




# A gamma generalised linear model as an alternative to log linear real estate price functions



## Authors:

Dane Bax<sup>1</sup>   
 Temesgen Zewotir<sup>1</sup>   
 Delia North<sup>1</sup> 

## Affiliations:

<sup>1</sup>School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

## Corresponding author:

Dane Bax,  
 danebax@gmail.com

## Dates:

Received: 28 Apr. 2019  
 Accepted: 02 Aug. 2019  
 Published: 05 Dec. 2019

## How to cite this article:

Bax, D., Zewotir, T. & North, D., 2019, 'A gamma generalised linear model as an alternative to log linear real estate price functions', *Journal of Economic and Financial Sciences* 12(1), a476. <https://doi.org/10.4102/jef.v12i1.476>

## Copyright:

© 2019. The Authors.  
 Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

## Read online:



Scan this QR code with your smart phone or mobile device to read online.

**Orientation:** Residential property markets play an important role in economies, informing policy development and decision-making. However, measuring quality-adjusted growth is difficult because of the heterogeneity of properties. Hedonic regression is frequently used in real estate econometric studies as a quality-adjusted technique to estimate residential property prices for the development of price indices. Log linear models are typically used to derive these hedonic price functions.

**Research purpose:** This article develops hedonic pricing functions using generalised linear models for South African residential property listings over a 5-year period.

**Motivation for the study:** A parametric alternative to the log linear model is investigated to address the limited studies conducted in South Africa. An important feature of this study is the inclusion of different property types and the geographic scope.

**Research approach/design and method:** The data set consisted of 415 200 residential properties from all over South Africa. The data spanned a period from January 2013 to August 2017. Several generalised linear models were developed and compared.

**Main findings:** The gamma generalised linear model provided the best overall fit, generalising well to the unseen validation data. An added benefit of this model is that the estimates were kept on the original scale, avoiding the need for back transformation which is an appealing feature of any model. A dummy locational variable was shown to account for the spatial dependency in the data.

**Practical/managerial implications:** This framework provides property market participants with the ability to quantify the utility derived over the marginal distribution of the physical characteristics of properties. This research presents the groundwork to create a property price index where index number theory could be applied to the counterfactual predicted values obtained from hedonic price models to measure price inflation over time

**Contribution/value-add:** This study analysed the South African residential property market based on an online company's data, purportedly covering the entire market. No real estate hedonic price studies have been identified in South Africa with this level of scope. The gamma generalised linear model is a novel candidate to develop parametric real estate hedonic price functions.

**Keywords:** generalised linear models; real estate economics; model comparison and evaluation; spatial modelling; hedonic price functions.

## Introduction

The importance of measuring residential property price inflation is paramount to households and economies; however, the heterogeneity of properties makes it difficult (De Haan & Diewert 2011). Log linear models have been used extensively in real estate economics to estimate property prices and inflation. This study investigates generalised linear models as an alternative to the typical log linear approach. Cross-sectional hedonic price functions are developed and compared for the South African residential property market over a 5-year period.

## Background and objective

Residential property is an important component of individual and national wealth where it is capitalised on household balance sheets, informing economic policy formulation (Hill 2013). Goodhart and Hoffman (2008) conducted a study providing evidence of relationships between house prices, credit and broad money. Using vector auto-regression fitted with ordinary

least squares, their research showed statistically significant relationships between home prices and the macro economy. Bordo and Jeanne (2002) found an increased likelihood of a financial crisis occurring when real estate prices reached a peak or shortly after a bust, in a study of advanced economies spanning from 1970 to 2001. This resonates with the views of De Haan and Diewert (2011) who assert that sharp declines in home prices can adversely affect the debt to equity ratio and credit ratings. Residential property has an important role in economies and understanding price inflation is imperative; however, measuring price inflation is difficult because of infrequent transactions and the heterogeneity of properties.

Hedonic regression is ubiquitous in the construction of residential property price indices where log linear models are commonly developed in the price estimation procedure (De Haan & Diewert 2011; Jiang et al. 2015). Hedonic regression has been found useful as a quality-adjusted methodology where pure price changes are measured and not simply changes in the composition of samples in different periods (Shimizu, Nishimura & Watanabe 2010). Hedonic pricing measures the price of an item through its utility bearing characteristics where the price of the item is determined by the vector of its characteristics (Rosen 1974). Hedonic pricing describes the functional relationship of a heterogeneous item and the implicit attributes:

$$P_j = P(Z_j) = P(Z_{j1}, Z_{j2}, \dots, Z_{jk}) \quad [\text{Eqn 1}]$$

where  $P_j$  is the price of the  $j$ th item which is a function of a set of characteristics  $Z_j$  (Goodman 1978). Hedonic pricing is useful when estimating the price of heterogeneous goods. Heterogeneous or differentiated goods are goods that differ in their respective composition of characteristics; however, consumers consider the set of characteristics closely related, defining it as a single item (Day 2003). Hedonic pricing mathematically models residential property prices as a function of structural and location characteristics (Lyons 2015). Ordinary least squares models are typically employed to estimate the marginal contributions of each characteristic, taking the form:

$$\theta(\mu) = X\beta \quad [\text{Eqn 2}]$$

where  $\beta$  is  $p < n$  unknown parameters, the matrix  $X_{n \times p}$  is a set of known independent variables and  $X\beta$  is the linear structure (Lindsey 2005). The implicit price for characteristic  $i$  of property  $j$  is calculated by taking the partial derivative. Because of the positive domain and positively skewed nature of residential property prices, log linear models are often adopted as a quality-adjusted technique that controls for changes in the quality of properties transacted in different periods, whilst reducing heteroscedasticity in the residuals (Silver 2016). Day (2003); Bourassa, Cantoni and Hoelis (2007); and Els and Von Fintel (2010) conducted separate hedonic price studies for different property markets using log linear models. Els and Von Fintel (2010) found that the assumption of the linear functional form was violated, finding quantile regression more appropriate to capture the hedonic price function. A potential problem with transforming

property prices to the log scale is that exponentiation of the fitted values produces geometric mean estimates and not arithmetic mean estimates (Olivier, Johnson & Marshall 2008). Another potential concern is the assumption that property prices are lognormal when a different distribution family may represent the data better. Generalised linear models incorporate exponential classes of distribution families, which facilitate modelling the response on the original scale. The objective of this study is to investigate generalised linear models as an alternative framework to the log linear model in the development of hedonic price functions for the South African residential property market.

## Generalised linear models

Generalised linear models are a natural extension of classical linear models where properties such as linearity and computing parameter estimates are similar (McCullagh & Nelder 1989). Generalised linear models are characterised by three components. Firstly, a stochastic or random component representing a response variable  $y$ , consisting of independent observations  $(y_1, y_2, \dots, y_n)$ , belonging to a class of an exponential family distribution in the form of:

$$f(y; \theta, \varnothing) = \exp \left\{ \frac{\theta y - b(\theta)}{\varnothing} + c(y, \varnothing) \right\} \quad [\text{Eqn 3}]$$

where  $\varnothing$  is a dispersion parameter and  $b(\cdot)$ ,  $c(\cdot)$  are known functions and the range of  $Y$  does not depend on  $\theta$  or  $\varnothing$ . For a random response variable  $Y$  with distribution of form  $3 E(y) = \mu$ . Secondly, a systematic component that consists of a set of covariates  $(x_1, x_2, \dots, x_p)$  which combine linearly with the coefficients to produce the linear predictor  $\eta$ . Therefore,  $\eta = \beta X$ . Finally, a link function that connects the stochastic and systematic components where  $\eta = \mu$ .

This generalisation takes the form:

$$\eta_i = g(\mu_i) \quad [\text{Eqn 4}]$$

where  $g(\cdot)$  denotes the link function and  $\eta = \mu$  through the link function. The link function relates the conditional mean to the systematic component, namely the covariates. This formulation allows for the exponential family of distributions including normal; however, the link function may become any monotonic differentiable function, which then allows extensions to distributions such as Poisson, binomial and gamma amongst others (McCullagh & Nelder 1989). This means that generalised linear models are suitable for modelling continuous data as well as count and binary data.

Generalised linear models obtain maximum likelihood estimates of parameters belonging to an exponential distribution family using the iterative reweighted least squared algorithm where the link function makes the systematic effects linear (Nelder & Wedderburn 1972). Maximum likelihood estimates are a vector of parameter estimates produced by a model function which makes the observed data probable given the model function (Lindsey 2005).

The primary goodness-of-fit measure for generalised linear models is called the deviance which is the logarithm of a ratio of likelihoods (McCullagh & Nelder 1989). The analysis of deviance makes model assessment and comparison possible in terms of the choice of covariates. Given a set of data, two extreme models are possible. Firstly, a null model with one parameter which represents a common  $\mu$  for all the  $y$ s. Secondly, a complete model where all the  $y$ s are different, matching the data completely. Fitting a model with more than one parameter represents a saturated model that can be compared to the null model (Dobson & Barnett 2018). The fitting of  $n$  parameters is performed by maximising the likelihood of matching the model to the likelihood of the data through the deviance that differs based on the distribution.

For the normal distribution, the deviance is simply the sum of squares just like ordinary least squares which means that fitting a normal or lognormal distribution with the identity link function, where the natural logarithm of the response is taken, is equivalent to fitting a linear or log linear ordinary least squares model. For generalised linear models, the saturated model should have a lower deviance than the null model, indicating that the inclusion of  $n$  parameters is a better fit. Guisan and Zimernam (2000) propose that variance reduction in model formulation is generally a desired characteristic of the goodness of fit as with generalised linear models, where deviance reduction can be converted to an equivalent  $R^2$  statistic:

$$D^2 = (\text{Null deviance} - \text{Residual deviance}) / \text{Null deviance} \quad [\text{Eqn 5}]$$

where  $D^2$  is the deviance explained or the amount of deviance accounted for by the model. Naturally, this leads to an understanding of the residuals of generalised linear models where the deviance residuals are reported as a measure of discrepancy. Deviance residuals are calculated as follows:

$$\text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i^2} \quad [\text{Eqn 6}]$$

This formulation shows that deviance residuals are calculated by taking the signed square root of the  $i$ th observation to the total model deviance. One can begin to understand the quality of fit that reflects the choice of the link function and linear predictor using deviance residuals (Nelder & Wedderburn 1972). McCullagh and Nelder (1989) state that through the appropriate link function and linearity of the systematic component, the desired error distribution of the deviance residuals can be achieved which should resemble normal theory residual plots, except for certain plots, like in the case of binomial errors. Standardised deviance residuals are approximately normal which is preferable to Pearson residuals that tend to reflect any skewness of the underlying distribution. Plotting the standardised deviance residuals against the fitted values can provide an informal check of the goodness of fit depending on the type of generalised linear model, where any curvature could suggest the incorrect choice of link function, omitted independent variables or the omission of quadratic terms in the independent variables (Davidson & Snell 1991).

The selection of generalised linear models in this study involved choosing the appropriate distribution of  $Y$  and choosing the relationship between  $\eta$  and  $\mu$ . Three candidate combinations of model families and link functions were fit to the data, specifically the gamma log model, the normal log model and the lognormal identity model. These models are hereafter referred to as the gamma model, the normal model and lognormal model.

## Spatial dependency

Prices of adjacent properties are often related which can lead to correlation in the residuals of regression models, violating the assumption of independence (Bourassa et al. 2007). Spatial autocorrelation or dependency is a challenging problem in real estate modelling where correlation manifests in two-dimensional space unlike serial correlation which is one dimensional. Bourassa et al. (2007) found that the inclusion of a submarket dummy variable accounted for spatial autocorrelation and outperformed geostatistical and lattice approaches. A similar approach was adopted in this study where a factor area variable was included to account for spatial dependence in listing prices.

Variograms were utilised to understand the spatial autocorrelation structure. Variograms display the dissimilarity of observations that vary in space as a function of the distance between them (Ploner 1999). The sill represents spatially autocorrelated sample locations, and the range is where the distance flattens out and the sample locations are no longer spatially autocorrelated. A variogram will be flat when no correlation or low correlation is present which indicates randomness in the structure (Chiles & Delfiner 1999). The nugget effect is an important concept in variograms and describes the variability between observations that are closely spaced which could be inherent in the data or because of the sampling component (Clark 2010). Therefore, in the context of this study, a large nugget effect could be the product of closely clustered properties with similarly signed and order of magnitude residuals that would overestimate the amount of spatial dependency. A prevalent test developed by Moran (1950) is a two-dimensional specification test for spatial autocorrelation, analogous to a test of univariate time series correlation (Anselin 2006):

$$I = \frac{e' W e / S_0}{(e' e) / n} \quad [\text{Eqn 7}]$$

where  $e$  represents the regression model residuals and  $w$  is the spatial weighting matrix and  $s_0$  is the standardisation factor that relates to the sum of weights for the nonzero cross-products. This test is applied to test for the presence of spatial autocorrelation.

## Research design

The open source programming language R was used to perform the statistical analysis in this research (R Core Team 2018). The data were provided by an online residential property portal that aggregates listings from real estate

**TABLE 1:** Description of the data.

Variable	Type	Description	Summary statistics (min; mean; max)
Listing price	Market price	The advertised price of the property	1,000; 2 461 210; 200 000 000
Size	Structural	The size of the structure in square meters	2; 260; 85 102
Lot	Structural	The land size in square meters	2; 1,163; 99 999
Bedrooms	Structural	The number of bedrooms	0; 3; 78
Bathrooms	Structural	The number of bathrooms	0; 2; 78
Property type	Structural	The type of property e.g. house	Not applicable
Suburb	Locational	The suburb the property is located	Not applicable
Province	Locational	The province the property is located	Not applicable
Listing date	Time	The date the property was advertised	Not applicable

Note: The variables have been rounded to the nearest whole number.

agencies throughout South Africa. The period of the data is from January 2013 to August 2017, and Table 1 describes the data set.

The data were inspected for consistency where categorical variables were standardised and data types were constrained to the correct type. Property listings were duplicated by different agencies resulting in many properties being captured more than once. This could result in biased estimates, and therefore duplicate properties were identified and removed using row-wise string matching. Missing values resulted in the removal of observations. The summary statistics show that the spread of the numeric characteristic variables was large with lower and upper bounds that are unlikely.

Real estate agents populate data into automated feeds which could result in incorrect data capturing and anomalous data. To deal with the anomalous data, an autoencoder was developed using the H2O open source machine learning framework (Ledell et al. 2019). An autoencoder is a deep learning neural network aimed at reducing the feature space which can be viewed as a non-linear alternative to principal component analysis (Hastie, Tibshirani & Wainwright 2015). Given enough data, the network will learn the identity of the data via non-linear reduced representation of the original data (Candel et al. 2018). A high reconstruction error for a data point indicates that the data point does not match the learned pattern and is anomalous. Lower limits were set on the listing price, size and lot variables. Listing price was set to  $\geq$  ZAR 200 000; size and lot were set to  $\geq$  35 m<sup>2</sup>. These figures were chosen based on the ABSA Bank property price index which was used as a guideline (Luus 2002). Properties with a reconstruction mean squared error  $\geq$  9.39e-07 were discounted. Therefore, based on the results of the autoencoder, properties in the top 5th percentile of the reconstruction error were treated as anomalous.

The final data set consisted of 415 200 properties, and the spread of the variable distributions was greatly reduced after the anomalous data points were removed, evident in Table 2.

Although the lot variable was used to detect anomalous data, it was discounted for modelling purposes as lot was not

**TABLE 2:** Final data summary statistics.

Variable	Listing price	Size	Lot	Bedrooms	Bathrooms
Minimum	200 000	35	35	1	1
Mean	2 159 173	231	752	3	2
Maximum	19 700 000	2080	10 365	13	12

Note: The variables have been rounded to the nearest whole number.

applicable for the property type apartment and often omitted. Lot was set to the size variable for apartments during the autoencoder learning stage.

The design approach chosen develops separate cross-sectional models for each year in the data which contrasts to building a single model using pooled cross-sectional data. The chosen framework will facilitate the development of a property price index that won't change previous estimates when future periods are introduced. A pooled period approach would result in new samples being added to the original sample, which would change previous estimations when constructing a residential property price index (De Haan & Diewert 2011). Therefore, for each year, various generalised linear models were developed where listing price was regressed on the physical and locational attributes of residential properties. Tabular and graphical summaries of the model fit for each of the candidate models are presented for comparative purposes. Thereafter, based on generalisability and goodness of fit, the best model is selected and expounded upon.

## Ethical considerations

Ethical clearance was obtained from the Research Ethics Committee of the University of KwaZulu-Natal, protocol reference number: HSS/0209/016M.

## Results and discussion

The data were split into two sets, training and validation, where 70% of the data were used for training and 30% of the data were used for validating the models. This was done to test model generalisability on unseen data for the development of future models. The holdout data for a given model provide a more robust estimate of the generalisation error compared to the training error (Blum, Kalai & Langford 1999). Partitioning the data into training and holdout sets for each year involved writing a function to ensure that the splits were random and that the distribution of the response was similar for each split and to the original data. The function ensured that each area factor level was present in each split. Model performance and generalisation was tested using the root mean squared error (RMSE) which is a measure of spread that compares the closeness of the model outcomes to the observed data (Gujarati 2004). A lower RMSE is indicative of less variability between model estimates and the observed data. The Akaike information criterion (AIC) statistics were also computed. When comparing models, the AIC is useful for model selection as it provides an assessment of the quality of different models given a set of data (Greene 2003). A lower AIC is indicative of better fit. Akaike information criterion concomitantly considers goodness of fit using the likelihood

function whilst penalising model complexity through the number of parameters. Model selection was based on a combination of reported statistics, namely deviance explained, holdout RMSE, AIC and model fit based on diagnostic residual plots. Table 3 details the results of each yearly model fit.

Each model produced consistent deviance explained statistics for each year respectively, where the gamma and lognormal models shared the highest amount of deviance explained. Moreover, the gamma and lognormal models appear very similar in terms of holdout RMSE and AIC statistics. The AICs produced by the lognormal models were not directly comparable to the other models as the response variable was on the logarithmic scale. The AICs of the lognormal models were made comparable by subtracting the sum of logarithms of the response variable from the likelihood. Based solely on the AICs, the lognormal models appear to fit the data the best, as they consistently produced the lowest AIC statistics. Considering only the holdout RMSE statistics, the normal model outperformed the two other models with consistently lower RMSE statistics each year. No evidence of overfitting is present as the training and holdout RMSEs are quite similar, indicating that the models generalise to unseen data. This suggests model robustness to the introduction of future periods.

Discerning the best model based solely on the goodness-of-fit measures reported above is difficult, and a graphical examination of the residuals is necessary. The goodness-of-fit residual diagnostic plots for each yearly model are illustrated in Figures 1 and 2 from left to right, beginning with the gamma model, followed by the normal model and finally the lognormal model. Figure 1 presents the residuals versus fitted values where the y-axis represents the deviance residuals and x-axis represents the fitted values. Figure 2 presents the quantile-quantile (Q-Q) plots for normality.

**TABLE 3:** Model summaries.

Year	Deviance explained	Training RMSE	Holdout RMSE	AIC
<b>Gamma model summary statistics</b>				
2013	0.89	708 816	719 286	665 059
2014	0.87	762 918	764 730	1 689 332
2015	0.87	768 004	772 905	1 808 202
2016	0.87	731 159	746 554	2 557 727
2017	0.88	723 854	724 390	1 779 451
<b>Normal model summary statistics</b>				
2013	0.83	666 427	704 715	692 589
2014	0.82	724 525	743 688	1 748 933
2015	0.83	721 649	743 687	1 866 417
2016	0.83	685 171	710 324	2 637 126
2017	0.84	682 036	693 513	1 835 507
<b>Lognormal model summary statistics</b>				
2013	0.89	709 765	716 993	664 303
2014	0.88	766 117	762 544	1 687 050
2015	0.87	767 251	774 243	1 805 578
2016	0.88	727 294	741 349	2 553 572
2017	0.88	724 097	726 167	1 777 340

Note: The deviance explained figures are rounded to two decimal places. The other figures are rounded to the nearest whole number.

RMSE, root mean squared error; AIC, Akaike information criterion.

The fitted versus residual diagnostic plots for the gamma and lognormal models are very similar and do not indicate any discernible pattern in the deviance residuals, one of the required assumptions. However, the normal model shows signs of heteroscedasticity at the upper quantiles, violating the assumption of constant variance.

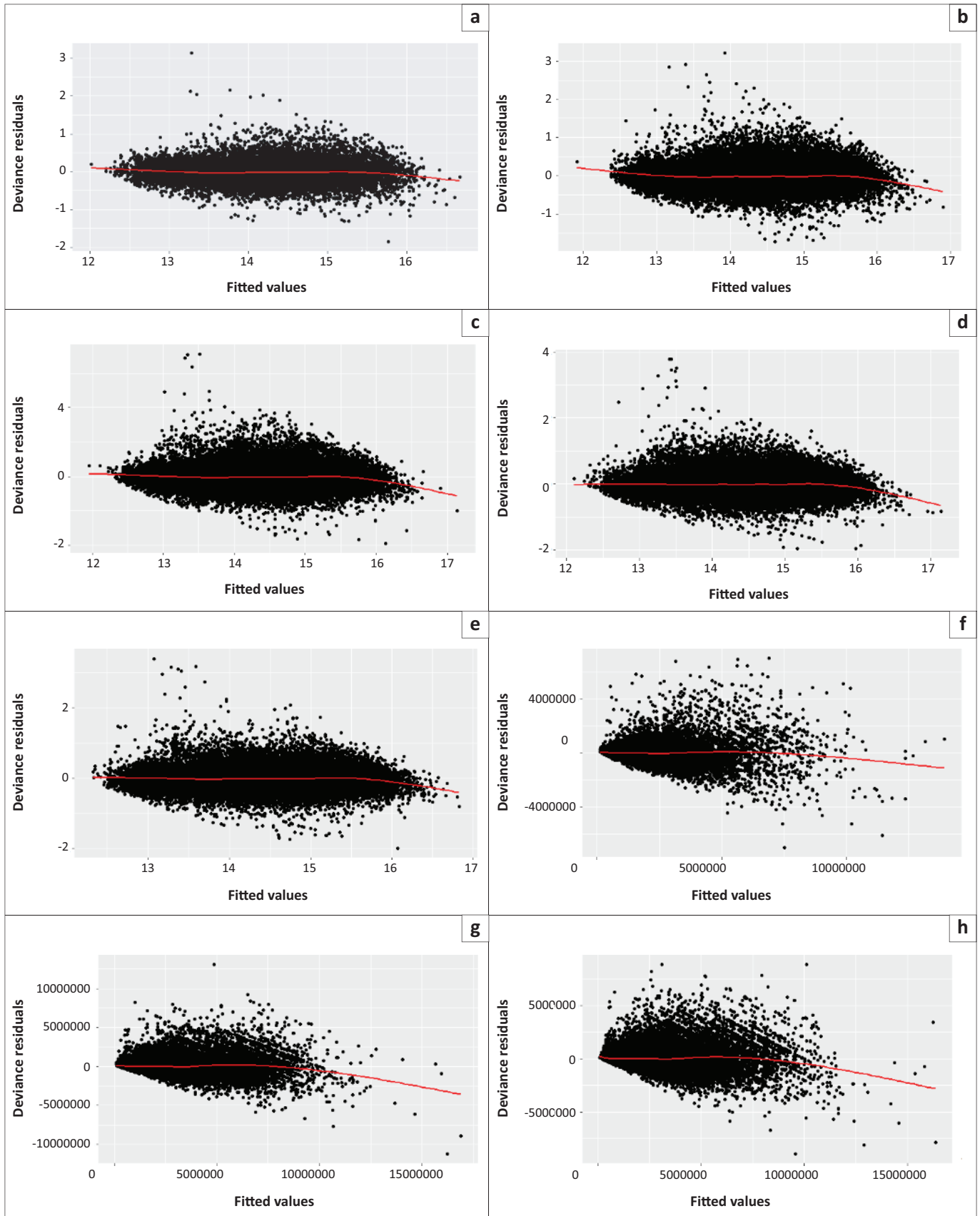
None of the plots is perfectly normal with deviation at the upper and lower quintiles; the S-shaped curves indicate heavy tailed residual distributions. The gamma and lognormal Q-Q plots appear the best behaved in terms of the normality assumption. The normal model appears to fit the data poorly in terms of the diagnostic plots, whilst the gamma and lognormal models appear to represent the data the best.

A possible caveat of using the lognormal model for modelling listing prices is that the expected values are on the log scale and back transformation is necessary. Transforming expected values from the log scale back to the original scale by means of exponentiation results in geometric mean estimates and not arithmetic mean estimates (Olivier et al. 2008). However, the natural logarithm is monotonic, and the back transformed estimates are equivalent to median estimates if the distribution of  $\log(x)$  is symmetric (Musset 2006). An appealing feature of the gamma and normal models is that expected values are kept on the original scale where arithmetic mean expected values are computed. For this reason, the lognormal models are discounted from the candidate model selection. The gamma models are chosen over the normal models based on the diagnostic plots, lower AICs and similar holdout RMSEs. A discussion of the gamma modelling results ensues.

The property type factor variable included six levels, namely apartment, cluster, duplex, house, simplex and townhouse. The property type apartment was used as the reference level, resulting in the other property types being compared to this level. Table 4 tabulates the beta coefficient estimates for each covariate along with the corresponding p-values. To make reporting succinct, Table 4 discounted the area (factor variable) coefficients as there were over 2000 factor levels present in the data that varied between years. The area variable was used as a control variable to account for variability amongst listing prices and to account for the spatial dependency in the data.

The coefficients given by  $\hat{\beta}$  are expressed as percentage effects. The covariates  $\log(\text{Size})$  and the number of bathrooms were consistently statistically significant for each year. The natural logarithm was applied to the size covariate to improve linearity where Figure 3 shows no discernible pattern in the plots indicating this transform was appropriate. The coefficients can be interpreted as follows:

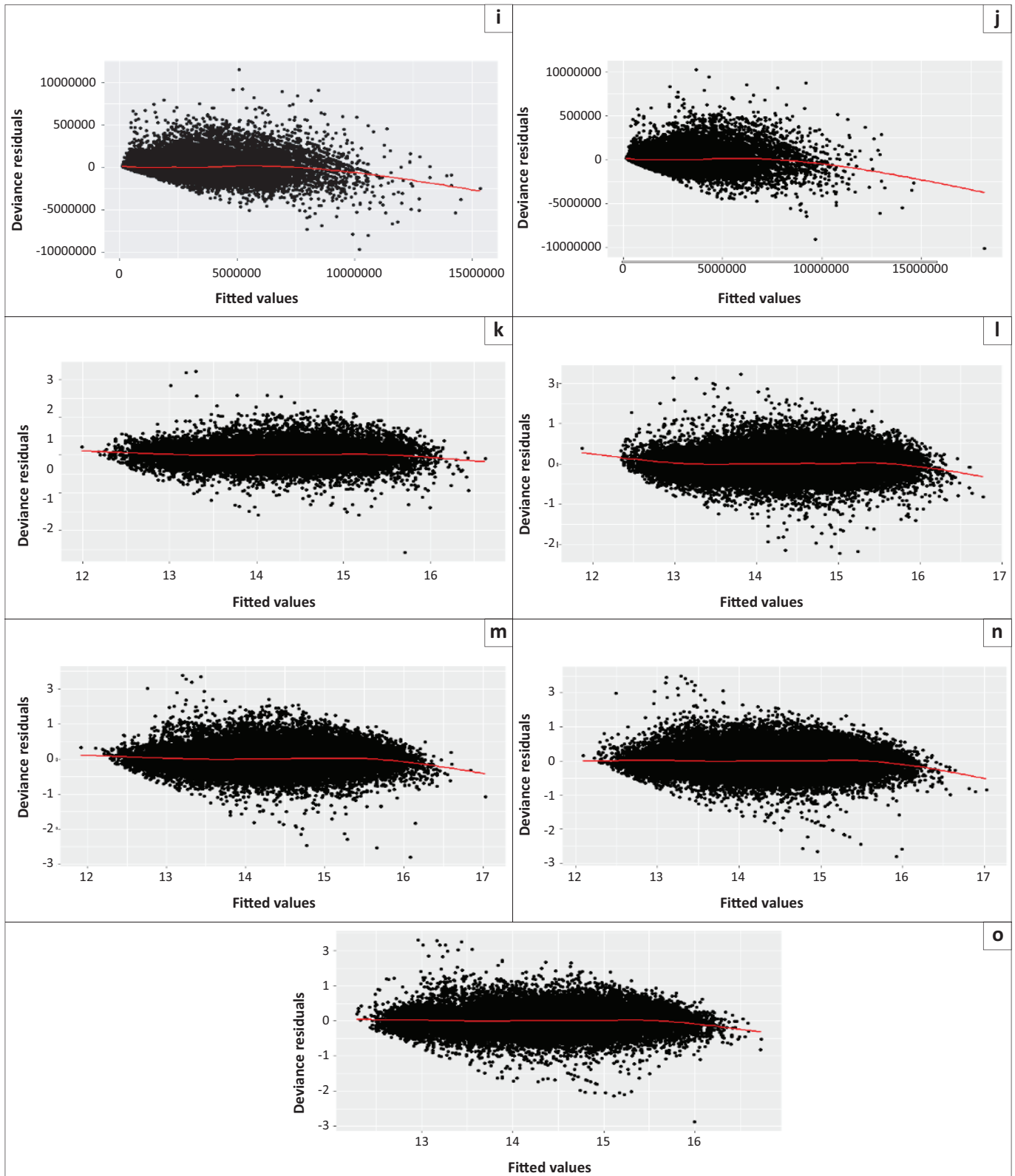
- A 1% increase in size (square metres), on average, increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.
- Each additional bedroom, on average, increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.



Note: Models from left to right: (a) gamma 2013, (b) gamma 2014, (c) gamma 2015, (d) gamma 2016, (e) gamma 2017, (f) normal 2013, (g) normal 2014, (h) normal 2015, (i) normal 2016, (j) normal 2017, (k) lognormal 2013, (l) lognormal 2014, (m) lognormal 2015, (n) lognormal 2016, (o) lognormal 2017.

**FIGURE 1:** Model fitted versus residual plots.

Figure 1 continues on the next page →



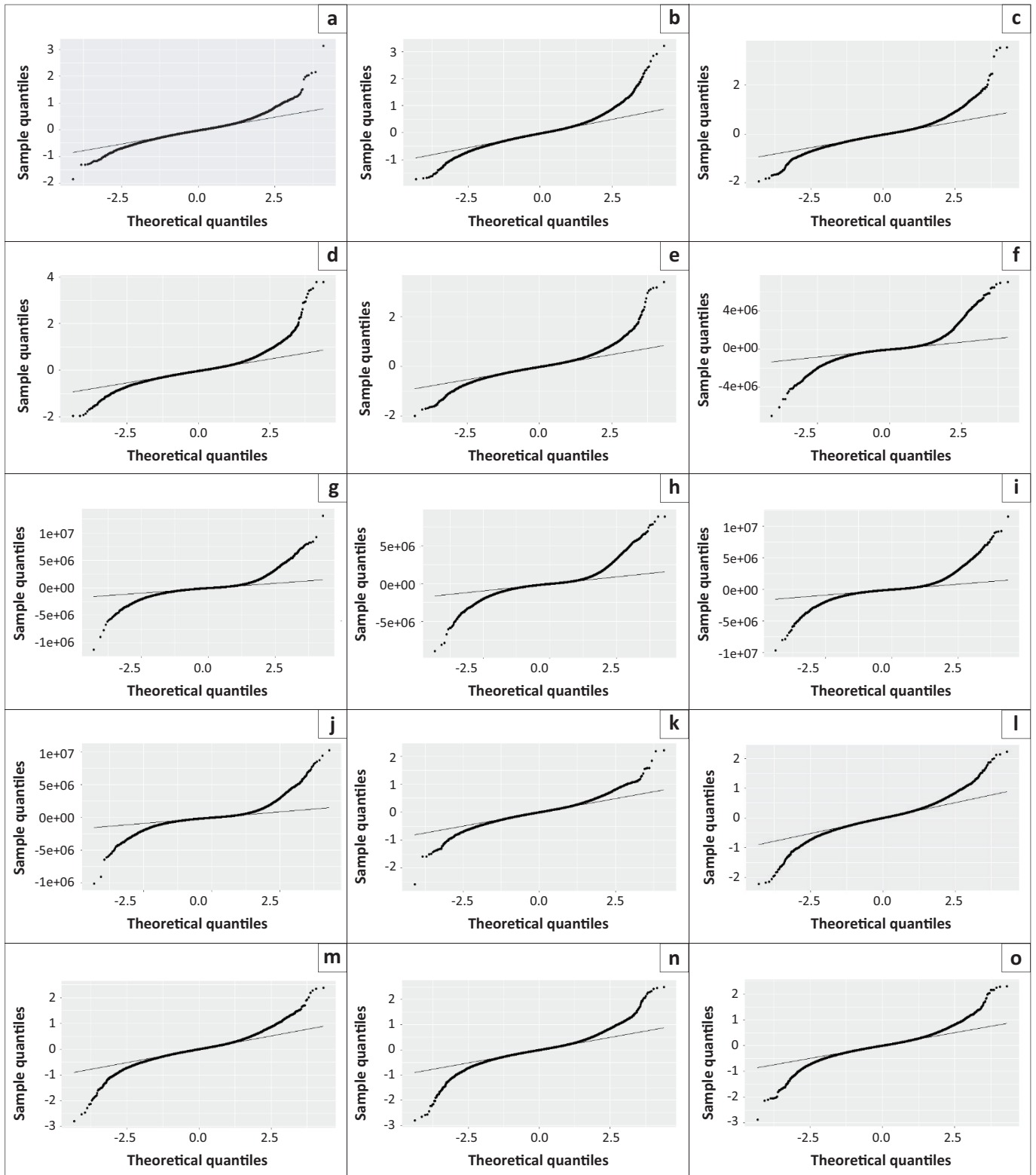
Note: Models from left to right: (a) gamma 2013, (b) gamma 2014, (c) gamma 2015, (d) gamma 2016, (e) gamma 2017, (f) normal 2013, (g) normal 2014, (h) normal 2015, (i) normal 2016, (j) normal 2017, (k) lognormal 2013, (l) lognormal 2014, (m) lognormal 2015, (n) lognormal 2016, (o) lognormal 2017.

**FIGURE 1 (Continues...):** Model fitted versus residual plots.

- Each additional bathroom, on average, increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.
- The property types in Table 4 are percentage difference comparisons between apartments where a property type

was  $\hat{\beta} \times 100$  (%) greater than or less than apartments (reference level) depending on the sign in front of the  $\hat{\beta}$ .

It is evident from Table 4 that each additional bathroom, on average, contributes more to the listing prices of



Note: Models from left to right: (a) gamma 2013, (b) gamma 2014, (c) gamma 2015, (d) gamma 2016, (e) gamma 2017, (f) normal 2013, (g) normal 2014, (h) normal 2015, (i) normal 2016, (j) normal 2017, (k) lognormal 2013, (l) lognormal 2014, (m) lognormal 2015, (n) lognormal 2016, (o) lognormal 2017.

FIGURE 2: Model quantile-quantile plots.

homes than each additional bedroom. An appealing feature of this parametric framework is the transparency and interpretability of the model coefficients. Property market participants are able to make informed decisions about renovating their homes or making comparative buying decisions by examining the marginal utility of different characteristics.

Plotting the residuals against individual covariates of the linear predictor should result in a null pattern, like the residual versus fitted values plot (McCullagh & Nelder 1989). The natural logarithm was applied to the size covariate to improve linearity where Figure 4 shows no discernible pattern in the plots indicating this transform was appropriate.



TABLE 4: Gamma model results summary.

Year	2013		2014		2015		2016		2017	
	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$
Intercept	9.913	2e-9	11.013	2e-9	10.932	2e-9	11.305	2e-9	11.148	2e-9
Log (size)	0.664	2e-9	0.626	2e-9	0.558	2e-9	0.479	2e-9	0.512	2e-9
Bedrooms	0.003	0.313	0.017	2e-9	0.021	2e-9	0.034	2e-9	0.025	2e-9
Bathrooms	0.111	2e-9	0.096	2e-9	0.112	2e-9	0.117	2e-9	0.112	2e-9
Cluster	0.090	2e-9	0.136	2e-9	0.146	2e-9	0.187	2e-9	0.187	2e-9
Duplex	0.003	0.874	0.025	0.104	0.035	0.024	0.086	2e-9	0.079	2e-9
House	0.027	3e-4	0.063	2e-9	0.103	2e-9	0.158	2e-9	0.141	2e-9
Simplex	0.061	4-e5	0.068	5e-7	0.078	2e-9	0.117	2e-9	0.087	2e-9
Townhouse	0.050	0.064	0.063	2e-9	0.077	2e-9	0.090	2e-9	0.099	2e-9

Note: Numbers were rounded to three decimal places and scientific notation was adopted for brevity.

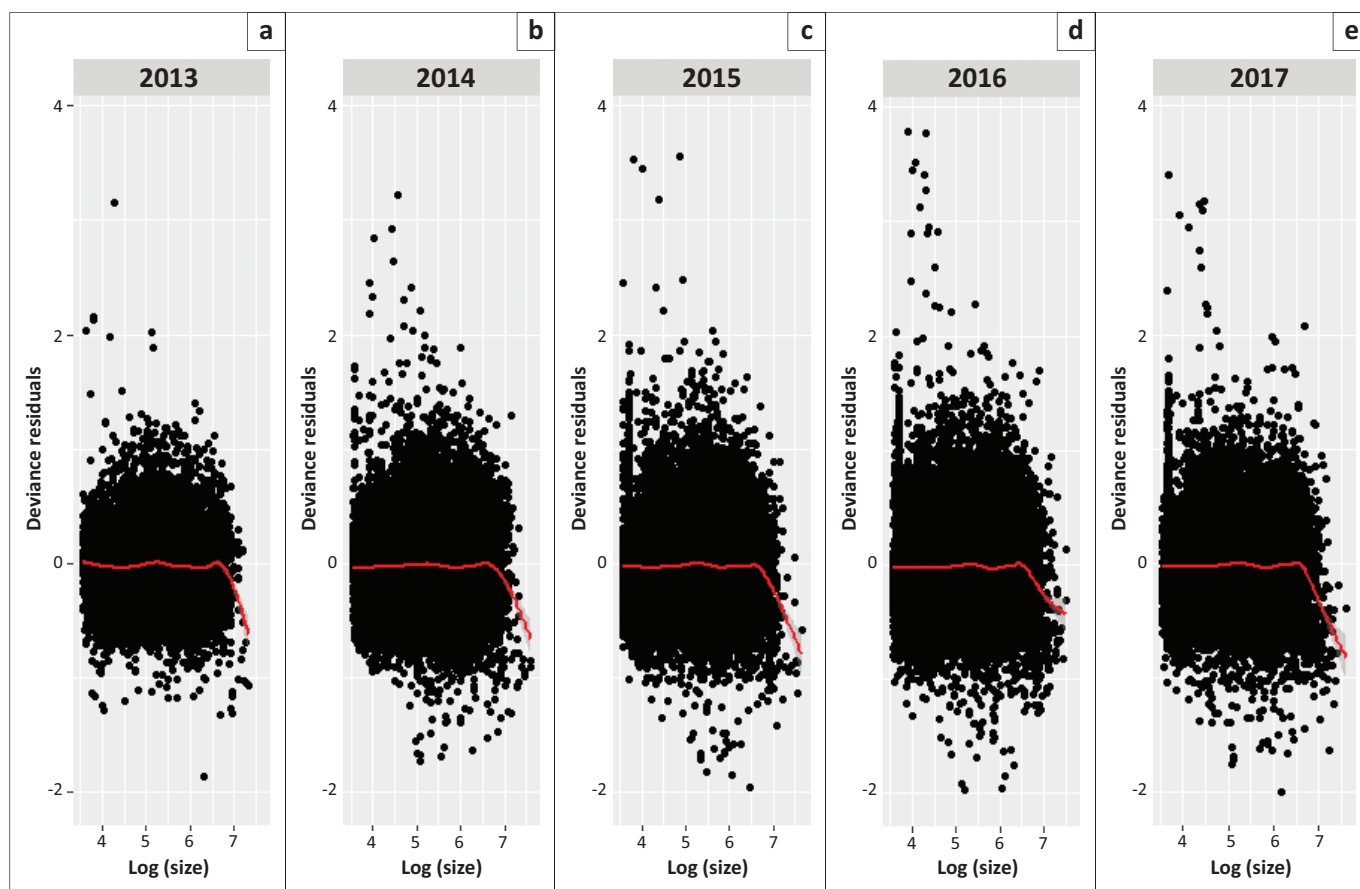


FIGURE 3: Gamma model residuals against transformed size covariate.

TABLE 5: Analysis of deviance.

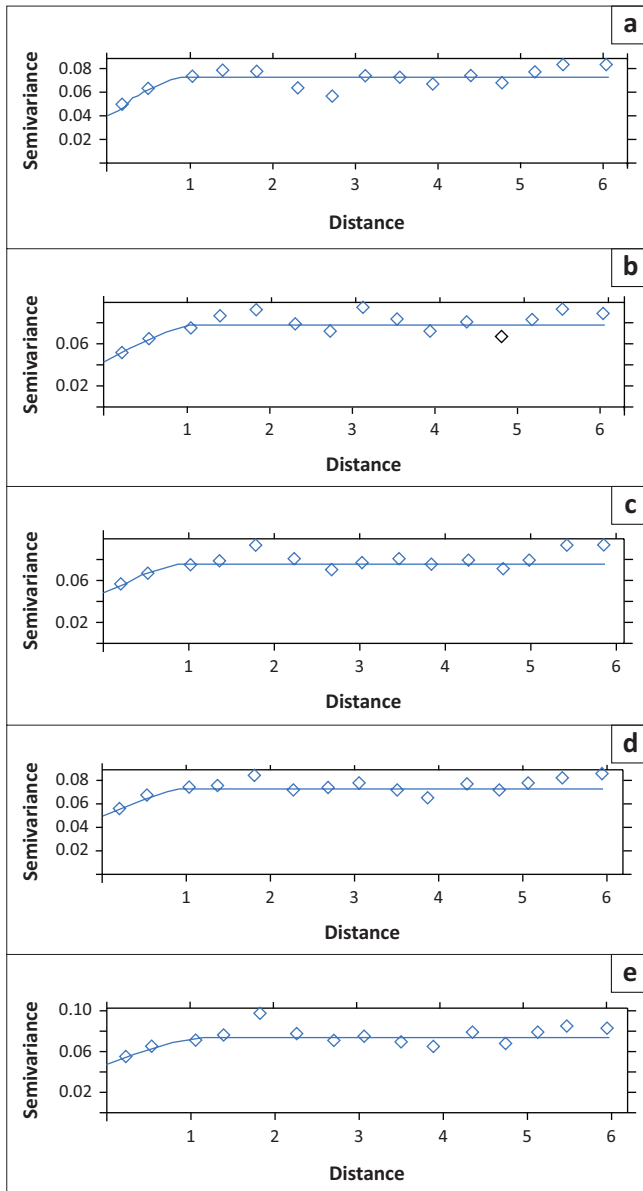
Year	Residual deviance	Null deviance
2013	1470.18	13 394.60
2014	4056.69	31 487.73
2015	4448.07	33 272.34
2016	6030.46	45 648.77
2017	4008.25	32 260.69

The analysis of deviance presented in Table 5 indicates that the residual deviance for each yearly gamma model was consistently lower than null deviance. This means that the covariates accounted for greater deviance explained than intercept only models and as such indicates a good fit.

The modelling of spatial data in this study required the assessment of the assumption of independence, which was

investigated using several plots and by performing hypothesis tests. Variograms quantify the spatial dependence in data by describing the spatial variance. The yearly gamma models’ residuals were plotted using spherical variograms and are presented in Figure 4 where similarities were found between all models. The ranges, distances beyond which the data are no longer correlated, are quite long, which suggests spatial autocorrelation is not an issue in the modelling results. The nugget effects as a percentage of the total sills are quite large which could indicate some variation at a small scale.

A permutation test for Morans I was applied to formally test for the presence of spatial autocorrelation where, under the null hypothesis, the data are randomly dispersed. The Morans I statistic or correlation coefficient ranges between -1 and 1, where -1 shows perfect negative spatial



**FIGURE 4:** Gamma model variogram plots. (a) variogram 2013 model, (b) variogram 2014 model, (c) variogram 2015 model, (d) variogram 2016 model, (e) variogram 2017 model.

**TABLE 6:** Permutation test for Morans I.

Year	Statistic	<i>p</i>
2013	-0.0312	0.999
2014	-0.0267	0.999
2015	-0.0129	0.999
2016	-0.0207	0.999
2017	-0.0345	0.999

autocorrelation and 1 shows perfect positive spatial autocorrelation. Hundreds of permutations were run, 999 in total, for each yearly gamma model. The results of the tests are presented in Table 6 which indicate a weak negative correlation. Formally, at an alpha of 0.05, there is not enough evidence to reject the null hypothesis of no spatial autocorrelation for each yearly gamma model. This coincides with the findings of Bourassa et al. (2007) where the addition of a location dummy variable accounted for spatial dependence adequately.

## Conclusion

Residential property is a barometer of individual and collective wealth and acts as measure of financial stability in an economy. Measuring residential property prices is difficult because of the heterogeneity thereof. The estimation of residential property prices using hedonic modelling is pervasive in real estate economic literature where log linear models are typically employed. This article investigated generalised linear models as an alternative to log linear models to develop hedonic price functions to estimate residential property listing prices in South Africa over a 5-year period. The gamma generalised linear model provided the best fit and good generalisability whilst keeping the expected values on the original scale, which is an appealing alternative to log linear models. The spatial dependence of residential properties was effectively accounted for by including an area factor variable, supported by variograms and Morans I permutation tests, showing no evidence to reject the null hypothesis of no spatial autocorrelation. This framework provides property market participants with the ability to quantify the utility derived over the marginal distribution of the physical characteristics of residential properties. This research presents the groundwork to create a property price index where index number theory could be applied to the hedonic price models to measure price inflation over time.

## Acknowledgements

### Competing interests

The authors have declared that no competing interests exist.

### Authors' contributions

D.B. contributed to the conceptual design, research methodology, cleaning and analysing the data, developing the models and visualisations using R, a statistical programming language. This article forms part of his PhD degree which he is currently pursuing. T.Z. contributed towards the conceptual design and research methodology as the supervisor of D.B. D.N. contributed towards the conceptual design and research methodology as the supervisor of D.B.

### Funding information

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 and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any affiliated agency of the authors.

## References

- Anselin, L., 2006, 'Spatial econometrics', in T.C. Mills & K. Patterson (eds.), *Palgrave handbook of econometrics Vol 1, econometric theory*, Palgrave Macmillan, New York, pp. 901–941.
- Blum, A., Kalai, A. & Langford, J., 1999, 'Beating the hold-out: Bounds for K-fold and progressive cross-validation', *COLT '99 Proceedings of the twelfth annual conference on Computational learning theory*, July 07–09, Santa Cruz, CA, pp. 202–208.
- Bordo, M.D. & Jeanne, O., 2002, *Boom-busts in asset prices, economic instability, and monetary policy*, CEPR Discussion Paper 3398, Centre for Economic Policy Research, London, viewed 12 February 2019, from <https://www.nber.org/papers/w8966>.
- Bourassa, S.C., Cantoni, E. & Hoesli, M., 2007, 'Spatial dependence, housing submarkets, and house price prediction', *Journal of Real Estate Finance and Economics* 35(1), 142–160. <https://doi.org/10.1007/s11146-007-9036-8>
- Candel, A., LeDell, E., Parmar, V. & Arora, A., 2017, *Deep learning with H2O*, H2O.ai Inc., CA, viewed 20 February 2019, from <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf>.
- Chiles, J. & Delfiner, P., 1999, *Geostatistics: Modeling spatial uncertainty*, p. 695, John Wiley & Sons, New York.
- Clark, I., 2010, 'Statistics or geostatistics? Sampling error or nugget effect?', *Fourth World Conference on Sampling and Blending*, vol. 110, Geostokos Ltd, Scotland, October 21–23, 2009, pp. 13–18.
- Davison, A.C. & Snell, E.J., 1991, 'Residuals and diagnostics', in D.V. Hinkley, N. Reid & E.J. Snell (eds.), *Statistical theory and modelling: In honour of Sir David Cox*, pp. 83–106, Chapman and Hall, London.
- Day, B., 2003, *Submarket identification in property markets: A hedonic housing price model for Glasgow*, Working Paper, The Centre for Social and Economic Research on the Global Environment, School of Environmental Science, and University of East Anglia, Norwich.
- De Haan, J. & Diewert, E., 2011, *Handbook on residential property indices*, Eurostat European Commission, viewed 12 February 2019, from <https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF>.
- Dobson, A. & Barnett, A., 2008, 'An introduction to generalized linear models', in P.C. Bradley, J.F. Julian, T. Martin & Z. Jim (eds.), *Texts in statistical science series*, vol. 77, 3rd edn., Chapman & Hall/CRC Press, Boca Raton, FL.
- 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. <https://doi.org/10.1111/j.1813-6982.2010.01244.x>
- LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C. et al., 2019, *h2o: R Interface for H2O*, R package version 3.22.1.1, viewed 01 February 2019, from <https://CRAN.R-project.org/package=h2o>.
- Goodhart, C. & Hofmann, B., 2008, 'House prices, money, credit, and the macroeconomy', *Oxford Review of Economic Policy* 24(1), 180–205. <https://doi.org/10.1093/oxrep/grn009>
- 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](https://doi.org/10.1016/0094-1190(78)90004-9)
- Guisan, A. & Zimmermann, N.E., 2000, 'Predictive habitat distribution models in ecology', *Ecological Modelling* 135, 147–186. [https://doi.org/10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9)
- Gujarati, D.N., 2004, *Basic Econometrics*, 4th edn., Tata McGraw-Hill, New York.
- Greene, W.H., 2003, *Econometric analysis*, 5th edn., Prentice Hall, Upper Saddle River, NJ.
- Hastie, T., Tibshirani, R. & Wainwright, M., 2015, *Statistical learning with sparsity: The lasso and generalizations*, CRC Press, Boca Raton, FL.
- Hill, R.J., 2013, 'Hedonic price indexes for residential housing: A survey, evaluation and taxonomy', *Journal of Economic Surveys* 27(5), 879–914. <https://doi.org/10.1111/j.1467-6419.2012.00731.x>
- 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, 121–131. <https://doi.org/10.1016/j.jbankfin.2015.08.026>
- Lindsey, J.K., 1997, *Applying generalized linear models*, Springer Science & Business Media, New York.
- Luus, C., 2002, *The ABSA Residential Property Market Database for South Africa—Key Data Trends and Implications*. BIS papers no 21.
- Lyons, R.C., 2015, *Measuring house prices in the long run: Insights from Dublin, 1900–2015*, viewed 29 April 2018, from <http://eh.net/eha/wp-content/uploads/2015/05/Lyons.pdf>.
- Mccullagh, P. & Nelder, J., 1989, *Generalized linear models*, vol. 37, CRC Press, London.
- Moran, P., 1950, 'A test for the serial independence of residuals', *Biometrika*, 37(1–2), pp. 178–181.
- Musset, L., 2006, *OECD environment health and safety publications series on testing and assessment*, no. 54 [pdf], viewed 12 January 2019, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2006\)18&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2006)18&doclanguage=en).
- Nelder, J.A. & Wedderburn, R.W.M., 1972, 'Generalized linear models', *Journal of the Royal Statistical Society Series A* 135, 370–384. <https://doi.org/10.2307/2344614>
- Olivier, J., Johnson, W. & Marshall, G., 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, 333–338, 625–626. [https://doi.org/10.1016/S1081-1206\(10\)60595-9](https://doi.org/10.1016/S1081-1206(10)60595-9)
- Ploner, A., 1999, 'The use of the variogram cloud in geostatistical modelling', *Environmetrics* 10(4), 413–437. [https://doi.org/10.1002/\(SICI\)1099-095X\(199907/08\)10:4%3C413::AID-ENV365%3E3.0.CO;2-U](https://doi.org/10.1002/(SICI)1099-095X(199907/08)10:4%3C413::AID-ENV365%3E3.0.CO;2-U)
- R Core Team, 2018, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, viewed n.d., from <https://www.R-project.org/>.
- 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>
- Shimizu, C., Nishimura, K. & Watanabe, T., 2010, 'Housing prices in Tokyo: A comparison of Hedonic and repeat sales measures', *Journal of Economics and Statistics* 230, 792–813. <https://doi.org/10.1515/jbnst-2010-0612>
- Silver, M., 2016, *How to better measure hedonic residential property price indexes*, IMF Working Paper, WP/16/213, IMF, Washington, DC.