**UNIT 13 & 14 House Prices Project**

**Patricia Attah  
Kristi Herman**

Contents

[Introduction 2](#_Toc47802990)

[Analysis Question #1 2](#_Toc47802991)

[Problem Statement 2](#_Toc47802992)

[Checking Assumptions 3](#_Toc47802993)

[Comparing Competing Models 4](#_Toc47802994)

[See Appedix A Interpretation 4](#_Toc47802995)

[Confidence Intervals 5](#_Toc47802996)

[Analysis Question #2 6](#_Toc47802997)

[Problem Statement 6](#_Toc47802998)

[Comparing Competing Models 6](#_Toc47802999)

[Conclusion 7](#_Toc47803000)

[Appendices 8](#_Toc47803001)

[Appendix A – Analysis 1 8](#_Toc47803002)

[Appendix B – Analysis 2 11](#_Toc47803003)

## Introduction

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or the presence of a white-picket fence. With 1460 houses in the dataset and 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the goal of this project is to predict the final price of each home.  
  
Data Description

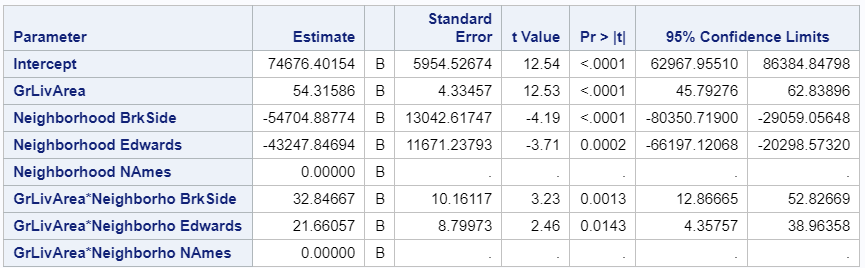
The data in this analysis is from Kaggle’s House Prices: Advanced Regressions Techniques competition. The full training dataset, test dataset, and explanation of variables is available here: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

* There are 1460 houses in the dataset with 79 explanatory variables and 1 response variable (SalePrice).
* The first analysis uses two explanatory, Neighborhood and Above grade/ground living area (GrLivArea), in relationship to sale price.
* The second analysis focuses on variable selection from all the explanatory variables to predict the SalePrice. The output of this analysis will be submitted to Kaggle for scoring.

## Analysis Question #1

Problem Statement  
Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in.  
Build and Fit the Model

***Predicted Sale Price = + + + \* GrLivArea) + \* GrLivArea)***  
Predicted (Sale Price | Neighborhood = NAmes) = +   
Predicted (Sale Price | Neighborhood = BrkSide) = + GrLivArea))   
Predicted (Sale Price | Neighborhood = Edwards) = + )  
  
Predicted (Sale Price | Neighborhood = NAmes) = 74,676 + 54.32  
Predicted (Sale Price | Neighborhood = BrkSide) = 19,971 GrLivArea))   
Predicted (Sale Price | Neighborhood = Edwards) = 31,429 )



### Checking Assumptions

**Addressing Outliers**There are two outliers in the dataset in the Edwards neighborhood. Both houses list over 4600 square feet of above ground living area with unusually low sales prices. Upon further investigation, both homes are listed with a sales condition of “partial.” These observations have been excluded from the analysis.

|  |  |  |
| --- | --- | --- |
| **With Outliers** | **Without Outliers** | |
|  |  | |
|  |  | |
|  | |  |

* **Linearity:** Judging from scatter plot, q-q plot, and histogram of residuals, there is not strong evidence against normality.
* **Constant Variance:** Checking pairwise scatter plots indicates a strong linear trend between GrLivArea and Sales Prices.
* **Normality:** There is little evidence from the residual plots of heteroscedasticity.
* **Independence:** The samples are from 381 houses after removing the two outliers. We will assume the observations are independent.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Residual Plots**

|  |  |
| --- | --- |
|  |  |

### Comparing Competing Models

### See Appedix A Interpretation

For every 100 square foot increase in living area, the increase in mean estimated sales price is $5,430 for houses in NAmes (p-value < 0.0001). While the mean sale prices of houses in BrkSide is estimated to be $54,704 less than mean sale prices in the NAmes, for every one hundred square foot increase in living area in BrkSide, the mean sale price is estimated to be $3,285 more than NAmes (p-value = 0.0013). The mean sale prices of houses in Edwards is estimated to be $43,248 less than mean sale prices in the NAmes, but for every one hundred square foot increase in living area, the mean sale price is estimated to be $2,166 more than NAmes (p-value = 0.0143).

### Confidence Intervals

95% confidence interval for the increase in sale price from NAmes to Brkside ($1,287, $5,283) when the living area increases 100 square feet.

95% confidence interval for the increase in sale price from NAmes to Edwards ($436, $3,896) when the living area increases 100 square feet.

Conclusion  
The evidence suggests that the sales price increases for additional living area in the BrkSide and Edwards neighborhoods compared to additional living area in the NAmes area. Because the sales prices are significantly higher in NAmes than Brkside (p-value = < 0.001) as well as Edwards (p-value = 0.0002), a variable other than living area may be associated with the difference.

## Analysis Question #2

### Problem Statement

With 1460 houses in the dataset and 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the goal of this project is to predict the final price of each home.  
Model Selection  
This analysis includes the following variable selection techniques for the models: Stepwise, Forward, Backward, and Custom.  
  
Checking Assumptions

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

Comparing Competing Models

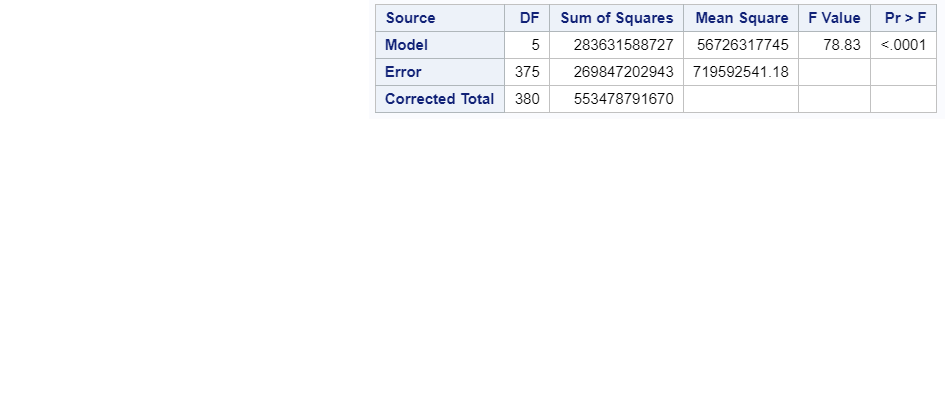
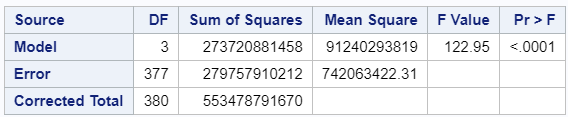
|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | 0.8694 | 1.33 E12 | TBD |
| Backward | 0.9679 | 9.32 E13 | TBD |
| Stepwise | 0.8703 | 1.42 E12 | TBD |
| CUSTOM | 0.9661 | 9.90 E13 | TBD |

Forward selection model variables:   
Intercept LotConfig Neighborhood BldgType OverallQual MSSubClass SaleCondition KitchenQual GarageCars BsmtQual BsmtFinType1 HeatingQC OverallCond CentralAir Electrical Foundation LotArea BsmtFullBath FullBath GrLivArea YearBuilt HalfBath KitchenAbvGr  
  
Backward selection model variables:   
Intercept MSSubClass GarageQual Exterior2nd GarageCond PavedDrive SaleType SaleCondition KitchenQual TotRmsAbvGrd GarageCars GarageType GarageFinish BsmtQual BsmtFinType1 TotalBsmtSF HeatingQC OverallCond CentralAir Electrical RoofStyle Exterior1st MasVnrType MasVnrArea ExterQual ExterCond Foundation MSZoning LotShape LotConfig Neighborhood BldgType HouseStyle OverallQual LotArea MoSold YrSold GarageArea BsmtFullBath BsmtFinSF1 YearRemodAdd FullBath BsmtUnfSF GrLivArea YearBuilt HalfBath BedroomAbvGr KitchenAbvGr  
  
Stepwise selection model variables:   
Intercept GarageCars GarageQual SaleType KitchenQual GarageFinish BsmtQual BsmtFinType1 CentralAir Foundation MSSubClass LotConfig Neighborhood BldgType OverallQual OverallCond GarageArea BsmtFullBath FullBath HalfBath KitchenAbvGr GrLivArea LotArea  
  
Custom Selection model variables:   
(Selected random variable with backward selection choice) Intercept MSSubClass GarageQual GarageCond SaleCondition KitchenQual TotRmsAbvGrd GarageType BsmtQual TotalBsmtSF OverallCond Electrical Exterior1st MasVnrType MasVnrArea ExterQual MSZoning LotShape Neighborhood BldgType HouseStyle OverallQual LotArea MoSold YrSold GarageArea BsmtFullBath BsmtFinSF1 YearRemodAdd FullBath GrLivArea YearBuilt HalfBath BedroomAbvGr KitchenAbvGr

### Conclusion

<tbd>

Appendices

Appendix A – Analysis 1  
  
**Extra Sum of Squares Test for Differences in Slope**  
  
Reduced Model  
  
Full Model  


**The evidence suggests that at least one of the slopes is different**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **SS** | **MS** | **F** | **P-Value** |
| **Model** | **2** | **9,910,707,269** | **4,955,353,634** | **6.89** | **0.0012** |
| **Error** | **375** | **269,847,202,943** | **719,592,541** |  |  |
| **Total** | **377** | **279,757,910,212** |  |  |  |

**Comparing Competing Models**

|  |  |
| --- | --- |
| **GrLivArea and Neighborhood With Interactions** | **GrLivArea and Neighborhood Without Interactions** |
|  |  |

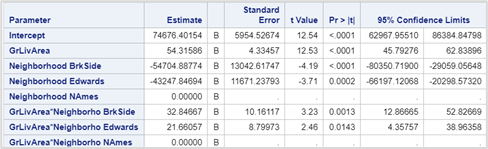
**Adj R2**Adj R2 is slightly better with interactions

|  |  |
| --- | --- |
| **GrLivArea and Neighborhood With Interactions** | **GrLivArea and Neighborhood Without Interactions** |
|  |  |

**Internal CV Press**No difference in variable selection with these variables

|  |  |  |
| --- | --- | --- |
| **Forward** | **Backward** | **Stepwise** |
|  |  |  |

**Parameters & Estimates**



/\* Analysis #1 Code \*/  
/\* Import data and sort it\*/

proc import OUT=WORK.TR

DATAFILE= "/home/u47487140/sasuser.v94/Bridge/train.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

/\* Subset the data \*/

data tr2;

set WORK.TR;

keep Neighborhood SalePrice;

where(Neighborhood in ('NAmes','BrkSide','Edwards'));

if GrLivArea > 4600 then delete;

run;

/\* Scatterplot \*/

title1 "House Data";

title2 "Living Room Area & Sale Price by Neighborhood";

axis1 label=(angle=90 "Sale Price") minor=(n=3);

axis2 label=("Living Room Area (Square Feet)") minor=(n=3);

proc gplot data = tr2;

plot SalePrice \* GrLivArea = Neighborhood /vaxis=axis1 haxis=axis2;

run;

quit;

/\* Matrix \*/

proc sgscatter data=tr2;

title "Scatterplot Matrix for Housing Data";

matrix SalePrice GrLivArea;

run;

title;

/\* Proc GLM with Interactions \*/

proc glm data = tr2 plot = all;

class Neighborhood (ref='NAmes');

model SalePrice = GrLivArea | Neighborhood / solution clparm;

run;

/\* Proc GLM without Interactions \*/

proc glm data = tr2 plot = all;

class Neighborhood (ref='NAmes');

model SalePrice = GrLivArea Neighborhood / solution clparm;

run;

/\* P value on 2 and 375 df \*/

data pval;

pvalue = 1-PROBF(6.89, 2, 375);

run;

/\* Forward Selection \*/

proc glmselect data = tr2;

class Neighborhood;

model saleprice\_log = grlivarea\_log Neighborhood / selection = forward;

run;

/\* Backward \*/

proc glmselect data = tr2;

class Neighborhood;

model saleprice\_log = grlivarea\_log Neighborhood / selection = backward;

run;

/\* Stepwise \*/

proc glmselect data = tr2;

class Neighborhood;

model SalePrice = GrLivArea Neighborhood / selection = stepwise;

run;

/\* Forward Selection \*/

proc glmselect data = tr2;

class Neighborhood;

model SalePrice = GrLivArea Neighborhood / selection = forward;

run;

/\* Backward \*/

proc glmselect data = tr2;

class Neighborhood;

model SalePrice = GrLivArea Neighborhood / selection = backward;

run;

/\* Stepwise \*/

proc glmselect data = tr2;

class Neighborhood;

model SalePrice = GrLivArea Neighborhood / selection = stepwise;

run;

Appendix B – Analysis 2