Kaggle House Price Analysis:

Finding a Correlation and Regression of Leading Contributing Factors

Mohammed Al-Rousan

Robert Lane

Southern Methodist University

Contents

[Q1: House Prices in Ames Iowa 3](#_Toc131854120)

[Q1 Assumptions 3](#_Toc131854121)

[Linear Relationship & Normality 3](#_Toc131854122)

[Linear Relationship 3](#_Toc131854123)

[Multivariate Normality 4](#_Toc131854124)

[No Multicollinearity 4](#_Toc131854125)

[No Autocorrelation 4](#_Toc131854126)

[Homoscedasticity 5](#_Toc131854127)

[Fit Analysis 5](#_Toc131854128)

# Q1: House Prices in Ames Iowa

This question focuses on 3 neighborhoods, marked in the dataset at Names, Edwards, and BrkSide.

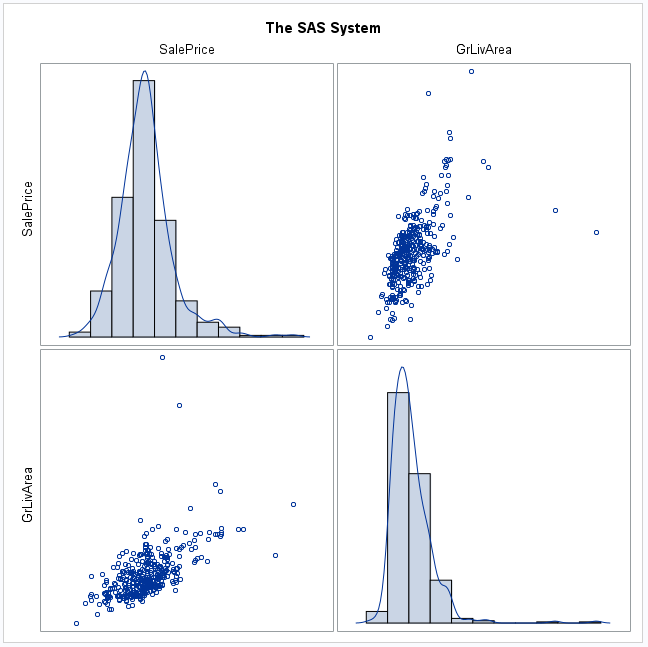
## Q1 Assumptions

### Linear Relationship & Normality

**proc** **sgscatter** data = neighborhoods;

matrix SalePrice GrLivArea / diagonal=(histogram kernel);

**run**;



Fail.

Both the independent and the dependent variables exhibit signs of right-skewness as well as increasing variance. Running a log-log model to attempt to correct for this. Also, coding neighborhoods as dummy variables for later use.

**data** loghood;

set neighborhoods;

logprice = log(SalePrice);

logarea = log(GrLivArea);

BrkSide = (Neighborhood = "BrkSide");

NAmes = (Neighborhood = "NAmes");

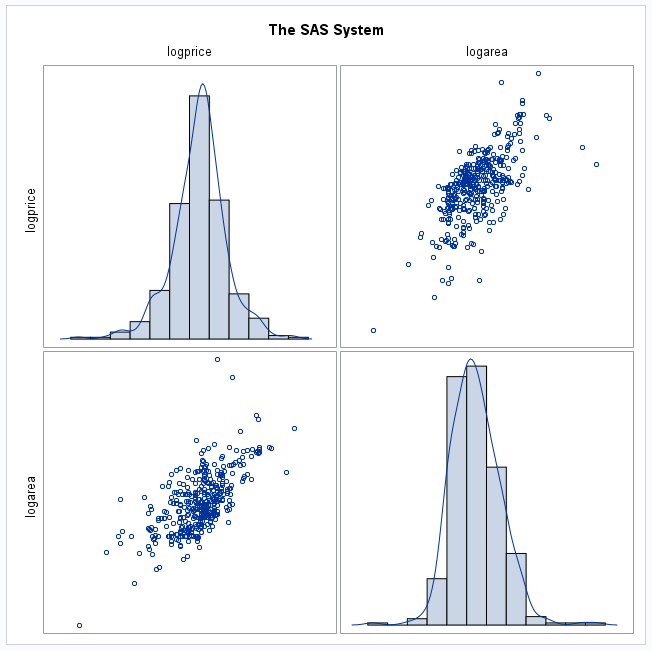
**run**;

### Linear Relationship

**proc** **sgscatter** data = loghood;

matrix logprice logarea / diagonal=(histogram kernel);

**run**;



Pass. The relationship between the log of the area and the log of the price does seem to exhibit a linear relationship.

### Multivariate Normality

Pass. Using the matrix above, it is evident that both inputs have normality.

### No Multicollinearity

Pass. Only one explanatory variable is used here.

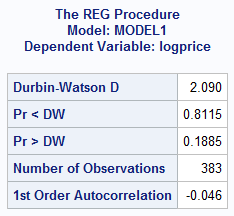
### No Autocorrelation

Running Durbin-Watson test for autocorrelation.

**proc** **reg** data = loghood;

model logprice = logarea BrkSide NAmes / dwProb;

run;



A Durbin-Watson score near 2 indicates that there is almost zero autocorrelation. Pass.

### Homoscedasticity

The variance visually appears to be pretty even at both low and high values, and for each axis. Pass.

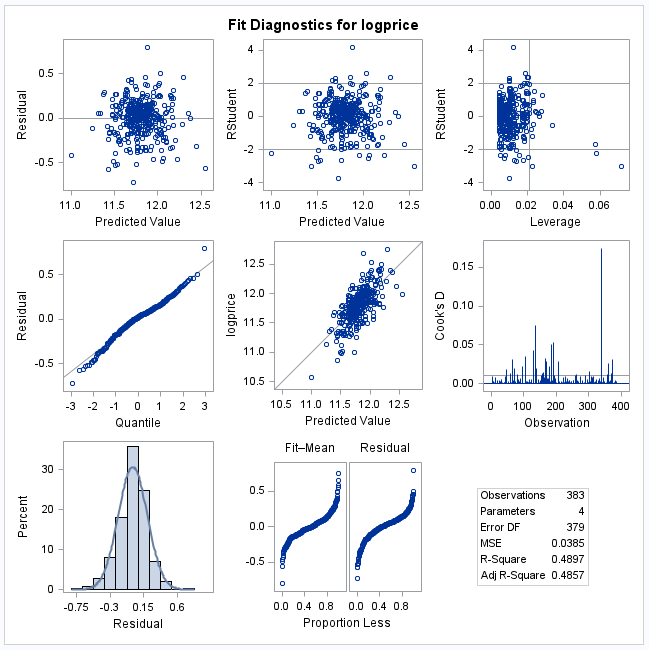
## Fit Analysis

**proc** **reg** data = loghood

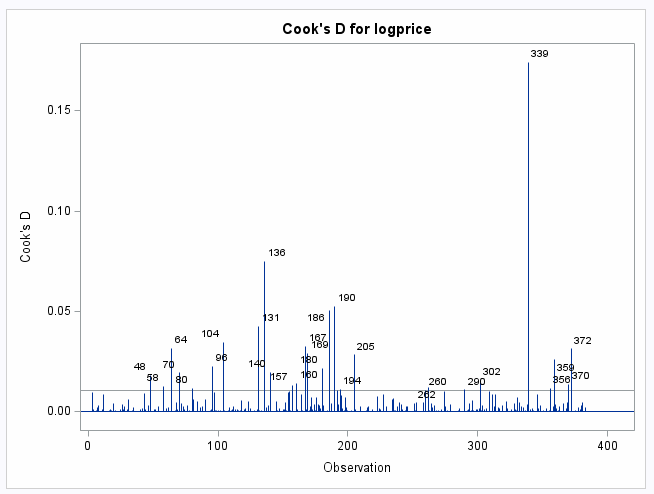
plots = (DiagnosticsPanel ResidualPlot(smooth));

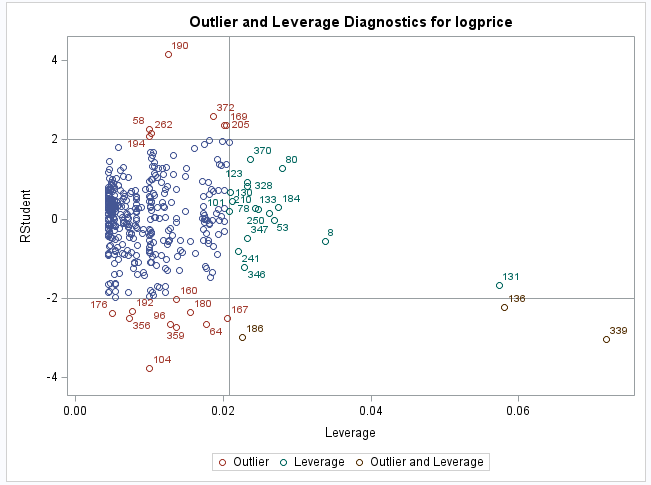
model logprice = logarea BrkSide NAmes;

**quit**;



The histogram and Q-Q plot both indicate that residuals are normally distributed, but the Leverage plot indicates that there are 3 influential outliers which should be checked.





186, 136, and 339 look suspect as influential values.

View just the relevant information for these houses:

**data** temp;

set loghood;

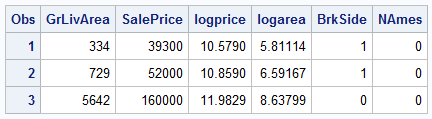
if \_n\_ in (**339**, **186**, **136**);

keep SalePrice GrLivArea logprice logarea BrkSide NAmes;

**run**;

**proc** **print** data=temp(obs=**3**);

**run**;



It seems unreasonable to make much of a prediction for the sale of a livable area of less than 500 square feet. For reference, that would be a living area off less than two typical parking spaces. It further seems unreasonable to expect a house larger than 5,000 square feet to sell at a price of $160,000. Therefore, of these three values, the first and last one (obs = 339 and 136) seem to be unlikely scenarios and may be removed. The middle one, while still an influential outlier, seems at least plausible.