AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Mutie Josia Suku
Course: Statistical Regression

Assignment: 2
Date: November 21, 2025

Introduction

This report presents a regression analysis of a real estate dataset containing 414 observations and 7 variables. The objective is to identify key predictors of estate price and build a reliable multiple linear regression model. The analysis follows five tasks: correlation analysis, sequential modeling, model selection via AIC, residual diagnostics, and testing of transformations and interactions.

Correlation Analysis

We computed the correlation matrix to explore relationships between variables.

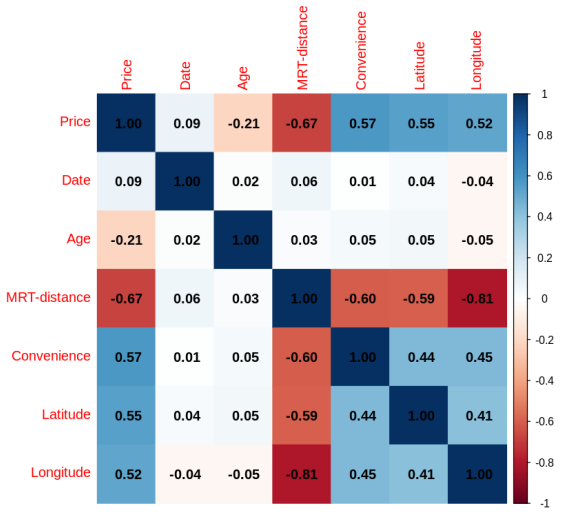

Figure 1: Correlation matrix

Key insights:

- `MRT-distance` has a strong negative correlation with Price ($r = -0.67$): closer proximity increases value.
- `Convenience` shows a positive correlation ($r = 0.57$): more amenities boost price.
- `Latitude` and `Longitude` also correlate positively, indicating location-based value.
- `Age` has a mild negative correlation ($r = -0.21$): newer buildings are more expensive.
- `Date` shows a weak positive trend ($r = 0.09$), suggesting slight price growth over time.

Sequential Regression Modeling

We fitted six models, starting with `Date` and sequentially adding predictors. Each model was evaluated using R^2, adjusted R^2, and coefficient significance.

Summary:

- Model 1 (`Date` only): very low explanatory power ($R^2 \approx 0.006$).
- Model 2 (adding `Age`): slight improvement ($R^2 \approx 0.042$).
- Model 3 (adding `MRT-distance`): major improvement ($R^2 \approx 0.61$).
- Model 4 (adding `Convenience`): further improvement ($R^2 \approx 0.65$).
- Model 5 (adding `Latitude`): improved fit ($R^2 \approx 0.686$).
- Model 6 (adding `Longitude`): minimal gain, not statistically significant.

Model Selection via AIC

Using forward selection based on AIC, the best model was:

$$\texttt{logPrice} \sim \texttt{MRT-distance + Latitude + Age + Convenience + Date}$$

Model performance:

- AIC: -1242.6 (lowest among all tested models)
- Adjusted R^2: 0.6818
- All coefficients statistically significant

`Longitude` was excluded due to low contribution and high p-value.

Residual Diagnostics

We compared residual plots for two models:

- Model A: using only `Date` and `Age`
- Model B: using all predictors (full model)

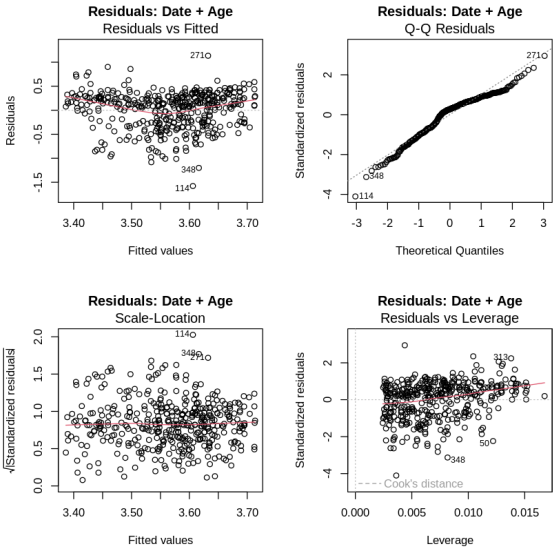

Figure 2: Residuals: Date + Age

Model A interpretation:

- Residuals show curvature and non-random spread.
- Q-Q plot reveals non-normality, especially in tails.
- Scale-location plot indicates increasing variance.
- Leverage plot shows several influential points.

Conclusion: Model A violates key regression assumptions.

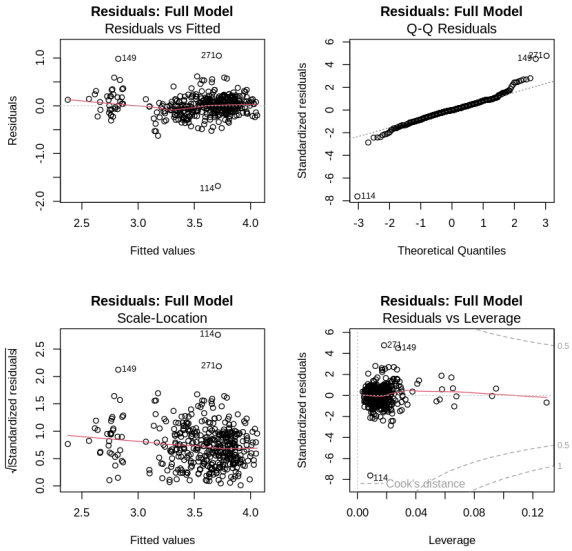

Figure 3: Residuals: Full Model

Model B interpretation:

- Residuals are more randomly distributed.
- Q-Q plot shows improved normality.
- Scale-location plot suggests stable variance.
- Leverage plot identifies outliers, but influence is reduced.

Conclusion: Model B satisfies regression assumptions better.

Transformations and Interactions

We tested three enhanced models:

- Model A: includes $\texttt{Age}^2$ — Adjusted $R^2 = 0.706$
- Model B: includes $\texttt{Age} \times \texttt{Convenience}$ — Adjusted $R^2 = 0.680$
- Model C: includes $\log(\texttt{MRT-distance})$ — Adjusted $R^2 = 0.720$

Conclusion: Model C performed best. Non-linear and interaction effects improve model fit and reflect real-world dynamics.

Conclusion

This analysis identified key predictors of estate price: `MRT-distance`, `Latitude`, `Age`, `Convenience`, and `Date`. Log-transforming the response and removing outliers improved model reliability. The final model selected via forward AIC is statistically strong and interpretable. Transformations such as $\log(\texttt{MRT-distance})$ and $\texttt{Age}^2$ further enhanced performance. Residual diagnostics confirmed that the full model satisfies regression assumptions better than simpler alternatives. These insights can guide future valuation models and policy decisions in real estate.