Comparative Analysis of Manual and Automated Variable Reduction Methods for Housing Price Regression

Viet Le

Abstract

This analysis compared two variable reduction techniques – manual analytical assessment and automated BIC selection – on a regression model predicting home prices. The model incorporated eight predictors like living space and location, achieving an R-squared of 0.57. Manual assessment reduced the model to four variables, retaining the R-squared of 0.56. BIC selection identified a six-variable model with R-squared of 0.569. ANOVA found the manual reduction significantly degraded fit while the BIC method did not. Overall, BIC selection efficiently reduced model complexity while maintaining predictive power. The analysis demonstrates automated BIC selection can improve variable reduction versus purely manual analytical approaches.

Introduction

Selecting predictive variables is crucial in regression modeling. Reducing variables decreases model complexity and multicollinearity. However, over-reduction can negatively impact model fit. Two common variable selection approaches are manual analytical assessment and automated methods like stepwise selection guided by information criteria.

This analysis compares these two techniques for reducing variables in a regression model predicting Seattle home prices. The model incorporates property attributes like living space, bedrooms, and location as predictors. An initial full model with all predictors is specified. The first reduction method uses manual analysis like residual plots and variance inflation factors. The second applies automated backward stepwise elimination based on the Bayesian Information Criterion (BIC).

The reductions are evaluated by comparing model fit statistics like R-squared. ANOVA tests assess significance of fit differences between the full and reduced models. The analysis aims to determine whether manual or automated BIC selection more efficiently reduces model complexity while preserving predictive ability. Findings will highlight best practices for variable selection in housing market modeling and regression analysis generally.

Data and Methods

1. Review the Data Used for the Experiment

The dataset contained 4691 historical home sales records from Seattle and surrounding towns. The target variable was sale price. Predictor variables included living space in square feet, parking lot size in square feet, floor count, condition rating (1-5 scale), construction year, renovation status, and zip code. Considerable positive skewness was observed in the sale price, living space, and parking lot size variables based on summary statistics and distributional examinations. Applying log-transformations helped normalize these variables to better meet linear modeling assumptions. Construction year and renovation status were encoded into binary dummy variables indicating whether the home was built or renovated after 2000. A location dummy variable represented proximity to the greater Seattle metropolitan area, specifically the cities of Bellevue, Sammamish, Kirkland, Seattle, Redmond, Shoreline, Issaquah, Bothell, Edmonds, and Renton.

An initial linear regression model was formulated incorporating all potential predictor variables:

The initial regression model took the following form:

$$\boxed{\text{Log(Price)} = \beta 0 + \beta 1\text{Bedrooms} + \beta 2\text{Log(LivingSpace)} + \beta 3\text{Log(LotSize)} + \beta 4\text{Floors} + \beta 5\text{Condition} + \beta 6\text{YearBuilt} + \beta 7\text{YearRenovated} + \beta 8\text{Location}}$$

Model 1. Initial Model

where Price is the response variable, LivingSpace is square feet of living area, LotSize is square feet of parking area, Floors is number of floors, Condition is quality rating, YearBuilt and YearRenovated are renovation status, and Location is a dummy variable for proximity to Seattle.

The R-squared value was 0.5698, indicating the model explained approximately 57% of the variance in home prices. The coefficient estimates are shown in the regression output as below:

Table 1. Initial Coefficients Estimation

Predictor	Estimate	Std Error	t value	p value
Intercept	5.536	0.117	47.16	<0.001***
Bedrooms	-0.071	0.007	-9.27	<0.001***
Log(LivingSpace)	0.929	0.019	48.99	<0.001***
Log(LotSize)	0.001	0.008	0.15	0.882
Floors	0.102	0.013	8.03	<0.001***
Condition	0.093	0.009	10.33	<0.001***
YearBuilt	-0.038	0.017	-2.26	<0.05*
YearRenovated	0.041	0.014	2.91	<0.01**
Location	0.37	0.012	30.66	<0.001***

2. Using Analytical Assessment for Variables Reduction

We move on examining the linearity between the target variable (Price) and its features. Figure 1 displays a scatter plot matrix, highlighting the relationships between the dependent variable and its predictor variables. The visual data suggests an approximate linear association among these variables, a crucial assumption in regression analysis.

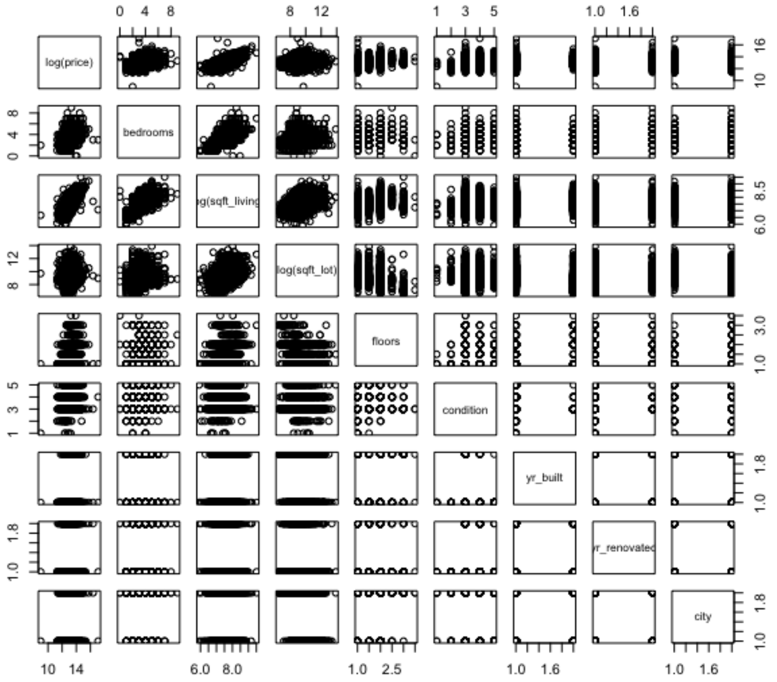


Figure 1. Scatter Plots Showing the Linearity between Price and Each Feature.

Following our examination of the relationships highlighted in the scatter plot matrix, it is essential to delve deeper into the potential issue of collinearity among our predictor variables. To assess this, we evaluated the Variance Inflation Factors (VIF). The VIFs for each predictor variable are as follows:

Table 2. Variance Inflation Factors

Predictor	VIF
Bedrooms	1.737458
Log(LotSize)	1.692643
Log(LivingSpace)	2.368026
Floors	1.6656
Condition	1.323138
YearBuilt	1.639316
YearRenovated	1.175637
Location	1.145977

Given that all VIF values are below the threshold of 5, multicollinearity does not pose a significant concern. This suggests our predictor variables exhibit a commendable degree of independence. Thus, model M1 appears largely free from multicollinearity issues.

Following our evaluation of the VIF values, we turn our attention to the plots of standardized residuals against each of the predictor variables, as depicted in Figure 2, Figure 3, and Figure 4. The scatter within the range of -4 to 4 on these plots suggests a random distribution of residuals. This indicates that model 1 is well-suited to the data, especially given the absence of a significant number of outliers.

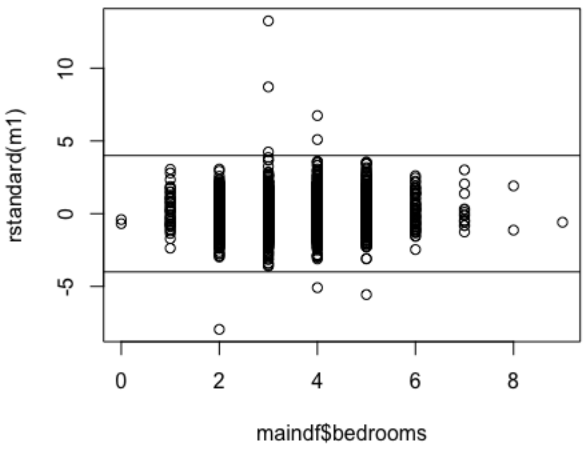


Figure 2. Number of Bedrooms vs Standard Residual

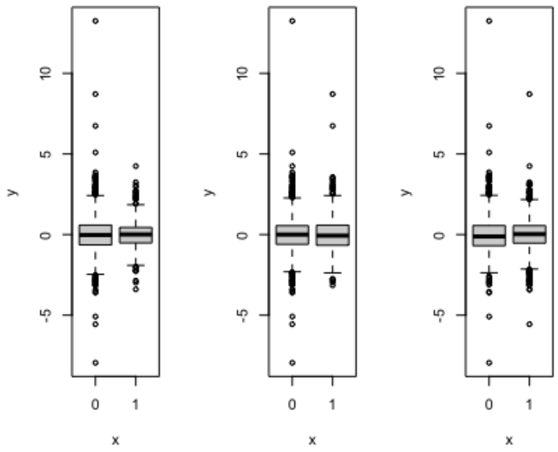


Figure 3. Residual Standard vs Dummy Variables

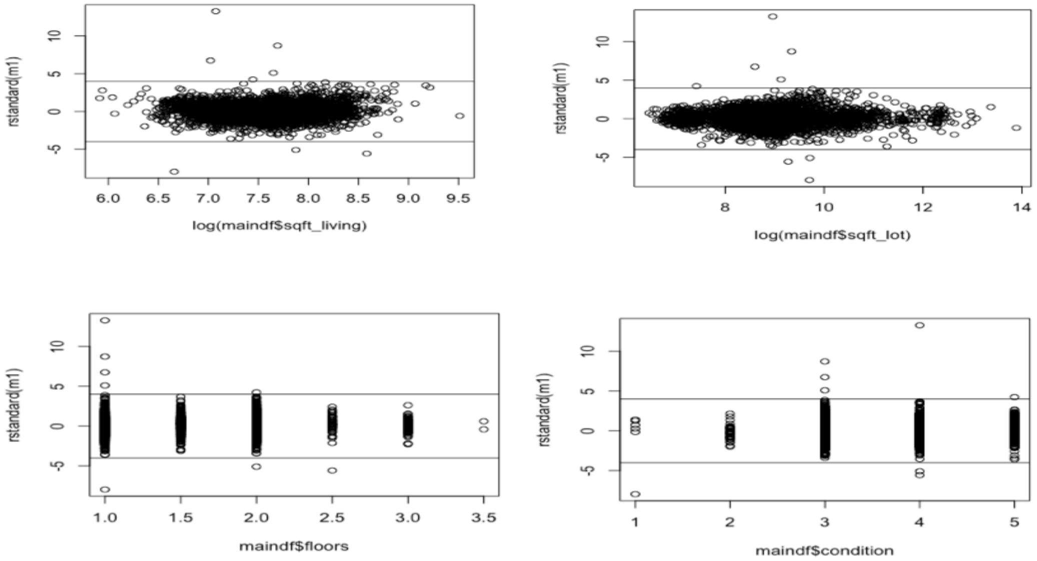


Figure 4. Standard Residual vs the Rest of Features

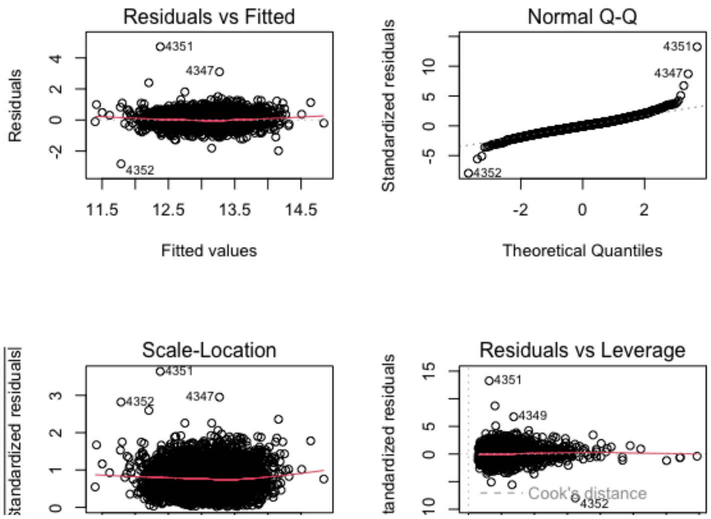


Figure 5. Diagnostic Plots

In Figure 5, we present various diagnostic plots for model 1, including the Standardized Residuals against Fitted Values, the Normal Q-Q plot, the Scale-Location plot, and the Cook's Distance.

Residuals vs Fitted: The distribution of datapoints around the zero line, without a discernible pattern, confirms the homoscedasticity of our model residuals. This suggests that model 1 satisfies the assumption of constant variance.

Normal Q-Q Plot: The data points closely following the straight line on the normal Q-Q plot indicate that the residuals are approximately normally distributed, reinforcing the assumption of normality for our model.

Cook's Distance: The absence of significantly high Cook's Distance values implies that there aren't notable leverage points adversely impacting our model. However, specific data points, namely 4347, 4349, 4351, and 4352, exert influence on the model and warrant further investigation.

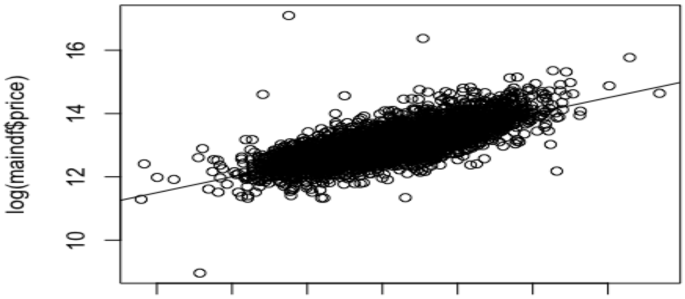


Figure 6. Fitted Prices vs Actual Prices

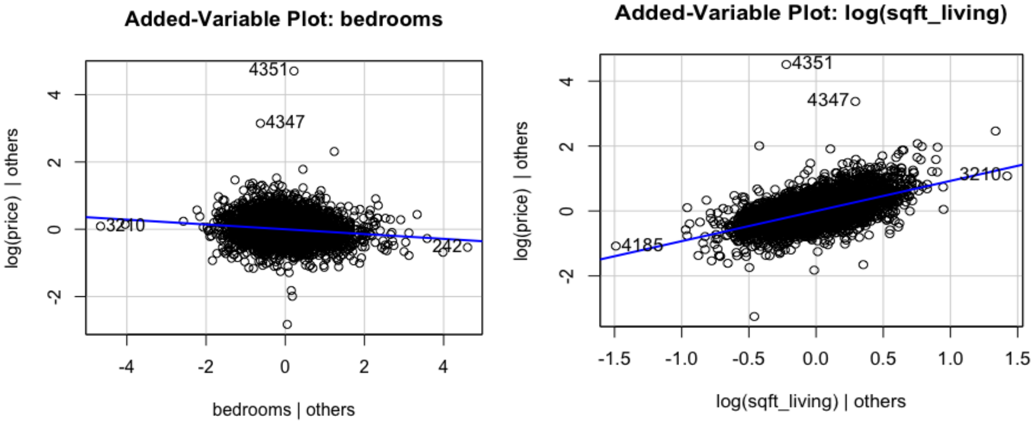
In Figure 6, we present a plot comparing the observations with their corresponding fitted values. The close alignment of the data points to the straight-line fit underscores the model's accuracy. This visualization further reinforces the credibility and aptness of Model 1 in representing the given data.

Moving forward, our attention is directed towards the correlation between selected feature variables and target, considering the influence of remaining features. The predictors 'bedrooms' and 'living area' demonstrate statistical significance in Model 1, as illustrated in Figure 7. Similarly, 'condition' and 'city' exhibit significant relevance, as depicted in Figure 8. On the contrary, the variables 'log(sqft_lot)', 'floors', 'year_built', and 'year_renovated' display a lack of statistical significance in Figure 8. Consequently, these features contribute minimally to the prediction of the target variable, Price.

In the Added Variables plot comparing 'Bedrooms' to 'Price', certain data points significantly influence the least squares estimate for the regression coefficient of 'Bedrooms'.

Specifically, cases 3210, 4185, 4347, 4351, and 242 stand out and warrant further examination. Likewise, the rightmost plot in Figure 7 pinpoints cases 3210, 4185, 4347, and 4351 as requiring additional scrutiny. Furthermore, in Figure 8, both cases 4347 and 4351 are highlighted for further investigation.

Figure 7. Added-Variables Plots



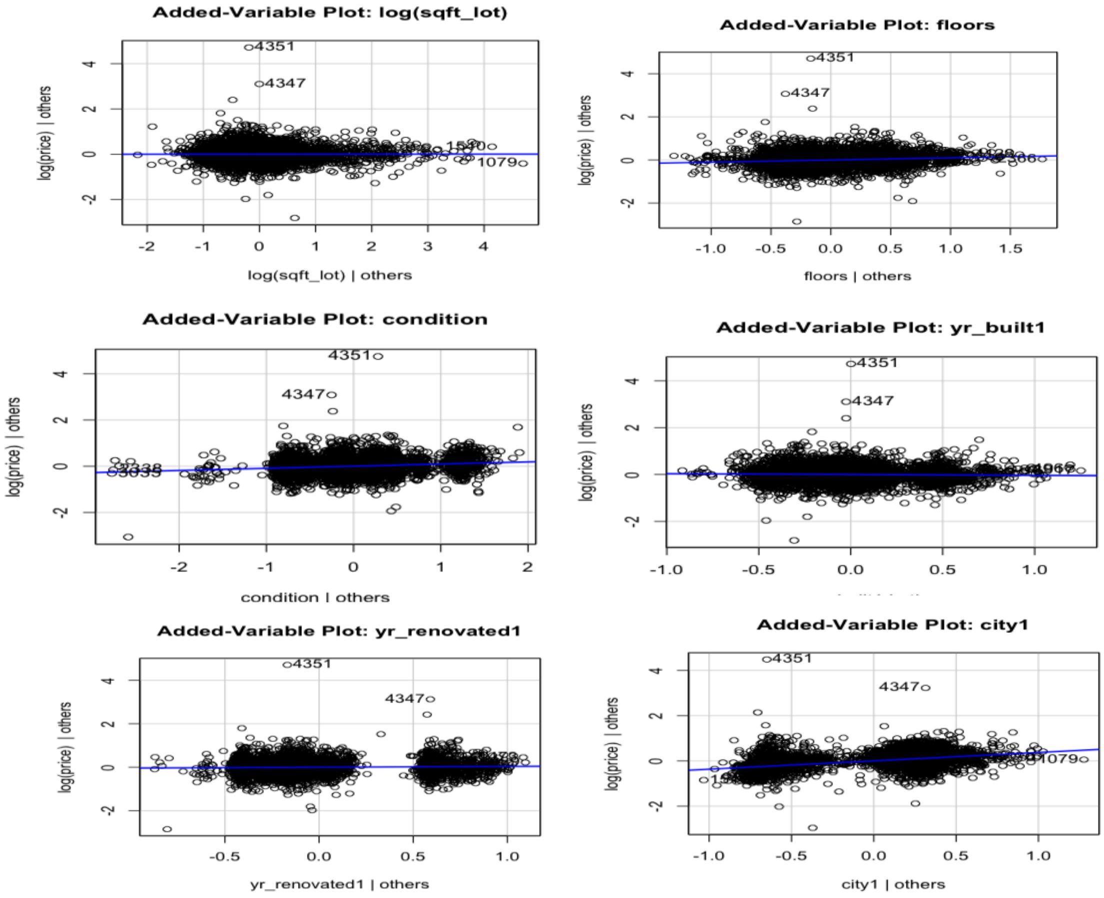


Figure 8. The Remaining Added Variables Plots

In our ongoing efforts to refine the predictive power and reliability of our regression models, we undertook a detailed analysis utilizing the added-variables plot. This analytical tool provided illuminating insights, leading to the formulation of a revised model:

$$\boxed{\text{Log(Price)} = \beta 0 + \beta 1\text{Bedrooms} + \beta 2\text{Log(LivingSpace)} + \beta 5\text{Condition} + \beta 8\text{Location}}$$

Model 2. New Model with Variables Reduction Using Analytical Assessment

In the Model 2, the R-squared value has decreased to 0.56. This suggests that the model now accounts for approximately 56% of the variance observed in home prices. Notably, while there is a slight decrease in the R-squared value, the benefits of model simplicity are realized through a reduction by four dimensions in our regression. The detailed coefficient estimates for this model can be found in the subsequent regression output:

Table 3. Coefficient Estimations of Model 2

Predictor	Estimate	Std Error	t value	p value
Intercept	5.48355	0.11436	47.95	<0.001***
Bedrooms	-0.072	0.007	-9.35	<0.001***
Log(LivingSpace)	0.96644	0.01627	59.39	<0.001***
Condition	0.07316	0.00792	9.241	<0.001***
Location	0.37662	0.01138	33.1	<0.001***

3. Using Information Criterion for Variables Selection

Given the observed Variance Inflation Factors (VIFs) are consistently below the threshold of 5, combined with the advantage of having a substantial number of observations in our dataset, we deem it appropriate to employ the backward elimination method using the Bayesian Information Criterion (BIC) for variable selection. Detailed progression and outcomes of this selection process are presented below.

Initial Model:

- Predictors: bedrooms, log(sqft_living), log(sqft_lot), floors, condition, yr_built, yr_renovated, location
- Residual Sum of Squares (RSS): 576.64
- BIC Value: -9326.01
- Most influential variable for removal: log(sqft_lot) (ΔBIC: -8.39)

Iteration 1:

- Predictors: bedrooms, log(sqft_living), floors, condition, yr_built, yr_renovated, location
- Residual Sum of Squares (RSS): 576.65
- BIC Value: -9334.41
- Most influential variable for removal: yr_built (ΔBIC: -2.48)

Iteration 2:

- Predictors: bedrooms, log(sqft_living), floors, condition, yr_renovated, location
- Residual Sum of Squares (RSS): 577.40
- BIC Value: -9336.89
- Variable of least significance: yr_renovated (Difference in BIC with and without the variable: 2.78)

Note: In each step, the variable causing the largest increase in BIC (or smallest decrease) was considered for removal. The iterative process aimed at optimizing the BIC value, signifying a balance between model fit and complexity.

We have refined our regression model through the Backward BIC method, resulting in the following optimized model, designated as model 3:

Log(Price) = β0 + β1Bedrooms + β2Log(LivingSpace) + β4Floors + β5Condition + β7YearRenovated + β8Location

Model 3. Model with Reduced variables by BIC

We conducted a linear regression analysis on our dataset, modeling the log-transformed apartment price as a function of various determinants. The regression output for model 3 is provided below:

Table 4. coeeficient estimations of model 3

Variable	Coefficient Estimate	Standard Error	t-Statistic	p-Value
(Intercept)	5.5367	0.1161	47.693	< 0.001
Bedrooms	-0.0705	0.0077	-9.186	< 0.001
Log(sqft_living)	0.9287	0.017	54.512	< 0.001
Floors	0.0893	0.0113	7.928	< 0.001
Condition	0.0988	0.0086	11.454	< 0.001
Yr_Renovated (Yes)	0.0465	0.0139	3.347	0.0008
Location (Central)	0.3683	0.0113	32.453	< 0.001

The R-squared value of model 3 is 0.5693, which means that approximately 56.93% of the variability in the log-transformed apartment prices is explained by the predictors in the model. That indicates the accuracy of model 3 does not change much compared to model 1; however, two variables are dropped, which can reduce the complexity of the computation.

Model Evaluation and Discussion

We conducted an ANOVA analysis to assess the differences in model fit between the comprehensive model 1 and its reduced counterparts, models 2 and 3.

Comparison between Model 1 and Model 2:

- Model 1: This model includes eight predictor variables: bedrooms, log(sqft_living), log(sqft_lot), floors, condition, yr_built, yr_renovated, and city.
- Model 2: This reduced model includes only four predictor variables: bedrooms, log(sqft_living), condition, and city.

From the ANOVA table:

- The residual sum of squares (RSS) for Model 1 is 576.64 and for M2 is 585.71.
- The difference in degrees of freedom between the two models is 4, indicating that four predictors were removed when forming model 2.
- The Sum of Squares (SS) difference is -9.0636, which signifies the variability explained by those four removed predictors.
- The F-statistic is 17.848 with a highly significant p-value of 1.501e-14 (far less than 0.05). This suggests that the predictors removed in Model 2 contributed significantly to the fit of Model 1.

In conclusion, the Model 2 significantly reduced the model's fit, as indicated by the highly significant F-value.

Comparison between Model 1 and Model 3:

- Model 1: As mentioned above, this model includes eight predictor variables.
- Model 3: This reduced model, based on the BIC criterion, has six predictors: bedrooms, log(sqft_living), floors, condition, yr_renovated, and city.

From the ANOVA table:

- The RSS for Model 1 is 576.64 and for Model 3 is 577.40.
- The difference in degrees of freedom between Model 1 and Model 3 is 2, indicating that two predictors were removed when forming Model 3.
- The SS difference is -0.75541, representing the variability explained by the two removed predictors.
- The F-statistic is 2.975 with a p-value of 0.05115, which is slightly above the common significance threshold of 0.05.

In conclusion, the reduction from Model 1 to Model 3 using the BIC criterion did not significantly degrade the model's fit. However, the p-value is marginally above the conventional threshold, suggesting that the predictors removed might have some minor significance, but it is not strong enough to be definitively considered as crucial contributors to the model's fit.

Conclusion

This analysis explored two methods for reducing variables in a pre-specified linear regression model predicting home prices. The initial full model incorporated eight predictors and achieved an R-squared of 0.57. The first reduction method used manual analytical assessment like residual plots and variance inflation factors. This resulted in a simplified model with four variables – bedrooms, living space, condition, and location. The R-squared slightly decreases, which is 56% (compared to 57%).

The second method applied stepwise elimination guided by the Bayesian Information Criterion (BIC). This produced a final model with six predictors – bedrooms, living space, floors, condition, renovation status, and location. The R-squared was 0.569, very close to the original model.

ANOVA tests showed the manual reduction significantly degraded model fit compared to the full model, with an F-statistic of 17.8. However, the BIC-guided reduction did not lead to a statistically significant loss in fit (F=3.0, p=0.05).

In conclusion, the automated BIC selection preserved model fit while reducing complexity. The manual approach led to over-reduction. For future variable selection, BIC-guided elimination appears preferable to purely manual analytical assessment. We can use Information Criterion solely to reduce the number of variables, while we can use the analytical assessment to tailor the dataset. Further refinement could involve cross-validation and regularization techniques.

Reference

[1] shree1992. (2017). House price prediction. from https://www.kaggle.com/datasets/shree1992/housedata/

[2] Sheather, S. (2009). A Modern Approach to Regression with R. Springer-Verlag New York. (Springer texts in statistics).