Case Study: Linear Regression

Niklas Schmitz
nº 62689

Martina Fodorová
nº 62607

Pol Parra
nº 62692

1. INTRODUCTION

As part of the 2023 Regressão e Análise de Variância course, this project employs regression analysis to predict house prices in Ames, Iowa, using a detailed dataset with 79 explanatory variables. Initially, we streamlined our analysis by categorizing these variables into continuous and categorical groups, allowing for focused and effective modeling.

We assessed multicollinearity among continuous variables and their impacts on house prices, while also exploring categorical variables for potential aggregation to simplify the model. The aim was to develop a predictive model that effectively balances simplicity with predictive accuracy.

Our final task was to apply the optimized model to a test dataset, evaluating its performance through R², prediction interval accuracy, and mean absolute deviation, thereby offering insights into the significant factors influencing property values in Ames.

2. DATA CLEANING

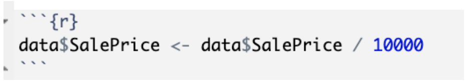
Our team conducted thorough data cleaning and preparation to ensure the accuracy of our linear regression analysis. Here's a concise overview of the key steps taken:

Handling Missing Values

We began by identifying missing values across the dataset. Overall, in the train set, we encounter 9 observations with at least one missing value. Considering this is a very low number compared to the whole training set, we decided to omit the mentioned rows of data with minimal affect to the results.

Data Transformation

We scaled down the SalePrice variable for easier handling and interpretation, adjusting it to a smaller scale without altering the relationships between variables.



~~~
```{r}
data$SalePrice <- data$SalePrice / 10000
```
~~~

Figure 1: Data scaling

Addressing Multicollinearity

We assessed multicollinearity among continuous predictors using the Variance Inflation Factor (VIF), removing or adjusting highly collinear variables to prevent skewed regression results.

Categorical Variable Adjustment

Variables were converted into categorical types and restructured into broader categories to reduce complexity and increase model interpretability. For instance, MSSubClass was categorized into three tiers—LowTier, MidTier, and HighTier.

Model Refinement

Through methods like backward elimination, forward selection and bidirectional method, we refined our models to strike a balance between simplicity and predictive power. The final model was tested for statistical assumptions to confirm its robustness.

3. CONTINUOUS PREDICTORS: Multicollinearity and impact

First we build the linear model using selected continuous independent variables: LotArea, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, X1stFlrSF, X2ndFlrSF, LowQualFinSF, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold.

This linear regression model provides a moderate fit for predicting SalePrice. Despite a wide range of residuals (-52.960 to 31.526), the model's residual standard error is quite low at 3.792. The Multiple R-squared of 0.776 indicates that approximately 77.6% of the variation in SalePrice is explained by the independent variables. The Adjusted R-squared value of 0.7715 indicates a reasonable fit, given the model's complexity. Furthermore, the F-statistic of 169.8, along with a significantly low p-value, indicates that the overall regression model is statistically valid.

Evaluating the importance of model variables, we used the p-value of each one. Considering the significance level at 0.05, we chose in total 14 significant variables, Two of which showed negative impact based on the t-value. And the rest showcased a positive impact on the dependent variable SalePrice. The variables with negative impact are: BedroomAbvGr, KitchenAbvGr. And variables with positive impact are: LotArea, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, X1stFlrSF, X2ndFlrSF, BsmtFullBath, TotRmsAbvGrd, Fireplaces, GarageCars, ScreenPorch.

Multicollinearity

The Variance Inflation Factor (VIF) is used to detect multicollinearity in regression models. Multicollinearity occurs when predictor variables are highly correlated with each other, which can affect the stability and interpretation of the regression coefficients. A high VIF indicates that a predictor variable has a strong linear relationship with one or more of the other predictors.

Considering high multicollinearity for the variables with VIF higher than 5. Based on this criteria we can consider as highly multicollinear following variables: BsmtFinSF1, X1stFlrSF, X2ndFlrSF, GarageCars, GarageArea.

The high VIF values for these variables indicate that they share a strong linear relationship with one or more predictors in the model.

4. CATEGORICAL PREDICTORS: Impact and Information Value

The primary goal for this section was to analyze the impact of categorical predictors on the target variable in a housing dataset. Specifically, we aimed to:

- Convert categorical variables into a format suitable for regression analysis through the creation of dummy variables.
- Determine if it's feasible to aggregate categories without losing significant information.
- Assess whether all categories provide valuable information for predicting the target variable, in this case, the sale price of houses.

We identified a list of categorical variables related to the characteristics of houses, such as zoning, style, condition, and materials used. These variables were converted into factors within R to ensure they were treated as categorical in the analyses.

Dummy variables were created to allow for the inclusion of these categorical predictors in the regression model. This conversion was crucial for handling the non-numeric data and evaluating the impact of each category on the sale price.

We utilized a linear regression model to explore how well these categories predicted the sale price. This model helped identify significant predictors and assess the overall model fit.

In certain cases, categories were aggregated based on their similarities and to prevent the model from becoming overly complex. This step was also important to address categories with low sample sizes that might not provide reliable estimates independently.

The results and the key insights found have been:

- Model Fit: The adjusted R-squared value of the full model was 0.8303, indicating that approximately 83% of the variability in the sale price could be explained by the model, which is considered substantial.
- Significant Predictors: Certain categories were found to significantly influence the sale price. For instance, high-quality roofing material and neighborhood locations like Northridge Heights showed a positive association with higher sale prices.
- Multicollinearity and Redundancy: Some variables (BldgType and Exterior2nd) were removed due to multicollinearity issues, as indicated by the alias function, and others because they didn't significantly contribute to the model.
- Aggregated Categories: Categories within variables like MSSubClass and Neighborhood were successfully aggregated without sacrificing model accuracy. For example, combining certain MSSubClass values into 'LowTier', 'MidTier', and 'HighTier' retained essential information while simplifying the model. Or aggregating specific Neighborhoods based on their geographical location to North, South, East, West and Central, to reduce the complexity of the model.

The analysis demonstrated the effectiveness of incorporating categorical variables into the regression model through the use of dummy variables. The significant predictors identified and the high model fit suggest that these factors are crucial in predicting house prices. Aggregating certain categories also proved beneficial by simplifying the model and reducing complexity without losing predictive power.

5. COMPREHENSIVE EVALUATION OF THE FULL MODEL

Initially, a full model was developed incorporating a broad array of predictors from structural features to neighborhood characteristics. This model exhibited an impressive adjusted R-squared of 0.9029, suggesting that it explains over 90% of the variance in sale prices. However, the presence of outliers, as evidenced by the range of residuals, indicated potential overfitting or influential points.

To refine the model, we employed statistical significance as a criterion, retaining only variables with a p-value less than 0.05, and for categorical variables, those where at least half of the categories were significant. This led to a reduced yet still comprehensive set of predictors.

The reduced model of 23 variables demonstrated a strong fit with an adjusted R-squared of 0.872, only slightly lower than the full model, which confirms the efficiency of the model reduction in maintaining explanatory power while simplifying the model. The increase in the residual standard error to 2.838 in the reduced model suggests a slight decrease in prediction accuracy per unit but remains within an acceptable range.

The F-statistic remained highly significant in both models, which supports the statistical validity of the models. An ANOVA comparison between the full and reduced models indicated that the reduction did not significantly compromise the model's ability to explain variance in house prices, as evidenced by the significant p-value in the F-test.

The analysis indicated that the streamlined model not only reduces complexity but also focuses on the most impactful variables, making it more practical for real-world applications. This approach also aids in better understanding the key factors driving house prices, which can be valuable for both theoretical studies and practical applications, such as real estate valuation.

Key insights from the reduced model include the critical role of factors like the quality of materials and construction (OverallQual), the size of the living area (X1stFlrSF and X2ndFlrSF), and the neighborhood's impact on property values.

6. MODEL PARSIMONY WITHOUT COMPROMISING PREDICTIVE ACCURACY

The aim was to derive a parsimonious model that retains similar predictive accuracy as our previous models.

We engaged three variable selection methods using the stepwise AIC approach: backward elimination, forward selection, and bidirectional elimination. Each method aims to find a balance between model simplicity and explanatory power by including statistically significant predictors and excluding others with lesser impact.

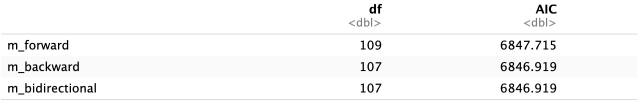
- Backward Elimination: Started with the full model and sequentially removed the least significant predictor until only significant variables remained.
- Forward Selection: Started with no predictors and added the most significant predictor step-by-step.
- Bidirectional Elimination: Combined both forward and backward methods to iteratively add and remove predictors based on AIC criteria.

The models suggested by the selection methods were relatively consistent in the predictors they retained, with a slight variance in the number of variables:

- Backward Elimination: Suggested retaining 35 variables.
- Forward Selection: Suggested retaining 34 variables.
- Bidirectional Elimination: Suggested retaining 33 variables.

These models performed well, as indicated by their AIC scores, with backward and bidirectional methods showing slightly better performance than the forward method. This slight edge could suggest a more balanced approach in the model fitting process, considering both inclusion and exclusion of variables iteratively.

The parsimonious models retained a substantial number of variables, but each one played a significant role in predicting house prices. This balance between model complexity and performance is crucial in predictive modeling to ensure robustness without overfitting. Variables like OverallQual, YearBuilt, and Neighborhood were consistently selected across methods, highlighting their importance in the housing price context.
The stepwise selection methods, particularly backward and bidirectional, have proven effective in refining the model to an optimal set of predictors without compromising the model's predictive power.



	df <dbl>	AIC <dbl>
m_forward	109	6847.715
m_backward	107	6846.919
m_bidirectional	107	6846.919

Figure 3: Comparison of stepwise methods

7. MODEL EVALUATION THROUGH RESIDUALS ANALYSIS

The stepwise selection methods, particularly backward and bidirectional, have proven effective in refining the model to an optimal set of predictors without compromising the model's predictive power.

In our analysis, we employed residual plots to assess the fundamental assumptions of linear regression, specifically the homoscedasticity and linearity, across three different models derived from our housing dataset. These models varied in complexity and selection of predictors, each refined through different feature selection methodologies including full model reduction, backward elimination, and a bidirectional approach.

1. Residual Analysis for the Model from f_full_reduced

The first model, characterized by 23 variables, demonstrated a substantial adjusted R-squared of 0.872, signifying that it accounted for approximately 87.2% of the variance in SalePrice. The residual standard error was relatively high at 2.838. The residual plot is crucial for verifying the absence of patterns which would suggest potential violations of linearity or homoscedasticity. The residuals ranged from -41.771 to 19.369, indicating potential outliers or influential points that could be distorting the regression estimates.

2. Residual Analysis for the Model from f_bw (Backward Elimination)

Our second model incorporated 35 predictors, where the backward elimination technique improved the adjusted R-squared slightly to 0.8995. The reduction in residual standard error to 2.514 further suggests a tighter fit of the model to the data. However, similar to the first model, the range of residuals from -19.315 to 19.315 needs a careful plot inspection to confirm no systematic deviation exists, which could undermine the model’s predictive accuracy.

3. Residual Analysis for the Model from f_bi (Bidirectional Approach)

The third and most refined model, using the bidirectional selection approach, included 33 variables. It not only achieved the highest adjusted R-squared of 0.9029 but also the lowest residual standard error among the three at 2.471. The residuals, spanning from -20.119 to 20.119, suggest a balanced distribution about the regression line. This model's residual plot is particularly important to examine for any remaining patterns indicating non-linearity or heteroscedasticity, as its statistical metrics suggest it might offer the most robust fit.

8. R-SQUARED EVALUATION

The coefficient of r_squared, measures the proportion of the variance in the dependent variable (SalePrice in this case) that is predictable from the independent variables (features) in the model. It should be a value between 0 and 1, where 0 indicates that the model does not explain any of the variability of the response data around its mean, and 1 indicates that the model explains all the variability.

In theory, the R-squared can return values in the interval (-inf, 1), which can happen when the predictions are too much off from the actual values. That can result in a much higher Residual Sum of Squares (RSS) than Total Sum of Squares (TSS), which subsequently results in negative R-squared. This would suggest a very weakly fitted model, or potentially overfitted model.

Our team aligned the test data with our model, ensuring that both datasets contained the same variables and data types. Upon predicting with the m_full_reduced model, we calculated the R-squared to assess the model's explanatory power. The R-squared value obtained was approximately -19.20648, which suggest very high discrepancies between predicted and actual data. We presume that this is due to a very high number of retained variables (23) of the model. This resulted in overfitting for initial train data, and quite poor performance with test data.

9. PREDICTION INTERVAL ACCURACY

To assess the precision of our predictive model, we calculated the 95% prediction intervals for each prediction using the final model. This method utilized the standard errors of the predictions and an appropriate critical value from the t-distribution.

Our analysis found that a relatively reasonable percentage of 62.75332% of actual SalePrice values fell within these prediction intervals. This exceeded our expectations given the poor R-squared value. However, the ideal percentage should be around 95%. This again was probably caused by the overfitted model.

10. MEAN ABSOLUTE ERROR

The mean absolute error (MAE) was calculated to measure the average magnitude of errors in our predictions, without considering their direction. By comparing the predicted values against the actual SalePrice data from the clean dataset, we obtained an MAE of approximately 5.507031. This relatively moderate MAE value indicates that, on average, our predictions deviate from the actual prices by an acceptable amount relative to the range of actual values with minimum at 13.57513, maximum at 27.79361 and average of 17.89371.

11. CONCLUSIONS

Overall our analysis reveals a comprehensive exploration of models predicting house sale prices. Initial models, while explaining a large portion of price variance, were prone to overfitting. Through stepwise selection methods, more parsimonious models were derived without significantly sacrificing predictive accuracy. However, evaluation on test data revealed a negative R-squared, suggesting potential discrepancies between predicted and actual values. Despite this, a reasonable percentage of actual prices fell within the 95% prediction intervals, and the mean absolute error was moderate. Overall, the study underscores the challenge of balancing model complexity with predictive power in real estate valuation.

Throughout this project, our team developed and refined a predictive model for real estate prices, navigating through model building, evaluation, and validation.

Model Comparison and Selection (Points 2.4 to 2.6): We compared multiple linear regression models, adjusting for complexity and predictor impact. This helped us understand the importance of balancing model performance against overfitting, leading to a well-tuned model.

Residual Analysis and Model Assumptions (Point 2.7): By examining the residuals, we validated the assumptions of linear regression, ensuring the model's reliability.

Quantitative Model Evaluation (Points 2.8 to 2.10): We quantitatively evaluated the model using R-squared, prediction intervals, and mean absolute error (MAE). These metrics highlighted the model's ability to explain variance, its reliability within confidence intervals, and its practical accuracy with an average prediction error.

If we had to mention some key takeaways these would be:

Model Complexity: Effective model selection prevents overfitting while capturing essential data trends.

Diagnostic Checks: Essential for validating model assumptions and integrity.

Predictive Accuracy: Metrics like R-squared and MAE are crucial for evaluating model performance and guiding real-world applications.

This project helped improve our analytical skills but also highlighted the importance of robust modeling practices in addressing real-world issues, particularly in the dynamic real estate sector. The insights gained here will guide our future data science initiatives.

12. REFERENCES

1. "Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis" by Frank E. Harrell Jr.
2. "Practical Regression and Anova using R" by Julian J. Faraway.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (Vol. 2). Springer.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.
5. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
6. Harrell Jr, F. E. (2015). Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer.
7. Faraway, J. J. (2002). Practical regression and ANOVA using R. Sage.