Predicting Mortgage Yield using Regression Analysis

Group 42

2025-04-02

1 Introduction

The study of A. H. Schaaf, 1966, "Regional Differences in Mortgage Financing Costs" investigates the existence and causes of regional differences in mortgage financing costs in the United States. While these differences in mortgage yields were decreasing in the early 20th century, they suprisingly remained stable after World War II. The paper explores two main explanations for this phenomenon:

1. Differences in investment value due to risk, terms, and liquidity.
2. Market imperfections such as legal barriers and information gaps.

The data used in this study comes from the Federal Home Loan Bank Board, which contains interest rates and fees in 18 SMSAs (Standard Metropolitan Statistical Areas). The findings suggest that distance from major financial centers, risk levels, and local demand for savings significantly affect mortgage yields. However, market structure and overall savings levels play a lesser important role.

The aim of this report is to analyze the data and develop a predictive model to predict Mortgage Yield (`mortYld` in %) based on 6 explanatory variables:

- **X1:** Loan-to-Mortgage Ratio, in % → High values indicate low down payments.
- **X2:** Distance from Boston, in miles → Measures regional proximity to financial centers.
- **X3:** Savings per New Unit Built, in $ → Indicator of regional credit demand.
- **X4:** Savings per Capita, in $ → Measures local savings levels (credit supply).
- **X5:** Population Increase, 1950-1960, in % → Proxy for housing demand growth.
- **X6:** Percentage of first mortgages from inter-regional banks, in % → Indicator of external financing reliance.

---

2 Exploratory Data Analysis (EDA)

2.1 Load Data and Libraries

Table 1: First few rows of the dataset

smsa	mortYld	X1	X2	X3	X4	X5	X6
Los Angeles-Long Bea	6.17	78.1	3042	91.3	1738.1	45.5	33.1
Denver	6.06	77.0	1997	84.1	1110.4	51.8	21.9
San Francisco-Oaklan	6.04	75.7	3162	129.3	1738.1	24.0	46.0

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

Here is a display of the data, on the first few rows of the dataset. It contains 8 columns. smsa is the Standard Metropolitan Statistical Area, which is the name of the city/region. Mortgage yield (mortYld) is the dependent variable, and X1 to X6 are the six variables. We can observe that all data are numerical values and there is no missing value for each region.

2.2 Univariate Analysis

2.2.1 Summary Statistics

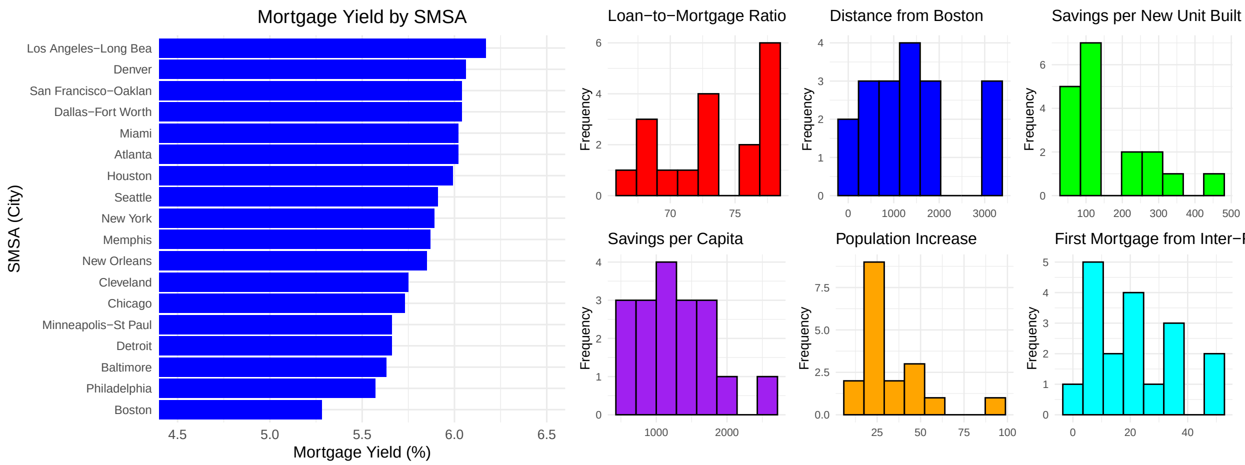
Table 2: Summary Statistics of Variables

mortYld	X1	X2	X3	X4	X5	X6
Min. :5.280	Min. :67.00	Min. : 0	Min. : 32.3	Min. : 582.9	Min. : 7.50	Min. : 2.00
1st Qu.:5.678	1st Qu.:70.03	1st Qu.: 648	1st Qu.: 85.9	1st Qu.: 792.9	1st Qu.:23.18	1st Qu.: 9.55
Median :5.880	Median :73.25	Median :1364	Median :122.2	Median :1161.3	Median :27.35	Median :18.70
Mean :5.841	Mean :73.38	Mean :1389	Mean :159.8	Mean :1245.9	Mean :33.03	Mean :20.95
3rd Qu.:6.020	3rd Qu.:77.22	3rd Qu.:1847	3rd Qu.:218.2	3rd Qu.:1556.6	3rd Qu.:44.10	3rd Qu.:30.43
Max. :6.170	Max. :78.10	Max. :3162	Max. :428.2	Max. :2582.4	Max. :88.90	Max. :51.30

Through this summary, we already observe that mortgage yields don't vary much across regions. Most values are between 5.2% and 6.2%, suggesting relatively stable mortgage rates.

Loan-to-Mortgage Ratios (X1) are concentrated in between 67% and 78.1%, indicating relatively consistent lending practices across regions. Distance from Boston (X2) has a vast range (0–3162 miles), highlighting geographical diversity and potential financial access disparities. Savings per New Unit Built (X3) and Savings per Capita (X4) are characterized by means bigger than medians, representing right-skewed distributions, thus suggesting regional imbalances in credit demand and supply/in housing affordability across regions. Population Increase (X5) from 1950 to 1960 varies widely (7.5–88.9%), reflecting differing housing market pressures. Lastly, Percentage of First Mortgages from Inter-Regional Banks (X6) spans from 2.0% to 51.3%, meaning that some areas depend heavily on external financing while others rely more on local institutions.

2.2.2 Graphical Representation



With deeper analysis, although the variation across SMSAs is small, we see that regional differences still exist in mortgage yields, possibly due to economic factors like savings, loan terms, and regional banking practices.

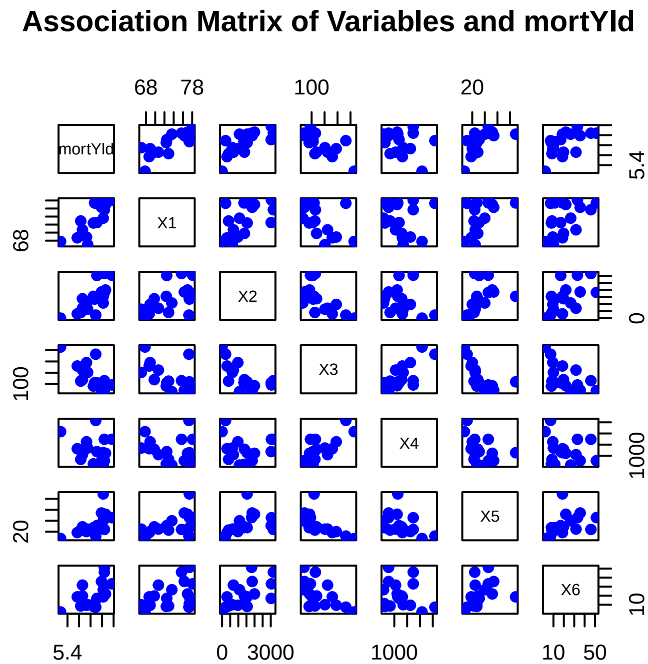
The histograms confirm the distribution of the explanatory variables.

The Loan-to-Mortgage Ratio (X1) shows low variance with most values concentrated between 67% and 80%, possibly indicating limited variability across regions. Distance from Boston (X2) displays a wide and almost homogeneous distribution, reflecting substantial geographic spread among SMSAs. Savings per New Unit Built (X3) and Savings per Capita (X4) both exhibit right-skewed distributions, suggesting that a few cities have notably higher savings levels. Population Increase (X5) is also highly right-skewed with one major outlier (increase of %~%25%), indicating that most regions had moderate growth, while a few experienced rapid expansion. Finally, the percentage of First Mortgages from Inter-Regional Banks (X6) is also right-skewed, with most cities relying minimally on external financing and a few showing heavy dependence.

Overall, the data suggests regional variation in housing finance conditions, credit accessibility, and mortgage market dynamics.

2.3 Bivariate Numerical Analysis

2.3.1 Association Analysis

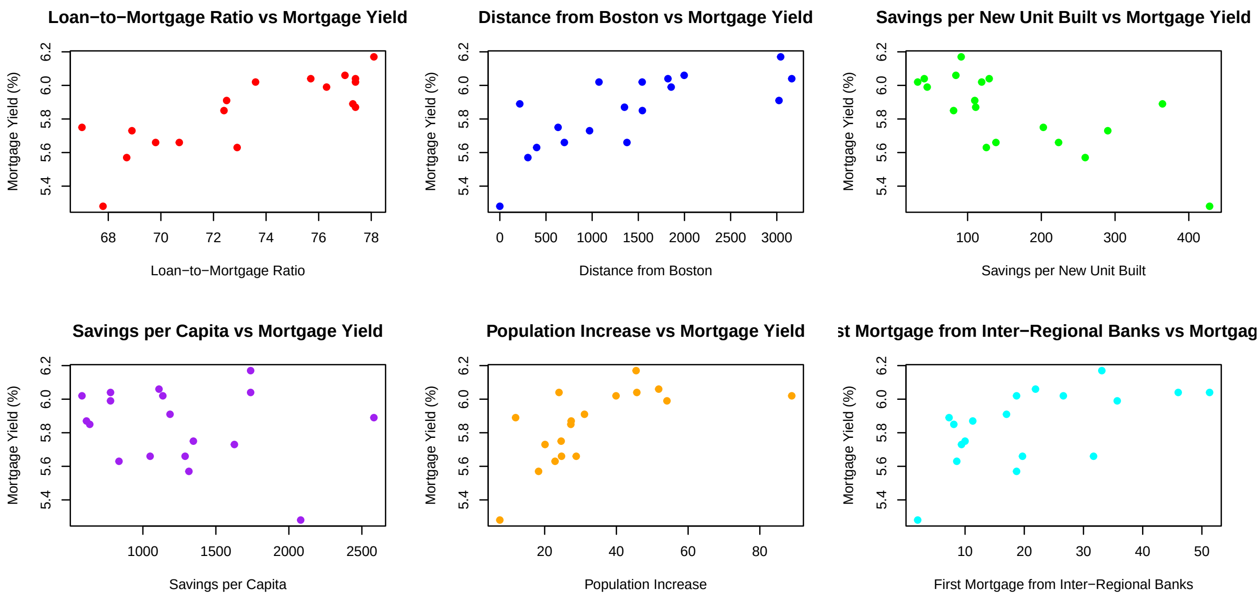


The Association Matrix provides a quick visual assessment of linearity, strength of linear associations among predictors, and outlier detection. It complements numerical analyses like the correlation matrix and VIF.

We visualize bivariate relationships, i.e. how each variable relates to the others and mortYld, and assess if a relationship looks linear, curved or weak, as well as positive or negative. We can also spot outliers or cities that don't follow the general trend.

We can see that most of the plots are random dispersion, while some are linear, and some are curved. X3 seems to be positively associated with X4 and negatively with X5. X2 and X3 seem exponentially associated. X6 seems to be negatively associated with X3.

Let's take a closer look into the Association Matrix, regarding the relationship between Mortgage Yield (%) and the explanatory variables (x-axis), representing the first row in the precedent figure.

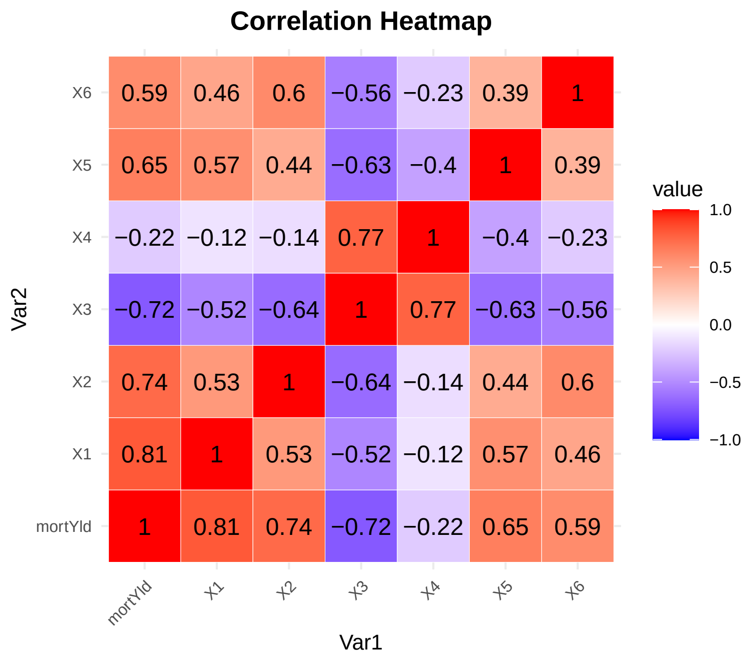


1. Loan-to-Mortgage Ratio: As this ratio increases, the Mortgage Yield increases. This suggests a positive correlation, and that higher Loan-to-Mortgage Ratios (more borrowed money relative to the property value) are associated with higher mortgage yields.
2. Distance from Boston: There is a positive correlation. Boston represents a major financial center with surplus capital. Regions further from Boston might have higher yields.
3. Savings per New Unit Built : There seems to be a negative correlation. This indicates that areas with more savings dedicated to new construction have better access to local financing, resulting in lower mortgage yields.
4. Savings per Capita: The relationship is less distinguishable but appears to be a weak negative correlation or a random dispersion.
5. Population Increase: There is a positive association which can be seen as a square-root relationship. High population growth may imply higher demand for housing, increasing mortgage yields due to heightened competition for available funds. We can observe a potential outlier at the right side of the plot.
6. First Mortgage from Inter-Regional Banks: No clear trend. It seems like the reliance on external financing (measured by the Percentage of First Mortgages from Inter-Regional Banks) does not significantly influence mortgage yields.

To resume:
- X1, X2 and X5 seem to be the most influential variables positively correlated with Mortgage Yield.
- X3 is the most influential variable negatively correlated with Mortgage Yield.
- X4 and X6 variables show weak relationships with mortgage yields.

These observations support the findings of Schaaf (1966) that distance from financial centers, risk factors, and local demand for savings contribute to yield variations.

2.3.2 Correlation analysis



X3 and X4 are strongly correlated (0.77).

X2 and X3 have a high negative correlation (-0.64).

X3 is also strongly negatively correlated with X5 at -0.63.

X1, X2 and X5 have a high positive correlation with mortYld, and X3 a strong negative one. In consequence, we can confirm our previous statements about the scatter-plots.

We can then think about removing one of the highly correlated predictors, if multicollinearity affects the regression model.

These correlations tell only about the if the variables are linearly associated. A low value doesn't mean that the variables are not correlated in another way.

We can also see that X2 is only weakly positively correlated ($r = 0.21$) to X6 after controlling for X3; compare this to the much higher simple correlation ($r = 0.60$). In other words, much of the apparent correlation between X2 and X6 can be explained by their mutual positive correlation with X3.

---

3 Model Fitting

In this analysis, all predictors are continuous variables and each observation corresponds to a unique SMSA. Since the dataset contains no grouping or categorical factors with unequal group sizes, this is a standard multiple regression model with one observation per row. Therefore, the design is not factorial and does not involve unbalanced group structures. As a result, the order in which predictors are entered into the lm() function does not influence the coefficient estimates, F-tests, or model interpretation.

3.1 Pairwise Simple Regressions

Table 3: Simple Linear Regressions: R^2 and p-values

Predictor	R_squared	p_value
X1	0.654	0.0000
X2	0.546	0.0005
X3	0.517	0.0008
X4	0.049	0.3763
X5	0.419	0.0037
X6	0.346	0.0103

The table summarizes the strength of individual linear relationships between each predictor (X1–X6) and the mortgage yield using simple linear regression.

- **X1** has the **strongest linear association** with mortgage yield, explaining approximately **65.4% of its variance** and is highly significant ($p < 0.001$)
- **X2** and **X3** also show strong and significant associations ($R^2 = 0.546$ and 0.517, respectively).
- **X5** and **X6** show moderate yet significant associations ($R^2 = 0.419$ and 0.346).
- **X4 (Savings per Capita)** does **not** show a significant relationship with mortgage yield ($R^2 = 0.049$, $p = 0.3763$), suggesting it may not be a strong individual predictor.

This preliminary analysis indicates that variables X1, X2, and X3 are the most promising candidates for predicting mortgage yield in a multivariate model.

3.2 Null Model vs Full Model Comparison

Table 4: Comparison of Null and Full Model (ANOVA)

term	df.residual	rss	df	sumsq	statistic	p.value
mortYld ~ 1	17	0.8485778	NA	NA	NA	NA
mortYld ~ X1 + X2 + X3 + X4 + X5 + X6	11	0.1098038	6	0.738774	12.3349	0.0002523

The ANOVA comparison between the null model and the full model reveals that the full model, which includes the predictors, significantly improves the model fit. The null model (intercept-only) does not explain much of the variation in mortgage yield.

The full model, provides a better explanation of the mortgage yield, as shown by the significant F-statistic and the p-value. This indicates that at least one of the predictors is significantly related to mortgage yield, and the explanatory variables are useful for improving the model.

The model explains approximately **87%** of the variance in mortgage yield, and after adjusting for the number of predictors, **80%** is still explained. This is a strong fit. The overall model is statistically significant, with a very low p-value. Once again, it means that at least one predictor contributes significantly to explaining the variation in mortYld. The intercept appears to be really significant to fit the model.

Table 5: Summary of Full Linear Model

term	estimate	std.error	statistic	p.value
(Intercept)	4.2852	0.6682	6.4127	0.0000
X1	0.0203	0.0093	2.1835	0.0515
X2	0.0000	0.0000	0.2896	0.7775
X3	-0.0016	0.0008	-2.1029	0.0593
X4	0.0002	0.0001	1.7944	0.1002
X5	0.0013	0.0018	0.7267	0.4826
X6	0.0002	0.0023	0.1024	0.9203

However, most variables do not show statistically significant individual contributions. Only X1 and X3 show strong significance ($p \approx 0.05$). Other variables, X2, X5 and X6, do not show strong individual effects. This suggests that a reduced model may be more interpretable.

3.3 Make stepwise regression to select the best model

Table 6: Stepwise AIC Steps

Step	Model	RSS	AIC
Start	X1 + X2 + X3 + X4 + X5 + X6	0.1098	-77.79
Step 1	X1 + X2 + X3 + X4 + X5	0.1099	-79.77
Step 2	X1 + X3 + X4 + X5	0.1109	-81.61
Step 3	X1 + X3 + X4	0.1159	-82.81

Table 7: Residual Summary of Stepwise Model

Statistic	Value
Min	-0.1723
1Q	-0.0189
Median	0.0061
3Q	0.0406
Max	0.1460

Table 8: Coefficients of Final Stepwise Model

term	estimate	std.error	statistic	p.value
(Intercept)	4.2226	0.5814	7.2629	0.0000
X1	0.0223	0.0079	2.8143	0.0138
X3	-0.0019	0.0004	-4.4599	0.0005
X4	0.0002	0.0001	3.0261	0.0091

(mettre en annexe ?)

The stepwise regression process identified X1, X3, and X4 as the most significant predictors of mortality yield, leading to the final model.

It's interesting to see that X4 appears among the 3 most significant predictoes although it shows the weakest correlation in the correlation matrix. Multiple regression measures the effect of each variable while holding all other constant. As X4 has very strong correlation with X3, holding X3 can make the unique contribution of X4 clearer.

The final model explains approximately **83.4% of the variance** in mortgage yield using only these three predictors.
The AIC isn't increased by a lot when keeping the other variables, which means that these are still statistically valid but not so useful. The final is simpler but still explain the data just as well or better.

The residual standard error (0.091) is low, and the overall model is highly significant, indicating a good fit.
We end up with : mortYld = 4.223 + 0.02229.X1 - 0.001863.X3 + 0.0002249.X4

Let's try a model with 2-way interactions.

Table 9: Coefficients of Interaction Model

term	estimate	std.error	statistic	p.value
(Intercept)	5.3710	2.0329	2.6421	0.0229
X1	0.0069	0.0266	0.2601	0.7996
X3	-0.0001	0.0096	-0.0108	0.9916
X4	-0.0009	0.0025	-0.3706	0.7180
X1:X3	0.0000	0.0001	-0.1582	0.8772
X1:X4	0.0000	0.0000	0.4641	0.6516
X3:X4	0.0000	0.0000	-0.1046	0.9185

Table 10: Residual Summary of Interaction Model

	Statistic	Value
0%	Min	-0.1876
25%	1Q	-0.0156
50%	Median	0.0072
75%	3Q	0.0309
100%	Max	0.1552

Table 11: Fit Statistics of Interaction Model

R^2	Adjusted R^2	Residual Std. Error	F-statistic	p-value	DF
0.8698	0.7988	0.1002	12.25	3e-04	6

The interactions increase the complexity of the model for an improvement that seems very small.

We decided not to include a 3-way interaction model in our analysis. Given the small sample size (18 observations), adding high-order interactions would significantly reduce degrees of freedom and increase the risk of overfitting. Moreover, 3-way interactions are often difficult to interpret meaningfully.

3.4 Model Comparison

Table 12: Comparison of Model Performance Metrics

Model	R2	Adj_R2	AIC	Residual_SE	F_statistic
Full Model	0.871	0.800	-24.708	0.100	12.335

Stepwise Model	0.863	0.834	-29.731	0.091	29.493
2-Way Interaction Model	0.870	0.799	-24.600	0.100	12.250

The **Stepwise Model** offers the best trade-off between simplicity and performance: It has the **lowest AIC**, indicating the best model fit among the three. Despite having a slightly lower $\mathbf{R^2}$ than the full model, it achieves the **highest Adjusted** $\mathbf{R^2}$.

It also has the **lowest residual standard error** and the **highest F-statistic**, confirming overall model significance and parsimony.

Table 13: ANOVA Comparison: Stepwise vs Interaction Model

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Stepwise model	14	0.1159	NA	NA	NA	NA
Interaction model	11	0.1105	3	0.0055	0.1813	0.9069

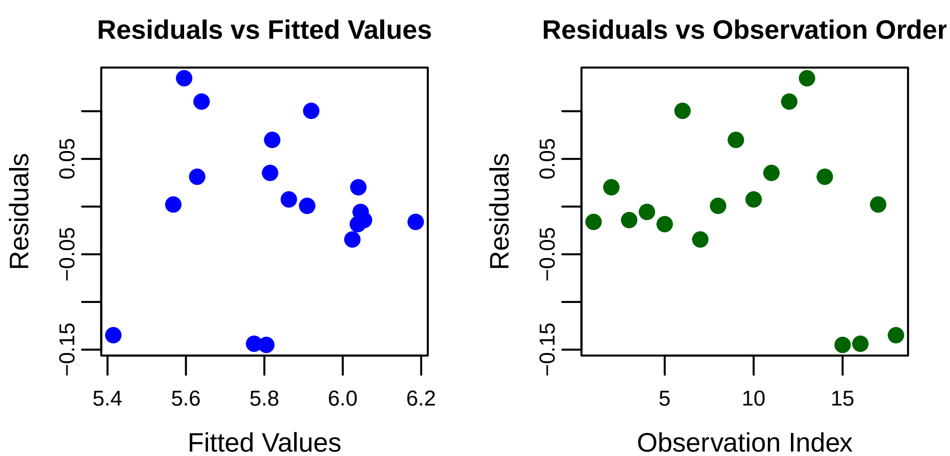
An ANOVA was conducted to assess whether including 2-way interaction terms significantly improved the model fit. The test yielded an F-statistic of 0.18 and a p-value of 0.91, indicating that the additional interaction terms did not meaningfully reduce the residual variance.

As a result, we retained the simpler model with only main effects (X1, X3, and X4), which offers comparable explanatory power and better interpretability.

---

4 Model assumptions and Diagnostics

4.1 Independence evaluation



The 1st graph shows that the residuals appear randomly scattered around 0. There's no clear pattern. This suggests the assumptions of linearity and constant error variance (homoscedasticity) are reasonably met.

The 2nd graph can help us conclude that there is no consistent trend so the independance is verified.

4.2 Multicolinearity diagnostic

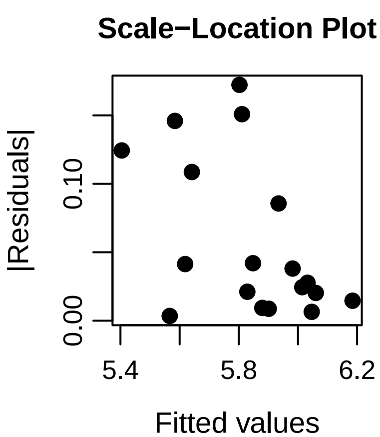
Table 14: Variance Inflation Factors (VIF)

	vif_values
X1	1.886802
X3	4.550125
X4	3.348330

All variables have a VIF value under 5 meaning that variables are not too highly related and that no variable should be eliminated. This confirms our choice of keeping X3 and X4 even if they showed a high correlation coefficient.

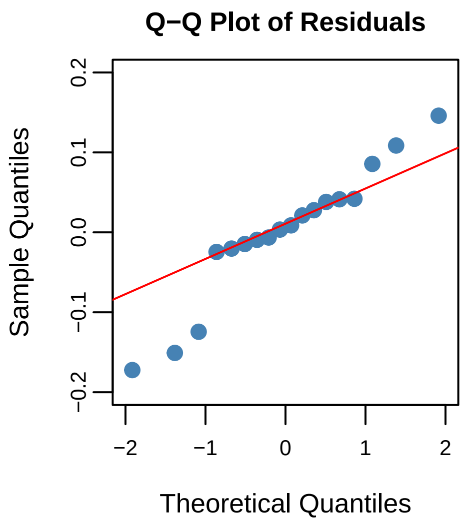
4.3 Homoscedasticity

Random scatter indicates good assumption of homeoscedasticity. If we can distinguish a clear pattern, then we have potential heteroscedasticity issue.



4.4 Normality Check

If points lie on 45 degrees line, it means the residuals are normally distributed. If we can see a curved pattern, then the normality assumption is violated



---

5 Final estimated Model

---

6 Conclusions

- The analysis showed that [mention significant predictors] have a strong relationship with mortgage yield.
- The assumptions of linear regression were [state if met or violated].
- The model provides [good/poor] predictive accuracy based on [R^2 and residual analysis].
- Future improvements could involve [mention possible improvements like transformations, additional predictors, etc.].