**ANALYSIS 1:**

**Introduction:**

Century 21 Ames (a real estate company) in Ames Iowa has commissioned us 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

**Data Description**:

The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for use in data science education. The data set ,train.csv describing the sale of individual residential property in Ames, Iowa from 2006 to 2010 contains 1460 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 21 continuous) involved in assessing home values , with a total of 81 variables. 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. Hence the specific variables are SalesPrice( dependent variable ) is explained with living area of the house (**GrLivArea)** and **SalesPrice** (and its relationship to square footage) depends on the **neighborhood** ( three neighborhood of interest )

Further information can be found at <http://jse.amstat.org/v19n3/decock.pdf>

**Restatement of Problem:**

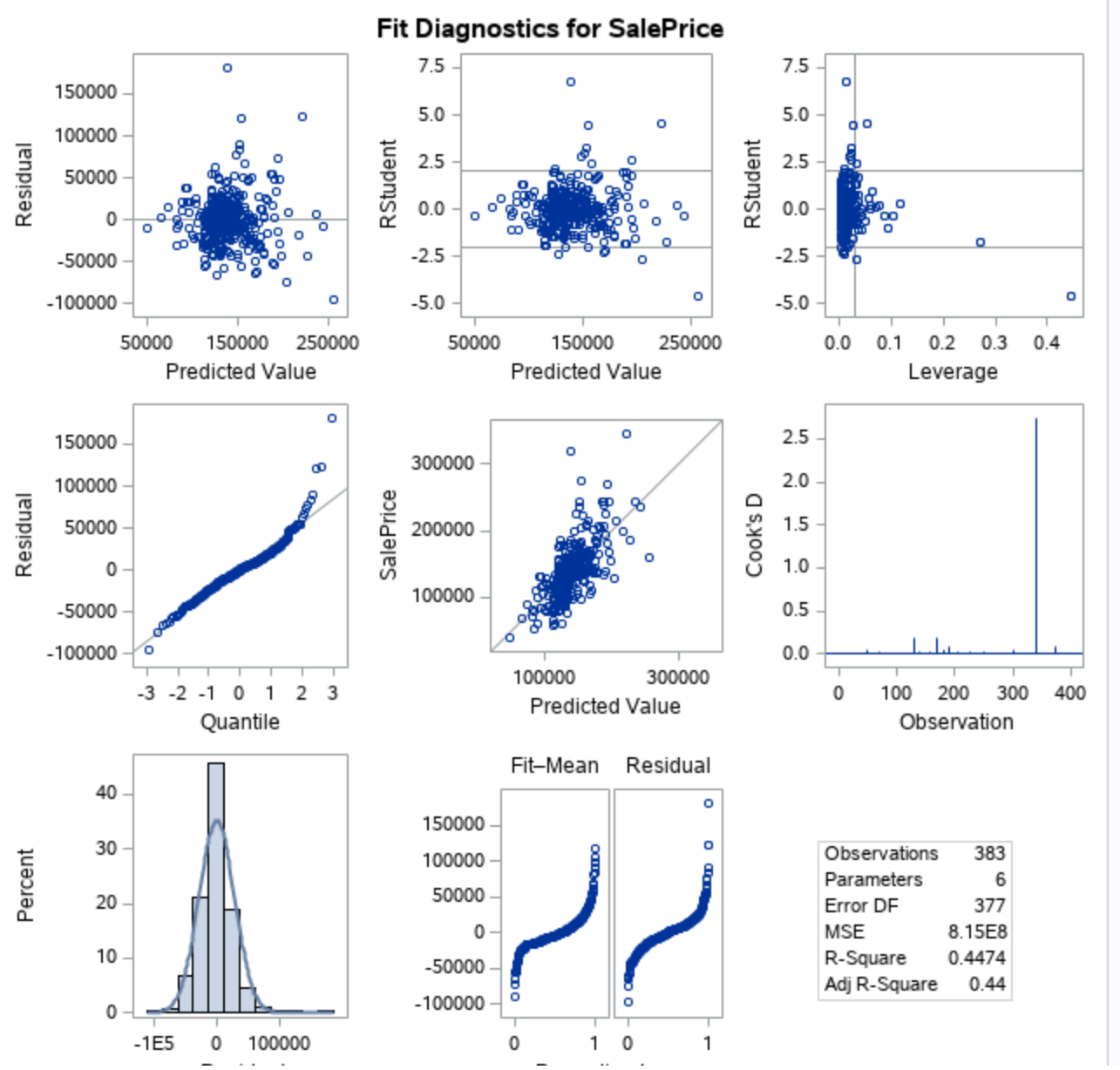
provide our client with a well written conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods

**Build and Fit the Model:**

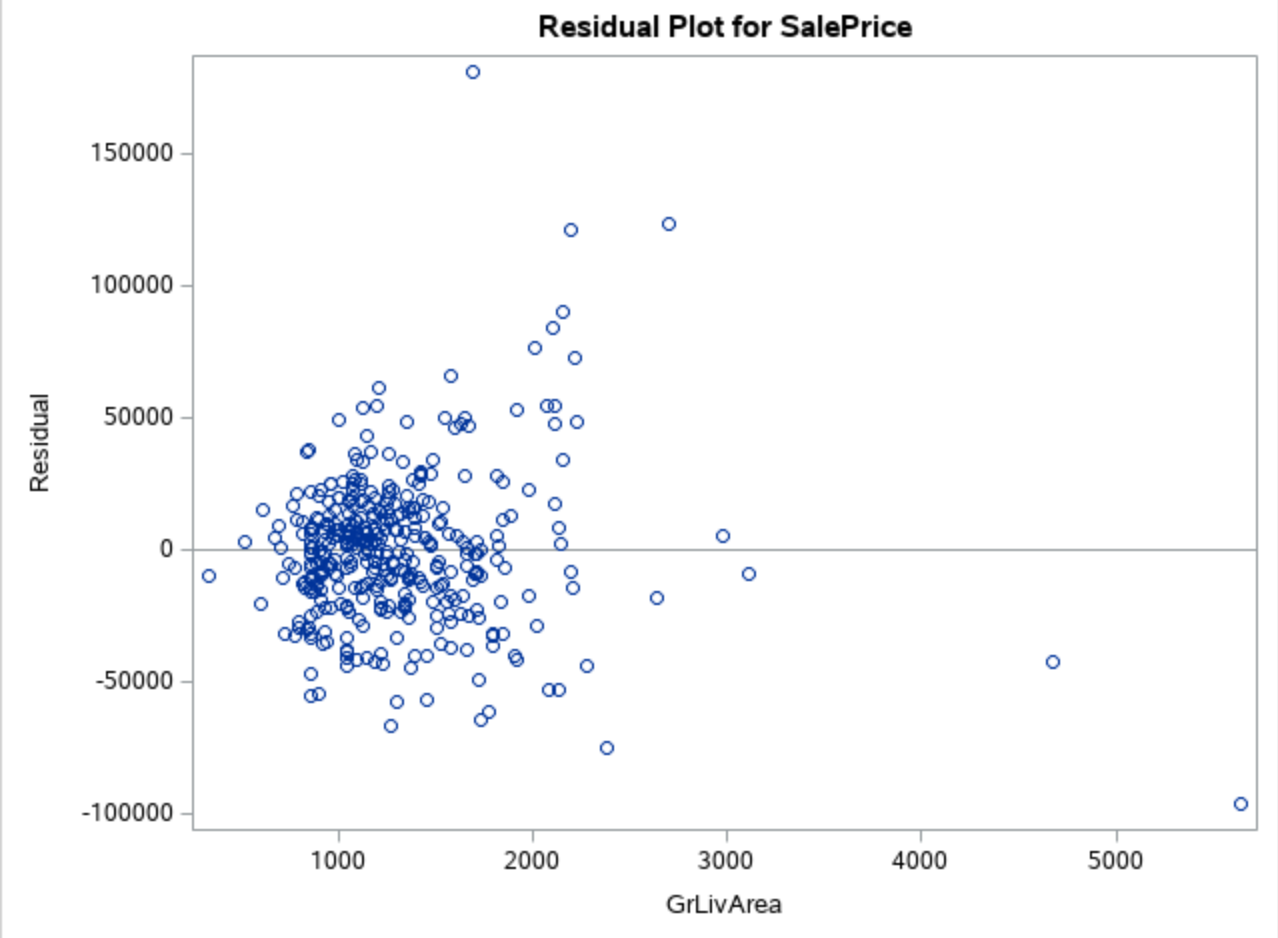
Build and fit a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide.

fit diagnostics for different cases

Linear – linear model



**The residual plot also presented as**

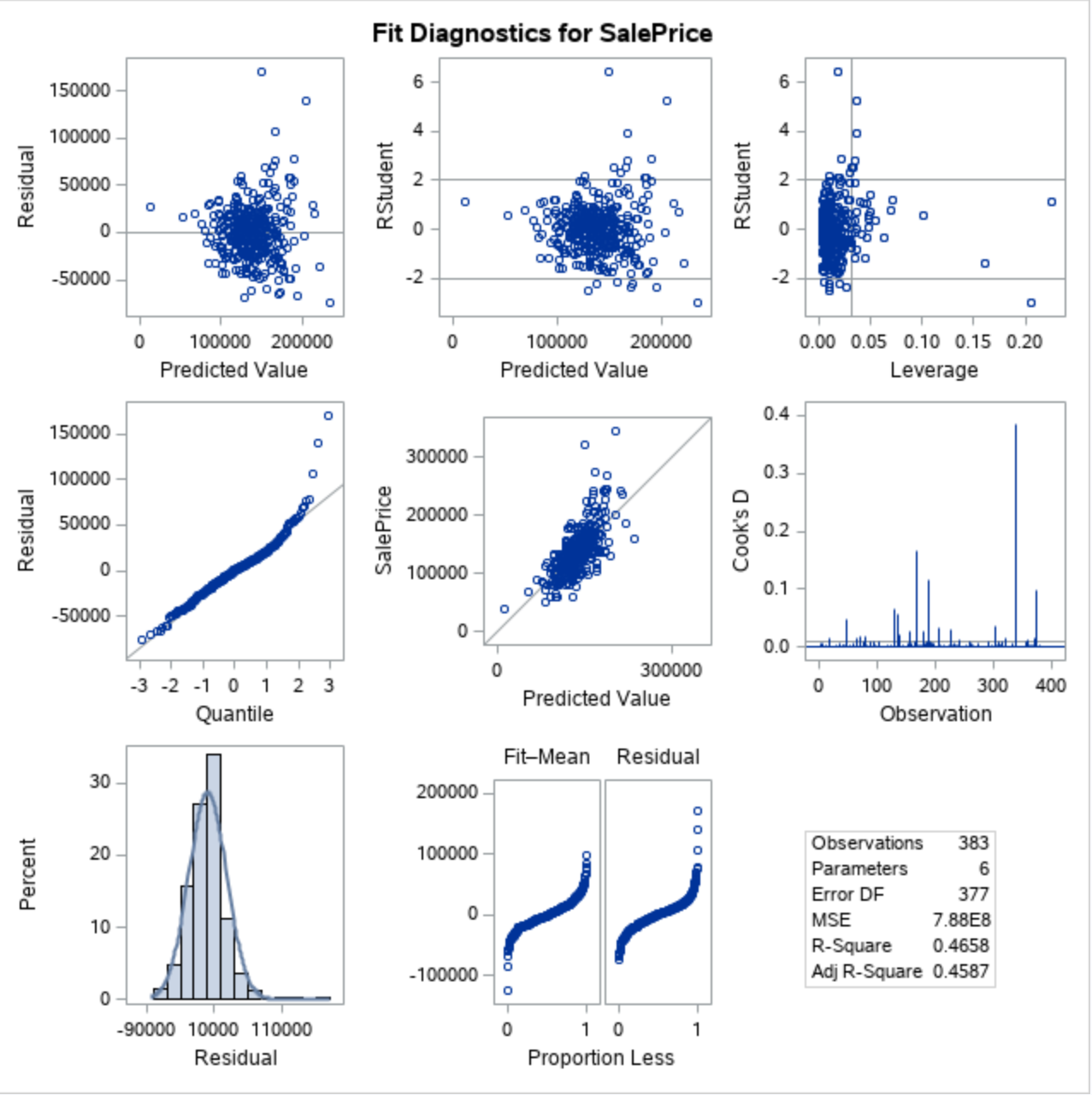


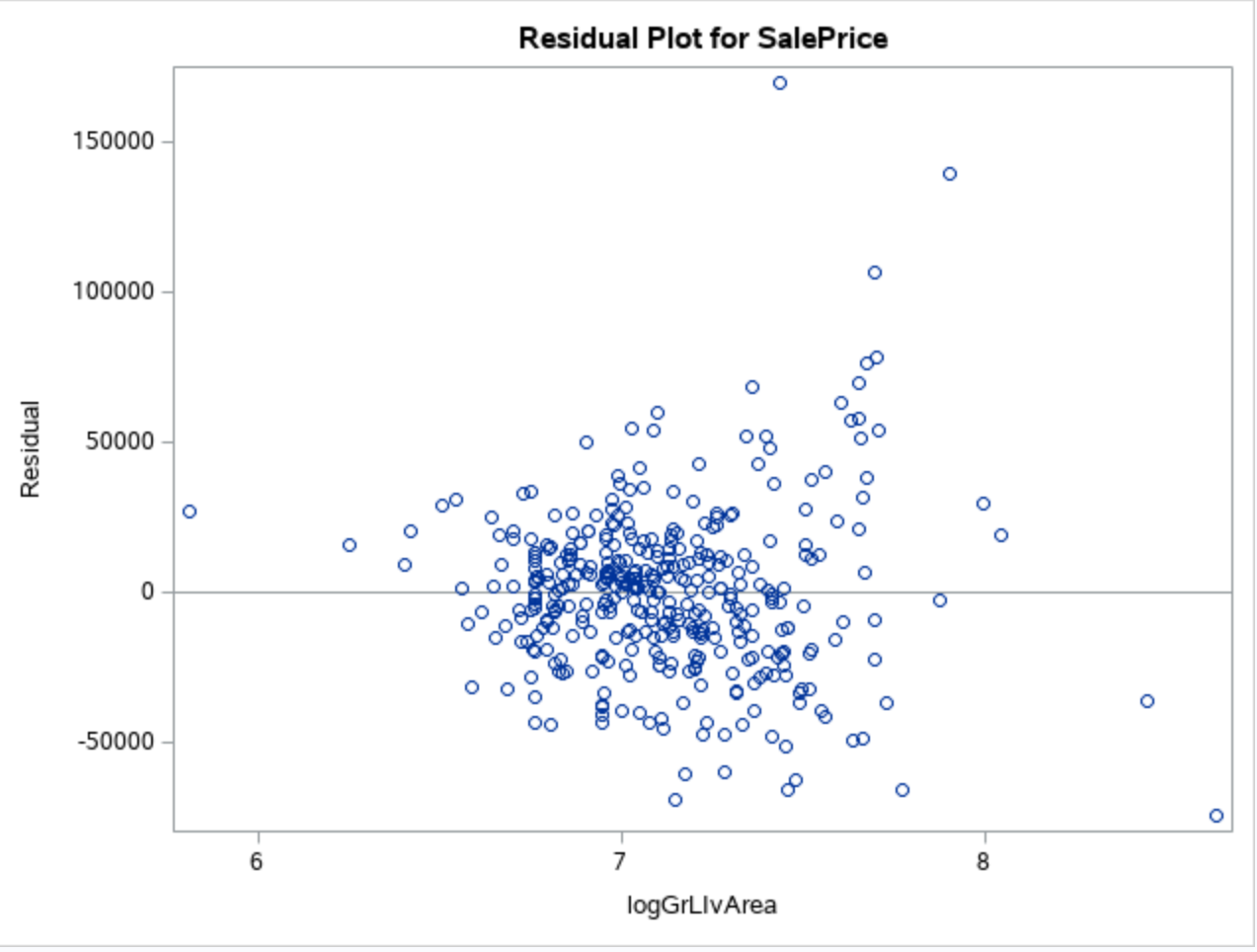
**Comparing Competing Models:**

Adj R2

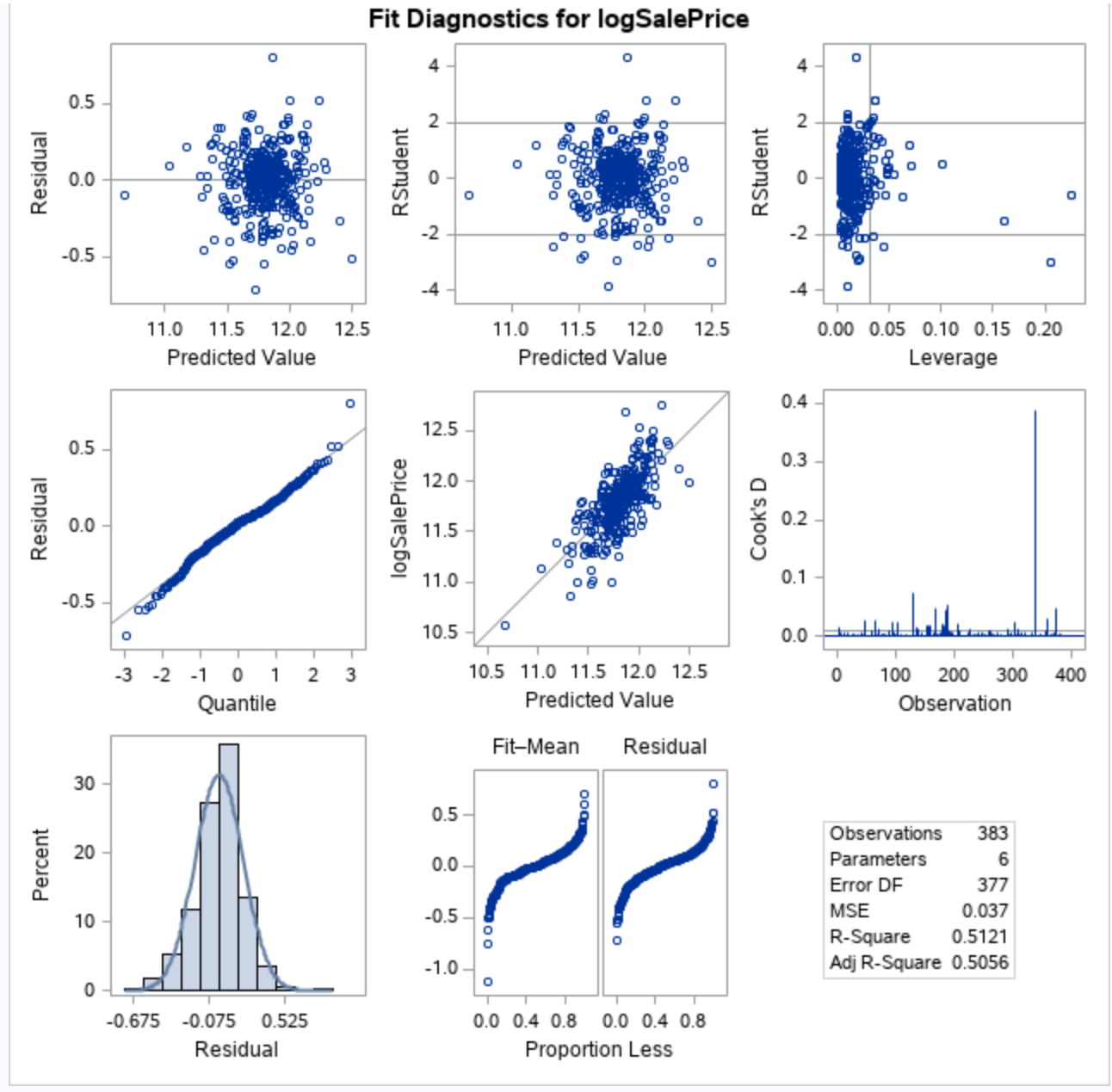
Internal CV Press

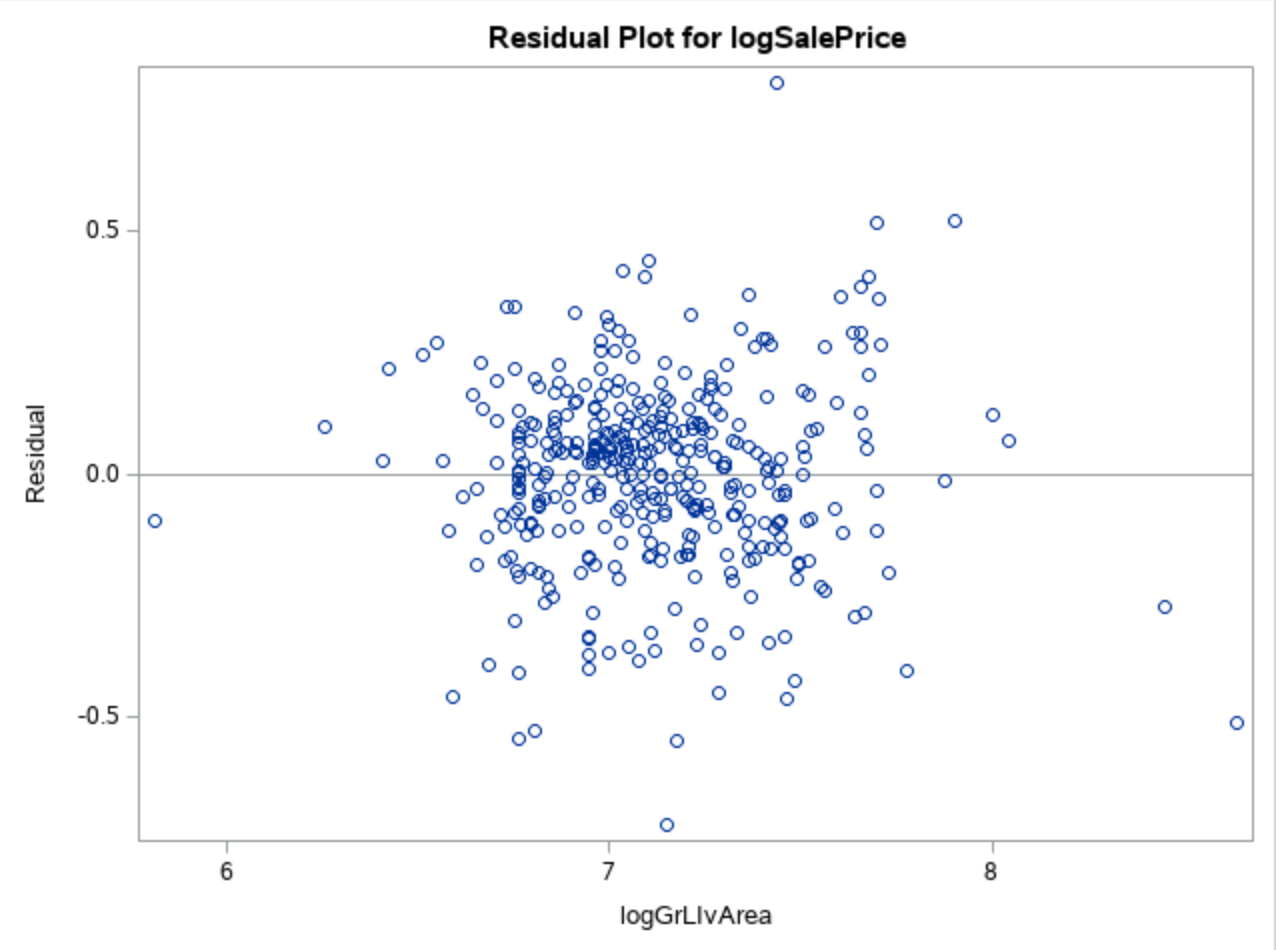
Linear – log





**Log log model**





**Checking Assumptions:**

* Residual Plot: The residual plot resembles somewhat of a random scatter of points around the 0 line, although there is a slight suspicion of non-constant variance judging from the dense cloud around.
* Studentized Residual Plot: This plot is very similar to the residual plot, although this plot identifies potential outlying observations. This plot identifies a potentially very outlying point with a predicted value of 15. This may provide some evidence against the normality assumption and this point should be examined further.
* Histogram of Residuals: The histogram of residuals displayed does not provide strong evidence that the residuals are not normally distributed.
* Q-Q Plot of Residuals: The Q-Q Plot of residuals provides no evidence against the residuals being normally distributed.
* The model is a reasonable fit with transformations, although transformations may be investigated to handle the possible problem with equal standard deviations.

**A discussion supporting the use of the model you chose (support that the assumptions are met).**

* Linearity: Met with original and log-log model. (view scatterplots)
* Normality: Log-log model looks slightly better. (view histograms)
* Equal standard deviations: Log-log model looks much better. (view residual scatter plots)
* Independence: We will assume independence, although not much is known about how these species were chosen.
* Outliers: There are some outliers at the tail end of the data set but the log log model looks better.
* We will proceed to make inferences on a log – log model.

Residual Plots

Influential point analysis (Cook’s D and Leverage)

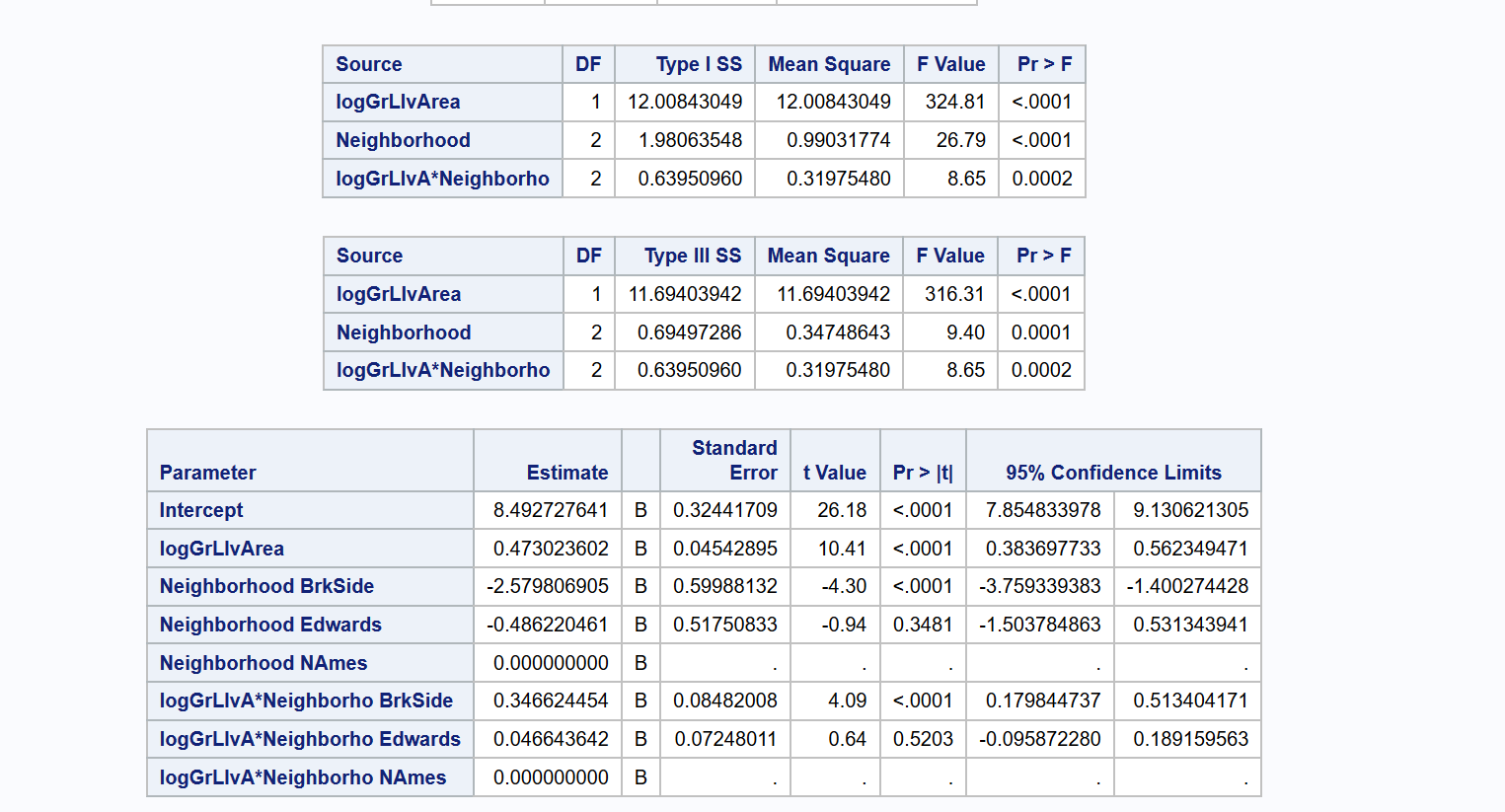
Observed significant Cook’s D observation on all models except the linear model which only has one spike at the tail end of the observations

Leverage seems normal for all the three models

**Parameters:**

Estimates

From the log – log model ;



Interpretation

Confidence Intervals

**Conclusion:**

A short summary of the analysis.

**Appendix:**

Code and additional screen shot of plots and graphs

proc import datafile='/folders/myfolders/testCleaned.csv'

DBMS=csv out=work.testCleaned replace;

run;

proc print data=work.testCleaned;

run;

proc import datafile='/folders/myfolders/train.csv'

DBMS=csv out=work.train replace;

run;

proc print data=work.train;

run;

\* Build a Model;

\*

\* Fit plot assesment on log log model ;

data logTrain;

set train;

if Neighborhood = 'BrkSide' OR Neighborhood ='NAmes' OR Neighborhood = 'Edwards';

logGrLIvArea= log(GrLIvArea);

logSalePrice = log(SalePrice);

run;

proc print data=logTrain;

run;

\*Linear - linear model ;

proc glm data = logtrain plot = all;

class neighborhood;

model SalePrice = GrLivArea |neighborhood / solution;

run;

\*log - log model ;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model logSalePrice = logGrLivArea |neighborhood / solution clparm;

run;

\* linear log model ;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model SalePrice = logGrLivArea |neighborhood / solution clparm;

run;

\*log - linear model;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model logSalePrice = GrLivArea |neighborhood / solution clparm;

run;

proc reg data = logTrain ;

model logSalePrice = logGrLivArea /VIF ;

run;

\*Coded scatter plot ;

proc sgplot data=train;

reg x=GrLivArea y=SalePrice / group=neighborhood clm cli;

proc sgplot data=train;

where neighborhood ne contains ('NAmes', 'Edwards' ,'BrkSide') ;

reg x=GrLivArea y=SalePrice / group=neighborhood clm cli;