भारतीय प्रौद्योगिकी संस्थान मुंबई
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
ज्ञानम् परमम् ध्येयम्

Anand Yadav

Project Report

Regression Analysis: Model Selection & Diagostics

Abstract

This project presents a rigorous, end-to-end regression analysis to predict housing sale prices using the Ames, Iowa dataset. Moving beyond a simple predictive model, the focus is on building a statistically defensible and interpretable linear model. A systematic **forward stepwise selection** process, guided by **Adjusted R^2, Mallows' C_p, and the AIC**, was employed to identify an optimal set of predictors from 79 potential variables. Comprehensive **diagnostics**, including multicollinearity checks (VIF) and influence analysis (Cook's Distance, DFFITS), were performed to validate the model's integrity. The analysis reveals that while removing influential outliers improves in-sample metrics (Adj. R^2 increased from 0.85 to 0.92), it degrades predictive accuracy on unseen data. The final model, therefore, highlights the crucial trade-off between statistical purity and real-world generalization.

Introduction

Beyond the curb appeal and the fresh coat of paint, what truly drives a home's value? This project moves past guesswork to answer that question with data-driven precision. Using the rich and complex Ames Housing dataset as our guide, we embark on a mission to build more than just a predictive algorithm—we construct a **statistically sound and interpretable model** that uncovers the most significant drivers of property value. This analysis is a deep dive into the science of valuation, demonstrating how a rigorous process can transform a vast array of housing features into a reliable and defensible price prediction tool.

Methodology

To ensure our model is both accurate and trustworthy, we employed a sophisticated and systematic methodology. The process began with a **rigorous forward stepwise selection**, using multiple criteria like **Adjusted R^2, Mallows' C_p, and AIC** to scientifically isolate the most impactful variables. To safeguard against statistical pitfalls, we conducted thorough **multicollinearity checks using VIF** to eliminate redundant predictors. Finally, the model was put under the microscope with a full suite of **influence diagnostics**, including **Cook's Distance and DFFITS**, to identify and analyze outlier data points. This multi-stage approach guarantees a final model that is not only highly predictive but also transparent and statistically robust.

Inferences

1. Inference from Multicollinearity Screening (VIF)

The first crucial step after initial data prep is checking for multicollinearity. The notebook wisely uses the **Variance Inflation Factor (VIF)** on a subset of the most correlated features.

- **Observation:** The VIF table shows that **GarageCars** (VIF = 5.51) and **GarageArea** (VIF = 5.22) have the highest scores, both exceeding the common threshold of 5.0. GrLivArea is also high (VIF = 5.42).
- **Inference:** This result tells a clear story: the number of cars a garage can hold is highly predictive of its total area in square feet. This is intuitive—a 3-car garage is almost always larger than a 1-car garage. Including both variables in a model is redundant. The model would struggle to distinguish the individual impact of each feature on the house price because they essentially carry the same information. The high VIF for GrLivArea (living area) and TotRmsAbvGrd (total rooms) tells a similar story about the relationship between house size and the number of rooms.
- **Conclusion:** The analysis correctly identifies that it would be statistically unsound to keep both GarageCars and GarageArea in the final model. One must be dropped to ensure the model's coefficients are reliable and interpretable. This is a critical insight that prevents a common modeling pitfall.

2. Inference from Forward Stepwise Selection (Adj R², Cp, AIC)

This is the core of the model-building process. Instead of guessing, the notebook uses three different statistical criteria to build a model one variable at a time.

- **Observation:** The history table shows the results for adding variables step-by-step.
  - **Adjusted R^2:** This value keeps increasing with each new variable added, suggesting that even the 10th variable is still adding some explanatory power.
  - **Mallows' C_p:** The C_p statistic measures the trade-off between model fit and complexity. It starts very high (bad) and drops significantly, hitting its lowest point around step 7 (MSSubClass, C_p = 1.12), after which it starts to rise again.
  - **AIC (Akaike Information Criterion):** Similar to C_p, this metric balances fit and complexity. The negative AIC value continues to get larger (which is better) up to the 10th step (WoodDeckSF, AIC = -1031.67).
- **Inference:** This step tells the story of diminishing returns. The first few variables (OverallQual, GrLivArea, YearBuilt) provide massive boosts in model performance. However, as more variables are added, the improvement gets smaller. Mallows' C_p suggests that the "sweet spot" is a 7-variable model, as

adding more variables after that introduces more complexity than is justified by the improvement in fit. AIC, being slightly less strict, continues to favor adding more variables up to the 10th step.

- **Conclusion:** All three criteria agree on the first several variables, confirming their importance. However, they diverge on the optimal model size. This is a common and important finding—it shows that there isn't always a single "perfect" model. The choice depends on whether you prioritize simplicity (C_p) or slightly higher explanatory power (AIC/Adjusted R^2). The notebook correctly highlights this trade-off.

3. Inference from PRESS Statistic Comparison

The **Predicted Residual Sum of Squares (PRESS)** is a fantastic out-of-sample validation technique. It assesses how well the model predicts new data by systematically leaving out each observation and predicting it.

- **Observation:** The PRESS comparison table shows the results for the top 15-variable models selected by Adjusted R^2, C_p, and AIC. Crucially, **all three criteria lead to the exact same 15-variable model**. They have identical metrics: Adjusted R^2 (0.854), AIC (-1076.9), and PRESS (30.22).
- **Inference:** This is a very strong and reassuring result. It tells us that even though the criteria disagreed on the "best" smaller models (like the 7-variable model from C_p), they ultimately converge on the same set of important predictors when the model size is extended. This convergence gives us high confidence in the stability and importance of the selected features.
- **Conclusion:** The PRESS statistic confirms that the chosen 15-variable model is robust. A low PRESS value indicates good predictive ability. Since all methods agreed on this model, it is a defensible and logical choice to move forward with for further diagnostics.

4. Inference from Diagnostics & Influence Plots

This section investigates the *health* of the chosen model by examining its errors (residuals) and identifying influential data points.

- **Observation:** The diagnostics table shows extreme values for several metrics. The maximum **Cook's Distance** is 4.17 (far exceeding the threshold of ~0.003), and the minimum **DFFITS** is -8.95 (far beyond the threshold of ~0.23). The (optional but shown) Influence Plot would visually confirm these points as major outliers. The Q-Q plot shows that the residuals deviate significantly from the normal line at the tails.
- **Inference:** This tells a critical story: a few specific houses in the training data are having a disproportionately large impact on the model. These are likely very unusual properties (e.g., extremely large for their neighborhood, sold under special conditions) that the model struggles to predict accurately. Their large residuals and high leverage are "pulling" the regression line towards them,

potentially making the model less accurate for more typical houses. The non-normal tails in the Q-Q plot are a direct symptom of these extreme errors.

- **Conclusion:** The model, while good, is being skewed by a handful of influential outliers. To build a more robust and reliable model, these specific data points must be investigated and likely removed. The analysis correctly identifies that simply accepting the model at this stage would be a mistake.

5. Inference from Remediation & Refit

Based on the diagnostics, the notebook removes the 96 identified influential points and refits the model on the cleaned data.

- **Observation:** The comparison table shows dramatic improvements in the model's fit on the *training data* after remediation. The **Adjusted R^2 jumps from 0.854 to 0.920**, the **AIC improves significantly from -1077 to -1804**, and the **PRESS statistic is cut by more than half, from 30.22 to 11.71**.
- **Inference:** This confirms our suspicion from the diagnostic step. By removing the problematic houses, the model can now fit the remaining, more "typical" houses much more accurately. It no longer has to compromise its fit to account for the extreme outliers. The massive drop in the PRESS statistic suggests this new, refitted model is much better at predicting houses *within this cleaned dataset*.
- **Conclusion:** Remediation was highly successful in improving the model's internal consistency and fit. The resulting model is statistically superior *on the data it was trained on*. The next crucial question, however, is whether this "cleaner" model is actually better at predicting genuinely new, unseen data.

6. Inference from Hold-Out Validation

This is the final and most important test. Both models (before and after remediation) are evaluated on the unseen validation set.

- **Observation:** The results here are fascinating. The **"Before" model** (trained on all data, including outliers) achieves a **higher R^2 (0.895)** and a **lower RMSE (0.133)** on the validation set. The **"After" model** (trained on the cleaned data) performs slightly **worse (R^2 = 0.881, RMSE = 0.141)**.
- **Inference:** This is the most profound insight of the entire project. It tells us that while removing outliers created a model that was statistically "prettier" on the training data, it actually made it **worse at generalizing to new data**. This is likely because the validation set also contains some unusual houses, and the "Before" model, having been exposed to outliers during training, was more robust and less surprised by them. The "After" model was overfitted to the clean, typical houses and performed poorly when faced with the messiness of the real world.
- **Conclusion:** The "Before" model, despite its less impressive training metrics and messy residuals, is the superior model for practical prediction. This is a classic

example of the difference between **explanatory** and **predictive** modeling. The refitted model might be better for *explaining* the relationships in typical houses, but the original model is better for the raw task of *prediction*.

Overall Conclusion

This project tells the complete story of a realistic data science project. It begins with a broad set of potential predictors, systematically narrows them down to a robust and defensible model, and then critically evaluates that model's flaws. Most importantly, it demonstrates that the "best" model on paper is not always the best model in practice, a crucial lesson for any data scientist. The final decision to prefer the model trained on the full (but messy) data is a sign of mature analytical judgment focused on real-world performance.