Kaggle Project - Team Fat Tails

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. However, it is essential to review the data because it proves that there are many other influences in price negotiations than the number of bedrooms or a white-picket fence.

Data Synopsis

The Ames House dataset was compiled by Dean De Cock and contains 79 explanatory variables describing almost every aspect of residual home in Ames Iowa from 2006 to 2010. The data set contains 2930 observations involved in assessing home values.

@import "../figs/procmeans.html"

- More data definitions

Analysis Question 1

Restatement of Problem

To build and fit a model, an analysis must be performed to identify features of the dataset that are statistically significant in their relation to, and prediction of, the sales price.

Build the Model

Interrogate the Data

To build and fit a model, an analysis must be performed to identify features of the dataset that are statistically significant in their relation to, and prediction of, the sales price.

- QOI: this is where century targets.

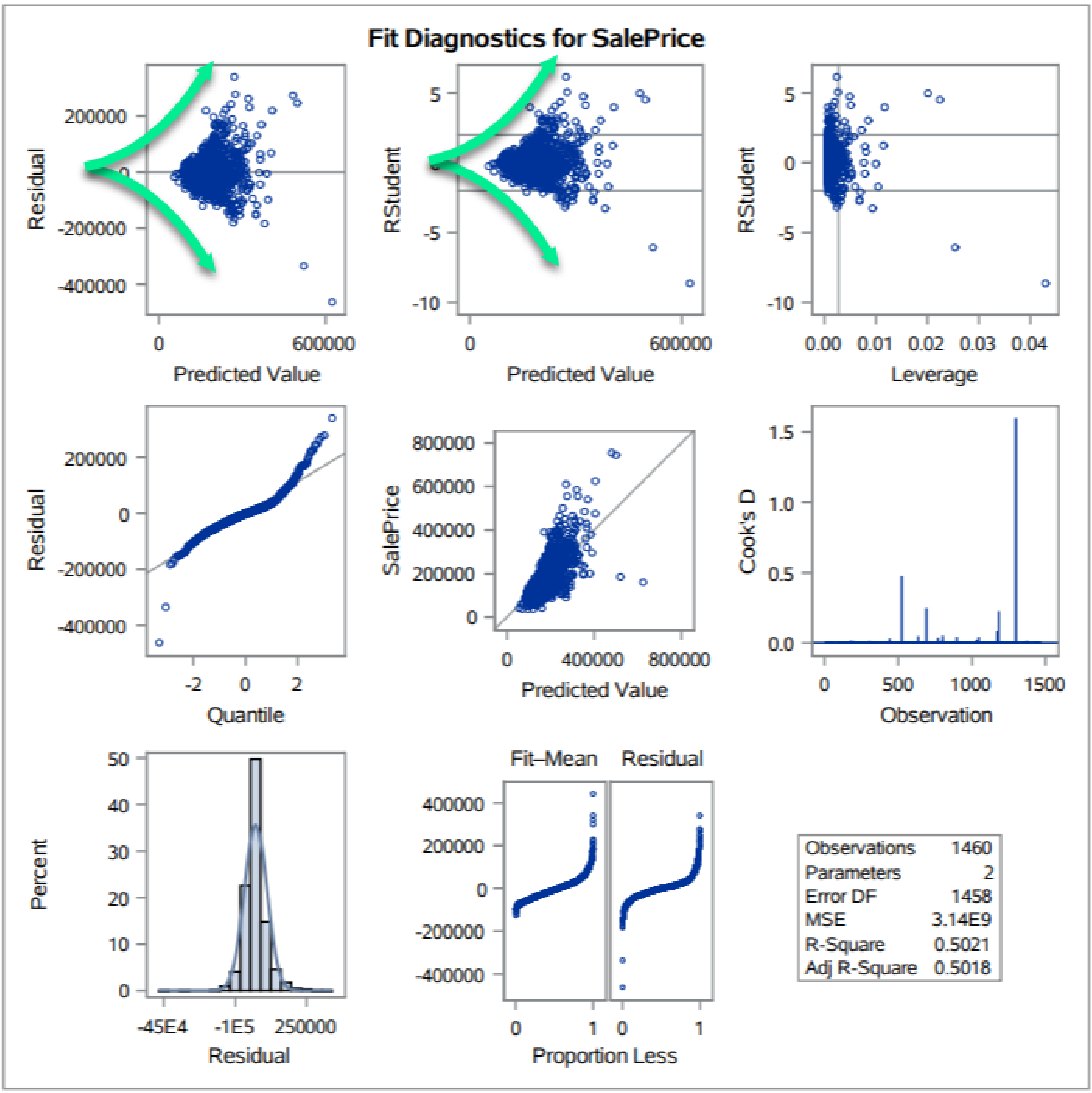
- We chose to log both variables because of the graphs below, and other words.

Fit the Model

@import "../figs/analysis1_solution.html"

Check Assumptions

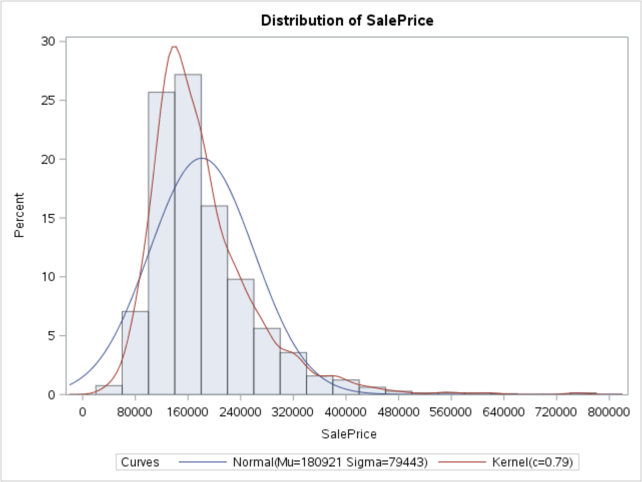
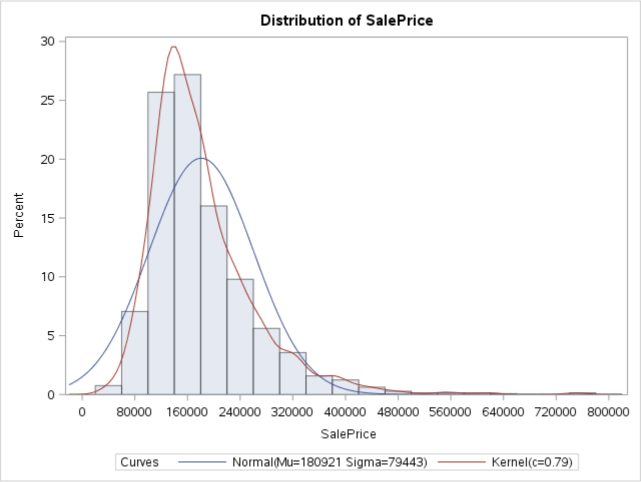
Homogeneity of Variances



Observations	1460
Parameters	2
Error DF	1458
MSE	3.14E9
R-Square	0.5021
Adj R-Square	0.5018

Normality

Solarized dark	Solarized Ocean

Solarized dark	Solarized Ocean
	

Residual Diagnostics

Outlier Analysis

Model Comparison

No Interactions

With Interactions

ANOVA Comparison

Adj R2

Parameters & Equations

```
    - Estimates
    - Interpretation
    - Confidence Intervals
```

Neighborhooods: $x_1 = BrkSide = Brookside $x_2 = NAmes$ = Ames $x_3 = Edwards = Edwards

Variables: SalesPrice = SP LivingArea = LA

General Formula:

$$ \hat\mu { {log(SP)} } ,=, \beta_0, +,\beta_1,Brookside, +,\beta_2,Edwards, +,\beta_3,Ames, +,\beta_4(log(LA),Brookside) +\beta_{5},(log(LA),Edwards) $$

Ames (North):

$$ \hat\mu { {log(SP_{Ames})} } ,=, \beta_0, +,\beta_1,Brookside, +,\beta_2,Edwards, +,\beta_3,Ames, +,\beta_4(log(LA),Brookside) +\beta_{5},(log(LA),Edwards) $$

Brookside:

$$ \hat\mu { {log(SP_{Brookside})} } ,=, \beta_0, +,\beta_1,Brookside, +,\beta_2,Edwards, +,\beta_3,Ames, +,\beta_4(log(LA),Brookside) +\beta_{5},(log(LA),Edwards) $$

Edwards:

$$ \hat\mu { {log(SP_{Edwards})} } ,=, \beta_0, +,\beta_1,Brookside, +,\beta_2,Edwards, +,\beta_3,Ames, +,\beta_4(log(LA),Brookside) +\beta_{5},(log(LA),Edwards) $$

Conclusion

Prediction mean sales price by neighborhood.

To interpret the model, a change in Living Room Square Feet is a doubled increase. For the neighborhood with approximately the same mass, it is estimate that a 10-fold increase in the Living Area Square feet is associated with a XX which is a 83.2% increase in the median Sales Price of the neighborhood. (P value < 0.001). At a 95% confidence intervals for the increase in sales price of XX = CI which equates to an estimated increase between X% and X%.

Analysis Question 2

Restatement of Problem

Model Selection

```
        Type of Selection
              Stepwise
```

Forward - Starts empty, incrementally adds variables Backward - Starts full, incrementally removes variables. Stepwise - ???, More random. Can add or remove until it finds the best score. Custom - ???

Checking Assumptions

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

  <!-- TODO: Scatterplot matrix of all variables used in any of the models. Add
these to the appendix. -->
```

Comparing Competing Models

```

                Adj R2
                Internal CV Press
                Kaggle Score


        Conclusion: A short summary of the analysis.

```

Check Assumptions

Homogeneity of Variances

Normality

Residual Diagnostics

Outlier Analysis

Kaggle Score for stepwise submission: 0.14880 Kaggle Score for forward elimination: 0.14880 Kaggle score for backward elimination: 0.21225

Appendix A

SAS Program

main.sas

@import "main.sas"

dataimport.sas

@import "dataimport.sas"

procmeans.sas

@import "procmeans.sas"

analysis1_model1.sas

@import "analysis1_model1.sas"

Appendix B

@import "../data/data_description.md"

Appendix C

Downloading from the Kaggle API

Using Code Blocks in Markdown

Using SAS in Markdown Code Blocks