Predicting Mortgage Yield using Regression Analysis

Group 42: Clara Delandre, Majandra Garcia, Paola Biocchi, Coline Leteurtre

2025-07-08

1 Introduction

The study of A. H. Schaaf, 1966, “Regional Differences in Mortgage Financing Costs” (Schaaf 1966), 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: differences in investment value due to risk, terms, and liquidity, and 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 model to predict Mortgage Yield (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)

Each SMSA in the dataset is described by its Mortgage Yield as the dependent variable, along with six explanatory variables (X1 to X6). These variables include financial ratios, regional distances, savings indicators, population growth, and bank origination shares. All variables are numerical, and a preliminary check confirms there are no missing values in any of the observations.

2.1 Univariate analysis

2.1.1 Numerical analysis

We begin with a numerical summary of each variable:

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

Table 1: Summary Statistics of all Variables

Through this summary, we already observe that Mortgage Yields (**mortYld**) 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%. 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. Population Increase (**X5**) from 1950 to 1960 varies widely (7.5–88.9%). 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.

2.1.2 Graphical analysis

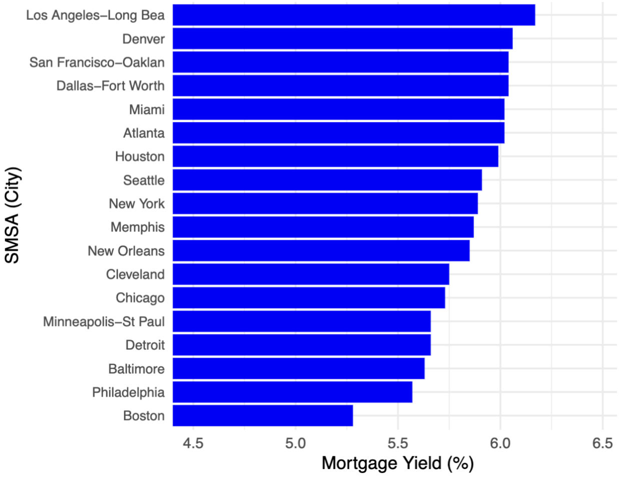


Figure 1: Histogram of Mortgage Yield across SMSAs (Standard Metropolitan Statistical Area)

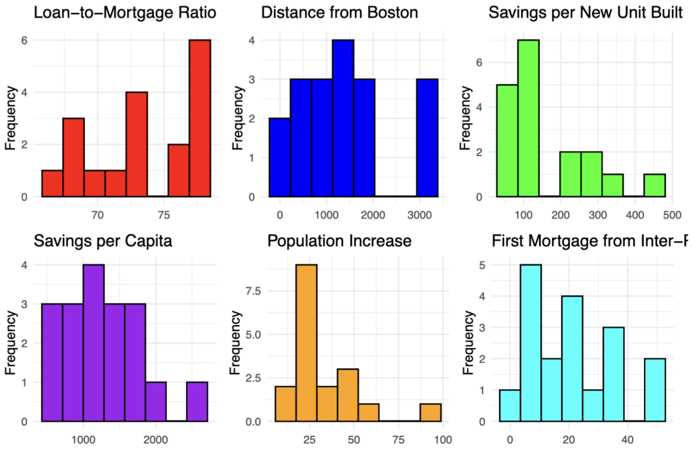


Figure 2: Histograms of Mortgage Yield across Predictor Variables

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, possibly indicating limited variability across regions. Distance from Boston (**X2**) displays a wide and almost homogeneous distribution, reflecting substantial geographic spread among SMSAs. The right-skewed distributions of Savings per New Unit Built (**X3**) and Savings per Capita (**X4**) suggest that a few cities have notably higher savings levels. Population Increase (**X5**) is also highly right-skewed with one potential major outlier, indicating that most regions had moderate growth, while a few experienced rapid expansion. Finally, the percentage of First Mortgages from Inter-Regional Banks (**X6**) show that 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.

2.2 Bivariate analysis

2.2.1 Graphical analysis

The Association Matrix provides a quick visual assessment of bivariate relationships (how each variable relates to the others and `mortYld`), of types of associations among predictors (if a relationship looks linear, curved or weak, as well as positive or negative), and of outlier presence. It complements numerical analyses like the correlation matrix. We can see that most of the plots are random dispersion, while some are linear, and some are curved. **X3** is positively associated with **X4** and negatively with **X5**. **X2** and **X3** are negatively exponentially associated. On another side, **X6** is negatively associated with **X3**.

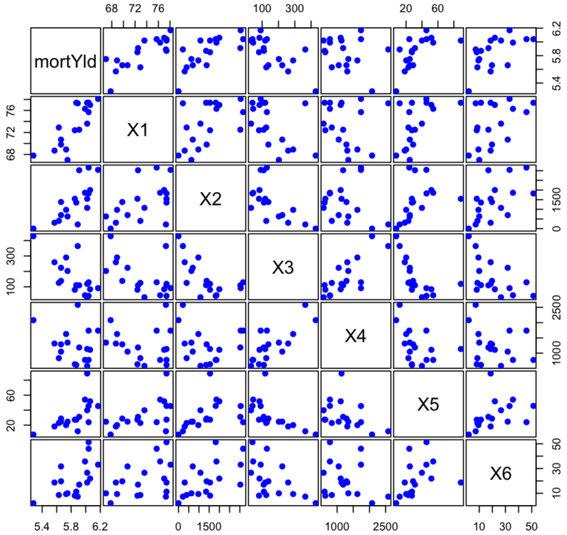


Figure 3: Association Matrix Between Mortgage Yield and Predictor Variables

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.

As **X1** 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. **X2** reveals a positive correlation with `mortYld`. Boston represents a major financial center with surplus capital.

Regions further from Boston might have higher Yields. We observe that **X3** is negatively correlated with `mortYld`. This indicates that areas with more savings dedicated to new construction have better access to local financing, resulting in lower Mortgage Yields. **X4**'s influence is less distinguishable but appears to be a weak negative correlation or a random dispersion. **X5** shows 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. **X6**'s variation shows no clear trend. We can interpret that the reliance on external financing does not significantly influence Mortgage Yields.

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

2.2.2 Numerical analysis

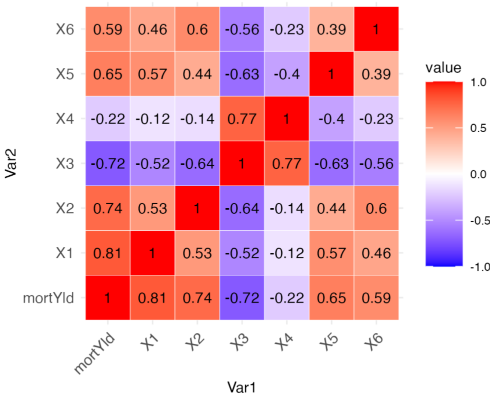


Figure 5: Correlation Heatmap of Mortgage Yield and Predictor Variables. Red = strong positive, Blue = strong negative, White = no correlation.

Now, let's take a look at the correlations between each variable and confirm our previous observations: **X3** is strongly positively correlated with **X4** (0.77) and negatively with **X2** (-0.64), **X5** (-0.63), and **X6** (-0.56). **X1** and **X2** exhibit strong positive correlation with `mortYld`, while **X5** shows moderate positive correlation, and **X3** a strong negative one. **X6** shows moderate positive correlation with `mortYld` as well. **X4** shows only weak correlation with `mortYld`.

This confirms what we saw earlier in the association matrix. We can then think about removing one of the highly correlated predictors, to see if multicollinearity affects the regression model. However, these correlations only indicate if two variables are linearly associated. Thus, a low value doesn't necessarily mean that the variables are not correlated in another way.

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 of the predictors for the linear regression model does not influence the coefficient estimates, F-tests, or model interpretation.

We aim to model the relationship between Mortgage Yield and a set of six explanatory variables using multiple linear regression.

The multiple linear regression model is defined mathematically as:

$$\mathrm{mortYld}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

where:

- $\mathrm{mortYld}_i$ is the mortgage yield for the i-th SMSA,
- X_{1i} to X_{6i} are the explanatory variables,

- β_0 is the intercept,
- β_1 to β_6 are the regression coefficients,
- ε_i is the error term for observation i.

We assume the classical linear regression assumptions:
1. Linearity: The relationship between each predictor and the outcome is linear.
2. Independence: The errors ε_i are independent across observations.
3. Homoscedasticity: The errors have constant variance: $\text{Var}(\varepsilon_i) = \sigma^2$.
4. Normality: The errors follow a normal distribution: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
5. No multicollinearity: The predictors are not perfectly linearly correlated.

The model is fitted using Ordinary Least Squares (OLS), which minimizes the sum of squared residuals:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{18} \left(\text{mortYld}_i - \beta_0 - \sum_{j=1}^{6} \beta_j X_{ji} \right)^2$$

3.1 Null Model vs Full Model Comparison

To test whether the explanatory variables significantly improve the model fit compared to the intercept-only model, we conduct an ANOVA comparing the null model and the full (alternative) model. In order to do so, we test the following hypothesis :

- **Null Hypothesis (H_0)**:

  $$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

  The hypothesis suggests the explanatory variables do not improve the model.
- **Alternative Hypothesis (H_1)**:

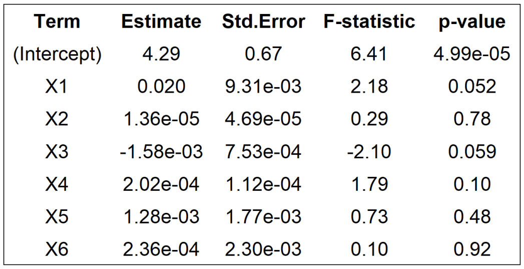
  $$\text{At least one } \beta_j \neq 0 \quad \text{for } j = 1, \ldots, 6$$

  The hypothesis suggests at least one explanatory variable significantly contributes to predicting mortgage yield.

Model	Residual_DF	RSS	DF	SS	F-statistic	p-value
mortYld ~ 1	17	0.85	NA	NA	NA	NA
mortYld ~ X1 + X2 + X3 + X4 + X5 + X6	11	0.11	6	0.74	12.33	2.52e-04

Table 2: ANOVA Comparison Between Null and Full Models. Residual degrees of freedom (Residual_DF) and residual sum of squares (RSS) show the unexplained variance. The degrees of freedom for the model (DF), F-statistic, and p-value test whether the full model significantly improves the fit compared to the null model. A significant p-value (typically $<$ 0.05) indicates that the additional predictors in the full model provide a better fit.

The ANOVA comparison between the null model (intercept-only) and the full model (including all predictors), reveals that the full model better explains the Mortgage Yield, as shown by the significant F-statistic and p-value (p < 0.001). This indicates that at least one of the predictors is significantly related to Mortgage Yield.



Term	Estimate	Std.Error	F-statistic	p-value
(Intercept)	4.29	0.67	6.41	4.99e-05
X1	0.020	9.31e-03	2.18	0.052
X2	1.36e-05	4.69e-05	0.29	0.78
X3	-1.58e-03	7.53e-04	-2.10	0.059
X4	2.02e-04	1.12e-04	1.79	0.10
X5	1.28e-03	1.77e-03	0.73	0.48
X6	2.36e-04	2.30e-03	0.10	0.92

Table 3: Summary of Full Linear Model

The Full model explains ∼87% of the variance in Mortgage Yield, and 80% after adjusting for the number of predictors, which highlights a strong fit. The Residual Standard Error is low, and the overall model is statistically significant, with a very low p-value ($p < 0.001$). Once again, it means that at least one term contributes significantly to explaining the variation in `mortYld`.

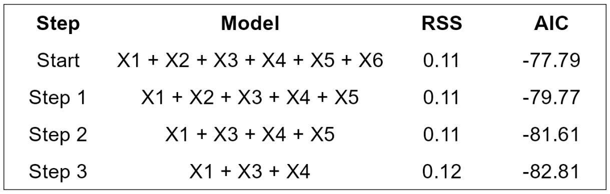
Adjusted_R^2	R^2	Std.Error	DF	F-statistic	p-value
0.80	0.87	0.10	6	12.33	2.52e-04

Table 4: Fit Statistics of Full Linear Model. R^2 shows the proportion of variance explained, Adjusted R^2 accounts for the number of predictors, and Standard Error measures the average distance of observed values from the fitted values.

The intercept appears to be strongly significant to fit the model (p < 0.001). On the other hand, most of the variables do not show statistically significant individual contributions: only **X1** and **X3** show weak significance (p ≈ 0.05), while the other variables, **X2**, **X5** and **X6**, do not show significant individual effects. This suggests that a reduced model may be more appropriate.

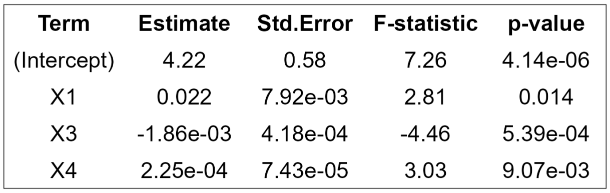
We end up with : $\hat{mortYld} = 4.29 + 0.020\cdot\mathbf{X1} + 1.36\cdot10^{-5}\cdot\mathbf{X2}$ - $1.58\cdot10^{-3}\cdot\mathbf{X3} + 2.02\cdot10^{-4}\cdot\mathbf{X4} + 1.28\cdot10^{-3}\cdot\mathbf{X5} + 2.36\cdot10^{-4}\cdot\mathbf{X6}$

3.2 Make stepwise regression to select the best model



Step	Model	RSS	AIC
Start	X1 + X2 + X3 + X4 + X5 + X6	0.11	-77.79
Step 1	X1 + X2 + X3 + X4 + X5	0.11	-79.77
Step 2	X1 + X3 + X4 + X5	0.11	-81.61
Step 3	X1 + X3 + X4	0.12	-82.81

Table 5: Stepwise AIC Process. AIC = Akaike Information Criterion. Lower AIC indicates a better trade-off between model fit and complexity.



Term	Estimate	Std.Error	F-statistic	p-value
(Intercept)	4.22	0.58	7.26	4.14e-06
X1	0.022	7.92e-03	2.81	0.014
X3	-1.86e-03	4.18e-04	-4.46	5.39e-04
X4	2.25e-04	7.43e-05	3.03	9.07e-03

Table 6: Summary of Final Stepwise Model

Adjusted_R^2	R^2	Std.Error	DF	F-statistic	p-value
0.83	0.86	0.091	3	29.49	2.62e-06

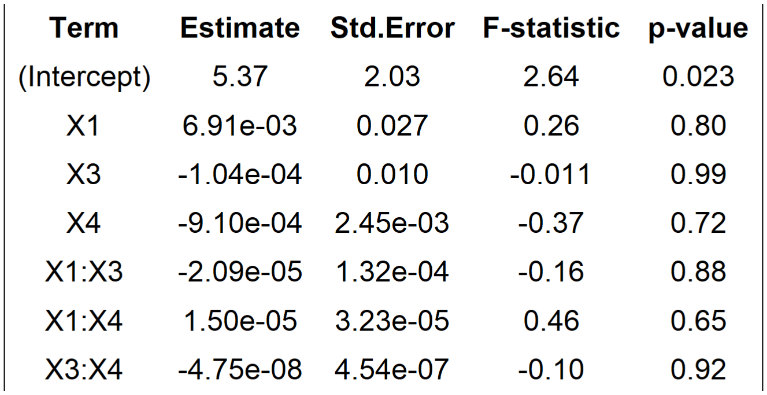
Table 7: Fit Statistics of Stepwise Model

The Stepwise regression process identifies **X1**, **X3**, and **X4** as the most significant predictors of Mortgage Yield, constituting the final model.

It is interesting to note that **X4** appears among the 3 most significant predictors although it shows very weak correlation in the Correlation Matrix. Multiple regression measures the effect of each variable while holding all others constant. As **X4** has very strong correlation with **X3** (0.77), holding **X3** can make the unique contribution of **X4** clearer.

The final Stepwise model explains approximately 83.4% of the variance in Mortgage Yield using only these three predictors. The AIC doesn't increases a lot when keeping more predictors, meaning that even if these predictors can still be statistically valid to keep, they are not so useful to the model. Though the final model is simpler, it explains the data just as well or better than more complex models. The Residual Standard Error (0.09) is low, and the overall model is highly significant ($p < 0.001$), indicating a good fit.
We end up with : $\widehat{mortYld} = 4.22 + 0.022 \cdot \mathbf{X1} - 1.86 \cdot 10^{-3} \cdot \mathbf{X3} + 2.25 \cdot 10^{-4} \cdot \mathbf{X4}$

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



Term	Estimate	Std.Error	F-statistic	p-value
(Intercept)	5.37	2.03	2.64	0.023
X1	6.91e-03	0.027	0.26	0.80
X3	-1.04e-04	0.010	-0.011	0.99
X4	-9.10e-04	2.45e-03	-0.37	0.72
X1:X3	-2.09e-05	1.32e-04	-0.16	0.88
X1:X4	1.50e-05	3.23e-05	0.46	0.65
X3:X4	-4.75e-08	4.54e-07	-0.10	0.92

Table 8: Summary of 2-way Interaction Model

The 2-way Interaction model, which is more complex than the Stepwise model, explains approximately 79.9% of the variance in Mortgage Yield. The Residual Standard Error (0.10) is low, and the overall model is highly significant ($p < 0.001$), indicating that at least one of the terms has a significant influence on Mortgage Yield.

Adjusted_R^2	R^2	Std.Error	DF	F-statistic	p-value
0.80	0.87	0.10	6	12.25	2.60e-04

~~Table 9: Fit Statistics of 2-way Interaction Model~~

None of the variables show statistically significant individual contributions: only the intercept appears to be moderately significant to fit the model ($p < 0.05$). This suggests that a reduced model may be more appropriate.
We end up with : $\widehat{mortYld} = 5.37 + 6.91 \cdot 10^{-3} \cdot \mathbf{X1} - 1.04 \cdot 10^{-4} \cdot \mathbf{X3} - 9.10 \cdot 10^{-4} \cdot \mathbf{X4} - 2.09 \cdot 10^{-5} \cdot \mathbf{X1}:\mathbf{X3} + 1.50 \cdot 10^{-5} \cdot \mathbf{X1}:\mathbf{X4} - 4.75 \cdot 10^{-8} \cdot \mathbf{X3}:\mathbf{X4}$

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.

3.3 Model Comparison

Model	Adjusted_R^2	R^2	RSE	AIC	F-statistic
Full Model	0.80	0.87	0.10	-24.71	12.33
Stepwise Model	0.83	0.86	0.091	-29.73	29.49
2-Way Interaction Model	0.80	0.87	0.10	-24.60	12.25

Table 10: Comparison of Model Performance Metrics. RSE (residual standard error) reflects the typical size of prediction errors; lower values indicate better fit.

The Stepwise model offers the best trade-off between simplicity and performance: it has the lowest AIC (~29.7), demonstrating the best model fit among the three. Despite having a slightly lower R^2 than the Full and 2-ways Interactions model, it achieves the highest Adjusted

R^2. It also has the lowest Residual Standard Error (0.09) and the highest F-statistic (~29.5). This confirms the overall model significance and parsimony.

Model	Residual_DF	RSS	DF	SS	F-statistic	p-value
Stepwise model	14	0.12	NA	NA	NA	NA
Interaction model	11	0.11	3	5.46e-03	0.18	0.91

Table 11: ANOVA Comparison Between Stepwise and Interaction Models

An ANOVA is then conducted to compare the Stepwise Model and the Interaction Model, which are nested — the Interaction Model extends the Stepwise Model by including additional two-way interaction terms. The test yields an F-statistic of 0.18 and a p-value of 0.91, indicating that the additional interaction terms do not significantly reduce the residual variance. As a result, the simpler model with only main effects (**X1**, **X3**, and **X4**) truly provides the best fit, as it also offers comparable explanatory power and better interpretability.

4 Model assumptions and Diagnostics

In order to trust the results of our regression model, we must ensure that the residuals satisfy the following assumptions:

1. The residuals have an expected value (mean) of 0: $\mathbb{E}[\varepsilon_i] = 0$
2. The residuals are homoscedastic (have constant variance): $\mathrm{Var}(\varepsilon_i) = \sigma^2$
3. The residuals are uncorrelated: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
4. The residuals are normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

We evaluate these assumptions using residual diagnostic plots.

4.1 Independence evaluation

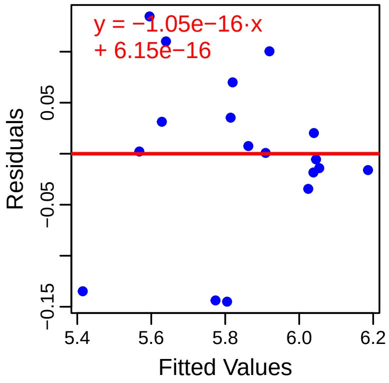


Figure 6: Residuals vs Fitted Plot. It displays residual spread to assess homoscedasticity; no clear pattern indicates constant variance.

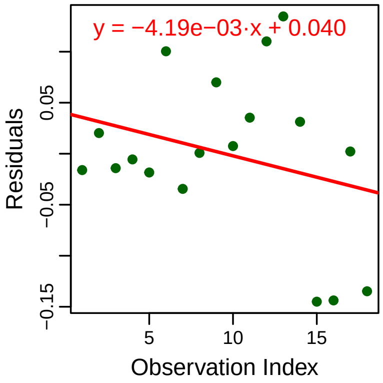


Figure 7: Residuals vs Observation Order Plot. It shows residuals over SMSAs to detect trends or autocorrelation; randomness suggests independence.

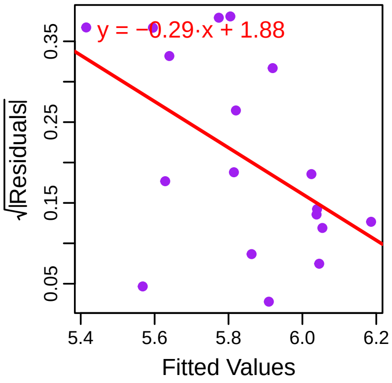


Figure 8: Scale-Location Plot. It shows the variance of residuals versus fitted values to check for homoscedasticity; a horizontal pattern suggests constant variance.

The `Residuals VS Fitted Values` plot on Figure 6 checks for linearity and constant variance. Ideally, residuals should be symmetrically scattered around zero with no clear pattern or funnel shape. In our case, the residuals are randomly dispersed with no visible trend, and the red regression line is nearly flat (slope $= -1.05 \times 10^{-16}$), suggesting that the residuals have approximately zero mean and constant variance. Therefore, assumptions 1 and 2 are satisfied.

The `Residuals VS Observation Order` on Figure 7 is used to detect autocorrelation in the residuals (assumption 3). A patternless distribution across observations indicates independence. The red regression line has a slope of -0.0042, which is close to zero, and residuals appear randomly scattered, suggesting that residuals are **not autocorrelated**. Hence, assumption 3 appears is verified.

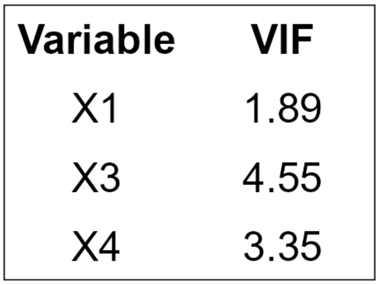
Lastly, the `Scale-Location` plot on Figure 8, shows the square root of standardized residuals vs fitted values. A line with constant spread indicates constant variance : although the slope is somewhat negative (-0.29), the spread remains relatively even. There is no clear increasing or decreasing funnel shape. This is a sign that our model doesn't suffer from heteroscedasticity and is likely a good fit : this supports the homoscedasticity assumption.

In conclusion, based on Figures 6–8, we find that the residuals have a mean close to zero, appear homoscedastic, and show no sign of autocorrelation. Therefore, assumptions 1, 2, and 3 are reasonably satisfied.

4.2 Multicolinearity diagnostic

We assess multicollinearity using the Variance Inflation Factor (VIF). All variables in the final model have VIF values below 5 (see Table 12), indicating that none of the predictors are highly correlated with each other.

Even though variables X_3 and X_4 had a pairwise correlation of 0.77, the VIF values of 4.55 and 3.35 respectively suggest acceptable collinearity. Therefore, these variables still provide enough unique, non-redundant information to justify keeping them in the model In conclusion, there is no evidence of problematic multicollinearity, and all explanatory variables contribute distinct information to the model.



Variable	VIF
X1	1.89
X3	4.55
X4	3.35

Table 12: Variance Inflation Factors (VIF). Values > 5: potential multicollinearity. Values > 10: strong multicollinearity.

4.3 Normality Check

The `Q-Q plot of Residuals` on Figure 9 checks whether the residuals follow a normal distribution (assumption 4). If residuals are normally distributed, the points should align closely with the 45-degree reference line. In our plot, most points fall near the line, particularly in the center, suggesting that the central portion of the distribution follows a normal pattern. However, several points at both tails deviate from the line (at the lower and upper ends of the theoretical quantiles), indicating potential departures from normality in the extremes : this could reflect outliers or heavy-tailed behavior. With only 18 observations, such deviations can be expected and are not strong evidence against normality.

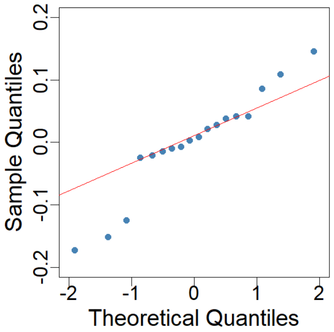


Figure 9: Q-Q Plot of Residuals. Deviations from the line suggest non-normality of model errors.

In conclusion, the residuals appear to be approximately normally distributed, and assumption 4 is reasonably met.

5 Conclusion

The final estimated model is : $\hat{mortYld}$ = 4.22 + 0.022·**X1** - 1.86·10^{-3}·**X3** + 2.25·10^{-4}·**X4** where **X1** is the Loan-to-Mortgage Ratio, **X3** is the Savings per New Unit Built, and **X4** is the Savings per Capita.

The analysis shows that these variables significantly impact Mortgage Yield. Mortgage Yield is positively influenced by the Loan-to-Mortgage Ratio, indicating that higher loan amounts relative to mortgages may lead to better returns for lenders. Conversely, Mortgage Yield is negatively impacted by Savings per New Unit Built, suggesting that more capital saved for construction could reduce reliance on mortgages, leading to lower returns. Finally, Savings per Capita has a positive, though small, effect on Mortgage Yield. As individual savings increase, it may signal a more financially stable environment, leading to slightly better mortgage performance.

While the assumptions of linear regression are generally satisfied, there are some minor deviations. The model shows strong predictive performance, accounting for 83.4% of the variance in Mortgage Yield, with homoscedasticity nearly achieved.

Future improvements could include exploring additional predictors, testing for non-linear relationships, or refining the model to better capture any residual heteroscedasticity.

References

Schaaf, A. H. 1966. "Regional Differences in Mortgage Financing Costs." *The Journal of Finance* 21 (1): 85–94. https://www.jstor.org/stable/2977600.