International Journal of Management Studies and Social Science Research

Assessing Accuracy in Property's Assessed Value Using Linear Regression Analysis

Bishwa Acharya[1*] and Derek Hostas[2*]

[1*]Whitacre College of Engineering, Texas Tech University, 902 Boston Ave, Lubbock, 79409, Texas, United States.
[2*]Department of Civil Engineering, Texas Tech University, 100 Engineering Center, Lubbock, 79409, Texas, United States.

DOI: https://doi.org/10.56293/IJMSSSR.2025.5516

IJMSSSR 2025
VOLUME 7
ISSUE 2 MARCH – APRIL **ISSN: 2582 – 0265**

Abstract: The purpose of this project is to assess accuracy in the appraisal value of property using statistical data analysis. Using regression analysis on data collected from residential properties of 88th Street in Lubbock, Texas, we determine whether the actual appraisal values of the properties are overestimation or underesti- mation. Our finding reveals several instances of overestimation and a few of underestimation.

Keywords: accuracy assessment, statistical analysis, regression analysis, property tax appraisal

1. Introduction

A property value assessment is a formal evaluation of a property's worth. It's typically conducted by a professional appraiser or assessor. For tax purposes, local governments employ appraisers to value properties. For other reasons including buying or selling a property, obtaining a mortgage, or settling an estate, private assessors are hired. Such assessments are necessary for multiple reasons such as property tax calculation, real estate transactions, mortgage lending, insurance purposes, or while settling legal disputes. Hence, having a accurate assessed value is significant.

Many works have been conducted focusing on providing accurate assessment or to assess their accuracy. Eckert et. al [1]. provided mass appraisal techniques enhancing the accuracy of residential property valuation models. McCluskey et. al [2] utilized mul- tivariate regression analysis in property valuation. Pagourtzi et. al. [3] comprehensively reviewed of real estate valuation methods, including regression analysis. Hendriks [4] focused on optimizing real estate value estimation using linear regression models. These researches predate modern statistical software, data visualizations techniques, and are limited to their contemporary relevance.

A recent focus of research has been the use of hedonic, large dataset-based mod- els. Gibbons et. al. [5] investigates the impact of environmental factors on property values using hedonic regression models. Some works have been focused to investigate factors outside of property's attributes. Opoku and Abdulai [6] examined the impact of infrastructure development on property values using regression analysis. Huang et. al. [7] made use of machine learning techniques to enhance the accuracy of property valuation models and compared them with traditional regression approaches. Ye and Wu [8] combined deep learning with regression models to predict housing prices, high- lighting data-driven approaches. Xie and Fang [9] combined regression analysis with geospatial data to enhance property value prediction accuracy, considering both spa- tial and temporal factors. The limitations of these research are that they need to large datasets, they are computationally intensive and requires more complex interpretation compared to simpler regression models.

We undertook this project to highlight the application of linear regression analysis to assess accuracy when data and the attributes are limited. This project was executed also because of the need of a quick and reliable analysis on property appraisal values. The data used for this project were obtained from the official website of Lubbock Central Appraisal District and the R software used was under Texas Tech University's license.

2. Methodology

This project makes use of linear regression analysis to model the assessed value of the property and then to make predictions with it to assess the accuracy of the valuations. The widespread appeal and utility of regression analysis stems from its straightforward approach of using an equation to represent the relationship between a response variable and a set of associated predictor variables [10]. The methods employed in this project are described briefly below.

2.1 Linear regression

A simple linear regression model, that is, a model with a regressor x that has a relationship with a response y that is a straight line. the mean of y is a linear function of x although the variance of y does not depend on the value of x.

2.2 Dredge

Dredge analysis is a statistical method used to explore and compare various models within a hierarchical framework. Dredge analysis helps one identify the "Best" model from a set of candidate models. "Best" can be defined in various ways, such as having the highest Akaike Information Criterion (AIC) [11].

2.3 ANOVA

ANOVA, Analysis of Variance, is a statistical technique used to compare the means of multiple groups. Here it used to test the significance of a model. ANOVA allows one to determine if a more complex model with additional predictors significantly improves the fit over a simpler model.

2.4 Stepwise regression

Stepwise regression is regression method used to select the" best" subset of predictors from a larger set for a regression model. It can be considered as a greedy algorithm that adds or removes predictors based on their statistical significance [12]. In this project, we have chosen p-value of F-statistic as significance criteria to add predictor in forward step-wise regression and remove predictor in backward step-wise regression.

2.5 Adequacy plots

Adequacy plots are graphical tools used to assess the fit of a statistical model to data [10]. The three adequacy plots employed in this project are:

2.5.1 Residuals vs. fitted values plot

It allows to inspect for heteroscedasticity, or unequal variance, in the residu-als. Randomly scattered residuals around the horizontal line at zero, means the variance is constant. Presence of pattern indicates heteroscedasticity.

2.5.2 Normal Q-Q plot

This plot allows to check for the normality of the residuals. The points falling approximately on a straight line suggests that the residuals are normally distributed. Deviations from the line indicate non-normality.

2.5.3 Residual vs leverage plot

It can be used to identify influential points. Points that are further away from the cloud of points in the center of the plot are potential high-influence points.

2.6 Box-Cox transformation

It is a statistical technique used to transform non-normal data into a more normal distribution. It is a data transformation technique, not a statistical test. The value of transformation parameter, λ is chosen to maximize the likelihood of the transformed data.

3. Case study

We chose to determine the bias in appraised value of property at 6321 88th Street, Lubbock, Texas was chosen using data of properties from 6309-6350 88th Street in Lubbock, Texas. For this we performed the following set of steps based upon need.

3.1 Data preprocessing

To begin the treatment of data, the house number column in the original data frame was assigned as the observation number for easier identification of the properties as well as deleting that column as it is not needed in the regression. Next, the houses at 6314 and 6322 were considered as outliers due to the fact that they were the only properties that included a pool house and this could potentially skew the regression. Finally, the "Homestead Cap Loss" predictor was removed as it can be considered irrelevant as it was for the current tax year and would only increase over time. Additionally, a large portion of observations have a recording of 0 for this predictor which will largely effect the regression. Homestead Cap Loss can further be defined, specifically in Texas, as a tax break given to homestead owners on taxes due on their property. This is calculated by limiting the value to at most 10% of the previous year's appraisal.

Table 1. Description of data

Parameter	Units	Variable	Predictor or Response
Market Value	USD	MV	Response
Total Improvement Market Value	USD	TIMV	Predictor
Total Land Market Value	USD	TLMV	Predictor
Total Main Area	Sq ft	TMV	Predictor
Main Area	Sq ft	MA	Predictor
Main Area Value	USD	MAV	Predictor
Garage Area	Sq ft	GA	Predictor
Garage Value	USD	GV	Predictor
Land	Sq ft	L	Predictor

Table 1 shows the nature of data used for further analysis, the response and predic- tor parameters, the units of their values, the abbreviation-based variables they were assigned to for convenience, and the nature of the each variable they represent.

The structure of data was observed and all predictors are being considered as integers. The predictors considered in this project are the aspects of each property that contribute or potentially contribute to the home's assessed value. These predictors will be considered in different forms and combinations in order to provide the best possible model(s) that shows the importance of certain predictors for future reference.

3.2 Initial regression

Initial raw regression using TMA (Total Main Area), MA (Main Area), GA (Garage Area), L (Land was performed and the summary of it was observed. The initial raw

first order regression without interactions showed significance in the predictors of TMA (Total Main Area) and L (Land) as the p-value of their F-statistic of were 0.01967 and 0.00318 respectively. This were significant at 95% confidence interval. The variables of TIMV (Total Improvement Market Value), TLMV (Total Land Market Value), MAV (Main Area Value), and GV (Garage Value) were not considered as predictors because they were not considered to be a direct contributor of the response variable.

Then, first order regression with interaction terms was performed. The summary of it showed no significance in any single predictor. All of the predictor had p-value of F-statistic more than 0.05. Although the multiple r-squared value increased from initial 0.7869 to 0.8005. This was due to multiple terms being added. But this resulted in the decrease of the overall p-value of F-statistic of the model, from initial 3.597e-12 to 6.729e-09. Initially, it was safe to say that interactions of the predictors will not be significant enough to be included in the model.

Initial regression with and without interaction suggested appropriate model to be

MV: lm(TMA + L).

We also considered additional model MV: lm(MA + GA + L). The additional model was also being considered at this point because TMA consists of both MA and GA. The difference between the two models is that the coefficients weigh slightly different in terms of which predictors effect the regression more. Both were being considered due to their similarity.

3.3 Dredge analysis

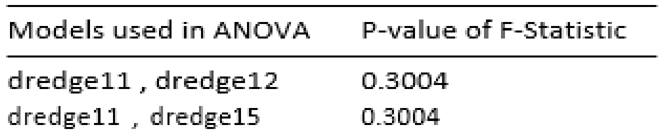
According to the dredge analysis, four models were shortlisted, based on their AIC values. The models shortlisted, their AIC values, their predictors, and the variables they would be represented are in Table 2.

Table 2. Results of dredge analysis

Variable	Predictors	AIC
dredge11	TMA, L	953.9
dredge12	TMA, GA, L	953.9
dredge15	TMA, MA, L	953.9

3.3.1 Comparing results of dredge analysis with ANOVA

Table 3. ANOVA on dredge results



Models used in ANOVA	P-value of F-Statistic
dredge11 , dredge12	0.3004
dredge11 , dredge15	0.3004

Observing the ANOVA results in Table 3, the shortlisted models according to the dredge analysis and confirming it via ANOVA are, MV: lm(TMA + L) and MV : lm(MA + GA + L). Simply, the market value of a home is solely dependent upon the land and total main area square footage.

3.3.2 Comparing results of dredge analysis with stepwise regression

For this project, since certain variables were only being considered to be predictors, only forward and backward step regression will be used. Backward step regression is more robust to multi-collinearity and forward is good for exploratory analysis.

Results from both of these confirmed model, model1 MV: lm(TMA+L) to be the best one so far. But, we also kept the additional model, model2 MV: lm(MA+GA+L) till then.

3.4 Visual inspection of adequacy plots

Three plots for each of the two shortlisted models were plotted to check the models' adequacy. They are residual vs fitted plot, normal Q-Q plot, residual vs leverage.After observing the adequacy plots of the two models, it is clear that there is an increasing pattern in the constant variance plots and skewness with light tails in the normal probability plots. Due to this observation in the adequacy plots, next the models were analyzed through Box-Cox Analysis to see if there is a need for a transformation.

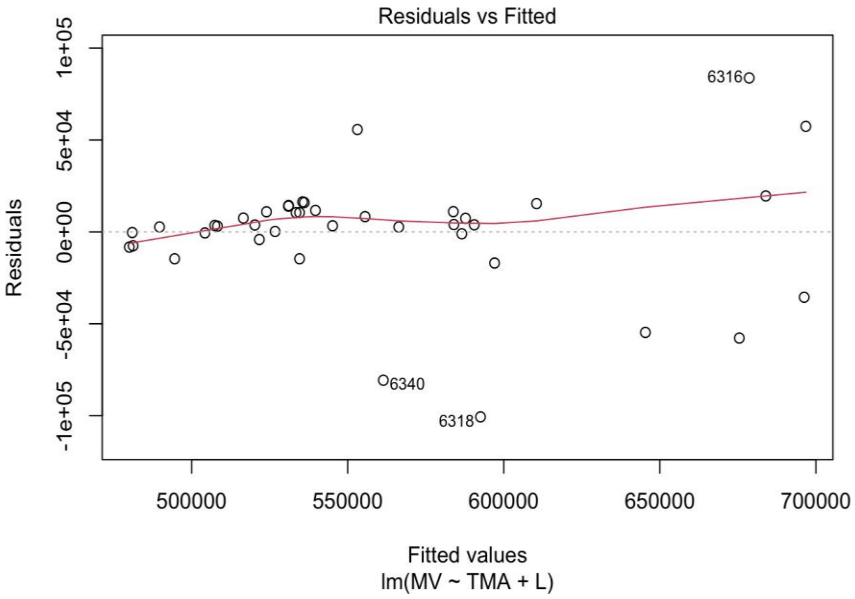


Fig. 1. Residual vs Fitted values plot for model1

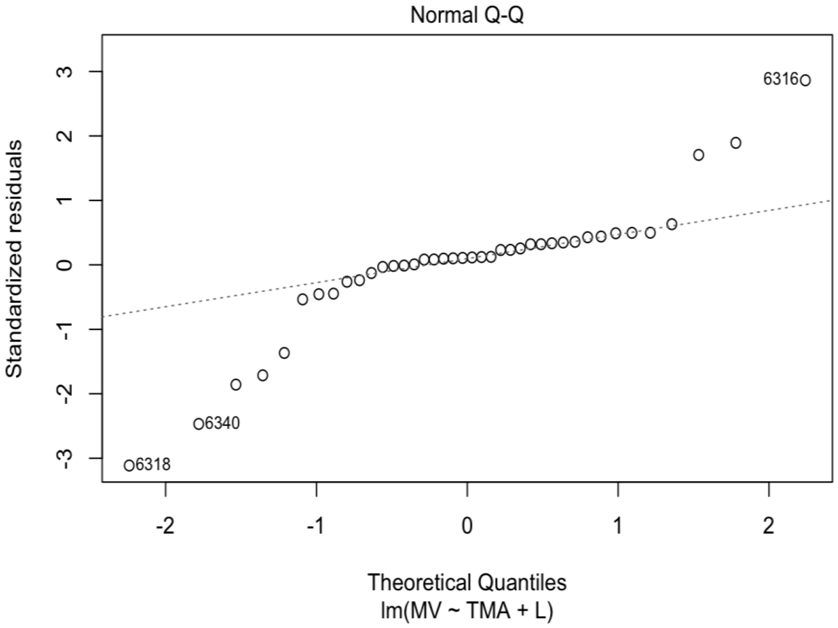


Fig. 2. Normal Q-Q plot of residuals for model1

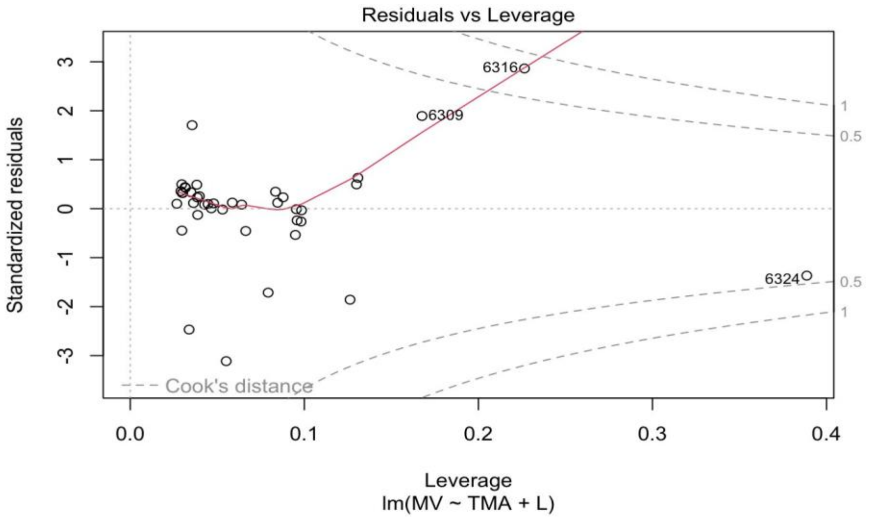


Fig. 3. Residual vs Leverage plot for model 1

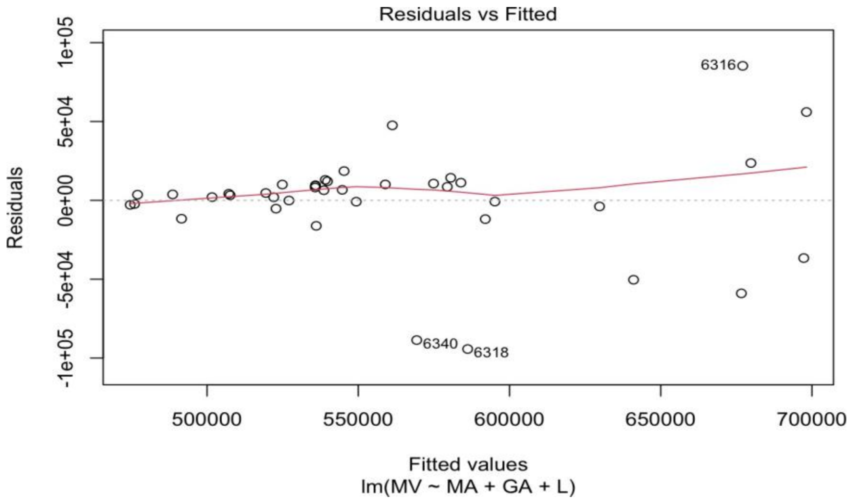


Fig. 4. Residual vs Fitted values plot for model2

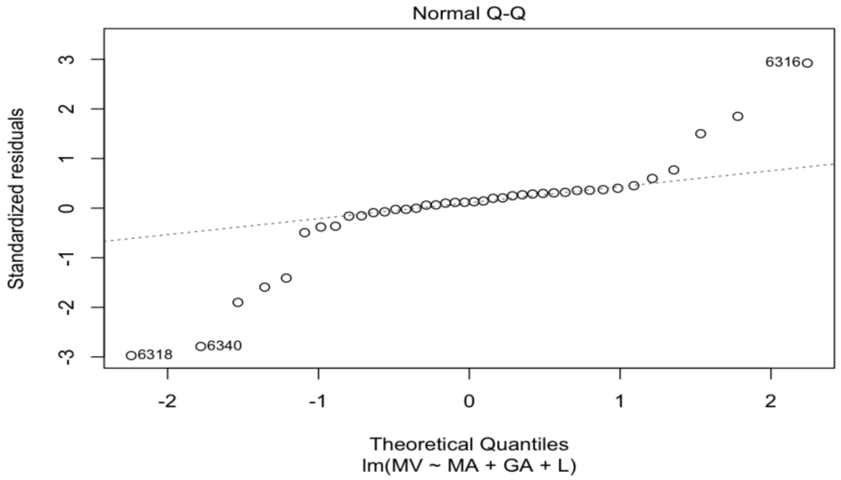


Fig. 5. Normal Q-Q plot of residuals for model2

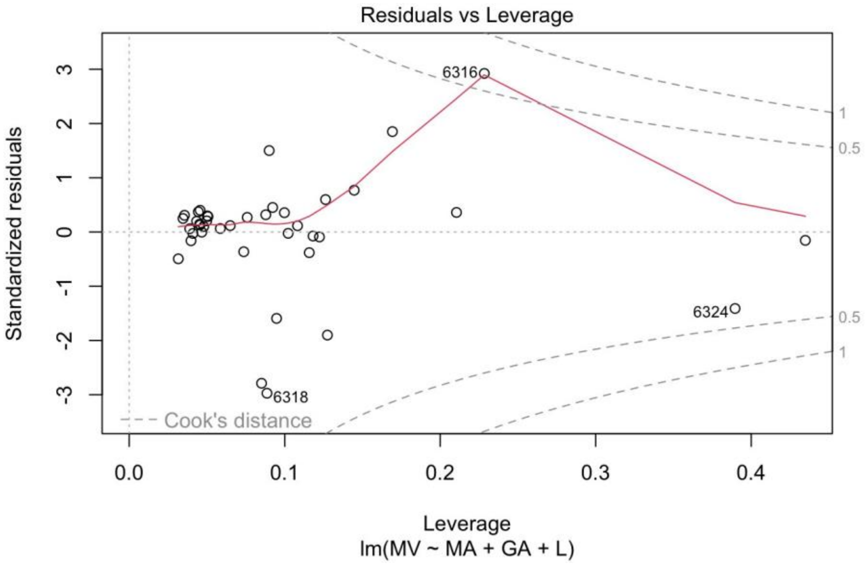


Fig. 6. Residual vs Leverage plot for model2

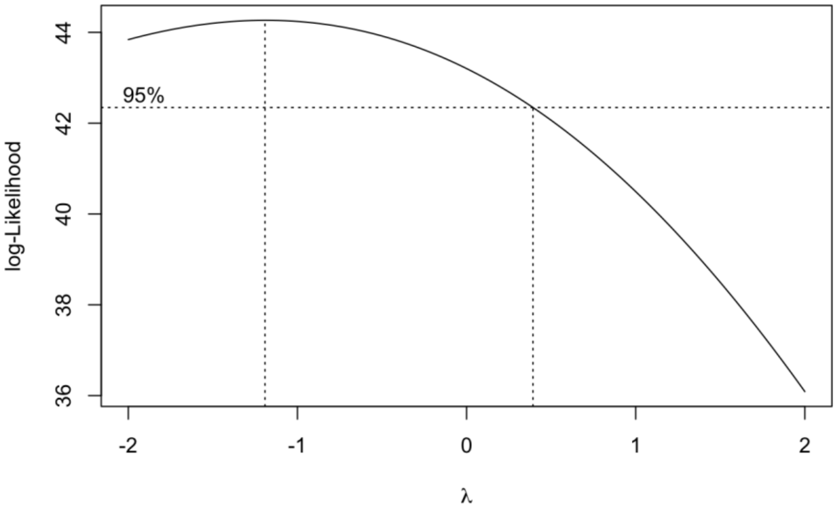


Fig. 7. Log-likelihood vs λ plot for model1

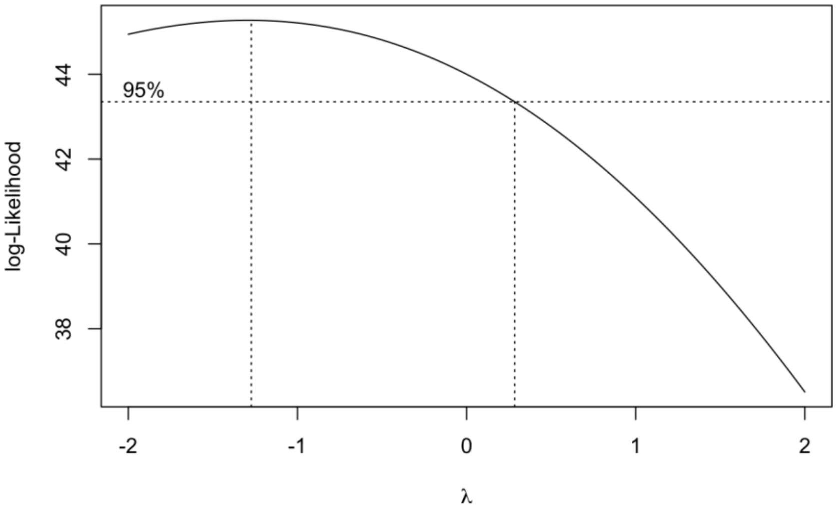


Fig. 8. Log-likelihood vs λ plot for model2

3.5 Box-Cox analysis

Log-likelihood vs λ plots for both models are shown in Figure 7 and 8. The value of λ = 1 is clearly not in the 95%confidence intervals for either model. Next, the value for lambda should be found at the maximum of the function in the Box-Cox plots to determine the power of the transformations.

3.5.1 Power transformation

The lambda value needed for the power transformation for model1 and model2 resulted in being -1.19 and -1.27 respectively. Using these powers, both models responses were transformed. Linear regression on each were performed and the adequacy observed and outliers were eliminated. It did not improve the significance of the models in terms of r-squared and p-values of F-statistic, so there is not a large concern in eliminating these observations. Table 4 summarizes the predictors and their corresponding p-value for F-statistic for both of these transformed models.

Table 4. Linear regression summary on transformed models

	Transformed Model 1		Transformed Model 2		
Predictors	TMV	L	MA	GA	L
P-value	1.73e-08	0.0186	2.58e-06	0.0176	0.0368

3.6 Discussion

The two transformed models we obtained are given by equations 1 and 2:

$$\text{tf MV 1} = 2.766e - 07 - 2.939e - 11(TMA) - 3.824e - 12(L) \qquad (1)$$

$$\text{tf MV 2} = 1.016e - 07 - 1.334e - 11(MA) - 7.314e - 12(GA) - 1.241e - 12(L) \ (2)$$

Here, tf MV 1 and tf MV 2 represent response-transformed model1 and model2.

Both of the original models, model1 and model2, are very similar in terms of adequacy, r-squared, and overall p-values, but produce slightly different responses. Both are under transformations to improve model adequacy, so when using either model, the response will need to be transformed back.

Fitting the data on the transformed model 1, and also calculating the confidence and prediction intervals, the calculations were transformed back. These transformed back intervals are shown in Figure 9.

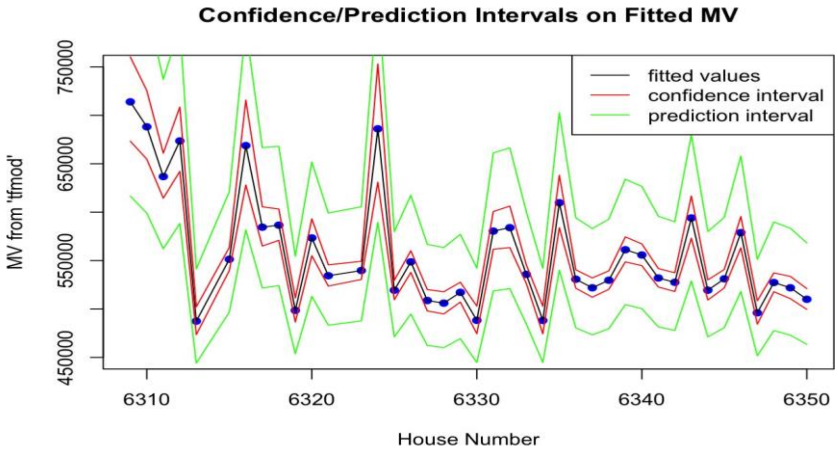


Fig. 9. Confidence and prediction intervals for model1

Observing the output above, it is clear to see based off the model regressed with the Total Main Area (TMA) and Land (L) predictors that the assessed value of the home at 6321 88th Street is greater than it should be by about $18,000. Observing the evaluation of the model 2 output, similarly, the model regressed with the Main Area (MA), Garage Area (GA), and Land (L) predictors shows that the assessed value of

the home is still greater than it should be by about $15,000. Furthermore, observing the entire neighborhood with the regression models concluded in the project, a total of 14 properties have appraisal values less than the evaluated regression values and 26 properties have appraisal values greater than the regressed values. This is true for both regression models being observed.

4. Conclusion

This project aimed to assess accuracy of assessed value of property employing statis- tical data analysis, using the neighborhood data. With these tools, it came up with two linear models, both of which concluded that the property's assessed value was an overestimation. The presence of overestimation was observed in most of the proper- ties in the neighborhood, while for some properties, the assessed value was concluded to be underestimation. This study could serve as a reference for similar investigation when needed to be carried out.The reasons for the overestimates or underestimates were not investigated as it was not within the scope of the project. But it could be possible area of research if more data of the predictors, such as age of the property are available. Further area of research could be use of hybrid or machine learning models using more data.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Data availability the data for this project will be available upon request.
Code availability the code for this project will be available upon request.

References

1. Eckert, J. K., & Gloudemans, R. J. (1990) Improving the Accuracy of Residential Valuation Models Assessment Journal, 6(5):25-32.
2. McCluskey, W. J., & Anand, S. (1999) The Application of Multivariate Regression Analysis in Property Valuations, Journal of Property Investment & Finance, 17(3), 323-334.
3. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003) Real Estate Appraisal: A Review of Valuation Methods Journal of Property Investment & Finance, 21(4): 383-401.
4. Hendriks, D. (2005) Optimizing Value Estimation Accuracy for Real Estate
5. Property Management, 23(2): 130-146.
6. Gibbons, S., Mourato, S., & Resende, G. M. (2014) The Amenity Value of English Nature: A Hedonic Pricing Approach Environmental & Resource Economics, 57(2): 175-196.
7. Opoku, R. A., & Abdulai, M. (2016). The Impact of Infrastructure Development on Property Values in the Accra Metropolis Urban Studies, 53(4): 723-740.
8. Huang, B., Wu, B., & Barry, M. (2016). Machine Learning Approaches to Improv- ing the Accuracy of Property Valuation Models. Journal of Real Estate Research, 38(1): 1-27.
9. Ye, Y., & Wu, Q. (2017). Predicting Housing Prices with Deep Learning and Traditional Regression Models. Journal of Urban Technology, 24(3):35-50.
10. Xie, F., & Fang, L. (2023). Geospatial and Temporal Regression Analysis for Property Value Prediction. Journal of Geographical Systems, 25(2): 247-266.
11. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012) Introduction to Linear Regression Analysis. Wiley.
12. Burnham, K. P., & Anderson, D. R. (2002) Model Selection and Multimodel Inference. Springer.
13. Draper, N. R., & Smith, H. (1998) Applied Regression Analysis. Wiley, 1998.