Linear Regression Modelling of Mortgage Yield

Cruz, Fennessy, Grobbelaar, Volpatti

April 10, 2025

Introduction

The role of a mortgage lender relies significantly on their mortgage yield, that is, the effective return earned from offering a mortgage for a house. Several factors play into forecasting this metric, not least the region that the house has been built in, and the financial profile of that area's residents. In this assignment, new home mortgage yields for 18 SMSAs (Standard Metropolitan Statistical Areas) are predicted from a series of financial and geographic features:

- Average Loan/Mortgage Ratio High values ==> Low Down Payments (X1)
- Distance from Boston in miles (X2)
- Savings per new unit built (X3)
- Savings per capita (X4)
- Pop inc 1950-1960 in
- Percent of first mortgage from inter-regional banks (X6)

Initially, the features and correlations between them are investigated, offering insights into the subsequent linear regression model. The exploratory data analysis will offer insights into interpreting the performance of the final model, which aims to accurately predict the mortgage yield (%).

Exploratory Data Analysis

The Table 1 summarizes the quantitative measures of the selected factors chosen to model the mortgage yield for new houses. This data contains the average monthly yields from March 1963 to April 1964 for the 18 SMSAs covered in this analysis.

The authors divide the causal factors in three categories: (a) Import needs and transfer cost, (b) Risks and (c) Local market structures.

Therefore, we will analyse this variables to reason the regression analysis that will follow.

A. *Import needs and transfer costs.* The uneven distribution of held savings, the differential demand for mortgage funds, and the cost of transferring funds represent the three main components of this category. Column 4 shows that savings per capita vary significantly across SMSAs. However, as illustrated in Fig. 1(d) by the non-linear (flat) relationship and reflected in an r^2 value of 0.049, this variable explains little of the regional yield variations.

More importantly, 42% and 52% of regional yield differences are explained by variations in demand, represented by population growth and savings per unit built, respectively.

For transfer costs, distance from Boston is used as a measure, given that the city records the lowest yields, and previous studies have shown that mortgage costs rise as one moves south and west. Both the graphical representation in Fig. 1(b) and the r^2 value of 0.546 highlight the significance of this variable in explaining yield variance.

B. *Risks*. Among all factors, the loan-to-value ratio exhibited the strongest relationship with mortgage yields ($r^2 = 0.654$), confirming expectations that higher risk levels correlate with higher yields.

C. *Local Market Structure*. The degree of competition is said to influence financing charges, however the authors found that there was not significance for explaining the mortgage yields, hypothesizing that this is reflected in other variables.

Figure 1 illustrates the scatter plots, revealing a non-linear relationship between the independent variables and mortgage yield, with most exhibiting a log-like pattern. This characteristic was considered in the model development to ensure appropriate transformations and improve model performance.

Univariate analysis

Table 1: Mortgage Yield and Explanatory Variables (Simple Regression Results and Summary Statistics)

SMSA	Mortgage Yield (%)	Avg Loan/ Mortgage Ratio	Distance from Boston (miles)	Savings per New Unit Built	Savings per Capita	Population Increase 1950-1960 (%)	% First Mortgage from Inter-regional Banks
Los Angeles-Long Beach	6.17	78.1	3042	91.3	1738.1	45.5	33.1
Denver	6.06	77	1997	84.1	1110.4	51.8	21.9
San Francisco-Oakland	6.04	75.7	3162	129.3	1738.1	24	46
Dallas-Fort Worth	6.04	77.4	1821	41.2	778.4	45.7	51.3
Miami	6.02	77.4	1542	119.1	1136.7	88.9	18.7
Atlanta	6.02	73.6	1074	32.3	582.9	39.9	26.6
Houston	5.99	76.3	1856	45.2	778.4	54.1	35.7
Seattle	5.91	72.5	3024	109.7	1186	31.1	17
New York	5.89	77.3	216	364.3	2582.4	11.9	7.3
Memphis	5.87	77.4	1350	111	613.6	27.4	11.3
New Orleans	5.85	72.4	1544	81	636.1	27.3	8.1
Cleveland	5.75	67	631	202.7	1346	24.6	10
Chicago	5.73	68.9	972	290.1	1626.8	20.1	9.4
Detroit	5.66	70.7	699	223.4	1049.6	24.7	31.7
Minneapolis-St Paul	5.66	69.8	1377	138.4	1289.3	28.8	19.7
Baltimore	5.63	72.9	399	125.4	836.3	22.9	8.6
Philadelphia	5.57	68.7	304	259.5	1315.3	18.3	18.7
Boston	5.28	67.8	0	428.2	2081	7.5	2
Mean ± SD		73.4 ± 3.8	1389.4 ± 975.0	159.8 ± 112.7	1245.9 ± 543.3	33.0 ± 19.1	20.9 ± 13.9
Coefficient of Determination (r^2) with MortYld		0.654	0.546	0.517	0.049	0.419	0.346

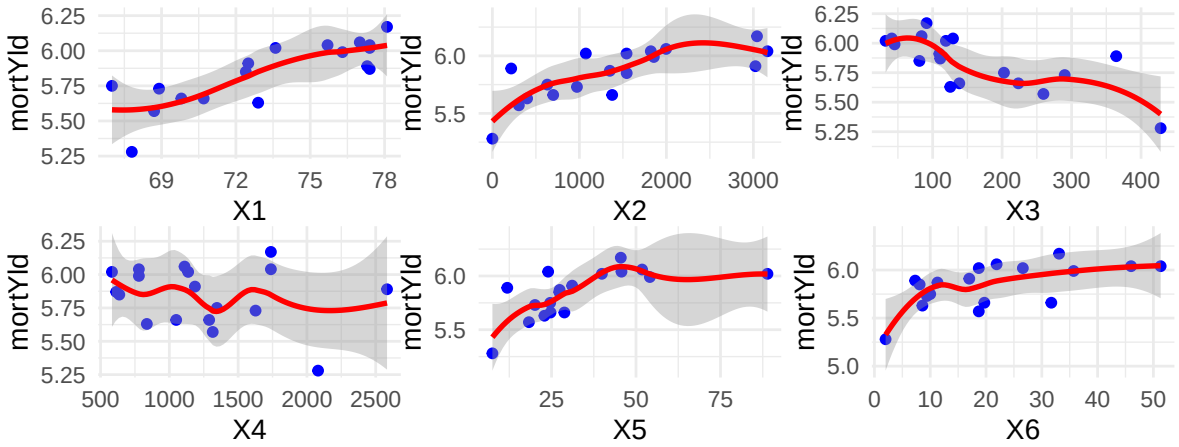

Figure 1: Scatter plots with trend lines

The histograms reveal that most variables exhibit varying degrees of right skewness. Thus, transformations of this variables may be necessary for statistical modeling to improve normality and interpretability, as discussed later.

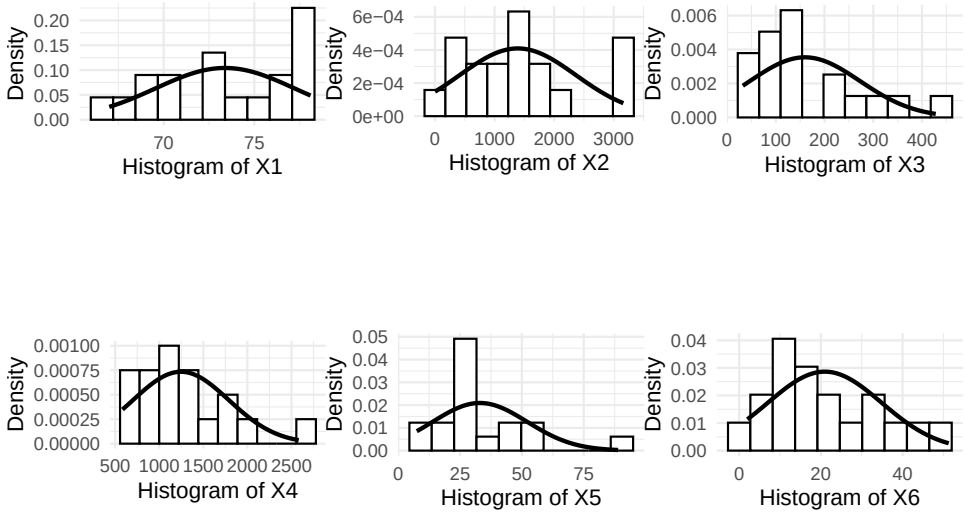

Figure 2: Univariate graphical analysis - Histograms

Correlation

The scatterplot matrix presents the partial correlations among the six explanatory variables (X1–X6), offering insights into their linear dependencies after controlling for other variables. The lower triangular panel displays bivariate scatterplots with fitted trend lines, illustrating potential nonlinear relationships. The upper triangular panel

contains partial correlation coefficients, where values exceeding an absolute value of 0.4 are highlighted in red to denote strong associations. Notably, X3 and X4 exhibit the highest partial correlation (0.9), suggesting a strong linear relationship after adjusting for other variables. This analysis aids in assessing multicollinearity and refining variable selection for subsequent regression modeling.

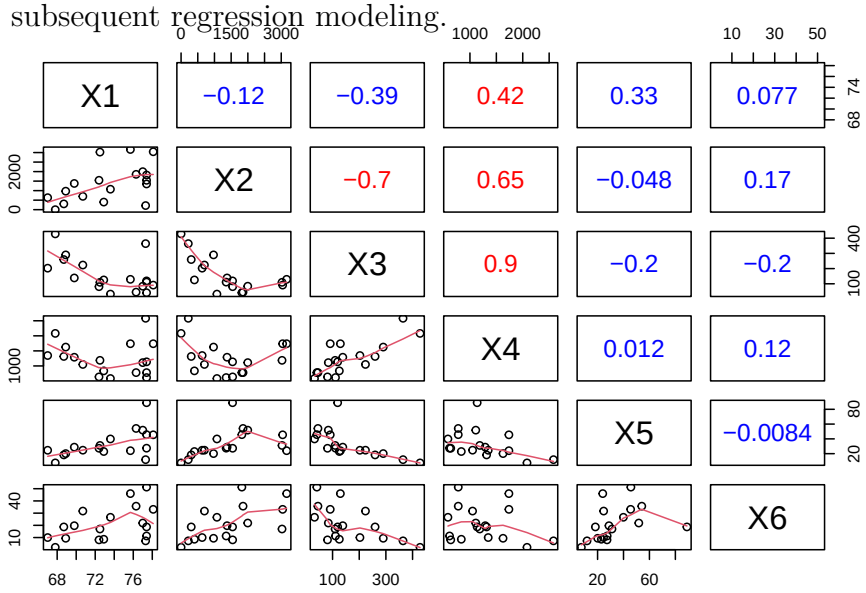

Figure 3: correlations

Model Fitting

Due to the high pairwise correlation that we observe for X3 and X4, we begin by initially considering the Variance Inflation factor (VIF) of a linear model that includes all predictors. VIF will be able to quantify how much inflation there is in our linear model due correlations between predictors

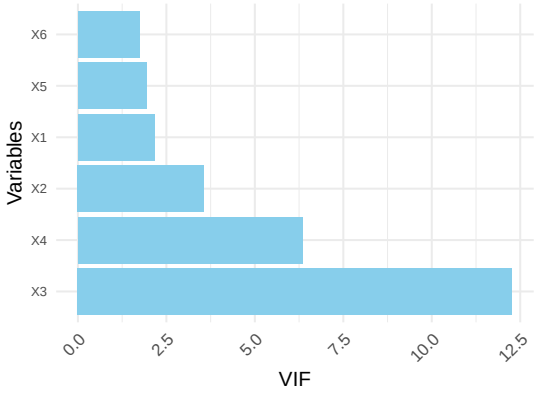

Figure 4: Variance Inflation Factor (VIF)

Step	Df	Dev	Resid.Df	Resid.Dev	AIC
1	NA	NA	12	0.14	−75.17
2 − X6	1	0.00	13	0.14	−77.04
3 − X5	1	0.01	14	0.15	−78.35

Table 2: Stepwise Model Selection Path

Figure 4

We see from Figure 4 that the VIF values for X3 and X4 are above the standardly accepted value of 5, supporting the indication from the correlation matrix, implying high multicollinearity. In Table 1, we have the comparisons of adjusted R^2, and from that

we see that X4 has a significantly lower R^2 than X3. Obersving the results from the correlation matrix, in conjunction with the high VIF values, and the large difference of adjusted R^2 for X3 and X4, we can conclude X4 has low explanatory power in our initial model. Meaning that it is in with in a reasonable assumption that we can remove X4 from our model and not have it be detrimental to the model's explanatory power.

The next step is to consider a linear model excluding X4:

$$\text{mortYld}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{5i} + \beta_5 X_{6i} \tag{1}$$

After selecting our new model, we aim to have a balance of our model's explanatory power and its complexity. To do this we run a stepwise regression on the model. When implemented in R, the process will iteratively add and substract predictors. When considering candidate models in this way, the stepwise regression will aim to minimise AIC, which guides its choice of predictors for the final model.

What we observe from Table 2, is that the stepwise regression process determined that removing X5 and X6 from our model, minimizes the AIC. This results in an optimized linear model of:

$$\text{mortYld}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \tag{2}$$

Now that we have our reduced model, we can focus in on the remaining predictors. As was shown previously in Figure 1, X1, X2, and X3 have seemingly non linear relationships with the mortgage yield. This would indicate that applying a non linear transform to our model, could potentially lead to improved fitting. We chose to consider three potential transforms, namely polynomial transform of the order 2:

$$\text{mortYld}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i}^2 + \beta_5 X_{2i}^2 + \beta_6 X_{3i}^2 \tag{3}$$

A Log of the response(mortgage yield):

$$\log(\text{mortYld}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \tag{4}$$

and lastly the log of the predictors:

$$\text{mortYld}_i = \beta_0 + \beta_1 \log(X_{1i}) + \beta_2 \log(X_{2i}) + \beta_3 \log(X_{3i}) \tag{5}$$

Prior to model comparisons, we can perform a Ramsey RESET test. This will allow us to determine whether additional polynomial terms will add explanatory power. In R implementation, the RESET test will implement a standard F-test, to determine whether the additional polynomial terms have significant value. Our RESET test resulted in a $p = 0.3$, indicating that an inclusion of polynomial terms offers no increased model fitting. This will allows to exclude the polynomial model in the further analysis. \ Now that we have restricted our models to the initial reduced linear model (Eq. 2), the log of the response (Eq. 4), and the log of the predictors (Eq. 5), we evaluate the performance of these models against each other. We do this by looking at the AIC, BIC, RMSE, and Adjusted R^2. We can see the comparisons in Figure. 5, where we observe that the Log-Predictors model out performs the other two models in majority of the metrics. From the AIC and BIC, the lower values for Log-Predictors indicate that it optimizes model fitting and complexity better than the reduced and Log-Response. For RMSE, Log-Response out performs the other two, while for Adjusted R^2 Log-Predictors performs the best again.

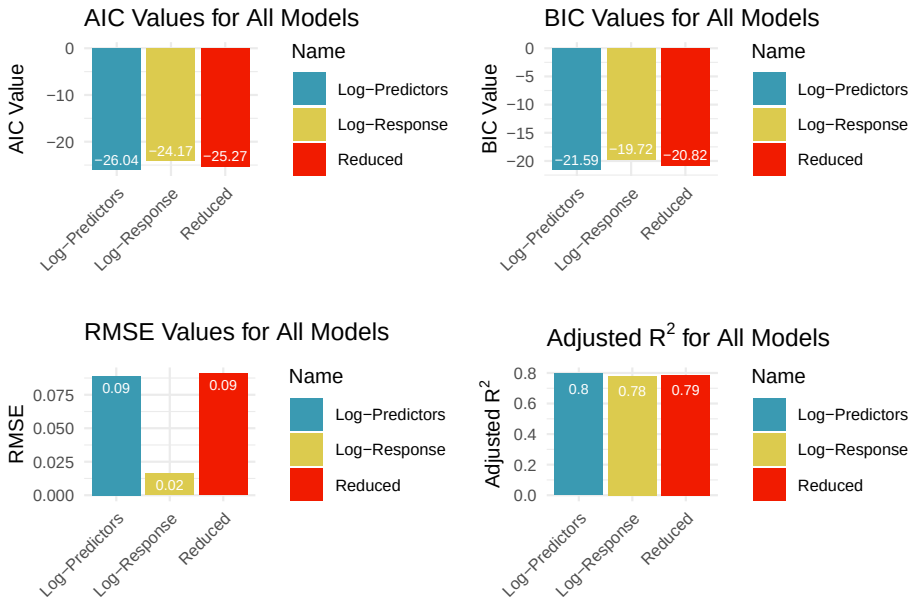

Figure 5: Metrics for Model Comparisons

From these metrics, we can see that the Log-Predictors model is most suitable for modeling Mortgage Yield. Log-Response might perform best in prediction power (RMSE), but the magnitude of difference when compared to Mortgage Yield values is low, supporting then the decision to go with Log-Predictors.

Now that we have selected the Log-Predictors as the best candidate model, we can evaluate it by looking at four diagnostic plots in Figure 6 (Residuals vs Fitted, Scale-Location, Q-Q Residuals, and Residuals vs Leverage). In the Residuals vs Fitted plot we want to look for a horizontal red line, indicating that as we add predictors we still remain with overall constant variance. In our model we can see some curvelinear trend, however, the scale of variation is relatively low around -0.2 to +0.1.

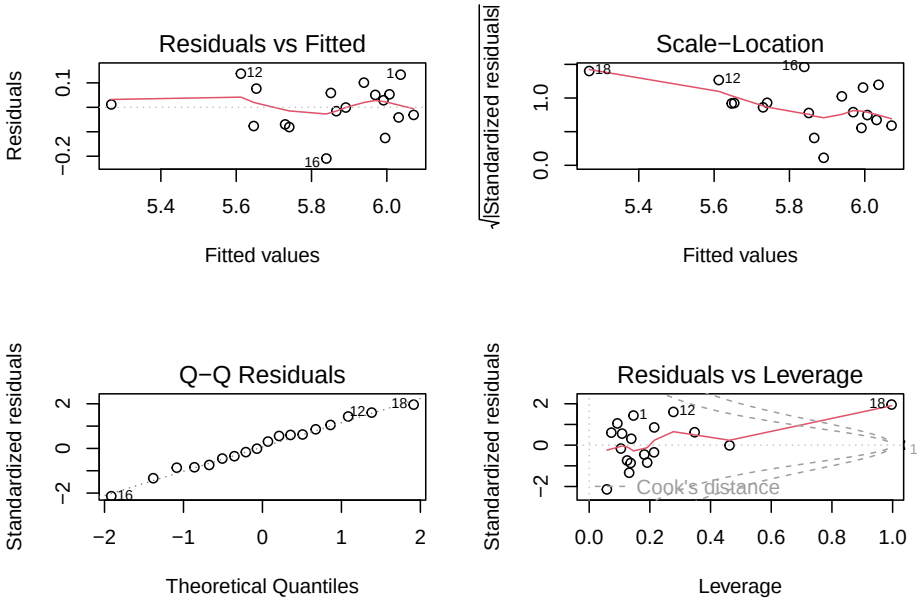

Figure 6: Diagnostic plots for the final model

Scale-Location aims to again asses the spread variance of with predictors. The downward trend of the plot could indicate variance spread. From the Q-Q plot, we can see that the residuals follow the reference line. This is a positive sign for our model, and indicates that the assumption that residuals should follow a normal distribution is satisfied for our model. This implies that hypothesis tests with our model should have high reliability. The Residuals vs Leverage plot is to investigate how influential certain observations are. We can check this by noting if the points are falling within the Cooks distance contour lines. What we observe is that while we have a single observation outside, the remaining points are all within a reasonable range, indicating general model stability. Due to some of homoscedasticity we saw, we decided to run a Breusch-Pagan test (BP test) to validate whether it should be of concern. From the BP test we obtained values of BP = 0.92096, and p-value = 0.8204, indicating that we do not have heteroscedasticity.

Discussion of any shortcomings of the final model

1. *Removal of X4 from consideration:* The introduction of X4 would have led to multicollinearity in the model, which itself leads to less meaningful and less accurate beta predictors. However, there is the possibility that X4 provided some additional information that X3 lacked; such an exclusion would lead to less predictive power and a less accurate model.

2. *The nature of the model training:* Standard practice when designing a model involves the testing of said model on a separate dataset. This allows us to conclude that the model is generalizable and is not simply overfit to the dataset used. As such, in the future, we would consider splitting the given data into testing and training sets so as to test and validate the model.

3. *Issues with the logarithmic model:* The logarithm of values less than or equal to zero is a forbidden operation. This excludes us from analysing data whose values are zero (e.g., houses in Boston, whose distance from Boston are zero). This resulted in us adding an infinitesimally small delta to allow us to perform this operation and consider these datapoints in our model without skewing the others much. In the future, we would consider what the best course of action would be between: 1. Adding the delta, 2. Removing Boston from consideration, 3. Modifying Boston's data by inputting the average distance from Boston from all other cities.

Conclusions

This report found that the log-transformed response model, using average loan-to-value ratio (X1), distance from Boston (X2), and savings per new unit built (X3), best explains mortgage yield variation across 18 SMSAs. This model was selected through stepwise regression and evaluated using AIC, BIC, RMSE, and adjusted R2, achieving an adjusted R2 of 0.80. While statistically sound, the model has limitations—most notably, the exclusion of X4 due to multicollinearity, the absence of a test set for validation, and the need to adjust zero values for log transformation. These limitations suggest areas for refinement. Despite this, the final model offers strong predictive utility and provides insight into the key economic and geographic factors affecting mortgage yields.