ORIGINAL RESEARCH ARTICLE

Application of regression techniques to capture value influences for mass valuation of residential property: *A case study of two residential estates in Nairobi*

Bernard Ochieng[1]

[1]*Department of Land Resource Planning & Management Jomo Kenyatta University of Agriculture & Technology*

Corresponding author email: *bernard.towe@gmail.com*

Abstract

This research project sets out to apply statistical techniques in the valuation of land and properties through various models. It focused on comparing predictive accuracies of mass valuation models with a dataset of 500 single-family property transactions in two neighbourhoods within Nairobi city. There are a number of statistical models that are used for mass valuation of properties. The first step in this study was to gather data on property sales used in the development of a base model and proposed model. Each of the property units in the database were geocoded and vectorized. The data was screened and visualized to investigate the nature of the potential association between the response, *Y*, and predictor variables, *X*. The predictor variables were tested for multicollinearity and a regression model developed based on hypothesized relationships. The model was tested for lack of fit by ordering the residuals using a residual scatterplots and histograms. Finally, the fitness statistics were reviewed by looking at the spread of the plot and evaluating observed values around the regression line, and examining how accurate the independent variables are in predicting the dependent variables. The results revealed an overall level of 0.96 for Komarock and 0.98 for Runda estate respectively. One measure of how well the model predicts is to compute the correlation between the actual values in the holdout sample and the predicted values. The correlation should be high when the model is valid. The correlation between the assessed value and the actual selling price is 0.71 and 0.98 for Komarock and Runda respectively. Determining the quality of the valuation output also requires measuring uniformity: uniformity between groups of properties and uniformity within groups (Abidoye, Huang, Amidu, & Javad, 2021). The coefficient of dispersion (COD) is the most used measure of valuation uniformity. The results show a COD of 18% and 10% for Komarock and Runda respectively.

Keywords: Valuation, mass valuation, computer assisted mass appraisal, regression analysis, heteroscedasticity, homoscedasticity

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

1.0 Introduction

Traditionally, valuations are considered accurate or inaccurate based on their simple comparison with actual transaction prices using error metric or the econometric techniques or both (Bogin & Shui, 2020). For a number of reasons, valuations and actual market prices (or market values) hardly coincide (Babawale, 2013). Empirical studies in the UK, USA, Australia and several other countries have confirmed that one-to-one relationship between valuation and actual transaction price is a rarity. Real estate is a special commodity. In order to really grasp the formation of its price, it is necessary fully to understand the theory of price formation, and for specific analysis. Real estate prices are the core of the real estate market. It reflects the needs of the relationship between the real estate investors and consumers. So assessment includes speculation and judgments of between people and real estate on real estate prices. The process of estimating the valuation, real estate valuers is essentially the process of economic studies.

Statistical standards of equity established by such organizations as the International Association of Assessing Officers (IAAO) give additional benchmarks by which modelers may test various approaches and methodologies. The need to achieve property valuation accuracy in any property market cannot be overemphasised due to the importance of the real estate section to the economic and household development of any nation (Abidoye & Chan, 2018). The automated techniques for mass valuation seem preferable for several reasons. First, spatial data analysis, *data from all sales* are utilized, rather than data from only three or four comparable properties that have sold recently. Appraiser bias with respect to choosing comparables or ‘‘comps,’’ therefore, would be eliminated. Although switching to a statistical analysis methodology would be costly and may require staff training, the ease of land valuation would largely overcome these costs in the long run.

In statistics, regression analysis consists of techniques for modelling the relationship between a dependent variable (also called response variable) and one or more independent variables (also known as explanatory variables or predictors) (Dielman, 2001). In regression, the dependent variable is modelled as a function of independent variables, corresponding regression parameters (coefficients), and a random error term which represents variation in the dependent variable unexplained by the function of the dependent variables and coefficients. In linear regression the dependent variable is modelled as a linear function of a set of regression parameters and a random error. The parameters need to be estimated so that the model gives the “best fit” to the data.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

Despite the issues regarding the validity of the assumption for using ordinary least squares (OLS), the basic linear model is the most widely used in empirical work with property transaction data (Ahn, Byun, Oh, & Kim, 2012). Model specification is the designing of models based on valuation theory and market analysis. It's the process of finding a unique set of model parameters that provide a good description of the system behaviour and can be achieved by confronting model predictions with actual measurements performed on the system (Bogdanovaa, Kamalovaa, Kravchenkoa, & Poltorak, 2020). Specifying a model includes selecting the variables to be considered and defining their relationships with each other and with market value. It is necessary to specify the factor that will determine the conformity of the estimated parameters with the market information before the estimated model is used for the real estate valuation. The accuracy of the selection of market information for estimating parameters of the valuation model should be greater than or equal to 0.95 (Adamczyk, Bieda, & Parzych, 2019). According to (Ireland, March 8, 2010), a model specification should be:

I. Accurate-the statistical quality results should demonstrate the quality of the model specification.
II. Rational-the model specification and calibration coefficients should be understandable
III. Explainable-the appraiser or model provider should be able to present (and explain) the model and its results
IV. Frugal-once the model calibration has reached its best quality statistics, the valuer should remove the least significant property characteristics one at a time until the quality statistics begin to noticeably decline.

Model specification issues fall into two broad categories for valuation purposes:

I. The functional form of the relationship between the dependent variable and the independent variables and
II. The choice of the variables to include in the model

Property valuation models seek to explain or predict the market value of properties from real estate data. Models are constructed to represent the operation of forces of supply and demand in a particular market and have evolved from three broad theories of value: the cost approach, the sales comparison approach, and the income approach. Model building requires good theory, data analysis, and research methods (Kauko & D'Amato, 2008). The best valuation models will be accurate, rational, and explainable. Appraisers are often too quick to use an already defined model without examining its assumptions and structure. In real estate valuation, multiple linear regression is more realistic representation of the interplay of the

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

variety of transactional and property characteristics that can affect the value of a predictor variable like price or rent (Ananth & Kleinbaum, 1997).

One important feature of using multiple regression analysis in valuation is the ability to include an expression of precision along with a point estimate value opinion (Bidanset & Lombard, 2014). Predictions concerning the response variable y based on given values for the explanatory variables fall into two categories: predictions concerning the mean of y and predictions concerning a single outcome of y. Generally speaking, Valuers are more interested in predicting the mean of y given specific values of the explanatory variables because definitions of market value and market rent focus on the most probable price and most probable rent rather than predicting a particular transaction price or rent.

The data set to be analysed included 500 observations on a variety of numerical and categorical variables. Numerical variables include sale price, sale date, number of baths, and number of bedrooms, built-up area, garage, extension, and location. Categorical variables include the nearness to a source of nuisance and site amenities. Although the data set contains a great deal of information, it is not formatted to facilitate multiple linear regression analysis. First, to the data was put into an amenable format, which at a minimum requires converting categorical variables into indicator variables. The reformatted house data set contains the following continuous and categorical variables:

Table 1: Continuous and categorical variables

Variable	Description	Conversion used in Modelling
Sale Price	Price in KES recorded	Used in the current format
Plot Area	Area in Acres	Used in the current format
No of Bedrooms	Number of bedrooms	Used in the current format
Built-up Area	Area in square feet	Used in the current format
Garage	Area in square feet	Used in the current format
Property Extensions	Area in square feet	Used in the current format
Staff Quarter	Area in square feet	Used in the current format
Site Amenities	Categorical variable	Converted to Dummy Variables where appropriate
External Nuisance	Categorical variable	Converted to Dummy Variables where appropriate

2.0 Materials and methods

2.1 Study area

Two residential neighbourhoods within the Nairobi Metropolitan Area formed the case study areas for this research. These areas were Runda and Komarock residential estates. These vary in location and character and allowed for model testing and validation. In addition to the transaction data, such as price, size, and quality, the data set was supplemented with data describing the exact spatial position (x and y coordinates). The property prices collected here were a mixture of both asking prices in certain cases and values exchanged between a willing

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

buyer and a willing seller in arms-length transactions. Second, information of sold and purchased residential properties (henceforth, firms' data) was collected from registered real estate firms operating in Nairobi city's property market. Nairobi is the capital and primate city of Kenya. It lies astride the Nairobi River at an altitude of about 1670 metres above sea level on coordinates 1°17′0″S, 36°49′0″E.

2.1.1 Runda Estate

Runda estate is categorized as high-income residential area. It is part of the larger Runda estate which is situated in is bounded by Limuru Road to the west, Ruaka Road to the south, Kiambu Road to the east, and other emerging residential areas to the north. While a significant number of the houses are owner-occupied, Runda is also a favourite with the expatriate community looking for rental housing, perhaps due to its proximity to the UNON Offices in Gigiri, and a number of Foreign Missions located in this side of Nairobi. The area is popular with humanitarian workers and diplomats, drawn to the wide, leafy streets lined with impressive houses, all in an acre or more of landscaped gardens. Single family dwelling units dominate this neighbourhood. The houses are large, modern and on big plots.

2.1.2 Komarock Estate

Komarock estate is a middle-class and a low income residential area, located 18 kilometres east of Nairobi CBD, in Embakasi Central subcounty, Nairobi County. It is predominantly residential with few doted commercial spots along frontages of the main collectors within the estate. It is covered in two parts, the old part popularly known as Sector which was put up between 70's and 80's. It consists of sections 1, 2 and 3. The other part is fairly new and popularly known as Phase.

Table 0: Zonal Classification

Zone	Area covers	GC	PR	Types of development allowed	Min area in ha	Remarks and policy issues
8	KOMAROCK			Residential-Mixed development • Flats • Maisonettes • Site-and-serviced Schemes condominiums (single rooms)	0.05ha	Comprehensive subdivision allowed Minimum to fit a house on type plan design.
	• RESIDENTIAL	50	70			
	• COMMERCIAL	80	150			

NB: *GC-Ground Coverage*
PR-Plot Ratio

2.2 Data description

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

The conclusions reached in empirical studies of multivariate market valuation models are, of necessity, tied inextricably to the data set used in their development (Borst, 2006). It is not possible to consider every variation in housing stock that might be encountered in one big city, such as Nairobi, let alone across several countries. Nonetheless, it is important that some variety is considered to make the conclusions reached in the study as robust as possible. To address this, sales transaction and property descriptive data were acquired from two different neighbourhoods within the city. The first step was to gather information from registry sales records, the classified ads in the local newspapers and from real estate companies. The available information was extracted as much from these sources, such as asking price, building size, lot size, number of bedrooms, and number of bathrooms, address, and age. Where information was missing, a site visit was done to obtain additional information. The final database contained 1000 properties which had sufficient information for analysis.
The sample size of 500 used in the study is in the range of what has been used in previous studies, for instance, the studies of (Abidoye & Chan, 2018) and (Tayani, Morano, Salvo, & Ruggiero, 2019). The thought process involved in making a decision regarding how many data observations are necessary for application of a regression model differs from the calculation of sample size for inferences about mean. In regard to a regression model, the measure of data sufficiency is based on degrees of freedom-i.e., the relationship between the number of observations (n) and the number of independent variables in the model (k). When the ratio of n to k is too low, the model is considered "overfitted" and the regression outcome is in danger of being data-specific, not representative of the underlying population.

For example, consider a ratio of n to k of 2:1. It is always possible to connect two points with a straight line. In this case the coefficient of determination, R^2, would always be equal to 1 in a simple linear regression model. However, the model may not actually explain anything. Since R^2 and the ability to generalise from a sample to a population are affected by the ratio of n to k, many researchers suggest that the minimum ratio should be in the range of 10 to 15 observations per independent variable, with a ratio of 4:1 to 6:1 as an absolute minimum. One indication of an overfit model due to a ration of n to k that is too low is an increase in adjusted R^2 as the least-significant variables are removed from the model.

To filter out non-arms-length sales, properties with sales prices under KES 3,000,000 for Komorock, KES 5,000,000 for Buruburu and KES 20,000,000 for Runda estate were omitted. The higher outliers were eliminated by excluding the observations with sales prices over KES20,000,000 for Komarock and Buruburu and KES 150,000,000 for Runda estate. The sample, as shown in Tables 1, the average sales price was approximately KES 8,400,000 for Komarock estate, KES 11,000,000 for Buruburu estate and KES 98,600,000 for Runda Estate.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

3.0 Methodology

Multivariate analysis begins with selection of an initial collection of the most important numerical variables, using the best subsets procedure. We can break down this procedure into two stages: analysis of the data described in the first table of the fieldwork data and investigation of nonlinear transformations including quadratics and reciprocals.

3.1 Data Exploration

Descriptive statistics for the continuous and numerical count variables are shown in the table 2-3. The "*describe ()"* function in the *'psych'* package can be used to generate basic descriptive information for the variables, including skewness and kurtosis.

Table 1: Komorock descriptive statistics derived in R Studio

Variable	Mean	SD	Median	Max	Range	Skew	Kurt
Lot Ac	0.04	0.04	0.03	0.33	0.32	5.71	33.09
BArea	970.61	1147.78	760	10302	10182	5.78	38.59
Bdms	2.64	0.76	3	4	4	-0.32	0.37
Exten	87.34	268.22	0	1461	1461	3.5	12.49
Staff Quar	49.07	157.85	0	1200	1200	4.83	26.86
Garage	0	0	0	0	0	NaN	NaN
Utilities	1	0	1	1	0	NaN	NaN
Mkt_Sgmt*	1	0	1	1	0	NaN	NaN
Site ameni	1	0	1	1	0	NaN	NaN
Extnl Nuis	0	0	0	0	0	NaN	NaN
Price	8,373,603.3	3,929,155.98	7,600,000	41,400,000	37,800,000	5.14	38.97

Sale price are centred on KES 8,300,000 and KES 98,633,000 for Komarock and Runda respectively based on the sample mean and median. Typical built-up area is about 120 to 10,302 square feet located on a roughly 0.04 of an acre lot in Komarock estate and 0 to 10,000 square feet located on a roughly 0.86 of an acre lot in Runda estate. The built-up area and the number of bedrooms are highly right skewed with each mean being much larger than the median. Central tendencies for these variables were best captured by their medians, at 760 square feet and 3 bedrooms for Komarock estate and 3,620 square feet and 4 bedrooms for Runda estate.

Table 2: Runda descriptive statistics

Variable	Mean	SD	Median	Max	Range	Skew	Kurt
Lot Acres	0.86	1.55	0.5	10.6	10.24	5.73	32.92
Built_Area	3565.49	2627.3	3620	10000	10000	0.42	-0.12
Bdms	3.49	2.03	4	7	7	-0.69	-0.61
Exten	157.69	267.8	0	760	760	1.22	-0.23
Staff_Quar	246.56	279.5	159	1044	1044	0.77	-0.24
Garage	131.33	213.14	0	670	670	1.15	-0.28

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

Utilities	1	0	1	1	0	NaN	NaN
Site ameni	1	0	1	1	0	NaN	NaN
Extnl Nuis	0.4	0.49	0	1	1	0.41	-1.87
Price	98,633,837	139,438,322	71,040,000	973,500,000	941,500,000	5.73	33

Comparisons was done between means and medians to identify excessive skewness (for example extensions and garage). The minimum and maximum values provided a range that can used to evaluate how well the data represented the market being studied. Finally, the presence of an external nuisance variable acted like a dummy variable when it was included in a model because the range is 0 to 1 in increments of 1(i.e., there are no partial nuisances). The standard deviations and medians were not included in the indicator variable table. These statistics provide no useful information for indicator variables. The means were, however, quite useful because the means of a dummy variable represent the proportion of the indicator variable set within a given category. For example, the mean for nearness to a source of nuisance for Runda was 0.4 meaning that 40% of the properties are situated near a source of nuisance. The table included minimum and maximum values for each variable. The inclusion of these statistics was to illustrate that these are dummy variables, and they provided were applied check for data entry errors. There were several observations about the data that highlight the differences in the housing stock of the three residential neighbourhoods:

I. The plot sizes in Runda estate were considerably larger than those of Komarock estate. This was consistent with the lower population density.
II. Runda estate had a mean selling price ten times the other two study areas.
III. Runda had more vacant parcels than the other study areas. This difference poses interesting challenges in model formulation as was presented in the model specification analysis for each estate.
IV. The houses in Komarock were smaller than that of the other two estates.

There is no information broken down about the number of half baths; instead, the total baths variable includes the sum of full plus half baths, so that a property with a value of 4 total baths may have 4 full baths or 3 full plus 2 half baths, for example. Observations with missing sales date or location information were dropped.

Sales that were not standard market transactions such as foreclosures, bankruptcies, land court sales, and intra-family sales were excluded. Further, for each year, observations with the bottom and top 1% sales prices were excluded to guard against non-arms-length sales and transcription errors. The data included typical house characteristics: living space, lot size, the number of bathrooms, bedrooms, and total rooms. The sample is limited to units with at least one bedroom and bathroom, 3 total rooms and 500 square feet of living space and no

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

more than 10 bedrooms and 10 bathrooms, 25 total rooms, and 8000 square feet of living space, or 10 acres.

3.2 Model development

The best subsets procedure limited to the data set's continuous variables identified the following six variable model as 'best' based on S_e, C_p, and R^2_{ADJ}

Price= f(Lot_Acres_ + Built_Area + Bdms + Exten + Staff_Quar + Garage + Extnl_Nuis)

Working with this base model, continuing studies included a series of stepwise models used to test how the base model held up as interaction variables nonlinear transformations were considered as explanatory variables.

The outputs in tables 3-5 and 3-6 also includes the *Adjusted* R^2 statistic, which was abbreviated as R^2_{ADJ}. The *Adjusted* R^2 for Komarock 0.599, which was slightly less than the R^2 *of 0.6161*. The importance of R^2_{ADJ} stems from the fact that addition of any random variable to a regression model will increase R^2, *whether* or not the new variable is truly relevant to systematic variation in the dependent variable. However, R^2_{ADJ} will increase with the introduction of an additional variable only when the increase is in excess of the expected increase in R^2 that would occur with the introduction of an irrelevant variable. R^2 is the proportion of SST represented by SSR, which can be expressed from the data as

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

Adjusted R^2 is calculated differently, accounting for the number of explanatory variables (k) in the model:

$$R^2_{ADJ} = \frac{\frac{SSE}{n+k-1}}{\frac{SST}{n-1}}$$

When a variable is added to the model, the divisor of SSE (n-k-1) decreases in size relative to the divisor of SST (n-1). The decrease in SSE from the addition of a new variable (i.e., the increase in SSR) must offset division by a smaller number relative to the divisor of SST. Otherwise, R^2_{ADJ} will remain constant or decrease when a new variable is added. When a variable is added to a model and R^2_{ADJ} *does* not increase, the new variable is explaining no

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

more than would be explained by adding any totally irrelevant random variable. For this reason, some analysts prefer to report R^2_{ADJ} instead of R^2 as a measure of explained variance absent the effect of the number of independent variables. In addition, the examination of the change in R^2_{ADJ} *as* variables are added to a model is a way of assessing whether or not an added variable is actually explanatory (Benjamin, Guttery, & Sirmans, 2004). Finally, R^2_{ADJ} *will always be less than R2,* and $R^2_{ADJ\ can}$ be negative when model fit is so poor that it explains less than a model including an equal number of totally random and irrelevant explanatory variables.

The adjusted R-squared that is printed in the output is designed to offset of the regression model over fitting the data when you have a larger number of independent variables and a smaller sample size. The adjustment to R-square is less in circumstances where you have a larger sample size and smaller numbers of independent variables. It is good practice to report on the adjusted R-square (along with R-square), especially in circumstances where there is more substantial difference between these values.

One added inferential measure is the model *F* statistic. The *F* statistic tests the null hypothesis that all of the independent variable coefficients are zero versus the alternative hypothesis that at least one is not zero. In other words, it assesses whether or not the model provides a better measure of the expected value of *y* and the mean of *y*. A variable will be significant if its p-value is less than 0.05. This result can be generated by the "*summary ()*" function in R studio.

```
# lm(formula = Price ~ Lot_Ac + Built_Area + Bdms + Exten + Staff_Quar, data = Komorock_Comps_CSV1)
```

Table 3: Output Variables, Komarock Preliminary Model

Intercept	Lot_Ac	Built_Area	Bdms	Exten	Staff_Quar
4,078,071.00	-2,696,201.00	1,935.00	698,915.00	4,721.00	3,842.00

Predictor	Coef	SE Coef	T	Pr(>|t|)
Constant	4,078,071.2	569,015.7	7.167	0
Plot Area	-2,696,201.1	3,494,680.0	-0.772	0.44200
Bedrooms	668,856.9	194,070.8	3.6012	0.00047
Built Area	1,935.2	191.1	10.125	0
Extension	4,721.3	621.4	7.592	0
Staff Quarter	3,842.2	924.0	4.158	0.000062

S	1,592,000.00
R-Sq	0.6161
R-Sq (adj)	0.5993
F-stat	36.6
p-value	0

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

The unstandardized regression slopes in the output are interpreted as the expected change in raw score units on the dependent variable between two cases differing by one raw score unit on the independent variable (holding the other independent variables constant). If the slope is positive, it indicates that higher values on the independent variables are associated with higher scores on the dependent variable (and lower values on the independent variable are associated with lower scores on the dependent variable). A negative slope indicates that higher (lower) values on the independent variables are associated with lower (higher) scores on the dependent variables. These coefficients are tested for significance using a t-test (Neter, Wasseman, & Kutner, 1990). The null hypothesis when testing the regression slope is that the population slope is zero. The alternative hypothesis is that the population slope is non-zero.

 # lm(formula = Price ~ Lot_Acres_ + Built_Area + Bdms + Exten + Staff_Quar + Garage + Extnl_Nuis, data = Runda_Comps_CSV)

Table 4: Output Variables, Runda Preliminary Model

Intercept	Lot_Ac	Built_Area	Bdms	Exten	Staff_Quar	Garage	Extnl_Nuis
7,890,650	80,539,963	3,156.8	2,048,969.	730.8	1,657.5	7,918.9	-3,270,760.7

Predictor	Coef	SE Coef	T	Pr(>|t|)
Constant	7,890,650.3	2,750,896.4	2.868	0.00714
Plot Area	80,539,963.6	2,706,885.0	29.754	0
Bedrooms	2,048,969.5	7,29,525.7	2.809	0.00829
Built Area	3,156.8	510.7	6.181	0
Extension	730.8	3,184.8	0.229	0.81992
Staff Quarter	1,657.5	3,725.7	0.445	0.65931
Garage	7,918.9	3,982.3	1.989	0.05510
External Nuisance	-3,270,760.7	1,849,827.8	-1.768	0.08628

S	4,583,000.00
R-Sq	0.9783
R-Sq (adj)	0.9737
F-stat	212.3
p-value	0

R-square of the model, which corresponds to the proportion of variance explained by the model, and it measures the strength of the relationship between the model and the dependent variable Y on a convenient 0 to 100% scale. The p-value and t-statistic for each regression coefficient in the model. These two metrics are considered to be the mirror of one another, and they measure the extent to which a given coefficient is statistically significant. The higher the t-statistic (which goes with the p-value near 0), the more significant the predictor is, meaning that this predictor should be kept in the model. On the other hand, a very low t-statistic (higher p-value) means that the predictor should be dropped.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

The final model includes most of the variables first considered after running best subsets on the continuous variables. To obtain standardized regression weights, we use the package lm.beta available within R Sudio. The standardized regression slope is equivalent to the unstandardized slope*(sdX/sdY). It is interpreted as the difference in standard deviation units on Y for a one standard deviation difference on X (holding the remaining IV's constant). Interpretation of the direction of an effect is the same as with unstandardized regression slopes.

4.0 Results and discussions

4.1 One variable profiles, charts, and graphs

These tools were used to show the distribution more clearly and completely than measures of central tendency and spread alone. An array is a listing of data from the lowest to the highest or vice versa. Table 3-1 and 3-2 below are examples.

Table 7: Array of comparable sales-Komarock & Runda estates

KOMAROCK ESTATE	Built	Lot	Bdms	Age	Exten	Staff	Price
Nairobi/Block 111/1929	1,038	0.0297	3	30	0	0	9,800,000.00
Nairobi/Block 134/35	1,278	0.0252	2	25	0	0	15,400,000.00
Nairobi/Block 111/1667	370	0.0385	2	30	1,560	0	11,200,000.00
Nairobi/Block 111/847	883	0.0334	4	30	0	0	9,800,000.00
Nairobi/Block 111/1839	1,146	0.0341	3	30	0	110	11,200,000.00
Nairobi/Block 111/504	1,137	0.0333	4	30	0	200	9,100,000.00
Nairobi/Block 111/1188	626	0.0519	3	30	437	0	7,000,000.00
RUNDA ESTATE	Built	Lot	Bdms	Age	Exten	Staff	Price
Nairobi/Block 99/41	2850	0.5040	5	15	0	800	60,000,000.00
7785/310 Runda Mae	3442	0.5001	5	20	320	520	100,000,000.0
14274/39	4200	5.4910	4	15	900	490	73,150,000.00
Nairobi/Block 99/211	3423	0.4942	4	15	0	572	79,680,000.00
14470/121	3750	0.4954	4	15	364	624	63,080,000.00
14470/233	6700	0.5033	6	15	700	325	111,220,000.0
Nairobi/Block 99/113	5725	0.5041	5	15	0	0	116,200,000.0
7785/1331 Runda Mae	4095	0.5429	4	20	660	375	83,000,000.00

Arrays can be useful because they show the entire sample sorted in order of magnitude. However, they are best suited to small samples in which direct visual analysis is feasible. Arrays form the basis for finding the minimum, maximum, median, and percentiles of data distribution and also make it easy to identify extremely small or large values or outliers.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

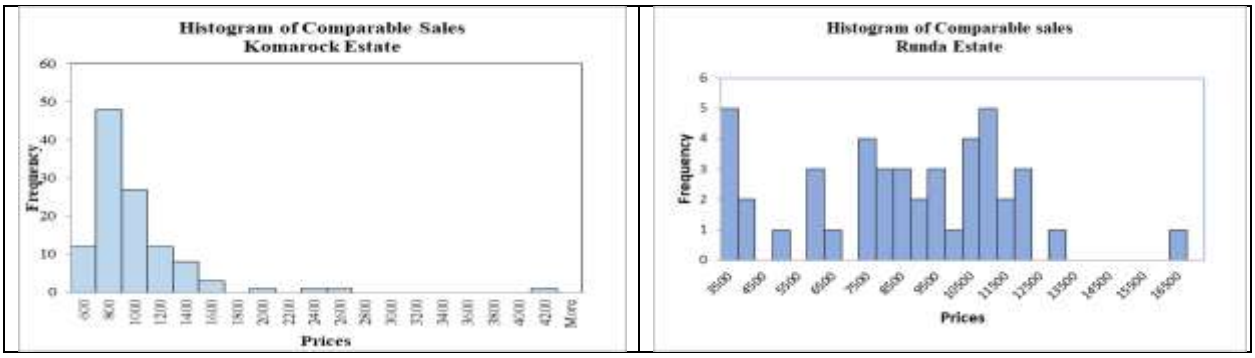


Figure 1: Histogram of comparable sales Komarock and Runda estates

There were 12 properties that were sold at between KES 4,000,000-6,000,000 and 48 properties were sold at between KES 6,000,000-8,000,000. Ten classes were constructed from the data gathered during fieldwork. The number of lots in each class is the frequency of occurrence (i.e., frequency is the count of observations in a class).

4.2 Two variable profiles, charts, and graphs

Each of the dots in the diagram represented a pair of observations about an individual unit-square feet of built-up area and sale price. The built-up area was measured on the horizontal axis and the corresponding price was measured along the vertical axis. The scatter plot demonstrated that these data are directly related. As the built-up area increased, price also increased. As the living area decreased, the price also decreased.

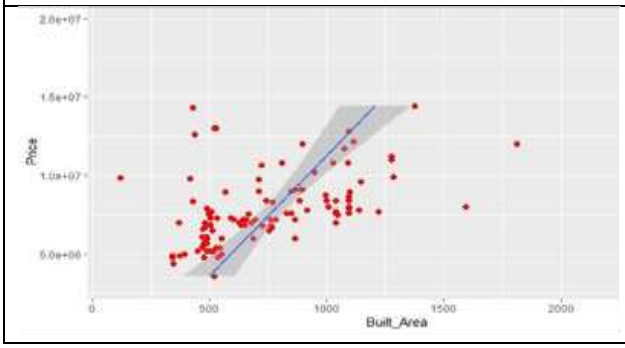
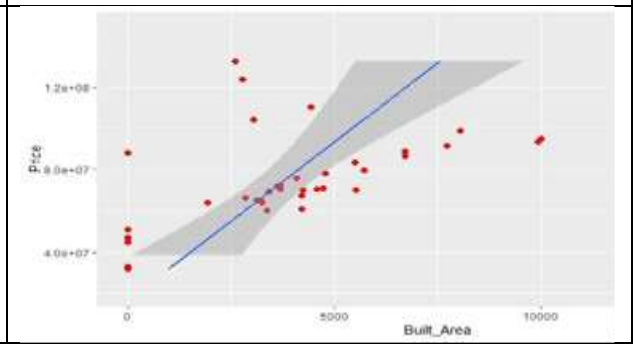
Komarock estate	Runda estate
	

Figure 2: Scatterplots for price vs. built-up area

Figure 3-3 shows the sale price with neighbourhood. The boxes represent the interquartile range (first to third quartile, or middle 50 percent) of the data. In mass valuation, it can be used to evaluate the relationship between a discrete variable, such as construction quality or neighbourhood, and sale price. In ratio studies box plots can be used to identify outlier ratios that may potentially distort sales ratio statistics.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

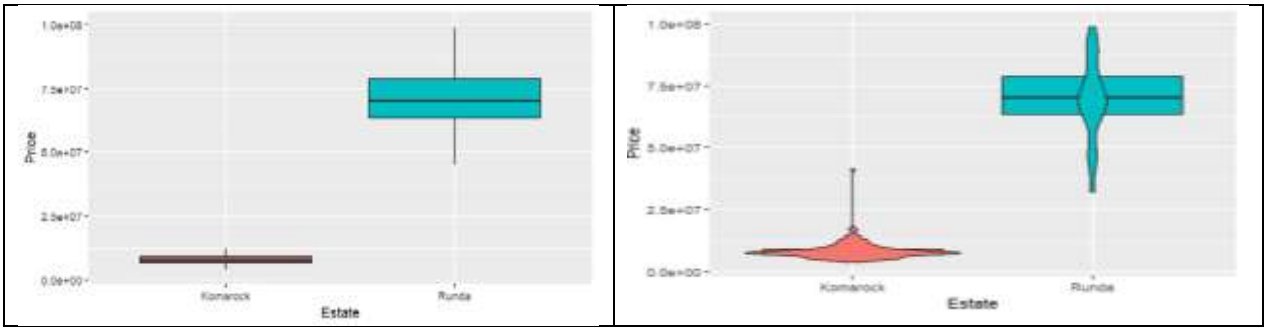


Figure 3: A histogram of price with neighbourhood

Figure 3.3 shows that the middle spread (the interquartile range) of price for Komarock was slightly left skewed and the upper quartile was highly right skewed. It appears that a few extreme values for prices, identified asterisks, were influencing the mean price calculation. The influence was not significant as these extreme cases were not many. This supported the conclusion that the median of KES 7,600,000 and mean of KES 8,300,000 were both better indicators of expected price in the market. This is confirmed by a skewness statistic value of 5.14 and 5.3 for Komarock and Runda estates respectively.

The figure 3.4 shows the comparison of price and lot area. The significant extent of linear correlation between price and lot area was quantified by the Pearson Correlation Coefficient, which is 0.07 and 0.99 for Komarock and Runda respectively. The coefficient of determination for these two variables were 0.82 for Komarock, indicating that 82% of price variability was associated with built-up area differences. The remaining variation in price variability can be attributed to other factors.

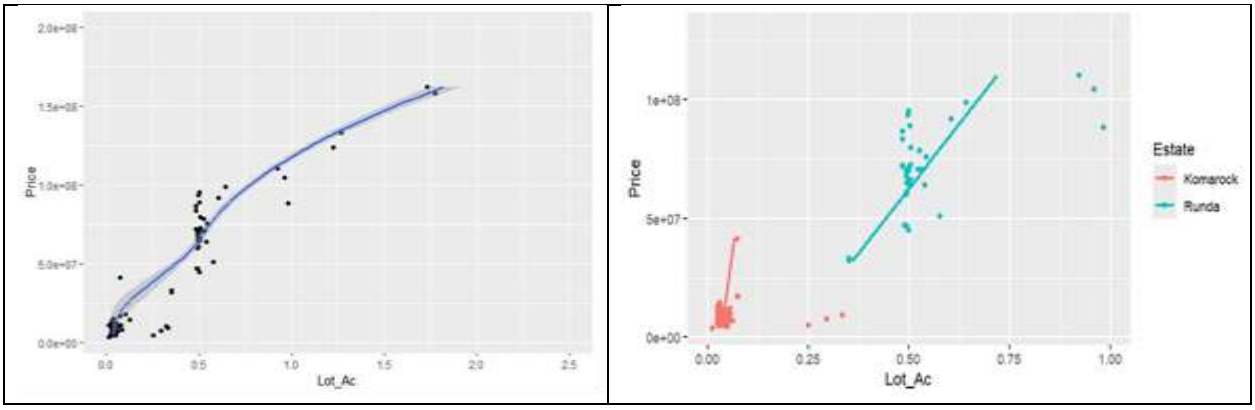


Figure 4: Scatterplots of price vs. lot area

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

From the figure3.4 it can be observed that a few portions of the residuals lie in a straight line. This indicates that the residuals of the model did not follow a normal distribution.

4.3 Correlations between the variables

Correlations between the response variable Sale Price and the continuous explanatory variables are presented in table 3-3. Correlations with Sale Price had the expected signs except for presence of an external nuisance, which was negative. For Runda estate, correlations with Sale Price and extension and staff quarter was also negative. This may be an indication that property owners prefer not to have an extension or a detached staff quarter, although more information would be required to confirm this conjecture. It could also be a data anomaly. If multicollinearity turns out to be an issue with a multiple regression model of these data, it is apt to reside with Built up Area, which was positively correlated with Bedrooms, Baths, Swimming Pool and Garage.

Table 8: Correlation matrix for Komarock estate

	Lot in Acres	Built up Area	Bedrooms	Extension	Staff Quarters	Price
Lot in Acres	1.00	0.10	0.03	0.08	-0.04	0.07
Built up Area	0.10	1.00	-0.01	0.33	-0.04	0.82
Bedrooms	0.03	-0.01	1.00	-0.16	0.05	0.07
Extension	0.08	0.33	-0.16	1.00	0.00	0.58
Staff Quarters	-0.04	-0.04	0.05	0.00	1.00	0.13
Price	0.07	0.82	0.07	0.58	0.13	1.00

The scatter approaches a straight line as the coefficient approaches –1 or +1, whereas there was no linear relationship when the coefficient is 0. A perfect correlation of –1 or +1 means that all the data points lie exactly on a straight line. Hypothesis tests and confidence intervals were used to address the statistical significance of the results and to estimate the strength of the relationship in the population from which the data.

Table 9: Correlation matrix for Runda estate

	Lot Acres	Built Area	Bedroom	Exten	Staff	Garage	Extnl Nuis	Price
Lot in Acres	1.00	0.01	0.07	-0.05	-0.13	-0.15	0.12	0.99
Built up Area	0.01	1.00	0.85	0.12	0.07	0.19	-0.20	0.11
Bedrooms	0.07	0.85	1.00	0.26	0.36	0.28	-0.07	0.15
Extension	-0.05	0.12	0.26	1.00	0.25	-0.04	-0.42	-0.02
Staff Quarters	-0.13	0.07	0.36	0.25	1.00	0.43	0.10	-0.10
Garage	-0.15	0.19	0.28	-0.04	0.43	1.00	0.25	-0.11
External	0.12	-0.20	-0.07	-0.42	0.10	0.25	1.00	0.10
Price	0.99	0.11	0.15	-0.02	-0.10	0.11	0.10	1.00

The scatter plot panels are shown in figure 3-6. Panel may be indicative of a nonlinear inverse relationship between sale price and lot area. Panel B is strongly indicative of a nonlinear

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

inverse relationship between sale price and age. Panel C suggests that the relationship between sale price and built up area is probably linear for these data.

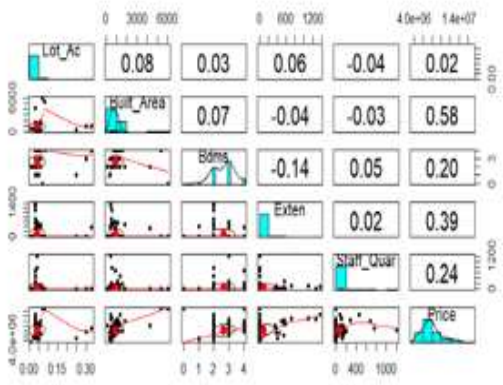
Komarock	Runda
	

Figure 5: Correlation scatterplot panels

The correlations among variables, histograms for the variables, and scatterplots were applied to identify potential non-linearities among the variables.

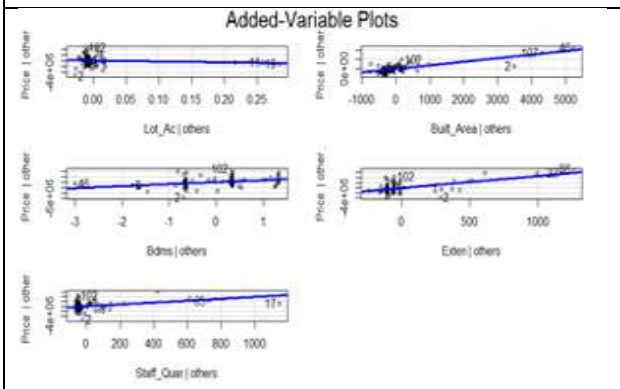
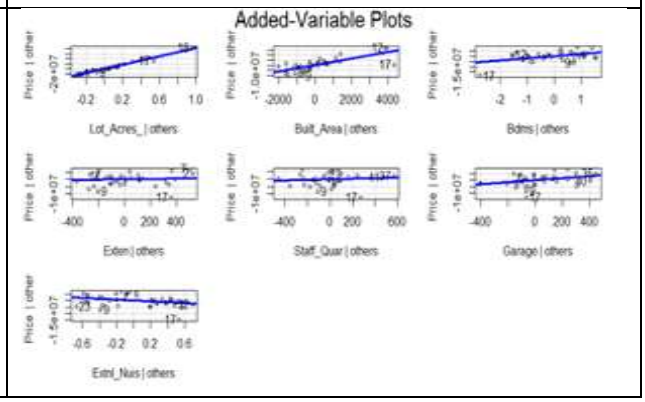
Komarock estate	Runda estate
	

Figure 6: Added Variables Plots

4.4 Final Model

The standardized regression slope is equivalent to the unstandardized slope*(sdX/sdY). It is interpreted as the difference in standard deviation units on Y for a one standard deviation difference on X (holding the remaining IV's constant). Interpretation of the direction of an effect is the same as with unstandardized regression slopes.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

4.4.1 Komarock estate

For Komarock estate, the residual standard error was 1,671,000, which gives an idea of the average distance that the observed values fall from the regression line. The R-squared value was 0.8251, indicating that approximately 82.5% of the variability in price is explained by the model. The adjusted R-squared was 0.8191, which took into account the number of predictors in the model and provided a more accurate measure of goodness-of-fit. This was quite high as it implied the model explained 81.9% of any variability in real estate price. The F-statistic was 136.8 with a p-value of 0, indicating that at least one predictor variable was statistically significant in explaining the variability in price.

```
# lm(formula=Price~Built_Area+ Bdms +Exten+ Staff_Quar, data Komorock_Comps_CSV1)
```

Table 10: Output variable-Final model for Komarock estate

Intercept	Built_Area	Bdms	Exten	Staff_Quar
3996598.8	2,415.10	661,871.10	5,475.40	3,881.40

Predictor	Coef	SE Coef	T	Pr(>|t|)
Constant	3996598.8	558146.1	7.160	0
Bedrooms	693162.8	193585.9	3.581	0.000503
Built Area	1923.5	190.2	10.113	0
Extension	4689.8	619.5	7.571	0
Staff Quarter	3872.0	921.6	4.201	0.00005

S	1,589,000.00
R-Sq	0.6141
R-Sq (adj)	0.6007
F-stat	45.76
p-value	0

This general model suggests that all the independent variables positively impacted real estate price, while number of bedrooms had the most significant positive impact overall, the other independent variables had a marginally positive impact. The model was statistically very significant as it explained more than 80% of the variance in real estate price, suggesting that other factors may also be influential. This model can be fine-tuned, but from the economic perspective, it passes the test.

4.4.2 Runda estate

In Runda estate, the residual standard error was 2,581,000, which gives an idea of the average distance that the observed values fall from the regression line. The R-squared value was 0.8251, indicating that approximately 82.5% of the variability in price is explained by the model. The adjusted R-squared was 0.9997, which took into account the number of predictors in the model and provided a more accurate measure of goodness-of-fit. This was quite high

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

as it implied the model explained 99% of any variability in real estate price. The F-statistic was 24,500.00 with a p-value of 0, indicating that at least one predictor variable was statistically significant in explaining the variability in price.

```
# lm(formula = Price ~ Lot_Acres_ + Built_Area + Exten + Staff_Quar + Garage, data = Runda_Comps_CSV)
```

Table 11: Output variable-Final model for Runda estate

Intercept	Lot_Ac	Built_Area	Bdms	Garage	Extnl Nuis
7,508,049	81,413,206	3,057	2,294,176	8,244	-3,410,071

Output of Variables

Predictor	Coef	SE Coef	T	Pr(>|t|)
Constant	7,508,049	2,625,625	2.860	0.007017
Plot Area	81,413,206	2,461,381	33.076	0
Built Area	3,057	480	6.369	0
Bedrooms	2,294,176	634,357	3.617	0.000908
Garage	8,244	3,622	2.276	0.028916
Extnl_Nuis	-3,410,071	1,678,724	-2.031	0.049653

S	4,494,000.00
R-Sq	0.9780
R-Sq (adj)	0.9749
F-stat	319.6
p-value	0

This general model suggested that all the independent variables positively impacts real estate price, while number of bedrooms had the most significant positive impact overall, the other independent variables had a marginally positive impact. The model was statistically very significant as it explained more than 80% of the variance in real estate price, suggesting that other factors may also be influential. This model can be fine-tuned, but from the economic perspective, it passed the test.

Several aspects of the final model are worthy of further consideration: First, all of the independent variables are significant at the 5% level except Bathrooms, which was significant at the 10% level. Second, price escalation was rapid during the time period covered by the data. Third, the variable age is significant in quadratic form with the coefficients of Age and Age-Squared both being highly significant. This was consistent with the valuation theory that depreciation is not usually a straight-line phenomenon. Finally, the sign on plot size is positive, as theoretically expected. Reversing the sign found in the correlation table once the effects of other price-determining variables had been accounted for.

Standardized slopes are often consulted to facilitate comparative judgments regarding which independent variables are contributing more (versus less) to the explained variation in the dependent variable. Downsides of these coefficients: (a) They are affected by the standard

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

deviations of the independent variables; (b) they are not particularly useful for interpreting effects with binary predictor variables, or when making comparative judgments of independent variables involving factors that have been recorded for inclusion in the analysis (to capture potential group effects). For more discussion, see (Darlington, 2017) and (Hayes, 2018). At least with respect to our independent variables which we are assuming to be continuous, built up area exerts a slightly stronger effect in the model than lot area for Komarock. The reverse is true for Runda estate where we see very little response by price to built-up area and quite significant positive response to lot size.

4.5 Checking for multicollinearity and heteroskedasticity

The final model selected, was checked with a scatter plot of standardised residuals versus price estimate for evidence of heteroskedasticity. The scatter plot below shows no apparent pattern of heteroskedasticity. The outliers may also reflect randomness in the market that cannot be fully accounted for without collecting more data. In the histogram below, the standard errors are slightly right skewed, which is consistent with the direction of the largest outliers in the accompanying scatter plot, but otherwise the distribution appears to be approximately normal.

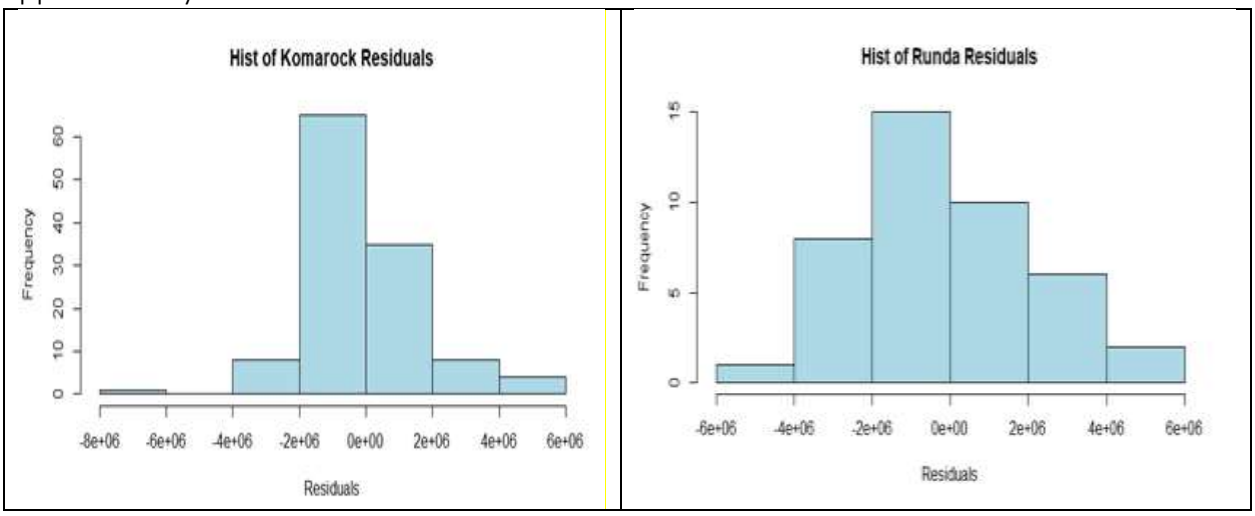


Figure 7: Histogram of residuals-response is the sale price

Residuals are the unexplained variance. They are not exactly the same as model error, but they are calculated from it, so seeing a bias in the residuals would also indicate a bias in the error.The most important thing to look for is that the black lines in figure 3-8 representing the mean of the residuals are all basically horizontal and centered around zero. This means there are no outliers or biases in the data that would make a linear regression invalid.

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

We can study residuals plots using the "*residualPlots*"() function from the 'car' package in R Studio. Figure 3-8 below shows plots of studentized residuals against fitted values on Y and against each independent variable. The fit lines included in the plots allow you to study where there may be non-linearities in the data. A non-significant test result suggests a non-significant deviation from linearity in terms of the relationship between the fitted values (blow) or IV's (above) and the studentized residuals. Non-linearity here indicates unmodeled non-linearities in our regression model.

```
# residualPlots(fit1_Koma, fitted=TRUE, type="rstudent", test=TRUE)
```

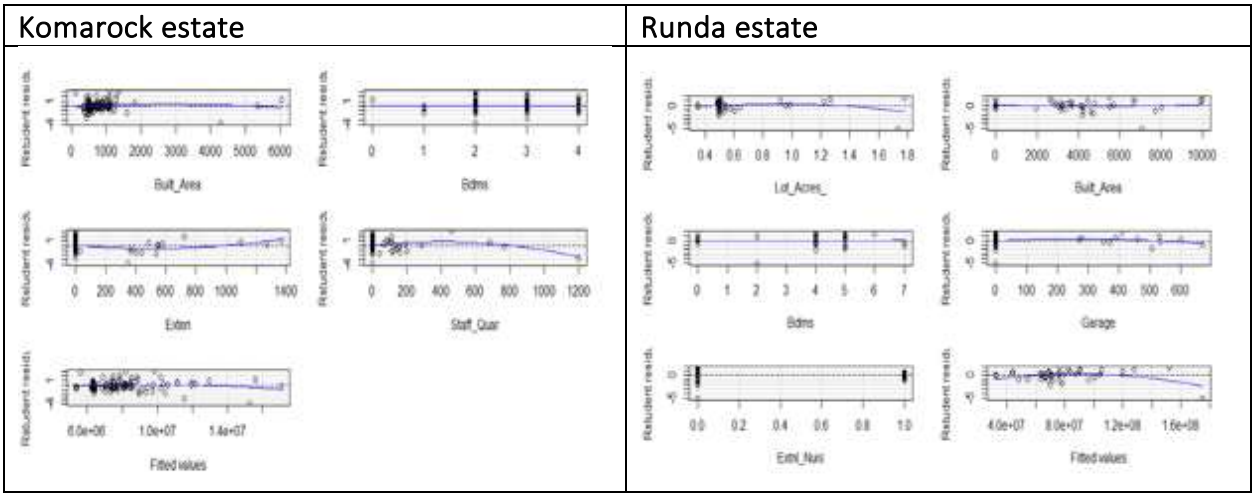


Figure 8: Scatterplots of standardized residuals versus fitted values (response is sale price)

These plots are also useful for identifying potential residual outliers (with studentized residuals>3; see & for detection of potential heteroskedasticity (non-constant variance) of residuals. Funnel shaped patterns often are a signal of a linear functional relationship between residuals and independent variables. Multicollinearity does not appear to be a serious problem, given the significant t statistics on the most-correlated independent variables (see the prior output of final model variables).

Table 12: Curvature tests for the model fit

Test stat and Pr(>|Test stat|) for the residuals				
	Komarock Estate		Runda Estate	
	Test stat	Pr(>|Test stat|	Test stat	Pr(>|Test stat|
Plot Area	-	-	-1.2220	0.229891
Bedrooms	-0.2969	0.767113	-0.3005	0.765589
Built Area	-1.2855	0.201214	1.3004	0.79441
Extension	2.4356	0.016414	0.6103	0.201969
Staff Quarter	-2.7265	0.007413	1.0503	0.30080
Garage	-	-	-1.1425	0.260991
Tukey test	-2.0370	0.041647	-3.1961	0.001393

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

VIF is a crucial metric that helps us understand the level of multicollinearity among predictors in a regression model. VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. It's a handy tool to identify collinearity issues in our model. From figure 3-9 below the variance inflation is averaging at 1 for Komarock while for Runda it varies from 1.5-3.5, the highest VIF is associated with built-up area, as expected. This bar plot will help us identify predictors that might be causing multicollinearity issues in our model.

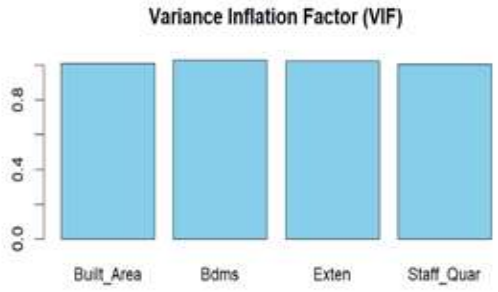
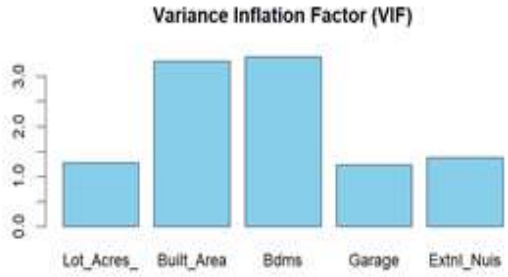
Komarock estate	Runda estate
	

Figure 9: Variance Inflation Factor (VIF)

With this price estimation model in hand, we can use it to generate market value estimates for similar properties within the selected neighbourhoods. Descriptions of three houses and their respective market value estimates are shown in the table below. Notice that all three of the prediction houses have independent variable values within the ranges of the independent variables in the data. If the model were to be applied for prediction outside of the range of the x variables, the tacit assumption would be that the relationship between price and the independent variables is unchanging outside of the range of the x variables. However, remember that the model cannot provide information to support this assumption. This does not mean that prediction outside of the range is unreliable. At times, the use of a regression model to predict values outside of the range of the independent variables may be justified if the analyst's market knowledge and experience is sufficient to support the decision.

4.6 Model Validation

Reference books on statistics offer several suggestions for regression model validation, including

I. collecting new data to assess the model's predictive ability on the new data
II. comparing results with theory and with previously published empirical studies
III. data splitting

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

Collecting new data is often not a practical option in applied valuation settings. Nevertheless, it is possible and recommended that analysis assess the signs of the variables in the regression equation and compare them with theoretical and intuitive expectations. Staying current on relevant published studies is an obvious priority and needs little discussion. The third option, data splitting, provides the most practical sample-specific and model-specific means of model validation and is worthy of further examination. Data splitting, which is also known as cross-validation, requires that the data set be divided into two subsets: (1) a model building set and (2) a validation set, usually referred to as a holdout sample. The holdout sample, which should be randomly chosen from the full data set, can be a small proportion of the full data set (e.g., 10% to 20%).

Two possible validation routines are recommended. The first routine is to compare the coefficients and significance levels derived from the model-building set with the coefficients and significance levels derived from a regression model using all the data. The results should be consistent, otherwise a small number of influential observations may be affecting the model disproportionately. The second routine is to use the regression model derived from the model-building set to predict the dependent variable values for the holdout sample. One measure of how well the model predicts is to compute the correlation between the actual values in the holdout sample and the predicted values. The correlation should be high when the model is valid. In tables 3-10 and 3-11 the correlation between the assessed value and the actual selling price is 0.71 and 0.99 for Komarock and Runda respectively.

Table 13: Ratio Studies for Komarock Estate

Sale Number	Property PIN	Assessed Value	Sale Price	Ratio
1	Nairobi Blk 111/1343	8,398,454.50	7,600,000.00	1.1051
2	Nairobi Blk 111/1843	8,072,680.20	7,600,000.00	1.0622
3	Nairobi Blk 134/35	8,534,320.20	11,000,000.00	0.7758
4	Nairobi Blk 134/872	8,091,915.20	7,500,000.00	1.0789
5	Nairobi Blk 111/1233	6,306,204.40	6,700,000.00	0.9412
6	Nairobi Blk 111/181	6,440,849.40	6,000,000.00	1.0735
	Mean	1.006		
	Weighted Mean	0.9880		
	STDEV	0.126		
	Range	0.329		
	COD	18%		
	COV	13%		
	CORREL	0.71		

Determining the quality of mass valuation also requires measuring uniformity: uniformity between groups of properties and uniformity within groups. The coefficient of dispersion (COD) is the most used measure of valuation uniformity. It's built on the average deviation, which measures the average absolute (sign-ignored) difference of the ratios from the median. Tables 3-10 and 3-11 give a coefficient of dispersion of 18% and 10% for Komarock and Runda respectively. The coefficient of variation (COV) is the standard deviation expressed as a

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

percentage. Like the standard deviation on which it is based, the interpretation of the COV depends on the extent to which the data are normally distributed. When the normality assumption is met, the COV is a powerful measure of uniformity.

Table 24: Ratio Studies for Runda Estate

Sale Number	Property PIN	Assessed Value	Sale Price	Ratio
1	14970/39	69,487,828.67	72,000,000.00	0.9651
2	7785/951	101,592,604.31	91,725,000.00	1.1076
3	112/193	895,968,594.49	973,500,000.00	0.9204
4	112/78	67,139,804.16	60,000,000.00	1.1190
5	112/9	47,319,106.73	46,900,000.00	1.0089
6	Nairobi Blk 99/113	82,467,051.14	79,600,000.00	1.0360
7	Nairobi Blk 99/123	47,742,455.41	47,000,000.00	1.0158
8	Nairobi Blk 99/210	73,134,844.76	65,000,000.00	1.1252
9	Nairobi Blk 99/211	71,175,510.41	69,280,000.00	1.0274
10	Nairobi Blk 99/217	69,633,741.41	65,000,000.00	1.0713
	Mean	1.0397		
	Weighted Mean	0.9675		
	STDEV	0.0672		
	Range	0.2048		
	COD	10%		
	COV	6%		
	CORREL	0.99		

When the distribution is approximately normal, as indicated by a histogram, or other statistics or graphs, the standard deviation and COV, the tests of valuation performance that assume normality are still useful. However, they should not be relied on if the distribution is clearly skewed or otherwise not normally distributed. For consistency, one can choose always to rely principally on the COD and nonparametric tests of appraisal performance. The standard on ratio studies (Gloudemans, R. J., 1999) calls for an overall level of appraisal of 0.90 to 1.10. The parametric t-test and the nonparametric binomial test are two good tests of the null hypothesis that properties in a jurisdiction, class, area, or other category are appraised at a specified percentage of market value. Both tests can be used to test the null hypothesis that the level of valuation equals a specified percentage, say 100 percent. The alternative hypothesis would be that the level of valuation is not 100 percent. In this case, the alternative hypothesis is two-tailed. In other words, we will reject H_0 in favour of H_1 if we can be confident the level of valuation is below or above 100 percent. If H_0 is rejected, the level of appraisal may still lie between 0.90 and 1.10.

Another, more reasonable approach consistent with the IAAO standard (International Association of Assessing Officers, 1990) is to test whether the level of valuation lies within the required range, say 0.90 to 1.10. In this case, the test will be one-tailed with the alternative hypothesis dependent on the calculated mean. If the calculated valuation level lies below the required level, say 0.9 the null hypothesis and alternative hypothesis would be:

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

H_0: level of valuation ≥ 0.90;
H_1: level of valuation < 0.90.

If the calculated valuation level were greater than 1.10, the null and alternative hypotheses would be:

H_0: level of valuation ≤ 1.10;
H_1: level of valuation > 1.10.

Note that if the calculated level meets the required standards (for example, 0.90 to 1.10), there is no need to test for compliance.

5.0 Conclusions

If the property tax is to be fair and provide a reliable revenue source, the valuation system must produce accurate and equitable value estimates. Despite the issues regarding the validity of the assumption for using OLS, the basic linear model is by far the most widely used in empirical work with property transaction data (Ahn, Byun, Oh, & Kim, 2012). To make sure that the obtained parameters are close to the market values, the constant terms determined in the parametric model and the conditional model cannot be greater than 25% of the corresponding transaction price. Real properties that do not satisfy this criterion may be burdened with gross errors. The information on these properties should be verified and, if no errors are identified in the data, they ought to be removed from the database for market analysis and valuation. However, even if the above condition is met, it is necessary to specify the factor that will determine the conformity of the estimated parameters with the market information before the estimated model is used for the real estate valuation.

The standard on ratio studies (Gloudemans, R. J., 1999) calls for an overall level of appraisal of 0.90 to 1.10. This study revealed an overall level of 0.96 for Komarock and 0.98 for Runda estate respectively. One measure of how well the model predicts is to compute the correlation between the actual values in the holdout sample and the predicted values. The correlation should be high when the model is valid. The correlation between the assessed value and the actual selling price is 0.71 and 0.99 for Komarock and Runda respectively. Determining the quality of mass valuation also requires measuring uniformity: uniformity between groups of properties and uniformity within groups. The coefficient of dispersion (COD) is the most used measure of valuation uniformity. The coefficient of dispersion of 18% and 10% for Komarock and Runda respectively.

6.0 Acknowledgements

This paper has benefited considerably from valuable suggestions from my colleagues Dr Mathew Kigomo and Dr Geoffrey Waweru both of Jomo Kenyatta University of Agriculture &

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

Technology. I am most grateful to the anonymous referee for his/her constructive comments and to Professors Awiti Kakumu and Agrey Thuo of Nairobi and Kenyatta Universities respectively for their excellent guidance during my research studies. Finally unending gratitude to my local supervisor Dr Mathew Gicheha for walking this academic journey with me. Any remaining errors are naturally my own responsibility.

7.0 References

Abidoye, B., & Chan, A. (2018). Achieving property valuation accuracy in developing countries the implication of data source. *International Journal of Housing Markets and Analysis, 11*(3), 573-585.

Abidoye, R., Huang, W., Amidu, A.-R., & Javad, A. A. (2021, February 15). An updated survey of factors influencing property valuation accuracy in Australia. *Property Management, 9*(3), 343-361.

Adamczyk, T., Bieda, A., & Parzych, P. (2019). Principles and criteria for using statistical parametric models and conditional models. *Real Estate Management and Valuation, 7*(2), 33-43. doi:10.2478/remav-2019-0013

Ahn, J. J., Byun, H. W., Oh, K. J., & Kim, T. Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, 8369-8379. doi:10.1016/j.eswa.2012.01.183

Ananth, C., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology, 26*(6), 1323–1333.

Babawale, G. (2013). Valuation accuracy – the myth, expectation and reality! *African Journal of Economic and Management Studies, 4*(3), 387-406.

Benjamin, J. D., Guttery, R. S., & Sirmans, C. F. (2004). Mass Appraisal: an introduction to multiple regression analysis for real estate valuation. *Journal of Real Estate Practice and Education, 7(1)*, 65-78.

Bidanset, P. E., & Lombard, J. R. (2014). Evaluating spatial model accuracy in mass real estate appraisal: A comparison of geographically weighted regression and the spatial lag model. *A Journal of Policy Development and Research, 16*(3), 169-182. Retrieved from https://www.jstor.org/stable/26326913

Bogdanovaa, T., Kamalovaa, R. A., Kravchenkoa, T. K., & Poltorak, A. I. (2020). Problems of modeling the valuation of residential properties. *Business Informatics, 14*(3), 7-23. doi:10.17323/2587-814X.2020.3.7.23

Bogin, A. N., & Shui, J. (2020). Appraisal accuracy and auto- mated valuation models in rural areas. *The Journal of Real Estate Finance and Economics, 60,*, 40-52. doi:https://doi.org/10.1007/s11146-019-09712-0

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11

Borst, R. A. (2006). The comparable sales method as the basis for a property tax valuations system and its relationship and comparison to geostatistical valuation models. *Delft University of Technology, OTB Research Institute for the Built Environment)*.

Darlington, R. &. (2017). *Regression Analysis and linear models: concepts, applications and implementation.* Newyork: Routledge: Guilford.

Dielman, T. (2001). *Applied Regression Analysis.* Pacific Grove, California: Duxbury.

Gloudemans, R. J. (1999). *Mass Appraisal of Real Property.* Chicago, IL: International Association of Assessing Officers.

Hayes, A. F. (2018). *Introduction to mediation, moderation and conditional process analysis: A regression-based approach (2nd ed.).* Newyork, NY: Routledge.

International Association of Assessing Officers. (1990). *Property Appraisal and Assessment Administration.* Kansas City: IAAO.

Ireland, M. a. (March 8, 2010). Introduction to valuation and spatial analysis,. *GIS/CAMA Technologies Conference.* URISA & IAAO.

Kauko, T., & D'Amato, M. (2008). *Mass appraisal methods: an international perspective for property valuers.* John Wiley & Sons. doi: https://doi.org/10.1002/9781444301021

Neter, J., Wasseman, W., & Kutner, M. H. (1990). Analysis of Variance, and Experimental Designs. In I. Homewood, *Applied Linear Statistical Models* (pp. 465-470). Irwin.

Tayani, F., Morano, P., Salvo, F., & Ruggiero, D. M. (2019). Property valuation: the market approach optimised by a weighted appraisal model weighted appraisal model. *Journal of Property Investment & Finance, Vol. 38*(No. 5, 2020), 399-418. doi:10.1108/JPIF-07-2019-0094

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v24i1.11