

# 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?

## Data Description

For this analysis, housing data was obtained between the years of 2006 and 2010 from 1,460 homes in the Ames area. Approximately 80 variables were collected for each home identifying various features that may contribute to a homes final sale price.

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

Many numerical variables in the available data showed high variance and were therefore given a log transformation to better perform in a linear regression model. The full list of variables that received a log transformation can be found below. It should be noted that sales price was among the variables.

- 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

### 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. The following linear forms below were identified to produce the lowest AIC for forward, backward, and step selection respectively. A custom model was also fit starting with the results of the forward selection and applying a leave one out cross validation technique. The resulting custom model was found to be identical to the forward selection.

Forward selection:

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:

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:

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

Custom:

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

### **Checking 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, Kaggle Score)

### **Conclusion**

(A short summary of the analysis)

### **Appendix**