**UNIT 13 & 14 House Prices Project**

**Patricia Attah  
Kristi Herman**

Contents

[Introduction 2](#_Toc47873686)

[Analysis Question #1 2](#_Toc47873687)

[Problem Statement 2](#_Toc47873688)

[Checking Assumptions 3](#_Toc47873689)

[Comparing Competing Models 4](#_Toc47873690)

[Confidence Intervals 5](#_Toc47873691)

[Analysis Question #2 5](#_Toc47873692)

[Problem Statement 5](#_Toc47873693)

[Comparing Competing Models 5](#_Toc47873694)

[Conclusion 7](#_Toc47873695)

[Appendices 8](#_Toc47873696)

[Appendix A – Analysis 1 8](#_Toc47873697)

[Appendix B – Analysis 2 10](#_Toc47873698)

## 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

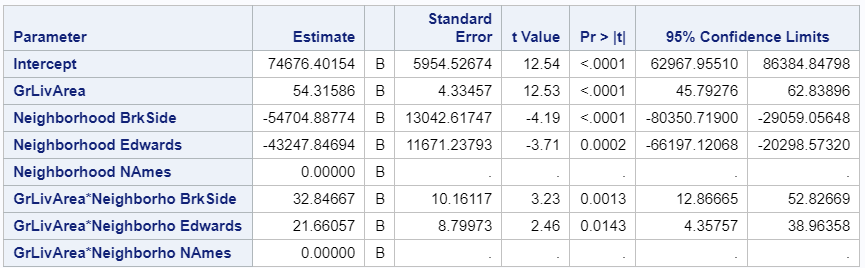
The data in this analysis is from Kaggle’s House Prices: Advanced Regressions Techniques competition. The full training dataset, test dataset, and explanation of variables is available here: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

* 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:** Checking pairwise scatter plots indicates a strong linear trend between GrLivArea and Sales Prices.
* **Constant Variance:** There is little evidence from the residual plots of heteroscedasticity.
* **Normality:** Judging from scatter plot, q-q plot, and histogram of residuals, there is not strong evidence against normality.
* **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 Appendix A   
Interpretation

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

### Confidence Intervals

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

95% confidence interval for the increase in sale price from North Ames 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 Brookside and Edwards neighborhoods compared to additional living area in the North Ames 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 overall estimated difference in mean prices.

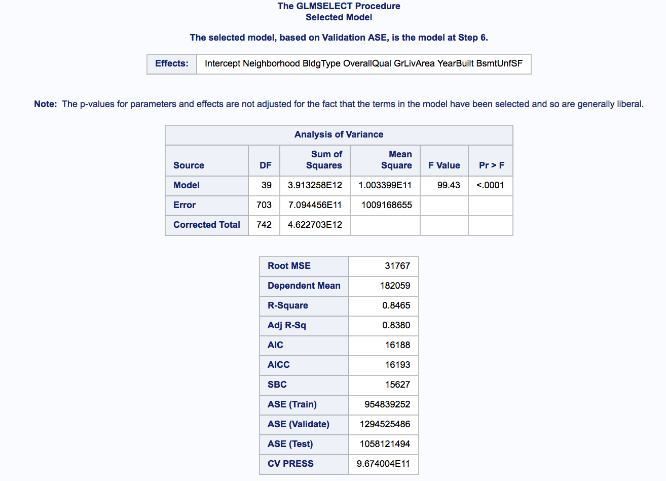
## Analysis Question #2

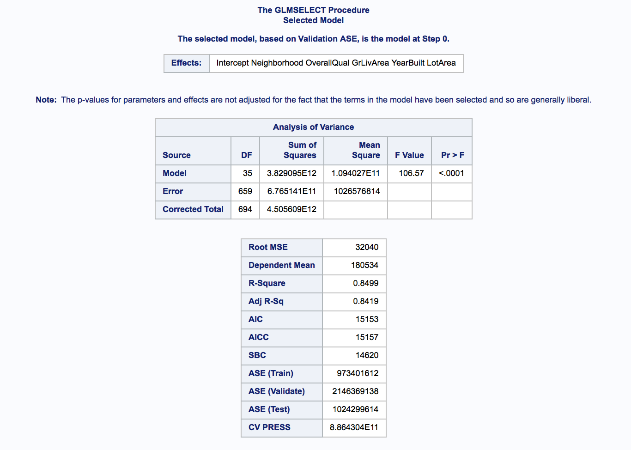
### Problem Statement

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   
See Appendix B.

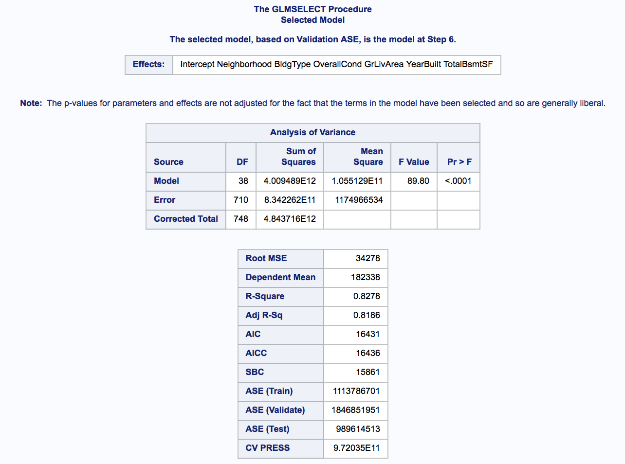
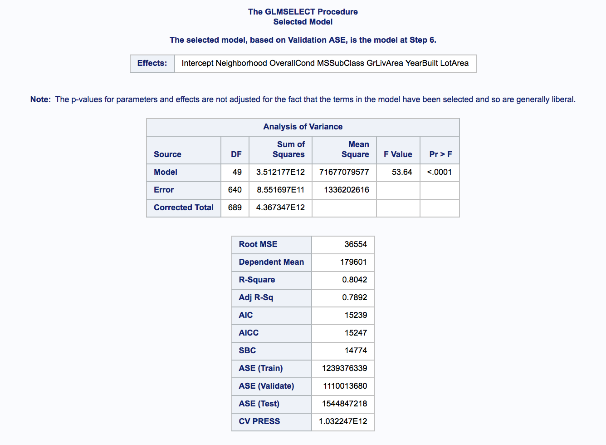
Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | 0.8380 | 9.67 E11 | 0.16847 |
| Backward | 0.8419 | 8.86 E11 | 0.19454 |
| Stepwise | 0.8186 | 9.72 E11 | 0.20957 |
| CUSTOM | 0.7892 | 1.03 E12 | 0.19188 |

**Forward selection model variables:**   
Neighborhood BldgType OverallQual GrLivArea YearBuilt BsmtUnfSF  
  
  
  
**Backward selection model variables:**   
Neighborhood OverallQual GrLivArea YearBuilt LotArea



**Stepwise selection model variables:**Neighborhood BldgType OverallCond GrLivArea YearBuilt TotalBsmtSF

  
  
**Custom Selection model variables:**   
Neighborhood OverallCond MSSubClass GrLivArea YearBuilt LotArea  


**Conclusion**

SAS results RMSE values for forward selection had the lowest value at 31767 however checking the AIC the backward selection was the lowest scored at 15153. Also, when looking at the adj R2 backward selection is the best with score of 0.842 while forward selection comes closely behind at 0.838 they are approximately the same. Results from SAS analysis of AIC, Adjusted R2 and RMSE supports forward selection variables as the most accurate. Stepwise and forward selection were most effective with picking the variables for Kaggle score using SAS, backward selection was least effective, it was noticed that no variable were removed using that method. Variables for forward selection are: SalePrice predictions: Neighborhood BldgType OverallQual GrLivArea YearBuilt BsmtUnfSF.

Appendices

Appendix A – Analysis 1  
  
**Comparing Competing Models**

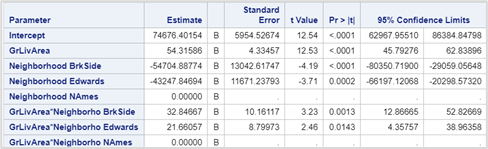
|  |  |
| --- | --- |
| **GrLivArea and Neighborhood With Interactions** | **GrLivArea and Neighborhood Without Interactions** |
|  |  |

**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;

|  |
| --- |
| /\* No2 Analysis\*/  /\* train test data set combined\*/  data test;  set test;  SalePrice = .;  ;  data train01;  set train test;  run;  /\* Stepwise selection\*/  %let categorical = Neighborhood BldgType OverallCond;  %let interval = GrLivArea YearBuilt LotArea TotalBsmtSF ;    PROC GLMSELECT DATA= train01 seed = 4  plots=all;  partition fraction(test=0.25 validate=0.25);  class &categorical / param=glm ref=first;  model SalePrice=&categorical &interval/ SHOWPVALUES  selection=stepwise(choose= cv stop =cv)cvdetails  select= cv  choose= validate;  output out = result\_stepwise p = predict;  run;  /\* Forward selection\*/  %let categorical = Neighborhood BldgType OverallQual;  %let interval = BedroomAbvGr GrLivArea YearBuilt LotArea BsmtUnfSF;  PROC GLMSELECT DATA= train01 seed = 1000  plots=all;  partition fraction(test=0.25 validate=0.25);  class &categorical / param=glm ref=first;  model SalePrice=&categorical &interval/SHOWPVALUES  selection=forward  select= cv  choose= validate;  output out = result\_forward p = predict;  run;  /\* Backward selection\*/  %let categorical = Neighborhood OverallQual ;  %let interval = GrLivArea YearBuilt LotArea;  PROC GLMSELECT DATA= train01 seed = 55  plots=all;  partition fraction(test=0.25 validate=0.25);  class &categorical / param=glm ref=first;  model SalePrice=&categorical &interval/SHOWPVALUES  selection=backward  select= cv  choose= validate;  output out = result\_backward p = predict;  run;  /\* Custom selection\*/  %let categorical = Neighborhood OverallCond MSSubClass MasVnrArea;  %let interval = GrLivArea YearBuilt LotArea;  PROC GLMSELECT DATA= train01 seed = 55  plots=all;  partition fraction(test=0.25 validate=0.25);  class &categorical / param=glm ref=first;  model SalePrice=&categorical &interval/SHOWPVALUES  selection=stepwise  select= cv  choose= validate;  output out = result\_custom p = predict;  run;  /\*kaggle results files\*/  data results\_backward;  set result\_backward;  if predict > 0 then SalePrice = predict;  if predict < 0 then SalePrice = 10000;  keep id SalePrice;  where id > 1460  ;  data results\_forward;  set result\_forward;  if predict > 0 then SalePrice = predict;  if predict < 0 then SalePrice = 15000;  keep id SalePrice;  where id > 1460  ;  data results\_stepwise;  set result\_stepwise;  if predict > 0 then SalePrice = predict;  if predict < 0 then SalePrice = 10000;  keep id SalePrice;  where id > 1460  ;  data results\_custom;  set result\_custom;  if predict > 0 then SalePrice = predict;  if predict < 0 then SalePrice = 10000;  keep id SalePrice;  where id > 1460  ; |

Appendix B – Analysis 2

* **Linearity:** Checking pairwise scatter plots indicates some linear trend between Sales Prices and the continuous variables.
* **Constant Variance:** There is some evidence from the residual plots of heteroscedasticity.
* **Normality:** Judging from scatter plot, q-q plot, and histogram of residuals, there is not strong evidence against normality.
* **Independence:** The samples are from 1460 houses. We will assume the observations are independent.
* **Influential Point analysis:** Leverage of the diagram there is no single point that has both high leverage and influence. From the Cook’s D chart there is no point that is considerably higher than all other points.

