MSDS 6371 Project Description (Weeks 13 and 14)

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 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

**Data and Description:**

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

**HOW TO KAGGLE VIDEO:**

**QOIs:** [**https://www.youtube.com/watch?v=0QJtczDPxZQ**](https://www.youtube.com/watch?v=0QJtczDPxZQ)

Read everything BEFORE you begin.

* Use SAS for this project.

Your team’s objective is to conduct 2 analyses:

1. ANALYSIS 1: Assume that Century 21 Ames (a real estate company) in Ames Iowa has commissioned you to answer a very important question with respect to their business. Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to simply 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 a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. It turns out that Century 21’s leadership team has a member that has some statistical background. Therefore, make sure and provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, of course, provide your client with a well written conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods. Remember that the company is only concerned with the three neighborhoods they sell in.

1. ANALYSIS 2: Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods. Your group is limited to only the techniques we have learned in 6371 (no random forests or other methods we have not yet covered). Specifically, you should produce 4 models: one from forward selection, one from backwards elimination, one from stepwise selection, and one that you build custom. The custom model could be one of the three preceding models or one that you build by adding or subtracting variables at your will. Generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa. In your paper, please include a table similar to the one below. The group with the lowest public Kaggle score will receive an extra 3 bonus points on the final exam!

Quick note on Kaggle completion: We only have one course under our belts so far (almost), but you can compete in this competition with the tools you have now (top 40th percentile or better!). After your next course (6372), you will really be able to do well (top 25th percentile or better!). With these skills as well as the skills you pick up in Data Mining and Quantifying the World, you will be able to compete with anyone!

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | .89 | 1272 | .721 |
| Backward | .78 | 1590 | .945 |
| Stepwise | .81 | 2001 | .888 |
| CUSTOM | .87 | 900 | .2345 |

NOTE 1: ALL ANALYSES MUST BE DONE IN SAS and all code must be placed in the appendix. Part of the grading process will be to run the code and verify the Kaggle score for each group.

Note 2: An extra 3 points on the final exam will be awarded to the team with the model with the lowest (best) Kaggle Score. In the unlikely event of a tie will split these points.

**Deliverables:**

Your group is to turn in a paper should be no more than 7 pages long (without the appendix). Please put your code in the appendix. If you are concerned with staying within the allotted 7 pages, put more screenshots and such in the appendix.

**Sample Format**

Required deliverables in the complete report:

The format of your paper (headers, sections, etc.) is flexible, although it should contain the following information.

Introduction

Total rows= 1460

Total variables = 81

Data Description

(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?)

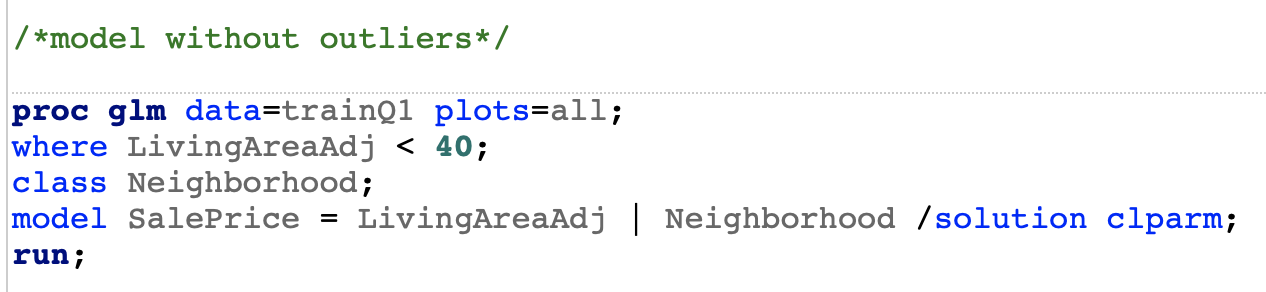
Analysis Question 1:

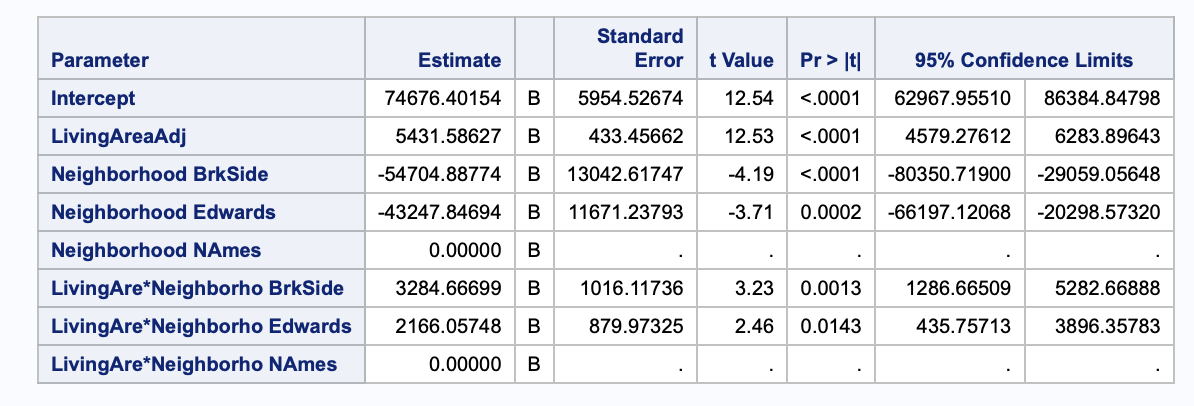
Restatement of Problem

How is the sale price of the house related to the square footage (in 100s of square feet) of the living area? Does the sale price of the house depend on the neighborhood the house is located in? (Three neighborhoods are Names, Edwards, and Brkside).

Build and Fit the Model



****

****

μ(SalePrice) = β0 + β1(LivingAreaAdj) + β2(Edwards) + β3(BrkSide) + β4(LivingAreaAdj\*Edwards) + β5(LivingAreaAdj\*BrkSide)

μ(SalePrice) = 74676.40 + 5431.57(LivingAreaAdj) – 43247.85(Edwards) -54704.88(BrkSide) + 2166.06(LivingAreaAdj\*Edwards) + 3284.67(LivingAreaAdj\*BrkSide)

Individual Equations:

NAmes:

μ(SalePrice) = 74676.40 + 5431.57(LivingAreaAdj)

In NAmes, for every additional increase in 100 square feet, the expected sale price increases by $5431.57. A 95% CI for this increase is between $4579.28 and $6283.90.

Edwards:

μ(SalePrice) = 31428.55 + 7597.63(LivingAreaAdj)

In Edwards, for every additional increase in 100 square feet, the expected sale price increases by $7597.63. A 95% CI for this increase is (5431.57 + 435.76, 5431.57+ 3896.36) = ($5867.33, $9327.93).

BrkSide:

μ(SalePrice) = 19971.52 + 8716.24 (LivingAreaAdj)

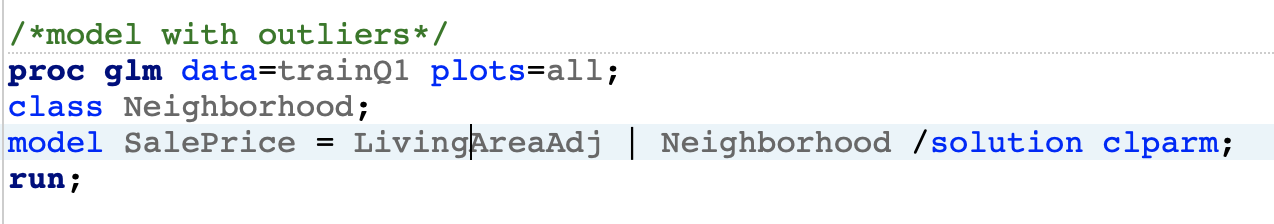
In BrkSide, for every additional increase in 100 square feet, the expected sale price increases by $8716.24. A 95% CI for this increase is (5431.57 + 1286.67, 5431.57 + 5282.67) = ($6718.24, $10714.24).

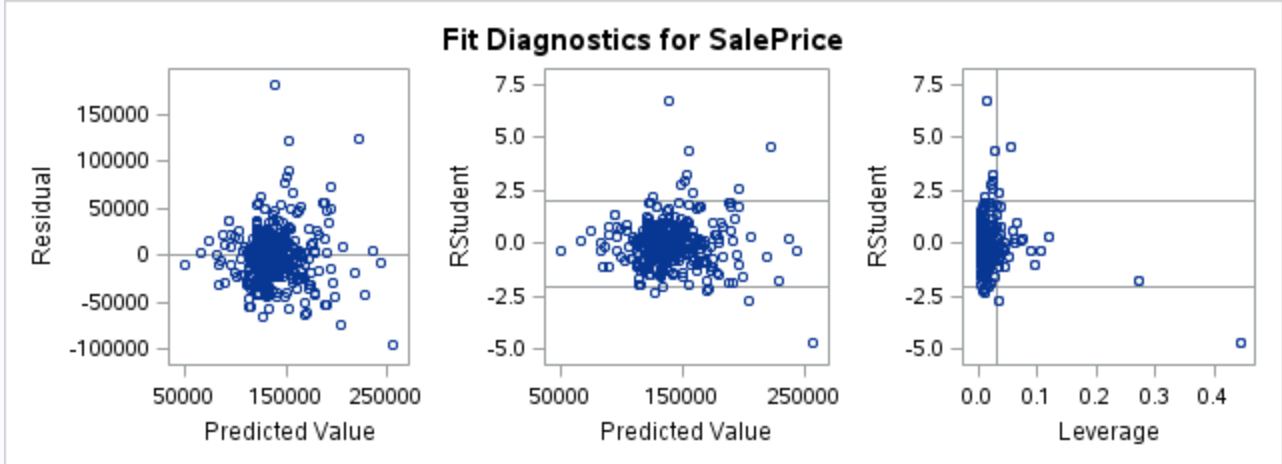
Checking Assumptions

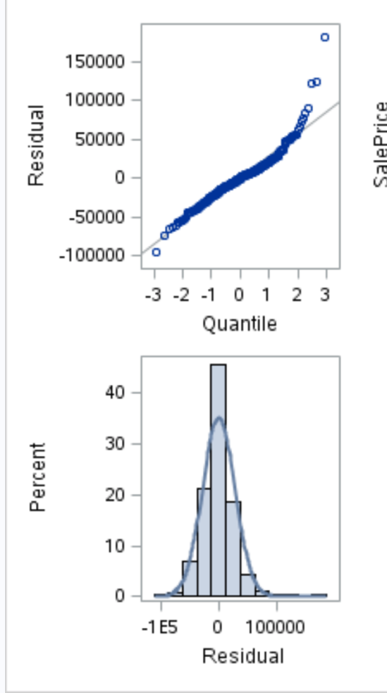
Residual Plots

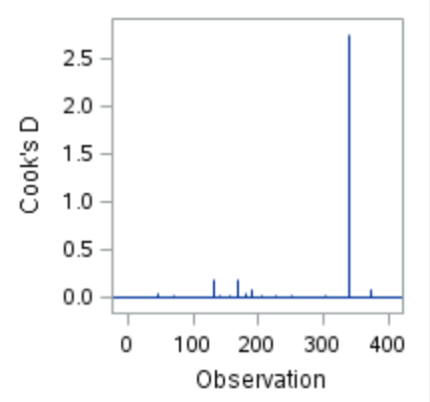
Influential point analysis (Cook’s D and Leverage)

Make sure to address each assumption.





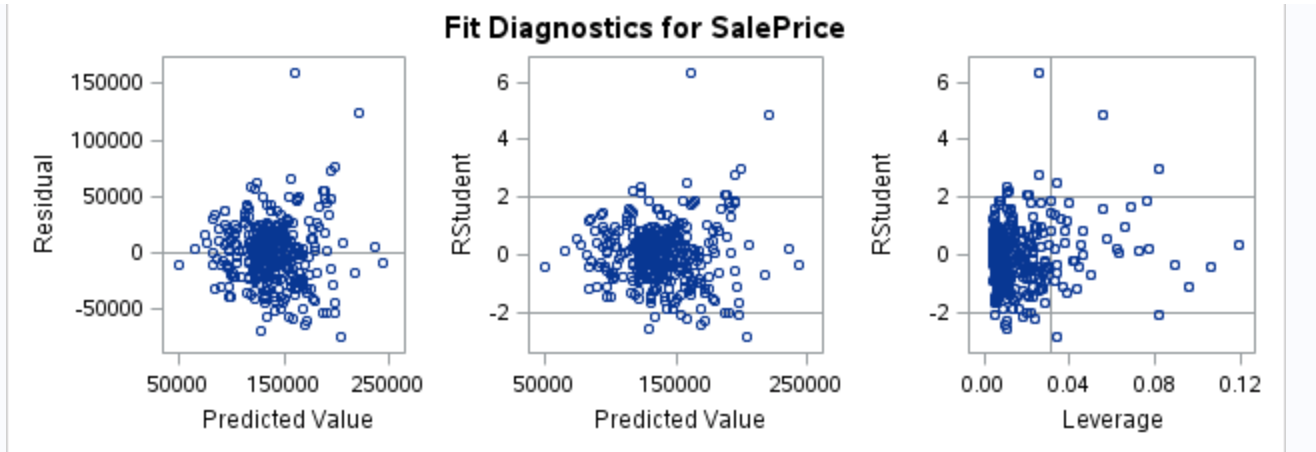


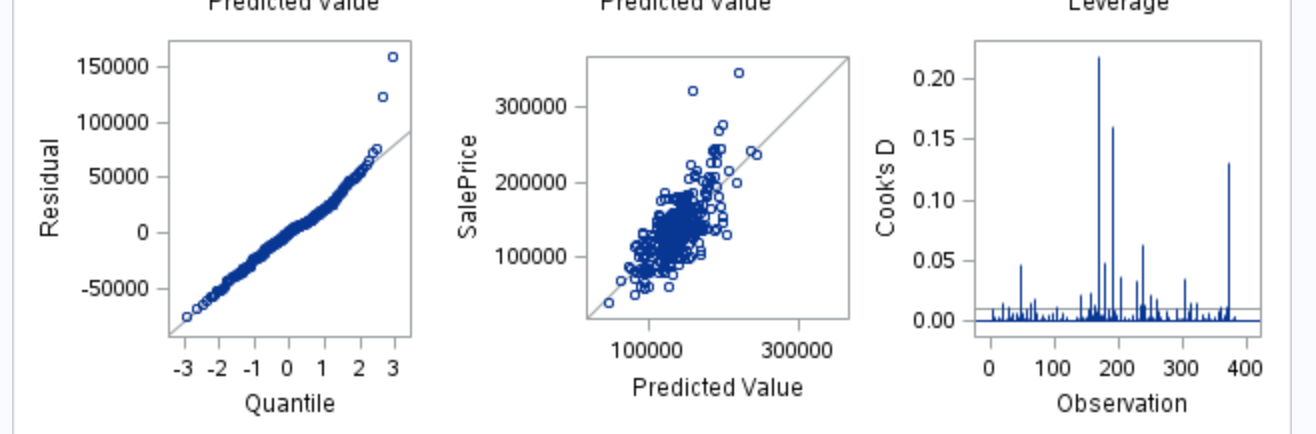


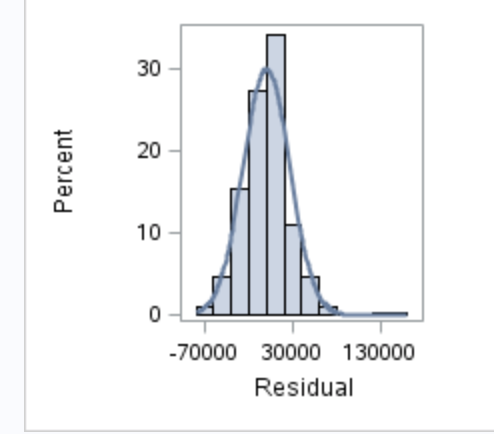
Outliers:

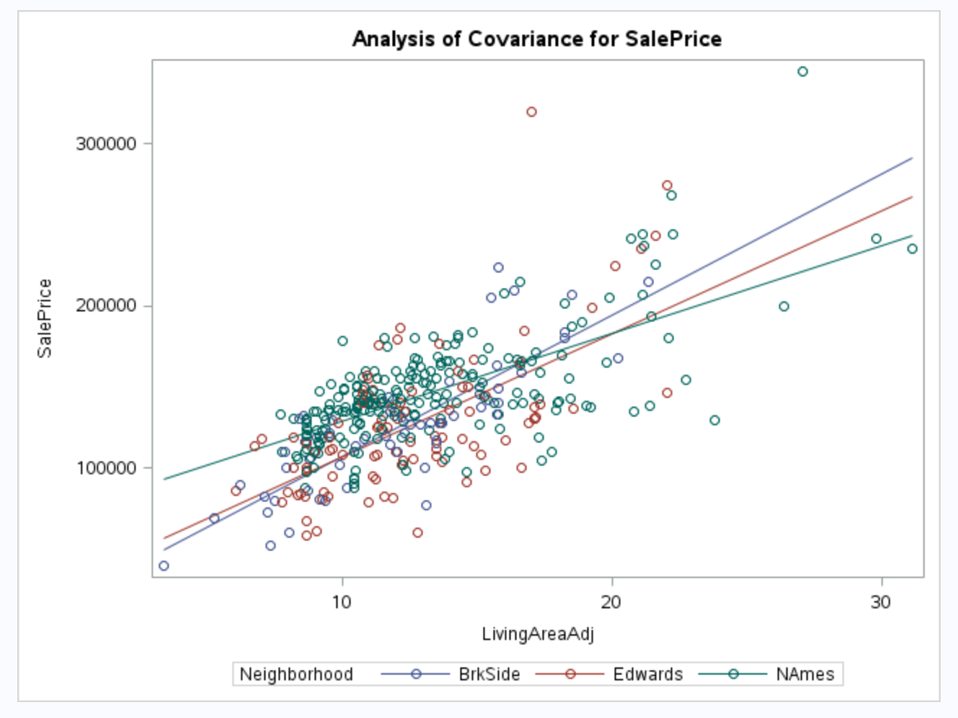
There is one extreme outlier with a very high studentized residual and leverage. It is identified by Cook’s D. There is another outlier with a less extreme studentized residual and leverage but it does not seem to be identified by Cook’s D. I decided to remove the two outliers and focus my analysis on where the LivingArea in hundreds of square feet is less than 40, which is justified because the outliers have high leverage.

After refitting my data:









Normality Assumption: Looking at the QQ plot and histogram, our residuals are normally distributed.

Constant Variance Assumption: Looking at the scatterplot and residual plot, there is a random cloud pattern, indicating constant variance.

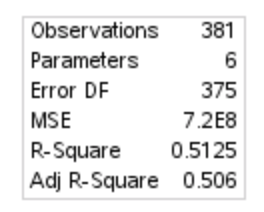
Linear trend assumption: There is definitely a linear trend between SalePrice and Living Area looking at the coded scatterplot.

Independence Assumption: We will proceed with caution because we know that the houses are clustered by neighborhood.

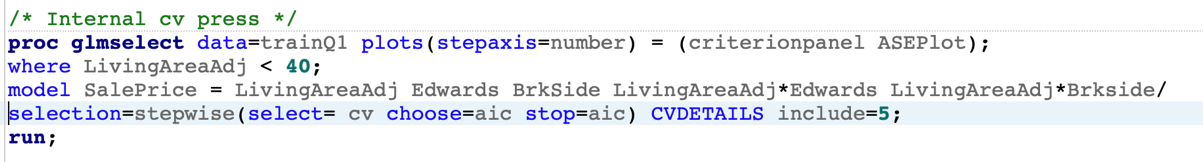
Model Metrics

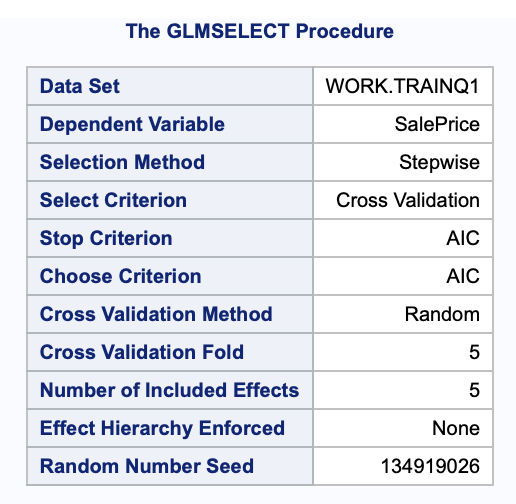
Adj R2

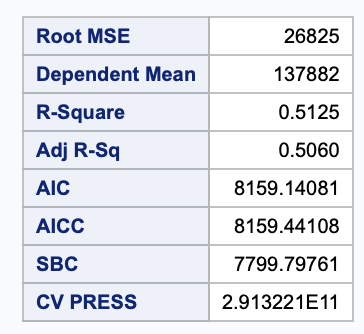
Internal CV Press



The adjusted R2 is 0.506.







The internal CV PRESS is 2.91 x 1011.

Parameters

Estimates

Interpretation

Confidence Intervals

Conclusion

A short summary of the analysis.

Analysis Question 2

Restatement of Problem

Model Selection

Type of Selection

Stepwise

Forward

Backward

Custom

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

Well commented SAS Code for Analysis 1 and 2

**Rubric:**

Presentation (30%):

Organized paper with title, headings, subheadings, etc.

Labeled plots, figures, tables and charts.

Every plot, figure, table and chart included is referenced in the paper and vice versa.

No spelling or grammatical errors.

Analysis Question 1: (35%)

Analysis Question 2: (35 %)