Journal of Economic and Financial Sciences
ISSN: (Online) 2312-2803, (Print) 1995-7076

Listing price estimation of apartments: A generalised linear model

Authors:
Dane Bax[1]
Mihalis G. Chasomeris[1]

Affiliations:
[1]Graduate School of Business and Leadership, University of KwaZulu-Natal, Durban, South Africa

Corresponding author:
Mihalis Chasomeris,
chasomerism1@ukzn.ac.za

Dates:
Received: 26 Mar. 2018
Accepted: 18 Jan. 2019
Published: 25 July 2019

How to cite this article:
Bax, D. & Chasomeris, M.G., 2019, 'Listing price estimation of apartments: A generalised linear model', *Journal of Economic and Financial Sciences* 12(1), a204. https://doi.org/10.4102/jef.v12i1.204

Orientation: Residential property is an important segment of the property market in South Africa. Residential property transactions are typically infrequent and relate to a highly differentiated set of items making measurement techniques complex and difficult.

Research purpose: The aim of this research was to develop a statistical model to estimate listing prices of apartments in KwaZulu-Natal, South Africa, and build a software application to disseminate the results thereof.

Motivation for the study: This study presents a novel alternative to the log linear (ordinary least squares) method of deriving a hedonic price function for residential property where the arithmetic mean is computed as the expected value and not the geometric mean.

Research design, approach and method: Using a data set of 1314 residential apartments provided by Private Property (Pty) Ltd, this research derives a hedonic price function for residential property using a generalised linear model based on the gamma distribution and log-link function.

Main findings: The results showed that floor area, number of bedrooms, number of bathrooms and a dummy variable for suburb (location) were statistically significant determinants of listing prices.

Practical/managerial implications: A software application, called the *listing price calculator*, was developed to disseminate the results of the model for commercial use by real estate buyers, sellers and agents, bridging the gap between academia and business.

Contribution/value-add: This study derives a hedonic price function for residential property using a generalised linear model based on the gamma distribution and log-link function, which is novel in South African research.

Keywords: residential property; listing price estimation; hedonic valuation; economics; geospatial modelling; generalised linear model.

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 policymakers closely monitor residential property price trends to gauge real house price growth and financial stability, as well as to monitor the activity and condition of the credit market (De Haan & Erwin 2011).

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's capitalised 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). The main objective of this research was to construct a hedonic pricing model to estimate listing prices for apartments within three KwaZulu-Natal coastal regions based on statistically significant structural and locational attributes. It appears that no similar study has been conducted on any property segment in KwaZulu-Natal province of South Africa. This study develops a hedonic price function using a generalised linear model based on the gamma distribution and log-link function, which has not been attempted before in South African research, and finding global research employing this methodology has proven difficult where no existing studies have been identified. This study presents an alternative to the log linear (ordinary least squares) method of deriving a hedonic price function for residential property where the arithmetic

mean is computed as the expected value and not the geometric mean. This study bridges the gap between academia and business by creating a software application that may be hosted online and used by real estate buyers, sellers and businesses to estimate listing prices of apartments.

Review of the literature: Hedonic valuation theory

Hedonic pricing theory

Prices of residential property are difficult to measure because of their heterogeneous nature, where it can be observed that dwellings are not identical even by the sole virtue of occupying different locations (Hill 2011).

Different methodologies exist to develop property price indices. The Organisation for Cooperation and Economic Development outlines several methodologies to construct residential property price indices (De Haan & Erwin 2011). Simpler methods, like the average or median mix adjustment approach, group properties into homogenous strata, calculating a central measure of tendency for each stratum. A weighted average is then applied to roll up the indices into a single index. The repeated sales approach performs regression on pooled property data for properties that have been transacted more than once in the estimation period. Hedonic price modelling is pervasive in economic literature and 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 & Werwatz 2004). Hedonic pricing is a mathematical technique used in economics that 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). The hedonic characteristics price approach runs separate regression models for each time period and calculates price inflation using index number theory (De Haan & Erwin 2011). This makes hedonic pricing a suitable approach to produce price estimates for cross sectional residential property data where properties have not been transacted frequently.

Residential property is a single class of good or commodity in the eyes of individuals, households and investors; however, it is differentiated or heterogeneous in nature (Hill 2011). 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 product's 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, supply and demand forces in the property market can determine each attribute's marginal contribution to the property's price implicitly (De Haan & Erwin 2011). 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 (1974) 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.

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

$$Zj = \left(Zj1, Zj2, Zj3, ...Zjk\right) \quad \text{[Eqn 1]}$$

Rosen (1974) defines hedonic pricing as the functional relationship between the price of a heterogeneous product and the associated attributes:

$$Pj = P\left(Zj\right) = P\left(Zj1, Zj2, Zj3, ..., Zjk\right) \quad \text{[Eqn 2]}$$

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). Notably, an increase in price is experienced by attributes that are more positive and a decrease in price is experienced by more negative attributes, *ceteris paribus* (Els & Von Fintel 2010). Regression analysis makes it possible to calculate the implicit price for attribute i of property j by taking the partial derivative, represented as follows:

$$\mathrm{P}i\left(\mathrm{Z}j\right) = \frac{\partial P}{\partial Zi} (i = 1 \text{ to } Z) \quad \text{[Eqn 3]}$$

This function describes the additional amount to be paid to obtain a marginally higher level of attribute Zi, *ceteris paribus* (Day 2003). Hedonic prices can be measured with the use of regression, a statistical technique that aims to establish the relationship between a set of property attributes and property prices by regressing property price on a set of property attributes.

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 because of data availability constraints and subjective judgements by the researcher where important variables are not included in the analysis (Jiang, Phillips & 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. Simplicity through parsimony of parameter selection is a desired feature of any model,

as complexity is reduced and predictive accuracy increased (McCullagh & Nelder 1989).

Residential hedonic models

Typically, international and local residential property hedonic price studies use ordinary least squares to derive hedonic pricing functions. Given a vector of a dependent variable and a matrix of independent variables, ordinary least squares make it possible to express the dependent variable as a linear combination of the independent variables (Greene 2003).

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. He applied the natural logarithm transform to the floor area variable in his study. 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 (Özyurt 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 et al. (2010) found that property price predictions were more accurate when submarket 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 submarket 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 & 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 it 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 pooled cross sectional hedonic analysis in the housing market of the Western Cape province, including Stellenbosch, Somerset West, Strand and Gordon's Bay, from 2004 to 2007, where they employed ordinary least squares and quantile regression techniques. Two models were derived using ordinary least squares, the first 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. 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. 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 the 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 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 its presence 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 & 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 2004).

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 11 structural variables and 1 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 because of an important omitted variable or multicollinearity as the model specified 12 independent variables in total. Multicollinearity is commonly experienced in ordinary least squares regression models which is caused by highly correlated independent variables. The variance inflation factor is a useful method for detecting the magnitude of multicollinearity (Chen & Rothschild 2010). Hedonic models are often subject to the presence of multicollinearity which can result in measurement errors and negative coefficients (Triplett 2005). Dodds (2011) found that heteroscedasticity was present in the linear hedonic model and applying a natural logarithm to the dependent variable made the error term less heteroscedastic.

Distribution and model selection

The gamma distribution provides a possible alternative to the commonly used log-linear approach to derive hedonic price functions for residential properties. A generalised linear model based on the exponential, 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 & Nelder 1989). 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 Bonyhady (1981) found that that the gamma distribution proved more effective than the log-normal distribution in dealing with the heteroscedasticity. Moran, Solomon, Peisach and Martin (2007) in a study of patients' intensive care units costs, found that cost models employing log-linear models were improved upon by the use of correctly specified generalised linear models, which more effectively modelled the error structure. Fu and Moncher (2004) conducted an analysis on insurance claims in an actuarial study where a generalised linear model was fit to the data. 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. The error structure of a model is an important consideration in modelling data. Examining the error structure through diagnostic plots provides guidance of how well a model fits the data (Murphy, Brockman & Lee 2000). An appealing feature of the use of generalised linear models is that estimates are kept in the natural units of measurement, producing estimations that are more attractive than transformed estimations through log-linear models (Jones 2010).

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 & Weisberg 2011). The bootstrap can be used as a general tool for assessing statistical accuracy (Hastie, Tibshirani, Friedman & Franklin 2005). 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. Wilcox (2008) used bootstrapping as a strategy to determine the correctness of hypothesis tests of coefficients in a multiple regression analysis that was found to be highly effective. Bootstrapping is used to validate the generalised linear model developed in this study.

The research methodology

Objectives of the study

The main objective of this study was to develop a hedonic pricing model for apartments located within three metropolitan KwaZulu-Natal coastal regions. The sub-objectives of this study were as follows:

- To determine an appropriate hedonic price model within these regions based on the distribution of apartment listing prices.
- To develop a model to estimate listing prices of apartments within these regions based on structural and locational attributes.
- To build a software application that facilitates the estimation of listing prices for apartments within these regions, given a set of structural and locational attributes.

The data

The location of the study involved three metropolitan coastal regions within KwaZulu-Natal, South Africa, namely Ballito, Umhlanga, and Durban Central which are subsets of the eThekweni Municipality. Within these regions, thirty-six suburbs were present in the data. Private Property (Pty) Ltd provided the data for the research. The data was a snapshot

of all the apartment listings for sale in Ballito, Durban Central and Umhlanga as at 23 February 2016. Rows with missing values were removed and rows with incorrect geographic coordinates were identified and removed by plotting the data on a map. Duplicate properties were removed based on having the same residential street address to avoid biasing the results. This resulted in a data set of 1314 observations for the study. The variables used this study are presented in Table 1 with a brief description and their respective use.

Listing price distribution and hedonic model

Identifying and testing the distribution of the dependent variable

Probability distributions serve as models for the mechanisms that create observed data (Greene 2003). Choosing the best estimator depends on the statistical properties of the sample distribution, efficiency, unbiasedness and consistency (Greene 2003). Based on this premise, identifying the correct distribution of the apartment listing prices in the data will be a fundamental feature of this study.

Gamma distribution

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

$$fx^{(x)} = \frac{\lambda^{(\alpha)}}{\Gamma(\alpha)} x^{\alpha-1} e^{\lambda x}, x > 0 \quad \text{[Eqn 4]}$$

where the two parameters of interest are the shape α and the scale λ (Kerns 2010). Evident from the gamma probability density function is that the gamma distribution extends for positive continuous variables greater than zero.

Villaseñor and González-Estrada (2015) devised a new goodness-of-fit test for the gamma distribution based on the ratio of the sample variance and the moment estimators. A Monte Carlo simulation provided evidence of the efficiency of the goodness-of-fit test. The Villaseñor and González-Estrada test was applied in this study to determine whether the gamma distribution was appropriate for the dependent variable.

TABLE 1: Study variables.

Variable	Description	Use
Listing price	The price for which an apartment is listed for sale on the Private Property website	Generalised linear model response variable
Bedrooms	The number of bedrooms for a given apartment	Generalised linear model independent variable
Bathrooms	The number of bathrooms for a given apartment	Generalised linear model independent variable
Size	The size in square metres for a given apartment	Generalised linear model independent variable
Suburb	The suburb in which a given apartment is located	Generalised linear model independent variable
Latitude	The latitude coordinate of the street address for a given apartment	Test for spatial autocorrelation
Longitude	The longitude coordinate of the street address for a given apartment	Test for spatial autocorrelation
Street address	The street address of a given apartment	Identification and removal of duplicate property entries

Generalised linear model using a gamma distribution and log-link function

Generalised linear models use the iterative reweighted least squares algorithm to obtain maximum likelihood estimates of model parameters for observations that belong to an exponential distribution family, where the systematic effects can be made linear through a link function (Nelder & Wedderburn 1972). Estimation and inference of generalised linear models are based on maximum likelihood estimation (McCallullagh & Nelder 1989). Generalised linear models are comprised of three components, namely a random or stochastic component, a systematic component and a link function (Nelder & Wedderburn 1972). The notation of the generalised linear model given by Lindsey (1997) is expressed as:

$$\eta_i = g_i(\mu_i) = x_i^T \beta \quad \text{[Eqn 5]}$$

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

A critical part of applied statistical modelling is checking model assumptions (Wood 2006). Residual analysis is paramount to assess how the model fits the data (Muchabaiwa 2013). For generalised linear models the checking of the assumed mean variance relationship is more difficult, which is why the raw residuals are not examined but rather standardised deviance residuals (Wood 2006). 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 standardised deviance residuals should be homogeneous (Murphy et al. 2000). Standardised deviance residuals are used to ascertain goodness of fit for generalised linear models where the standardised deviance residuals plotted against the fitted values should be homogenous (Carruthers et al. 2008). A lack of homogeneity of the standardised deviance residuals may arise if one or more important covariates are not accounted for or if an incorrect error distribution is specified due to an inappropriate link function (Carruthers et al. 2008).

Generalised linear models compare the saturated model with n parameters to the null or intercept only model through the deviance, an important measure of goodness of fit (Mc Cullagh & Nelder 1989). The analysis of deviance compares the null deviance to the residual deviance where a lower residual deviance is evidence of a better fit. This translates to whether the model with n parameters reduces the deviance more than a model with a single parameter.

Spatial autocorrelation

The presence of spatial autocorrelation in the residuals of a statistical model has the propensity to increase Type I 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 standardised Mantel statistic similar to the Pearson correlation coefficient (Borcard & Legendre 2012). The Mantel test is formulated as follows:

$$Z_m = \Sigma_{i=1}^{n} \Sigma_{i=1}^{n} g_{ij} \times d_{ij} \quad \text{[Eqn 6]}$$

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.

Multicollinearity

Multicollinearity 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 multicollinearity, where a value greater than 10 is indicative of multicollinearity (Kennedy 1985). By testing for the presence of multicollinearity, correct model specification and results can be obtained, which was a primary initiative of all the modelling done in this study. The variance inflation factor is calculated as follows:

$$VIF = \frac{1}{r_{(x1)}^2} \quad \text{[Eqn 7]}$$

Whereas $r^2_{(x1)}$ tends towards 1, the variance inflation factor (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 multicollinearity between $X2$ and $X3$ (Gujarati 2004).

Bootstrapping

Bootstrapping accounts for variance in the parameters estimated by drawing many repeated samples (Greene 2003). This technique was used as a general tool for assessing statistical accuracy in this study. The notation for bootstrapping is as follows:

$$\overline{T}^* = \widehat{E^*}\left(T^*\right) = \frac{\sum_{b=1}^{R} T_b^*}{R} \quad \text{[Eqn 8]}$$

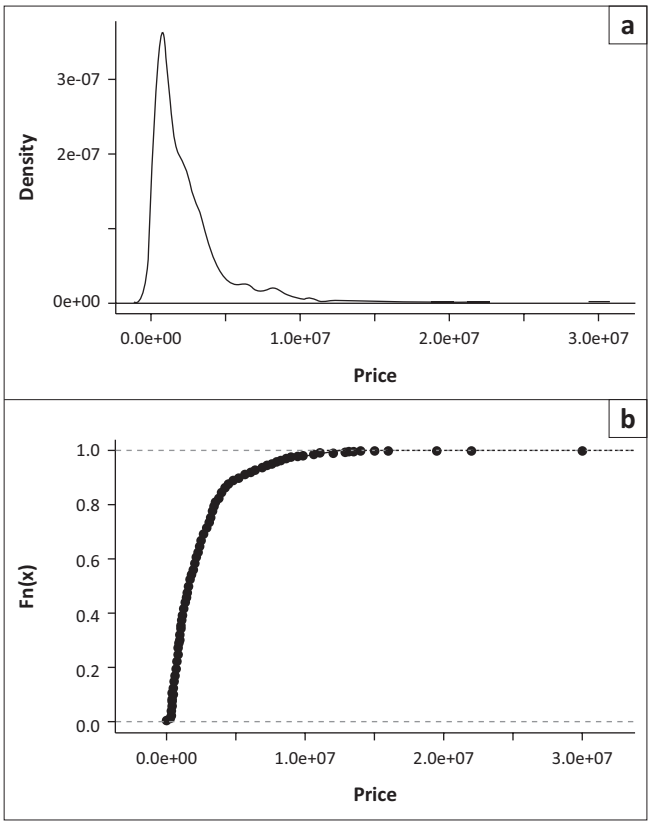
where $\overline{T}$* is the estimator or averaged bootstrapped estimate derived by $\frac{\sum_{b=1}^{R} T_b^*}{R}$, where R is the number of bootstraps applied (Fox & Weisberg 2011).

Key findings and discussion

Gamma distribution

The gamma distribution is suitable for non-negative continuous data. Figure 1 illustrates the kernel density estimator and cumulative density function.

The moment-generating function facilitates the accurate identification of a distribution and computes the respective moments (Kerns 2010). To ascertain whether the distribution of the listing prices is indeed gamma, the shape and scale parameters presented in Table 2 are determined using the matching moments method.



Note: N = 1314; Bandwidth = 3.81e+05.

FIGURE 1: (a) Kernel density estimator and (b) empirical cumulative density function of apartment prices.

TABLE 2: Shape and scale parameters for the apartment price distribution.

Variables	Estimate
Shape	0.9094943994262
Scale	0.0000003774355

To test that the apartment prices were gamma distributed, the Villaseñor and González-Estrada (2015) test was applied using the shape and rate parameters obtained from the matching moments function. The p-value was 0.3857, indicating that there was sufficient evidence not to reject the null hypothesis that the apartment prices follow a gamma distribution.

Variance inflation factor

The VIF presented in Table 3 was computed to test for the presence of multicollinearity in the independent variables which will adversely affect the model results. Evident from the VIF results is that there was no multicollinearity between the set of independent variables used in this study with the highest score being approximately 3.59.

The results of the generalised linear model based on the gamma distribution and log-link function are presented in Table 4. The results show that all the coefficients are statistically significant including all the levels of the dummy locational variable (suburbs). Moreover, the residual deviance

is less than null deviance indicating that the saturated model is a better fit than the null model.

Interpretation of the coefficients of the generalised linear model:

- A 1% increase in square metres or size of an apartment increases the price by approximately 0.733%.
- A one-unit increase in the number of bathrooms of an apartment increases the price by approximately 20.3%.
- A one-unit increase in the number of bedrooms of an apartment increases the price by approximately 3.85%.

TABLE 3: Variance inflation factor results.

Variables	VIF
Size	3.23
NumBeds	2.89
NumBaths	3.59
Suburb	1.09

NumBeds, number of bedrooms; NumBaths, number of bathrooms; VIF, variance inflation factor.

TABLE 4: Generalised linear model output.

Coefficients	Estimate	SE	Pr(> |t|)
(Intercept)	9.46220	0.15252	***
log(Size)	0.73296	0.03107	***
NumBaths	0.20298	0.01878	***
NumBeds	0.03851	0.01830	*
Suburb_Ballito	1.16161	0.10722	***
Suburb_Beachfront	0.58267	0.10742	***
Suburb_Berea	0.74698	0.10904	***
Suburb_Brettenwood Coastal Estate	1.17038	0.23149	***
Suburb_Carrington Heights	0.48521	0.17941	**
Suburb_Congella	-0.43305	0.17914	*
Suburb_Dunkirk Estate	0.80940	0.14942	***
Suburb_Durban CBD	0.23417	0.11202	*
Suburb_Esplanade	0.27414	0.11193	*
Suburb_Essenwood	0.83692	0.13930	***
Suburb_Glenwood	0.54902	0.10918	***
Suburb_La Lucia	1.33553	0.11921	***
Suburb_Morningside	0.71846	0.10923	***
Suburb_Mt Edgecombe	1.07865	0.14735	***
Suburb_Musgrave	0.71563	0.11761	***
Suburb_New Town Centre Gateway	1.28471	0.11707	***
Suburb_Overport	0.41942	0.13919	**
Suburb_Palm Lakes Estate	0.76619	0.15888	***
Suburb_Point Waterfront	1.14713	0.10878	***
Suburb_Prestondale	0.95626	0.23121	***
Suburb_Salt Rock	0.98652	0.14048	***
Suburb_Seaward Estates	0.55325	0.18071	**
Suburb_Shakas Rock	1.19157	0.11020	***
Suburb_Sheffield Beach	0.89679	0.11114	***
Suburb_Sherwood	0.53323	0.23222	*
Suburb_Simbithi Ballito	0.86659	0.11347	***
Suburb_Sunningdale	0.75334	0.23265	**
Suburb_Sydenham	0.43097	0.16861	*
Suburb_Tinley Manor and surrounds	0.85732	0.23219	***
Suburb_Umbilo	0.32250	0.12199	**
Suburb_Umhlanga Ridge	1.38917	0.11254	***
Suburb_Umhlanga Rocks	1.71270	0.10895	***
Suburb_Westridge	0.58978	0.14650	***
Suburb_Windermere	0.77823	0.16723	***
Suburb_Zimbali	1.09251	0.12327	***

NumBeds, number of bedrooms; NumBaths, number of bathrooms; CBD, central business district, SE, standard deviation.

***, $p < 0.001$;**, $p < 0.01$;*, $p < 0.05$.

Null deviance: 1091.47 on 1313 degrees of freedom.

Residual deviance: 104.44 on 1275 degrees of freedom.

- Each suburb coefficient is the percentage difference between the reference suburb.

The signs for all the coefficients are as expected based on *a priori* expectations, unlike what was experienced in the study by Dodds (2011). The number of bedrooms is statistically significant, which was not the case in the study of Els and Von Fintel (2010). The generalised linear model results show that there are 35 suburb coefficients, yet there is a total of 36 in the data. This is because one of the suburbs was withheld from the model output that all the other suburbs were compared to. The suburb that was withheld is Albert Park, in Durban Central, a comparatively low-priced suburb. The suburb coefficients presented in the results of the model will always be in comparison with the Albert Park suburb.

The first plot in Figure 2 is the jacknife standardised deviance residuals against the fitted values, indicating homogeneous

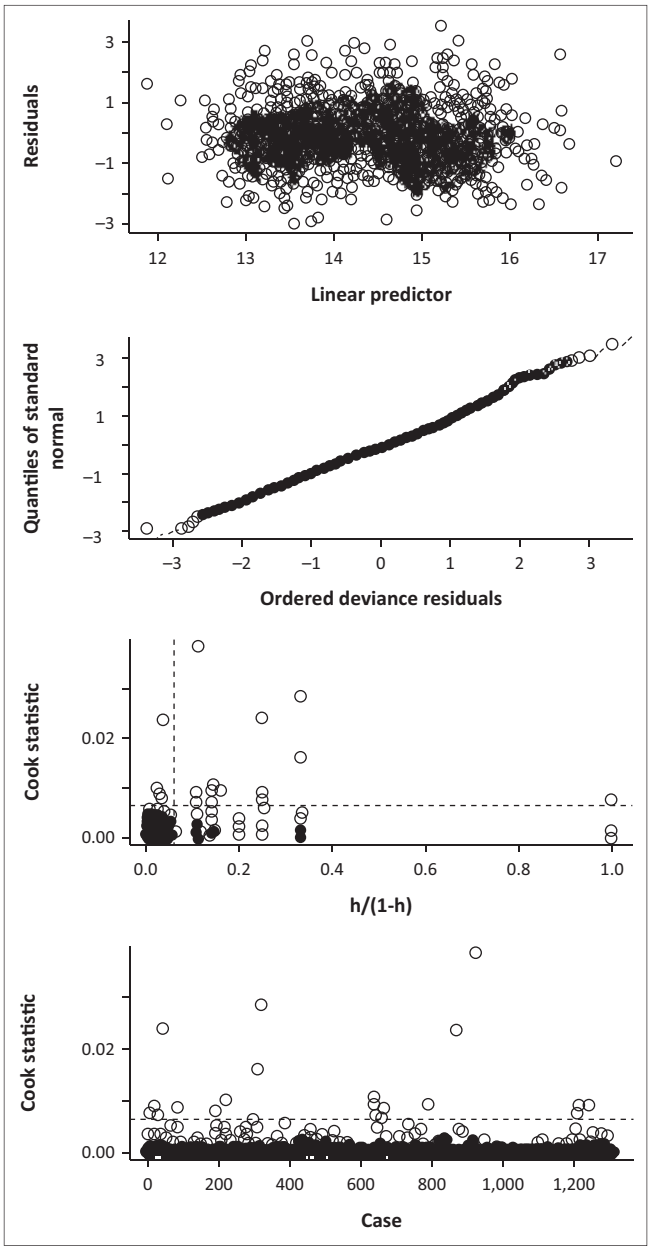


FIGURE 2: Generalised linear model residual plots.

variance and no curvilinear pattern of the standardised deviance residuals. The second plot is of the standardised deviance residuals, indicating normality thereof. The bottom two plots relate to observations with high influence. Upon inspection of these data points, no clear outliers are evident, indicating that these points may have high degrees of leverage or influence on the model.

Bourassa et al. (2010) found that property price predictions were more accurate when submarket dummy locational variables were used and the presence of spatial autocorrelation was mitigated. In order to formally test for the presence of spatial autocorrelation the Mantel test was computed. Distances between the apartments were computed in kilometres and saved as a distance matrix. The distance matrix was then passed to the Mantel test function along with a matrix of the residuals from the generalised linear model. The Mantel test then tested the null hypothesis of no relationship between the residuals and the distance matrix, providing a test of spatial autocorrelation. The correlation between the distance matrix and the residual matrix was −0.02411 and the *p*-value was 0.999, providing sufficient evidence not to reject the null hypothesis of no spatial autocorrelation. The use of the Mantel test to detect the presence of spatial autocorrelation in this study was consistent with the assertion of Borcard and Legendre (2012) and Diniz-Filho et al. (2013).

Reliability of results: Bootstrapping

Bootstrapping was used as a means of model validation where 5000 random samples were drawn with replacement. A model was then developed for each sample and the average of the 5000 models computed was performed. The results indicate that bootstrapped coefficients are similar to the generalised linear model coefficient which is evident by comparing the column headed 'original' with the column headed 'bootMed' in Table 5. Simulating many random sampling distributions provides a measure of reliability for

TABLE 5: Bootstrapped generalised linear model.

Variables	Original	BootBias	BootSE	BootMed
(Intercept)	9.462201	-0.008291175	0.139005	9.449233
log(Size)	0.732960	0.001169352	0.033635	0.734793
NumBaths	0.202984	0.000409610	0.021085	0.203209
NumBeds	0.038513	-0.000670899	0.019742	0.038008
Suburb_Ballito	1.161614	0.003347619	0.068875	1.164911
Suburb_Beachfront	0.582668	0.002622434	0.076079	0.584300
Suburb_Berea	0.746979	0.003865832	0.069661	0.749166
Suburb_Brettenwood Coastal Estate	1.170375	0.002221222	0.071806	1.172913
Suburb_Carrington Heights	0.485213	0.003705877	0.067818	0.490015
Suburb_Congella	-0.433046	-0.013349784	0.190698	-0.437814
Suburb_Dunkirk Estate	0.809397	-0.000937352	0.099751	0.813618
Suburb_Durban CBD	0.234167	0.002723281	0.072944	0.237637
Suburb_Esplanade	0.274145	0.002687679	0.067512	0.276546
Suburb_Essenwood	0.836923	0.001812316	0.080513	0.835307
Suburb_Glenwood	0.549024	0.003786160	0.068967	0.553202
Suburb_La Lucia	1.335534	0.003588477	0.088316	1.341386
Suburb_Morningside	0.718460	0.003579625	0.068544	0.721921
Suburb_Mt Edgecombe	1.078650	0.001625004	0.147574	1.085874
Suburb_Musgrave	0.715626	0.000504024	0.081211	0.717503
Suburb_New Town Centre Gateway	1.284705	0.003353466	0.072259	1.287007
Suburb_Overport	0.419416	0.001347456	0.101720	0.423324
Suburb_Palm Lakes Estate	0.766187	0.001666709	0.087186	0.769822
Suburb_Point Waterfront	1.147131	0.003243269	0.069162	1.152322
Suburb_Prestondale	0.956261	0.003711508	0.068127	0.957237
Suburb_Salt Rock	0.986517	-0.004324199	0.152297	0.987189
Suburb_Seaward Estates	0.553246	-0.000642886	0.099687	0.554347
Suburb_Shakas Rock	1.191571	0.002634724	0.070973	1.192938
Suburb_Sheffield Beach	0.896794	0.003826300	0.069548	0.901282
Suburb_Sherwood	0.533233	0.001232588	0.105212	0.533948
Suburb_Simbithi Ballito	0.866593	0.002369934	0.072558	0.869840
Suburb_Sunningdale	0.753337	0.004092389	0.073277	0.758305
Suburb_Sydenham	0.430975	-0.003935970	0.162656	0.430606
Suburb_Tinley Manor and surrounds	0.857316	0.003783832	0.066366	0.861646
Suburb_Umbilo	0.322495	0.001902980	0.083675	0.324751
Suburb_Umhlanga Ridge	1.389172	0.002586008	0.071693	1.392348
Suburb_Umhlanga Rocks	1.712698	0.002458290	0.074615	1.715394
Suburb_Westridge	0.589776	0.002838880	0.086824	0.594515
Suburb_Windermere	0.778229	-0.001156938	0.097210	0.782636
Suburb_Zimbali	1.092508	0.000032188	0.099351	1.091839

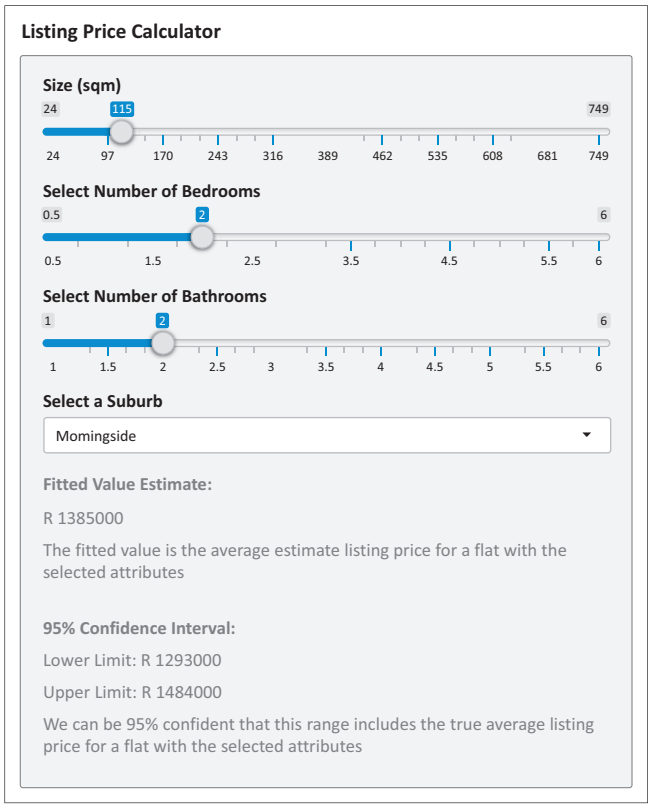
NumBeds, number of bedrooms; NumBaths, number of bathrooms; BootBias, represents the difference between the average bootstrapped value of the statistic and its original sample value; BootSE, are the bootstrap estimates of the standard errors which are computed as the standard deviation of the bootstrap replicates; BootMED, bootstrap estimate; CBD, central business district.

the generalised linear model parameter estimates where the results are consistent and represent a valid reflection of the data. These results coincide with the views of Carruthers et al. (2008) and Hastie et al. (2005) where the bootstrap methodology can be used as a general tool for assessing statistical accuracy.

Software application

A software application was built as the final objective of this study to render the results of the generalised linear model through a user interface, named the listing price calculator. The input parameters on the listing price calculator are dynamic so that the user can select the size of the apartment, the number of bedrooms, the number of bathrooms and the suburb. Each of these input parameters allows for values to be selected that are within the range of the independent variables used to build the model. The software application then calls the generalised linear model and calculates the average listing price and the 95% confidence interval. Figure 3 details the average price for a flat in Morningside in the Durban Central region that is 115 square metres in size (floor area) with two bedrooms and two bathrooms as 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 we can be 95% confident that this range includes the true average listing price.

Location is indeed an important determinant of listing prices of residential apartments. For example, using the listing price calculator one can estimate the listing price of an apartment in Umhlanga Rocks, within the Umhlanga submarket, that is 115 square metres in size (floor area) with two bedrooms and two bathrooms as R3 744 000; this is 170% more than an apartment with similar attributes located in Morningside (R1 385 000). This listing price calculator can add value to households wishing to sell their apartments, where they can obtain an understanding of market pricing dynamics and obtain listing price estimates. Furthermore, credit providers such as banks can use this tool to assess how an apartment is priced relative to the market to determine the fair market value of the asset. Real estate agencies could use this software application as a tool to value new apartments and determine listing prices that are congruent to the general market consensus.



Note: All figures are rounded to the nearest thousand.

FIGURE 3: Listing price calculator.

Conclusion and recommendations

This study bridges the gap between academia and business by creating a software application that may be used by real estate buyers and sellers to estimate listing prices of apartments. A data set of 1314 residential apartments in KwaZulu-Natal, South Africa, was used to develop an econometric model to estimate listing prices. This study develops a generalised linear model based on the gamma distribution and log-link function to derive a hedonic price function for residential apartments. Size, number of bedrooms, number of bathrooms and a dummy variable for suburb (location) are statistically significant. The reliability of the models' results was tested using non-parametric bootstrapping which provided a good measure of model validation by introducing variance through re-sampling with replacement. The coefficients of the bootstrapped model were similar to the original model, indicating that the original model results were reliable. Further research is required to determine how generalisable the statistical framework presented in this study is. Future research should include the use of the statistical framework propounded in this study across different geographic regions and across different residential property types. Further research could attempt to link the modelling framework presented in this study with pooled cross sectional data to develop a residential property price index.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

D.B. contributed to the conceptual design of the study objectives and research methodology, interpretation of results and writing of the article. D.B. cleaned the data, using the R language to analyse the data and write the software application. This article is compiled from D.B.'s MBA dissertation. M.G.C. contributed to the conceptual design of the study objectives, research methodology and software

application, as well as interpretation of results, supervision of the study, writing and editing of the article. M.G.C. was D.B.'s dissertation supervisor.

Ethical considerations

Ethical clearance for this study was obtained from the Humanities Social Sciences Research Ethics committee at the University of KwaZulu-Natal, protocol reference number: HSS/0209/016M.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Disclaimer

The views expressed in this article are the authors' own and not an official position of the University of KwaZulu-Natal.

References

Anselin, L., 2013, *Spatial econometrics: Methods and models*, p. 4, Springer Science & Business Media, New York.

Battese, G.E. & Bonyhady, B.P., 1981, 'Estimation of household expenditure functions: An application of a class of heteroscedastic regression models', *Economic Record* 57(1), 80–85, viewed 13 February 2016, from http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/.

Borcard, D. & Legendre, P., 2012, 'Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study', *Ecology* 93(6), 1473–1481, viewed 13 February 2016, from http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/.

Bourassa, S., Cantoni, E. & Hoesli, M., 2010, 'Predicting house prices with spatial dependence: A comparison of alternative methods', in University of Louisville and University of Geneva (eds.), *15th Conference of the Pacific Rim Real Estate Society Proceedings*, Sydney, Australia, January 18–21, Journal of Real Estate Research, Clemson, SC.

Bromideh, A.A. & 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), viewed 05 March 2016, from http://www.pjsor.com/index.php/pjsor/article/view/487.

Carruthers, E., Lewis, K., Mccue, T. & Westley, P., 2008, 'Generalized linear models: Model selection, diagnostics, and overdispersion', Memorial University of Newfoundland, Unpublished, viewed 13 February 2016, from http://www.mun.ca/biology/dschneider/b7932/B7932Final4Mar2008.pdf.

Chen, C.F. & Rothschild, R., 2010, 'An application of hedonic pricing analysis to the case of hotel rooms in Taipei', *Tourism Economics* 16(3), 685–694, viewed 13 February 2016, from http://ir.lib.ncku.edu.tw/.

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*, viewed 07 February 2016, from http://www.cserge.ac.uk/.

De Haan, J. & Erwin, D., 2011, *Handbook on residential property price indices*, Eurostat European Commission, viewed 12 February 2016, from http://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF.

Diniz-Filho, J.A.F., Soares, T.N., Lima, J.S., Dobrovolski, R., Landeiro, V.L., Telles, M.P.D.C. et al., 2013, 'Mantel test in population genetics', *Genetics and Molecular Biology* 36(4), 475–485, viewed 02 April 2016, from http://www.scielo.br/pdf/gmb/v36n4/v36n4a02.pdf.

Dodds, R.S., 2011, *An investigation into the hedonic price analysis of the structural characteristics of residential property in the West Rand*, viewed 02 April 2016, from http://wiredspace.wits.ac.za/.

Dormann, C.F., Mcpherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G. et al., 2007, 'Methods to account for spatial autocorrelation in the analysis of species distributional data: A review', *Ecography* 30(5), 609–628, viewed 02 April 2016, from http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/.

Els, M. & 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), 418–436, viewed 30 January 2016, from http://onlinelibrary.wiley.com.ukzn.idm.oclc.org/.

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, viewed 02 April 2016 from http://cran.rproject.org/doc/contrib/Fox-Companion/appendix-bootstrapping.

Fox, J. & Weisberg, S., 2011, *An R companion to applied regression*, 2 edn., Sage, Thousand Oaks, CA.

Fu, L. & Moncher, R.B., 2004, *Severity distributions for GLMs: Gamma or lognormal? Evidence from Monte Carlo simulations*, Casualty Actuarial Society Discussion Paper Program 149–230, viewed n.d., from https://www.casact.org/pubs/dpp/dpp04/04dpp149.pdf.

Gandy, A. & Kvaløy, J.T., 2013, 'Guaranteed conditional performance of control charts via bootstrap methods', *Scandinavian Journal of Statistics* 40(4), 647–668. https://doi.org/10.1002/sjos.12006

Goodman, A.C., 1978, 'Hedonic prices, price indices and housing markets', *Journal of Urban Economics* 5(4), 471–484. https://doi.org/10.1016/0094-1190(78)90004-9

Greene, W.H., 2003, *Econometric analysis*, 5th edn., Prentice Hall, Upper Saddle River.

Gujarati, D.N., 2004, *Basic econometrics*, 4th edn., Tata McGraw-Hill, New York.

Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J., 2005, 'The elements of statistical learning: Data mining, inference and prediction', *The Mathematical Intelligencer* 27(2), 83–85. https://doi.org/10.1007/BF02985802

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

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

Jiang, L., Phillips, P.C. & Yu, J., 2015, 'New methodology for constructing real estate price indices applied to the Singapore residential market', *Journal of Banking & Finance* 61, S121–S131. https://doi.org/10.1016/j.jbankfin.2015.08.026

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

Kennedy, P.E., 2005, 'Oh no! I got the wrong sign! What should I do?', *The Journal of Economic Education* 36(1), 77–92. https://doi.org/10.3200/JECE.36.1.77-92

Kerns, G.J., 2010, *Introduction to probability and statistics using R*, 1st edn., UPSUR, Youngtown, viewed 02 February 2016, from https://cran.rproject.org/web/packages/IPSUR//vignettes/IPSUR.pdf.

Liao, W.C. & Wang, X., 2012, 'Hedonic house prices and spatial quantile regression', *Journal of Housing Economics* 21(1), 16–27. https://doi.org/10.1016/j.jhe.2011.11.001

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

Lyons, R.C., 2015, *Measuring house prices in the long run: Insights from Dublin, 1900–2015*, viewed 29 April 2016, from http://eh.net/eha/wp-content/uploads/2015/05/Lyons.pdf.

Malpezzi, S., 2003, *Hedonic pricing models: A selective and applied review*, Section in Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan, Auckland.

Mccullagh, P. & Nelder, J.A., 1989, *Generalized linear models*, vol. 37, CRC Press, London.

Moran, J.L., Solomon, P.J., Peisach, A.R. & Martin, J., 2007, 'New models for old questions: Generalized linear models for cost prediction', *Journal of Evaluation in Clinical Practice* 13(3), 381–389. https://doi.org/10.1111/j.1365-2753.2006.00711.x

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

Murphy, K.P., Brockman, M.J. & Lee, P.K., 2000, 'Using generalized linear models to build dynamic pricing systems', in D.L. Lange (ed.), *Casualty actuarial society forum*, pp. 107–139, Winter, Colortone Press, Landover, MD.

Nelder, J.A. & Wedderburn, R.W.M., 1972, 'Generalized linear models', *Journal of the Royal Statistical Society Series A* 135, 370–384.

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

Özyurt, S., 2014, *Spatial dependence in commercial property prices: Micro evidence from the Netherlands*, viewed 17 April 2016, from https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1627.pdf?f4b7432ba10e5d1553a255b587a09d23.

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

Rosen, S., 1974, 'Hedonic prices and implicit markets: Product differentiation in pure competition', *Journal of Political Economy* 82(1), 34–55. https://doi.org/10.1086/260169

Shulz, R. & Werwatz, A., 2004, 'A state space model for Berlin house prices: Estimation and economic interpretation', *The Journal of Real Estate Finance and* Economics 28(1), 37–57.

South African Reserve Bank (SARB), 2015, *South African Reserve Bank Quarterly Bulletin*, South African Reserve Bank, viewed 12 February 2016, from https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6649/01Full%20Quarterly%20Bulletin%20%E2%80%93%20March%202015.pdf.

Stohldreier, M.T., 2012, *The determinants of house prices in Chinese cities*, viewed 16 February 2016, from http://www.econ.uzh.ch/ipcdp/theses/MA_MarieStohldreier.pdf.

Thayn, J.B. & 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), 47–66. https://doi.org/10.1080/00045608.2012.685048

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, Paris.

Villaseñor, J.A. & González-Estrada, E., 2015, 'A variance ratio test of fit for gamma distributions', *Statistics & Probability Letters* 96, 281–286. https://doi.org/10.1016/j.spl.2014.10.001

Wilcox, R.R., 2008, 'Post-hoc analyses in multiple regression based on prediction error', *Journal of Applied Statistics* 35(1), 9–17. https://doi.org/10.1080/02664760701683288

Wood, S., 2006, *Generalized additive models: An introduction with R*, CRC Press, Boca Raton, FL.