MSDS 6371 Project - Fall 2022

Andrew Yule, Krithika Kondakindi

Introduction

This analysis focuses on applying multi-linear regression techniques to predict the sales price of homes in the Ames, Iowa area. Two specific questions of interest (QOI) were identified:

1. Focusing on 3 specific neighborhoods in Ames, what is the best linear regression model that can be constructed solely utilizing the square footage of the living room and what can be interpreted from that model?
2. Utilizing all available variables and focusing on every neighborhood, what are the best linear regression models that can be constructed? For this QOI, submissions will be judged based on a Kaggle competition, using produced model predictions to achieve the lowest root mean squared error.

Data Description

For this analysis, housing data from the Ames area was supplied between the years of 2006 and 2010 from 1,460 homes. Approximately 80 variables were collected for each home identifying various features that may contribute to a homes final sale price. The variables ranged in nature from quantitative values like square footage of certain parts of the home, to qualitative values like quality rankings and building materials.

Upon analysis of the data, numerous variables were found to have a high number of missing values and were subsequently removed from the analysis.

Many numerical variables in the available data demonstrated high variance. Log transformations were used to reduce the variance and ultimately improve the performance of linear regression models. The full list of variables that received a log transformation can be found below. It should be noted that sales price (the variable of interest) was among the variables recieving a log transformation.

- LotArea

- FirstFlrSF
- SecondFlrSF
- GrLivArea
- WoodDeckSF
- OpenPorchSF
- EnclosedPorch
- ThirdSsnPorch
- ScreenPorch
- PoolArea
- MiscVal
- SalePrice

Analysis

Question 1

(Brief introduction to the questions of interest and the setting of the problem.)

Restatement of the Problem

(Where did the data come from? How big is it? How many observations? Where can we find out more? What are the specific variables that we need to know with respect to your analysis?)

Build and Fit the Model

Checking the Assumptions

(Residual Plots Influential point analysis (Cook's D and Leverage) Make sure to address each assumption.)

Comparing Competing Models

(Adj R2 Internal CV Press)

Parameters

(Estimates Interpretation Confidence Intervals)

Conclusion

(A short summary of the analysis.)

R Shiny: Price v. Living Area Chart

A web application was created to provide access to the model and can be found at the link below:

www.google.com

Question 2

For the second question of interest, the objective was to utilize all available data and variables to create the best linear regression model possible for predicting home sales prices. There are numerous techniques available for creating multi-linear regression models. This QOI focused on applying forward selection, backwards elimination, and stepwise selection to produce the highest performing models. Additionally, a final custom model was fit using cross validation techniques.

Model Selection

The olsrr package in R was utilized to perform each of the 3 stepping regression fittings. Minimizing the Akaike Information Criterion (AIC) was used as the model criteria. The following linear forms below were identified to produce the lowest AIC for forward, backward, and step selection respectively. A custom model was also attempted, however, it was unable to achieve higher accuracy than the forward selection model and thus was discarded.

Forward selection:

SalePrice = OverallQual + GrLivArea + Neighborhood + OverallCond + HouseStyle + YearBuilt + LotArea + RoofMatl + KitchenAbvGr + SaleCondition + Condition2 + Foundation + Fireplaces + Heating + ExterQual + Condition1 + PoolArea + ScreenPorch + WoodDeckSF + HeatingQC + CentralAir + BedroomAbvGr + FirstFlrSF + `2ndFlrSF` + Street + LandSlope + HalfBath + EnclosedPorch + SecondFlrSF + MiscVal + PavedDrive + `1stFlrSF` + BldgType + ExterCond + YearRemodAdd

Backwards elimination:

SalePrice = LotArea + Street + LotConfig + LandSlope + Neighborhood + Condition1 + Condition2 + BldgType + OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofMatl + ExterQual + ExterCond + Foundation + Heating + HeatingQC + CentralAir + `1stFlrSF` + `2ndFlrSF` + GrLivArea + HalfBath + BedroomAbvGr + KitchenAbvGr +

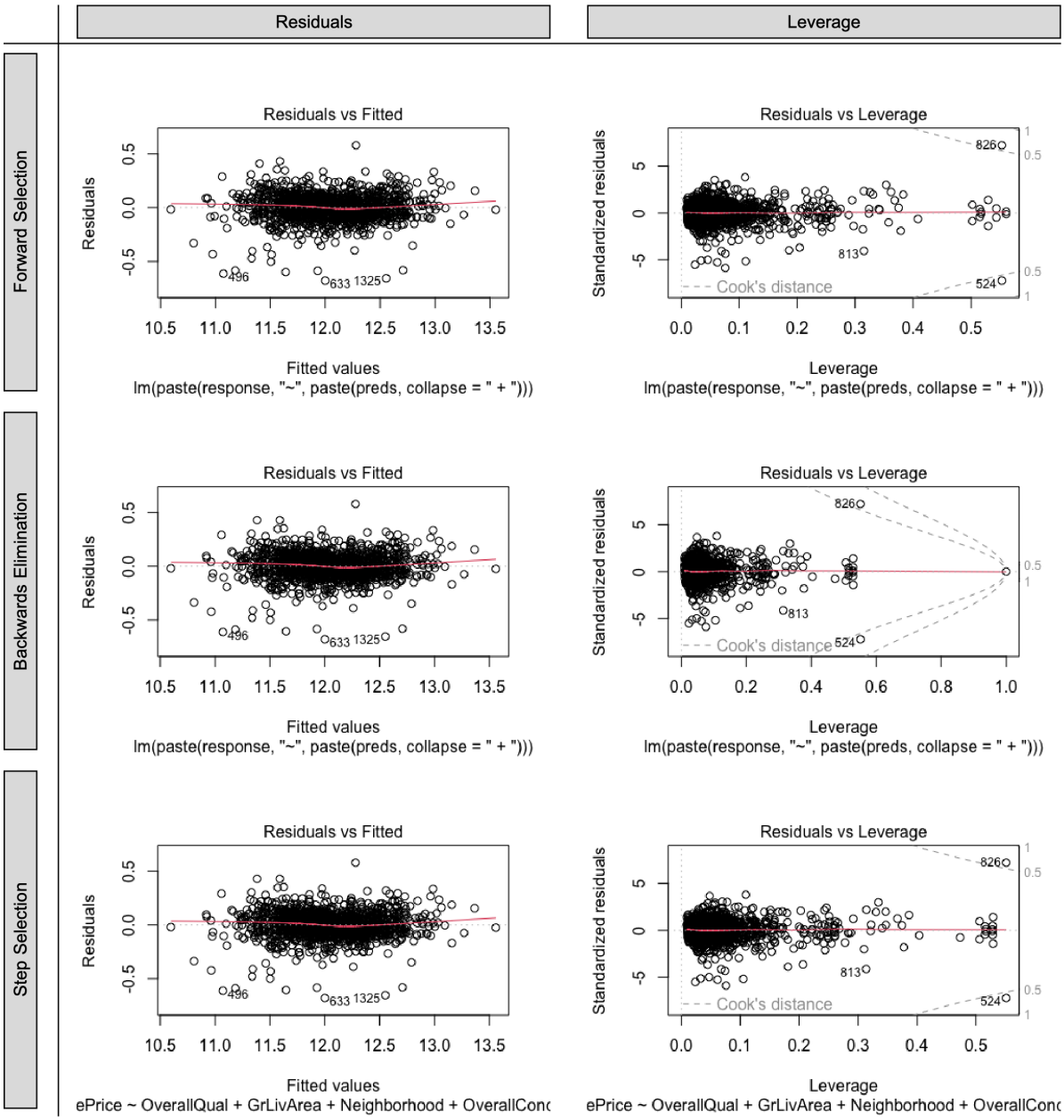
Fireplaces + PavedDrive + WoodDeckSF + EnclosedPorch + ScreenPorch + PoolArea + MiscVal + SaleCondition + FirstFlrSF + SecondFlrSF

Stepwise selection:

SalePrice = OverallQual + GrLivArea + Neighborhood + OverallCond + YearBuilt + LotArea + RoofMatl + KitchenAbvGr + SaleCondition + Condition2 + Foundation + Fireplaces + Heating + ExterQual + Condition1 + PoolArea + ScreenPorch + WoodDeckSF + HeatingQC + CentralAir + BedroomAbvGr + FirstFlrSF + `2ndFlrSF` + Street + LandSlope + HalfBath + EnclosedPorch + SecondFlrSF + MiscVal + PavedDrive + BldgType + `1stFlrSF` + ExterCond + LotConfig + YearRemodAdd

Checking Assumptions

(Residual Plots Influential point analysis (Cook's D and Leverage) Make sure to address each assumption)



Comparing Competing Models

The table below summarizes the key performance metrics associated with each of the three models produced. Each model was found to have the same Adjusted R Squared metric within rounding errors. When used to predict unknown home sales prices as part of the Kaggle competition, the Forward Selection model was found to achieve the lowest (best) Kaggle score of 0.13496.

Model	Adjusted R2	CV Press	Kaggle Score
Forward Selection	0.9094	X	0.13496
Backward Elimination	0.9094	X	0.13554
Stepwise Selection	0.9094	X	0.13554
Custom	X	X	X

Conclusion

In this analysis, a data set of home sales prices from Ames, Iowa was explored using multi-linear regression techniques. A model was produced to predict the sales price of homes specifically in the North Ames, Edwards, and Brk Side neighborhoods using only the greater living area square footage. That model achieved an RSquared metric of XX, meaning XX% of the variance in home prices in those 3 neighborhoods can be explained soley by the living room squared footage.

Finally, 4 additional models were produced to predict homes sales prices based on all variables available. Various multi-linear regression stepping approaches were used and the forward selection method was found to perform the best overall with an adjusted RSquared metric of 0.9094 meaning over 90% of the variance in home prices could be explained by the variables and associated model. Additionally, a Kaggle competition score of 0.13496 was achieved by the model on a blind test.

Appendix

The entire code for this analysis can be found below: