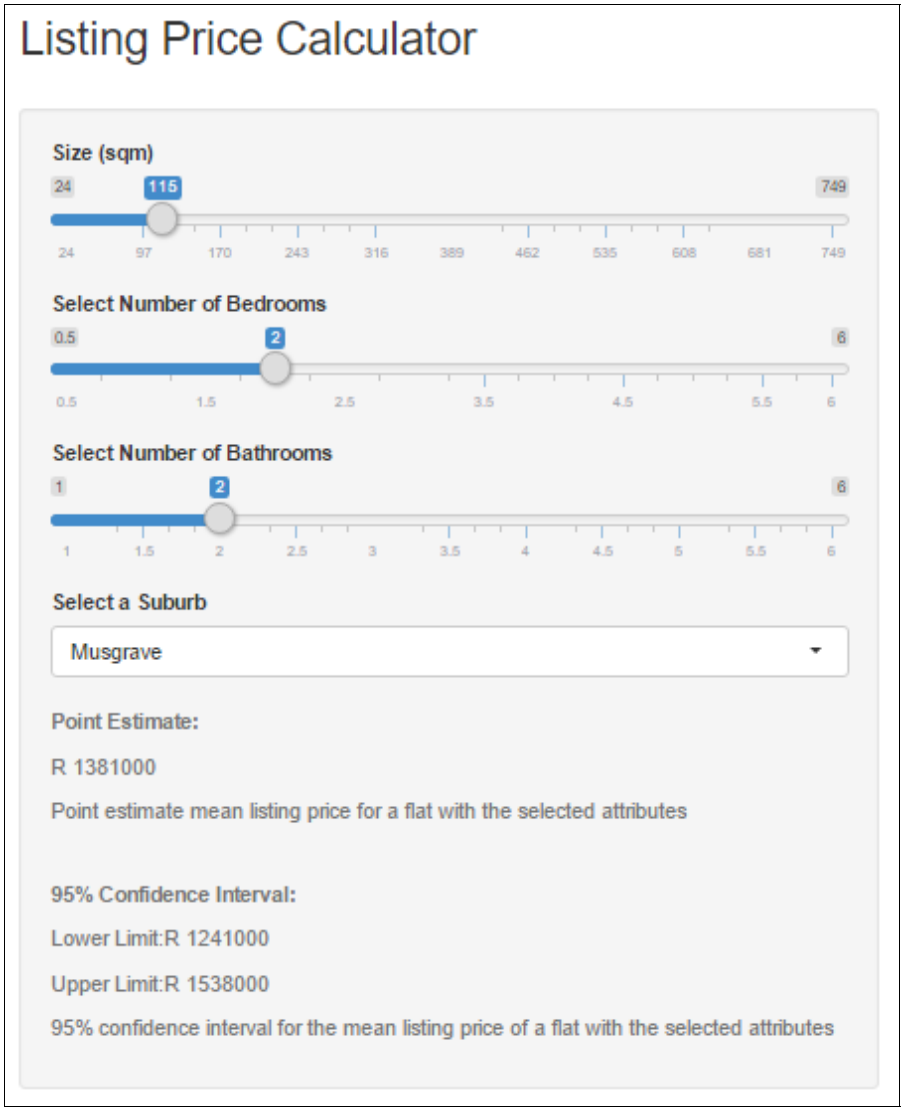


Figure 5.1: Software Application Concluding Example

The point estimate for a flat in Musgrave that is 115 square meters in size (floor area) with 2 bedrooms and 2 bathrooms is R1 381 000. The lower and upper limits of the 95% confidence interval are R1 241 000 and R1 538 000 respectively, which means that the true average listing price for a flat with the user selected attributed falls within this range.

5.3 Limitations

The study involved flats in the three KwaZulu-Natal sub-markets, therefore, further research is required in order to determine how generalisable the framework presented in this study to other segments of the South African residential property market such as different property

types and different suburbs. Furthermore, four independent variables were used to estimate the hedonic price function of flats in this study. This was based solely on the data availability. However, there could be additional covariates that could help explain the variation in the hedonic price function of flats for the three KwaZulu-Natal sub-markets.

Correlation research merely demonstrates that we can predict the behaviour of one variable from the behaviour of other variables. However, the presence of a confounding factor or confounding variables may make it hard to establish, in the findings or output, the outcome as being a direct consequence of the independent variables measured.

A clear limitation of the generalised linear model in the context of this study is the lack of South African or global empirical research to benchmark the findings against. However, the meticulous methodology and rigorous set of statistical tests applied in order to determine the suitability thereof provides a high degree of confidence that the generalised linear model is correctly specified and a good approximation of the data. Maintenance of the software application involves skilled expertise of statistical modelling and programming knowledge as the model will need to be updated periodically with new data. Furthermore, extending this novel econometric model and software application to different residential property markets in South Africa will require the same level of statistical and programming domain knowledge.

5.4 Recommendations

The recommendations that follow will be presented in the context of the practical use of the research for Private Property (Pty) Ltd and suggestions for future research.

5.4.1 Recommendations for Private Property (Pty) Ltd

Based on choice of using the ordinary least squares model or the generalised linear model methodology to develop a framework to commercialise a solution for downstream users such as households and investors, this study finds that the generalised linear model based on the gamma distribution and log-link function is a more suitable model as it meets all the requisite statistical axioms and parametric assumptions.

5.4.2 Recommendations for Future Research

Being the first generalised linear model used to develop a hedonic price function for residential property in South Africa and perhaps globally, further research is required in order to determine how generalisable the econometric framework presented in this study is.

Future research should include the use of the econometric framework propounded in this study across different geographic regions and across different residential property types such as houses and complexes.

A final research initiative could be to attempt to link the modelling framework presented in this study with time series data to develop a residential property price index.

5.5 Concluding Remarks

This aim of this study was to derive a hedonic price function for flats within KwaZulu-Natal coastal sub-markets based on statistically significant structural and locational attributes. The research objectives have been accomplished through the set of research hypotheses that were formulated and tested through rigorous statistical tests and techniques A generalised linear model based on the gamma distribution and log-link function was developed as a novel alternative to derive a hedonic price function for a segment of the residential property market in KwaZulu-Natal and proved to be a better model for this research problem than the traditional ordinary least squares modelling approach. Based on the results, a software application was developed to disseminate the results of the generalised linear model for potential commercial use in South Africa, bridging the gap between academia and business.

References

Anselin, L., 2013. *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Bank, S.A.R., 2015. *South African Reserve Bank Quarterly Bulletin*. [pdf] South African Reserve Bank. Available at: <https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6649/01Full%20Quarterly%20Bulletin%20%E2%80%93%20March%202015.pdf> [Accessed date: 12 February 2016].

Battese, G.E. and Bonyhady, B.P., 1981. Estimation of household expenditure functions: an application of a class of heteroscedastic regression models. *Economic Record*, [e-journal] *57*(1), pp.80-85. Available through: <http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/> [Accessed 13 February 2016].

Borcard, D. and Legendre, P., 2012. Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study. *Ecology*,*93*(6), pp.1473-1481. Available through: <http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/> [Accessed 13 February 2016].

Bordo, M.D. and Jeanne, O., 2002. *Boom-busts in asset prices, economic instability, and monetary policy* (No. w8966). [pdf] Massachusetts: National Bureau of Economic Research. Available at: <http://www.nber.org/papers/w8966> Accessed date [13 February 2016].

Bourassa, Steven, Eva Cantoni, and Martin Hoesli. Predicting house prices with spatial dependence: A comparison of alternative methods. In: University of Louisville and University of Geneva, *15th Conference of the Pacific Rim Real Estate Society*, Sydney, Australia, 18th -21st January 2009. Clemson: Journal of Real Estate Research

Bromideh, A.A. and Valizadeh, R., 2013. Discrimination between Gamma and Log-Normal Distributions by Ratio of Minimized Kullback-Leibler Divergence. *Pakistan Journal of Statistics and Operation Research*, *9*(4). Available at: <http://www.pjsor.com/index.php/pjsor/article/view/487>. Date accessed: 5 March 2016.

Cai, L. and Hayes, A.F., 2008. A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, *33*(1), pp.21-40. Available through: <http://online.sagepub.com.ukzn.idm.oclc.org/> [Accessed 13 February 2016].

Carruthers, E., Lewis, K., McCue, T. and Westley, P., 2008. Generalized linear models: model selection, diagnostics, and overdispersion. [pdf] Memorial University of Newfoundland, unpublished. Available at : <http://www.mun.ca/biology/dschneider/b7932/B7932Final4Mar2008.pdf> [Accessed Date: 13 February 2016].

Chang, W., et al., 2016. shiny: Web Application Framework for R. R package version 0.13.1. http://CRAN.R-project.org/package=shiny

Chen, C.F. and Rothschild, R., 2010. An application of hedonic pricing analysis to the case of hotel rooms in Taipei. *Tourism Economics*, *16*(3), pp.685-694. Available at : <http://ir.lib.ncku.edu.tw/> [Accessed Date: 13 February 2016].

Coenders, G. and Saez, M., 2000. Collinearity, heteroscedasticity and outlier diagnostics in regression. Do they always offer what they claim. *New Approaches in Applied Statistics*, *16*, pp.79-94. Available at : <http://www.stat-d.si/index.php?lang=en> [Accessed Date: 6 March 2016].

Corcoran, C. and Liu, F., 2014. Accuracy of Zillow's Home Value Estimates. *REAL ESTATE ISSUES*®, p.201445. Available at : <https://www.thelibrarybook.net/pdf-global-journal-of-business-research-volume-8-number-2-2014.html> [Accessed Date: 6 March 2016].

Creswell, J.W., 2009. *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications, Incorporated.

Dawson, E.L., Beasley, T.M. And Redden D.t., Performance of OLS and HCCM Estimators in Heteroscedastic ANCOVA Models. *Generalized Linear Model Journal*, [pdf] Available at: <http://www.glmj.org/archives/articles/Dawson_v41n2r.pdf> [Accessed Date: 6 March 2016].

Day, B., 2003. Submarket identification in property markets: a hedonic housing price model for Glasgow. *Centre for Social and Economic Research on the Global Environment,* [pdf] Available at: <http://www.cserge.ac.uk/> [Accessed Date: 7 February 2016].

DeBenedictis, L.F. and Giles, D.E., 1998. Diagnostic testing in econometrics: variable addition, RESET, and Fourier approximations. *Handbook of Applied Economic Statistics,* [pdf] Available at: <http://www.uvic.ca/socialsciences/economics/assets/docs/frese t.pdf> [Accessed: 2 April 2016].

de Haan, J. and Erwin, D., 2011. *Handbook on Residential Property Price Indices*. [pdf] Eurostat European Commission. Available through: <http://ec.europa.eu/eurostat/documents /3859598/5925925/KS-RA-12-022-EN.PDF> [Accessed date 12 February 2016].

Delignette-Muller, M.L., Dutang, C., 2015. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. URL Available at: <http://www.jstatsoft.org/v64/i04/ [Accessed 2 April 2016].

Demortier, L., 2007. P Values: *What They Are and How to Use Them*.[pdf] The Collider Detector at Fermilab. Available at: <http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf> [Accessed 2 April 2016]

Diniz-Filho, J.A.F., Soares, T.N., Lima, J.S., Dobrovolski, R., Landeiro, V.L., Telles, M.P.D.C., Rangel, T.F. and Bini, L.M., 2013. Mantel test in population genetics. *Genetics and Molecular Biology*, *36*(4), pp.475-485. [pdf] Available at: <http://www.scielo.br/pdf/gmb/v36n4/v36n4a02.pdf> [Accessed Date: 2 April 2016].

Dodds, R.S., 2011. *An investigation into the hedonic price analysis of the structural characteristics of residential property in the West Rand.* [pdf] Available at: <http://wiredspace.wits.ac.za/> [Accessed Date: 2 April 2016].

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Daniel, K, W. and Kühn, I., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), pp.609-628. Available through: <http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/> [Accessed 2 April 2016].

du Preez, M., Balcilar, M., Razak, A., Koch, S.F. and Gupta, R., 2014.*House Values and Proximity to a Landfill: A Quantile Regression Framework.* [pdf] University of Pretoria. Avaliable at <http://www.up.ac.za/media/shared/61/WP/wp_2014_42.zp39444.pdf> [Accessed 13 February 2016].

Duttagupta, R. and Fernandez, G., 2004. From fixed to float: operational aspects of moving toward exchange rate flexibility. *Social Science Research Network,* [online]. Available at: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=878950> [Accessed 3 March 2016].

Egghe, L. and Leydesdorff, L., 2009. The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, *60*(5), pp.1027-1036. [pdf]. Available at: <https://arxiv.org/ftp/arxiv/papers/0911/0911.1318.pdf> [Accessed 3 March 2016].

Els, M. and Von Fintel, D., 2010. Residential property prices in a submarket of South Africa: Separating real returns from attribute growth. *South African Journal of Economics*, *78*(4), pp.418-436. Available through: <http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/> [Accessed 30 January 2016].

Friedman, M. and National Bureau of Economic Research, 1957. *A theory of the consumption function* (Vol. 63). Princeton: princeton university press.

Friedman, M., 1997. *John Maynard Keynes.* [pdf]. Available at: <https://core.ac.uk/download/files/153/6993481.pdf> [Accessed 3 March 2016].

Fox, J., 2002. Bootstrapping regression models. An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book. Sage, Thousand Oaks, CA. URL http://cran. rproject. org/doc/contrib/Fox-Companion/appendix-bootstrapping. pdf.

Fox, J. and Weisberg, S., 2010. *An R companion to applied regression*. Sage Publishers.

Fu, L. and Moncher, R.B., 2004. Severity Distributions for GLMs: Gamma or Lognormal? Evidence from Monte Carlo Simulations. *Casualty Actuarial Society Discussion Paper Program*, pp.149-230.

Gandy, A. and Kvaløy, J.T., 2013. Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian Journal of Statistics*,*40*(4), pp.647-668.

Girouard, N. and Blöndal, S., 2001. *House prices and economic activity.* [pdf]. Available at: <https://www.oecd.org/eco/monetary/1888662.pdf> [Accessed 27 March 2016].

Glick, H., 2008. Methods for cost estimation in CEA: the GLM approach. *Academy Health, issues in Cost-Effectiveness Analysis. Washington, DC*.

Goodhart, C. and Hofmann, B., 2008. House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, *24*(1), pp.180-205.

Goodman, A.C., 1978. Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, *5*(4), pp.471-484.

Google Maps, 2016. *Map of Ballito, Umhlanga and Durban Central.* [online]. Google. Available from: <https://www.google.co.za/maps/@-29.845516,31.0457477,9.5z> [Accessed 23 February 2016].

Greene WH. *Econometric analysis*. India: Pearson Education

Gujarati, N.D., 2005. Basic Econometrics, Tata McGraw.

Harvey, Philip Scott, Henri P. Gavin, and Jeffrey T. Scruggs. "Probability Distributions." (2011).

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), pp.83-85.

Hill, R., 2011. *Hedonic price indexes for housing* (No. 36). [pdf] OECD Statistics Working Papers. Available at: <http://www.oecd-ilibrary.org/docserver/download/5kghzxpt6g6f.pdf?expires=1475604114&id=id&accname=guest&checksum=6907754F3DC94316455C91DC2665E9D0> Accessed date [02 October 2016].

Hinkins, S., Mulrow, E. and Scheuren, F., 2009. Visualization of complex survey data: Regression diagnostics. *2009 Proceedings of the Section on Survey Research Methods*, pp.2206-2218.

Jones, A.M., 2010. *Models for health care*. University of York., Centre for Health Economics.

Jordà, Ò., Schularick, M. and Taylor, A.M., 2015. Interest Rates and House Prices: Pill or Poison?. *FRBSF Economic Letter*, *25*.

Jiang, L., Phillips, P.C. and Yu, J., 2015. New methodology for constructing real estate price indices applied to the Singapore residential market. *Journal of Banking & Finance*, *61*, pp.S121-S131.

Kahle, D., Wickham, H. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Keller, G., 2012. *Managerial statistics Abbreviated*. South-Western Cengage Learning.

Kennedy, P.E., 2005. Oh no! I got the wrong sign! What should I do?. *The Journal of Economic Education*, *36*(1), pp.77-92.

Kennedy P. 1985. A guide to econometrics. 2nd ed. The MIT Press, Cambridge, UK.

Kerns, G.J., 2010. *Introduction to probability and statistics using r First Edition.* [e-book] : Youngtown: UPSUR. Available through: <https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf> [Accessed 2 February 2016]

Kothari, C.R., 2004. *Research methodology: Methods and techniques*. New Age International.

Kuhn, T.S., 2012. *The structure of scientific revolutions*. [pdf]. London: University of Chicago press. Available through: <http://projektintegracija.pravo.hr/_download/repository/Kuhn_Structure_of_Scientific_Revolutions.pdf> [Accessed 19 February 2016].

Kumo, W.L., 2015. Inflation Targeting Monetary Policy, Inflation Volatility and Economic Growth in South Africa . *African development Bank Group Working Paper Series*, (216).

Lee, S.L., 2006. *Property funds: how much diversification is enough?* [online]. Available at: <http://centaur.reading.ac.uk/20742/> [Accessed 7 May 2016].

Liao, W.C. and Wang, X., 2012. Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, *21*(1), pp.16-27.

Lindsey, J.K., 1997. *Applying generalized linear models*. Springer Science & Business Media.

Long, J.S. and Ervin, L.H., 2000. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*,*54*(3), pp.217-224.

Lyons, R.C., 2015. *Measuring house prices in the long run: Insights from Dublin, 1900-2015*. [pdf]. Available at: <http://eh.net/eha/wp-content/uploads/2015/05/Lyons.pdf> [Accessed 29 April 2016].

Malpezzi, S., 2003. Hedonic pricing models: a selective and applied review. *Section in*

Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan.

Manning, W.G., Basu, A., Mullahy, J. and Manning, W., 2002. Modeling costs with generalized gamma regression. *RO1 AA12664-01 A*, *2*.

McCue, T., Carruthers, E., Dawe, J., Liu, S., Robar, A. and Johnson, K., 2008. Evaluation of generalized linear model assumptions using randomization. *Unpublished manuscript. Retrieved fro m http://www. mun. ca/biology/dschneider/b7932/B7932Final10Dec2008. pdf*.

McCullagh, P. and Nelder, J.A., 1989. *Generalized linear models* (Vol. 37). London: CRC press.

Monson, M., 2009. Valuation using hedonic pricing models. *Cornell Real Estate Review*, *7*(1), p.10.

Moran, J.L., Solomon, P.J., Peisach, A.R. and Martin, J., 2007. New models for old questions: generalized linear models for cost prediction. *Journal of evaluation in clinical practice*, *13*(3), pp.381-389.

Muchabaiwa, H., 2013. *Logistic regression to determine significant factors associated with share price change*. [pdf]. Available at: <http://uir.unisa.ac.za/bitstream/handle/10500/13229/Final%20Desertation_46265147.pdf?sequence=1> [Accessed 29 April 2016].

Müller, M., 2012. Generalized linear models. In *Handbook of Computational Statistics* (pp. 681-709). Springer Berlin Heidelberg.

Murphy, K.P., Brockman, M.J. and Lee, P.K., 2000. Using generalized linear models to build dynamic pricing systems. In *Casualty Actuarial Society Forum, Winter* (pp. 107-139).

Musset, L., 2006. *OECD Environment Health and Safety Publications Series on Testing and Assessment No. 54.* [pdf]. Available at: <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2006)18&doclanguage=en [Accessed 12 May 2016].

Naimi, B., 2015. usdm: Uncertainty Analysis for Species Distribution Models. R package version 1.1-15. http://CRAN.R-project.org/package=usdm.

Nelder, J.A. and Wedderburn, R.W.M., 1972. Generalized linear models. *Encyclopedia of tatistical Sciences*.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), pp.673-690.

Oksanen, J., *et al*, 2016. vegan: Community Ecology Package. R package version 2.3-3. http://CRAN.R-project.org/package=vegan.

Olivier, J., Johnson, W.D. and Marshall, G.D., 2008. The logarithmic transformation and the geometric mean in reporting experimental IgE results: what are they and when and why to use them?. *Annals of Allergy, Asthma & Immunology*, *100*(4), pp.333-337.

Oredein, A.I., Olatayo, T.O. and Loyinmi, A.C., 2011. On validating regression models with bootstraps and data splitting techniques. *Global Journal of Science Frontier Research*, *11*(6).

Özyurt, S., 2014. *Spatial dependence in commercial property prices: micro evidence from the Netherlands*. [pdf]. Available at: <https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1627.pdf?f4b7432ba10e5d1553a255b587a09d23 [Accessed 17 April 2016].

Peng, R.D., 2015. *R Programming for Data Science*. [e-book]. Lean Publishing. Available at: <https://leanpub.com/rprogramming> [Accessed date: 12 February 2016].

Private Property, 2015. *Who is Private Property?* [online]. 24 June 2015. Available at: <http://www.privateproperty.co.za/advice/property-tv/articles/who-is-private-property/3391> [Accessed date: 03 April 2016]

R DEVELOPMENT CORE TEAM, 3rd May, 2016-last update, R: a language and environment for statistical computing [Homepage of R Foundation for Statistical Computing], [Online]. Available: <http://www.R-project.org>.

Rawlings, J.O., Pantula, S.G. and Dickey, D.A., 1998. *Applied regression analysis: a research tool*. Springer Science & Business Media.

The Research Advisors, 2006. *Sample Size Table*. [online]. Available at: <http://www.research-advisors.com/tools/SampleSize.htm> [Accessed date: 3 April 2016].

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, *82*(1), pp.34-55.

Saiz, A., 2012. *Interest Rates and Fundamental Fluctuations in Home Values*. Working Paper, Urban Economics Lab, MIT.

Sapra, S., 2005. A regression error specification test (RESET) for generalized linear models. *Economics Bulletin*, *3*(1), pp.1-6.

Schulz, R., 2003. *Valuation of properties and economic models of real estate markets.* [pdf]. Available at: <http://edoc.hu-berlin.de/dissertationen/schulz-rainer-2003-02-05/PDF/Schulz.pdf> [Accessed 11 February 2016].

Statistics South Africa, 2011. *Ethekweni*. [online]. Available at: <http://www.statssa.gov.za/page_id=993&id=ethekwini-municipality> [Accessed date: 28 April 2016].

Stohldreier, M.T., 2012. *The Determinants of House Prices in Chinese Cities*. [pdf]. Available at: <http://www.econ.uzh.ch/ipcdp/theses/MA_MarieStohldreier.pdf> [Accessed 16 February 2016].

Tan, Y.K., 2011. An Hedonic Model for house prices in Malaysia. In *International Real Estate Society Conference* (Vol. 15, No. 1, pp. 12-15).

Thayn, J.B. and Simanis, J.M., 2013. Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors. *Annals of the Association of American Geographers*, *103*(1), pp.47-66.

Thisted, R.A., 1998. *What is a P-value?, pp.1-6*. [pdf]. Available at: <https://f80de1cf-a-62cb3a1a-s-sites.googlegroups.com/site/hemilio/pvalue.pdf?attachauth=ANoY7coBwXUjjQmcbVHhESLswX_0XEWKOlo2fLEpF7H8leEe9VNpw2OKJCZPZ7iLM47JL0m8BLREofrZD_nXug0DWnPLRoHKb1L6TBAJ8oA3lscLdh4E8t1gK8YnWRxsCF1NifxquPn-LZYQNJ1hBtAjr2r8aCrXfhgHPxr4zm_7IY3eofmc2eQeLZe7AKmlIvnqUmXxXcBbLT6ly_XVG3NiLMi6MbjIA%3D%3D&attredirects=0> [Accessed 28 Appril 2016].

Triplett, J., 2005. *Handbook on hedonic indexes and quality adjustments in price indexes: special application to information technology products, V OECD Science,* Technology and Industry Working Papers, 2004/9, OECD Publishing.

Tserkezos, E., Temporal Aggregation and the Ramsey's (RESET) Test for Functional Form: results from Monte Carlo experiment.

Van der Merwe, E.J., 2004. *Inflation Targeting in South Africa*. Pretoria, South Africa: South African Reserve Bank.

Vastrad, C., 2013. Performance Analysis Of Neural Network Models For Oxazolines And Oxazoles Derivatives Descriptor Dataset. *arXiv preprint arXiv:1312.2853*.

Vavrek, M.J., 2011. fossil: palaeoecological and palaeogeographical analysis tools. Palaeontologia Electronica, 14:1T. http://palaeo-electronica.org/2011_1/238/index.html.

Villaseñor, J.A. and González-Estrada, E., 2015. A variance ratio test of fit for Gamma distributions. *Statistics & Probability Letters*, *96*, pp.281-286.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pp.817-838.

Wilcox, R.R., 2008. Post-hoc analyses in multiple regression based on prediction error. *Journal of Applied Statistics*, *35*(1), pp.9-17.

Williams, R., 2012. *Heteroscedasticity.* [pdf]. Available at: <https://library.saylor.org/handle/1/10742> [Accessed 19 May 2016].

Williams, M.N.,Grajales, C.A.G. And Kurkiewicz, D., 2013. Assumptions of Multiple Regression:Correcting Two Misconceptions. *Practical Assessment, Research and Evaluation* 18(11), p.11.

Wooldridge, J.M., 2010. *Econometric analysis of cross section and panel data*. MIT press. [pdf]. Available at: <http://bb.shufe.edu.cn/bbcswebdav/institution/%E7%BB%8F%E6%B5%8E%E5%AD%A6%E9%99%A2/teacherweb/2002000062/AdvEconiXi/%E5%A4%8D%E4%BB%B6%20Econometric%20Analysis%20of%20Cross%20Section%20and%20Panel%20Data.pdf> [Accessed 22 May 2016].

Zeileis, A., Hothorn,T., 2002. lmtest: Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL http://CRAN.R-project.org/doc/Rnews/

Appendix One

Ethical Clearance

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

15 June 2016

Mr Dane Gregory Bax 214580191
Graduate School of Business and Leadership
Wesrville Campus

Dear Mr Bax

Protocol reference number: HSS/0209/016M
New project title: Listing Price Estimation of Flats along the KwaZulu-Natal Coastal Sub-markets: A Novel Econometric Model

Approval notification – Amendment Application

This letter serves to notify you that your application for an amendment dated 10 June 2016 has now been granted **Full Approval**.

- **Change in Title**
- **Change of Objectives**
- **Change of wording of Research Hypotheses**

Any alterations to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study must be reviewed and approved through an amendment /modification prior to its implementation. In case you have further queries, please quote the above reference number. PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years

The ethical clearance certificate is only valid for a period of 3 years from the date of issue. Thereafter Recertification must be applied for on an annual basis.

Best wishes for the successful completion of your research protocol.

Yours faithfully

[redacted]

Dr S
Humanities Social Sciences Research Ethics

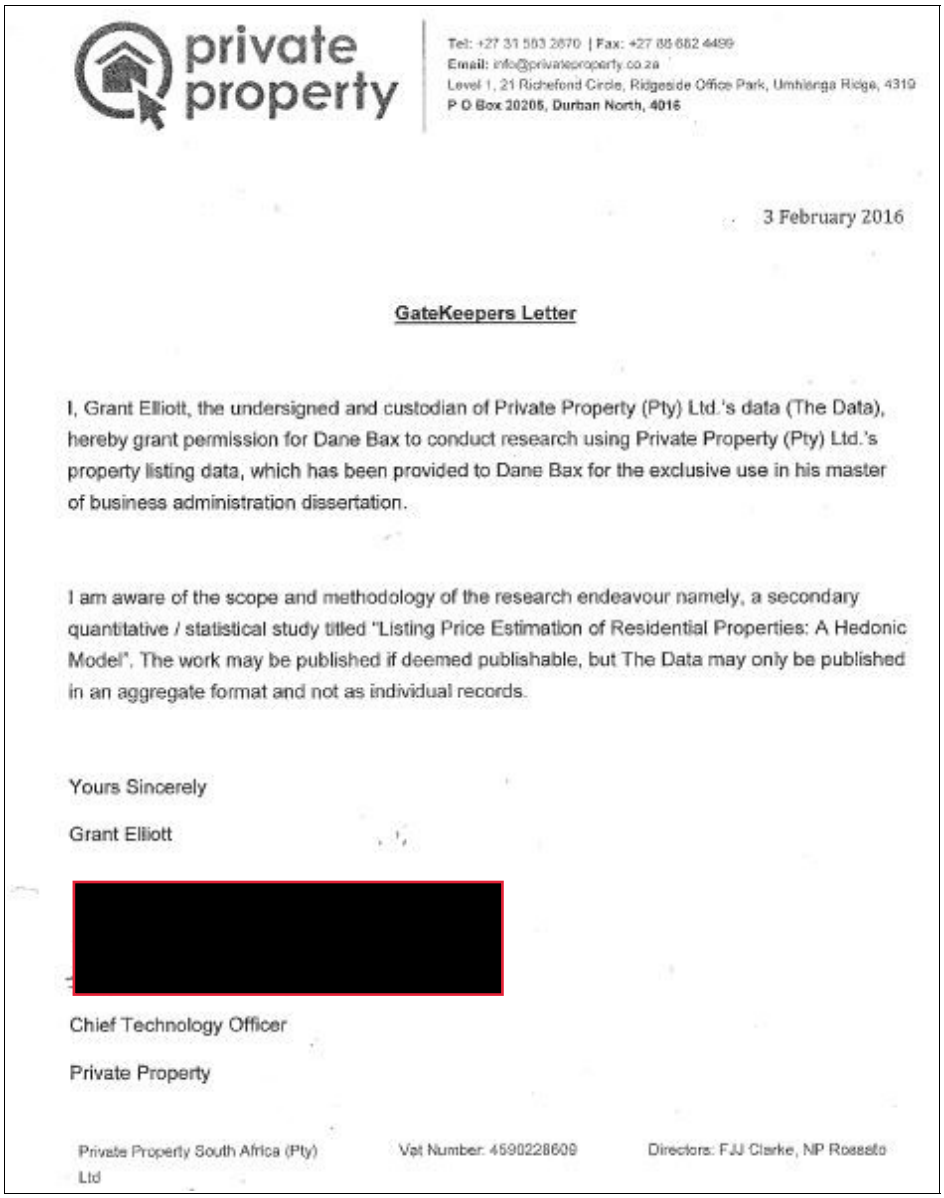
/pm

Supervisor: Dr Milhalis Chasomeris
Academic Leader Research: Dr M Hoque
School Administrator: Ms Zarina Bullyraj

Humanities & Social Sciences Research Ethics Committee
Dr Shenuka Singh (Chair)
Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Telephone: +27 (0) 31 260 3567/8350/4557 **Facsimile:** +27 (0) 31 260 4609 **Email:** ximbap@ukzn.ac.za / snymanm@ukzn.ac.za / mohung@ukzn.ac.za
Website: www.ukzn.ac.za

Appendix Two

Gatekeepers Letter



private property

Tel: +27 31 583 2870 | Fax: +27 86 682 4499
Email: info@privateproperty.co.za
Level 1, 21 Richefond Circle, Ridgeside Office Park, Umhlanga Ridge, 4319
P O Box 20205, Durban North, 4016

3 February 2016

GateKeepers Letter

I, Grant Elliott, the undersigned and custodian of Private Property (Pty) Ltd.'s data (The Data), hereby grant permission for Dane Bax to conduct research using Private Property (Pty) Ltd.'s property listing data, which has been provided to Dane Bax for the exclusive use in his master of business administration dissertation.

I am aware of the scope and methodology of the research endeavour namely, a secondary quantitative / statistical study titled "Listing Price Estimation of Residential Properties: A Hedonic Model". The work may be published if deemed publishable, but The Data may only be published in an aggregate format and not as individual records.

Yours Sincerely

Grant Elliott

[redacted]

Chief Technology Officer

Private Property

Appendix Three

Turnitin Report