Data Description

# Analysis Question 1:

## Question of Interest

How does Sale Price vary with Ground Living Area of homes in Iowa? Does neighborhood affect the Sale Price?

## Checking Assumptions

### Original Data

In the original data, constant variance is violated as seen in Code and Plots section “Original Diagnostics” Residual Plot. There are a few points whose residuals are visually far from rest. Furthermore, the quantile plot of residuals show a right skew. From a scatter plot there is a strong positive correlation between GLA and SalePrice.

### Log Transform

A log transform seemed appropriate due to the right skew QQ plot and many outliers from residual plot.

After log transform, qqplot is normally distributed and residual plot has no outliers. But the Cook’s D still shows an observation with potentially high influence.

### Influential Observation Removed

After removing Id 131 which had a Ground Living Area of 4676, Cook’s D of 5, we observe the following diagnostic plots in Code and Plots Section. There is still an observation 339 with Cook’s D of 2.5.

After removing Id 338 which had a Ground Living Area of \_, Cook’s D of 2.5, we observe the following diagnostic plots in Code and Plots Section.

Because we deleted the two observations who had Ground Living Area of 4676 and above , we must limit the conclusions of this report to houses with Ground Living Area below 4676 square feet.

## Parameters

Mean Log SalePrice = B0 (Brk Side) + B1\*GrLivArea + B2\* Names + B3\*Names\*GrLivArea + B4\*Edwards +

B5\*Edwards\*GrLivArea

### Estimates

Mean Log SalePrice = 10.79159+ 0.00073822\*GrLivArea + 0.65175\* Names + -0.00041410\*Names\*GrLivArea + 0.23389\*Edwards + -0.00019957\*Edwards\*GrLivArea

### Interpretation and Confidence Intervals

For houses in BrkSide, every 100 square feet increase in ground living area there will be a 7.7% increase in sale price (95% Confidence Interval: 6.3% to 9%). For houses in Names, 3.3% (95% CI 1.8% 4.8%).

For houses in Edwards, 5.4% (95% CI 3.7% to 7%). Houses with 0 ground living area in BrkSide are expected to have $48,533 Sale Price. Houses in Names are expect to have $1.9 more Sale Price at 0 Ground Living Area. Houses in Edwards are expected to have $1.3 more sale price at 0 ground living area.

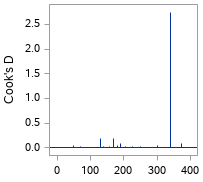
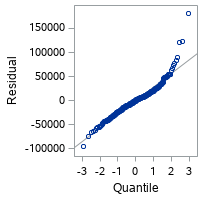
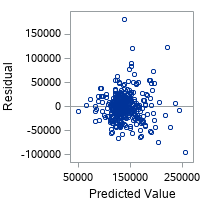
## Conclusion

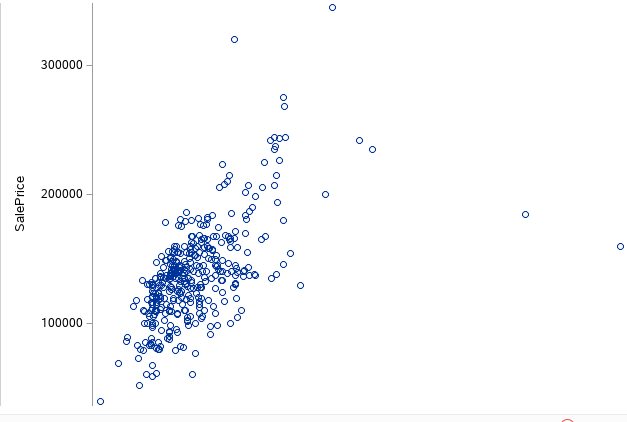
Increasing ground living area will have both a significant and practical effect on Sale Prices. In all 3 neighborhoods, increasing living area by 100 square feet will increase the Sale Price anywhere from 1.8% to 9% depending on the neighborhood. When a typical house sale price is in the $100,000 order of magnitude, then this percentage due to ground living area will be very practical.

Additionally, the increase in Sale Price due to Living Area differs between each neighborhood. Names has the lowest increase in Sale Price for every additional 100 square feet. Brk Side has the highest increase in Sale Price for every additional 100 square feet.

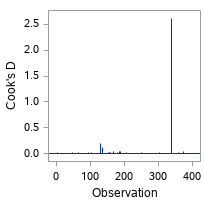
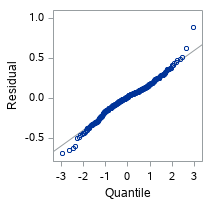
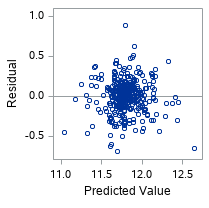
### Code and Plots

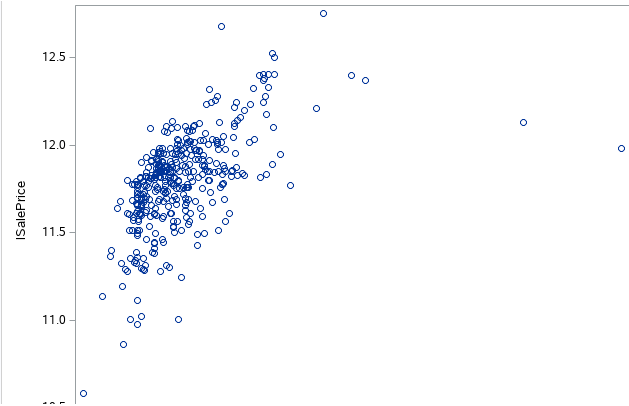
Original Diagnostics





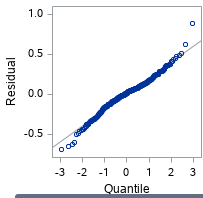
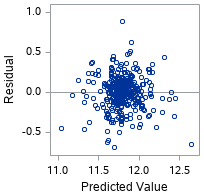
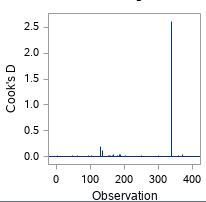
Transformed Diagnostics



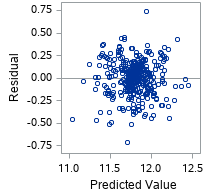
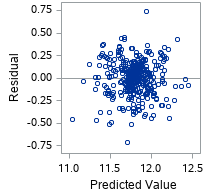
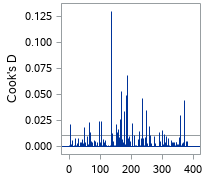


Diagnostics with Removed Influential Points

Observation 131 removed



Observation 131 and 339 removed



# Analysis Question 2

## Restatement of Problem

What is the Sale Price of a house based on any combination and transformation of the 81 attributes of the house?

## Model Selection

### Stepwise

### Checking Assumptions

#### Residual Plots

The residual plot shows a random cloud centered at 0 and constant variance. The residuals based on QQ plot look normally distributed with a few outliers on the left tail.

#### 

There do not seem to be points that have Cook’s D that are substantially higher than others. There are points with high leverage but their Cook’s D are not large.

### Forward

### Checking Assumptions

#### Residual Plots

Residual plots show random cloud centered at 0 with constant variance. Qq plot shows normally distributed residuals.

#### Influential point analysis (Cook’s D and Leverage)

After removing the house with Id of 411 and 1000 we see that there are evenly distributed Cook’s Ds.

### Backward

#### Residual Plots

From the residual plots we observe a random cloud centered at 0 with constant variance. From the qq plot we see that the residuals are normally distributed.

#### Influential point analysis (Cook’s D and Leverage)

There is one observation with a Cook’s D of around 0.125 which is above the 0.0027 threshold for Cook’s D using 4/number of observations. But there are no observations with both high leverage and high studentized residuals so we can leave it alone.

### Custom

## 

## Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Adj R2 | Internal CV Press | Kaggle Score |
| Stepwise | 0.9366 | 18.27958 |  |
| Forward | 0.9333 | 17.20477 |  |
| Backward | 0.9162 | 20.60715 |  |
| Custom |  |  |  |

Conclusion:

## Code and Plots

### Date Cleaning and Preparation

data train;

infile "/folders/myfolders/Kaggle/train.csv" dlm="," firstobs=2;

input Id MSSubClass MSZoning $ LotFrontage LotArea Street $ Alley $ LotShape $ LandContour $ Utilities $ LotConfig $ LandSlope $ Neighborhood $ Condition1 $ Condition2 $ BldgType $ HouseStyle $ OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle $ RoofMatl $ Exterior1st $ Exterior2nd $ MasVnrType $ MasVnrArea ExterQual $ ExterCond $ Foundation $ BsmtQual $ BsmtCond $ BsmtExposure $ BsmtFinType1 $ BsmtFinSF1 BsmtFinType2 $ BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating $ HeatingQC $ CentralAir $ Electrical $ X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual $ TotRmsAbvGrd Functional $ Fireplaces FireplaceQu $ GarageType $ GarageYrBlt GarageFinish $ GarageCars GarageArea GarageQual $ GarageCond $ PavedDrive $ WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC $ Fence $ MiscFeature $ MiscVal MoSold YrSold SaleType $ SaleCondition $ SalePrice ;

data test;

infile "/folders/myfolders/Kaggle/test.csv" dlm="," firstobs=2;

input Id MSSubClass MSZoning $ LotFrontage LotArea Street $ Alley $ LotShape $ LandContour $ Utilities $ LotConfig $ LandSlope $ Neighborhood $ Condition1 $ Condition2 $ BldgType $ HouseStyle $ OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle $ RoofMatl $ Exterior1st $ Exterior2nd $ MasVnrType $ MasVnrArea ExterQual $ ExterCond $ Foundation $ BsmtQual $ BsmtCond $ BsmtExposure $ BsmtFinType1 $ BsmtFinSF1 BsmtFinType2 $ BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating $ HeatingQC $ CentralAir $ Electrical $ X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual $ TotRmsAbvGrd Functional $ Fireplaces FireplaceQu $ GarageType $ GarageYrBlt GarageFinish $ GarageCars GarageArea GarageQual $ GarageCond $ PavedDrive $ WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC $ Fence $ MiscFeature $ MiscVal MoSold YrSold SaleType $ SaleCondition $ ;

/\*Prepare the train and test sets for prediction\*/

data test;

set test;

SalePrice=.;

;

data trainTest;

set train test;

run;

\*Find the columns with NAs or .;

proc format;

value $missfmt 'NA'='Missing' other='Not Missing';

value missfmt . ='Missing' other='Not Missing';

run;

proc freq data=trainTest;

format \_CHAR\_ $missfmt.; /\* apply format for the duration of this PROC \*/

tables \_CHAR\_ / missing missprint nocum nopercent;

format \_NUMERIC\_ missfmt.;

tables \_NUMERIC\_ / missing missprint nocum nopercent;

run;

/\* If BsmtCond is NA that means there is no basement. So put 0 square feet

for TotalBsmtSquareFeet. If MasVnrType is NA that means no Veneer so put 0 for MasVnrArea.

Etc... \*/

data trainTest;

set trainTest;

if MasVnrType ="NA" AND MasVnrArea =. THEN MasVnrArea=0;

if BsmtCond="NA" AND BsmtFinSF1=. THEN BsmtFinSF1=0;

if BsmtCond="NA" AND BsmtFinSF2=. THEN BsmtFinSF2=0;

if BsmtCond="NA" AND BsmtUnfSF=. THEN BsmtUnfSF=0;

if BsmtCond="NA" AND TotalBsmtSF=. THEN TotalBsmtSF=0;

if BsmtCond="NA" AND BsmtFullBath=. THEN BsmtFullBath=0;

if BsmtCond="NA" AND BsmtHalfBath=. THEN BsmtHalfBath=0;

if PoolQC="NA" AND PoolArea=. THEN PoolArea=0;

if GarageCond="NA" AND GarageArea=. THEN GarageArea=0;

if GarageCond="NA" AND GarageCars=. THEN GarageCars=0;

if FireplaceQu="NA" AND Fireplaces=. THEN Fireplaces=0;

/\*LotFrontage is simply missing values so impute the MEDIAN\*/

proc stdize data=trainTest out=trainTest missing=MEDIAN reponly;

var LotFrontage GarageYrBlt; /\* you can list multiple variables to impute \*/

/\*Find value of highest frequency for categorical variables with missing values\*/

/\*only for variables whose "NA" does not mean "None"\*/

PROC FREQ data=trainTest;

TABLES MSZoning Exterior1st Exterior2nd

Electrical KitchenQual Functional FireplaceQu

MiscFeature SaleType;

data trainTest;

set trainTest;

IF MSZoning = "NA" THEN MSZoning = "RL";

IF Utilities = "NA" THEN Utilities = "AllPub";

IF Exterior1st = "NA" THEN Exterior1st = "VinylSd";

IF Exterior2nd = "NA" THEN Exterior2nd = "VinylSd";

IF MasVnrType = "NA" THEN MasVnrType = "None";

IF Electrical = "NA" THEN Electrical = "SBrkr";

IF KitchenQual = "NA" THEN KitchenQual = "TA";

IF Functional = "NA" THEN Functional = "Typ";

IF SaleType = "NA" THEN SaleType = "WD";

/\*The only values with missing variables should be Alley MasVnrType BsmtQual BsmtExposure BsmtFintype

GarageTypeQualFinishCond PoolQC FireplaceQu Fence\*/

proc freq data=trainTest;

format \_CHAR\_ $missfmt.; /\* apply format for the duration of this PROC \*/

tables \_CHAR\_ / missing missprint nocum nopercent;

format \_NUMERIC\_ missfmt.;

tables \_NUMERIC\_ / missing missprint nocum nopercent;

run;

data trainTest;

set trainTest;

lSalePrice=log(SalePrice);

\*From question 1 there were influential points for GrLivArea;

data trainTest;

set trainTest;

if \_n\_ = 1299 then delete;

data trainTest;

set trainTest;

if \_n\_ = 524 then delete;

proc glm data=traintest;

where Neighborhood in ("NAmes","Edwards","BrkSide");

class Neighborhood;

model lsaleprice=grlivarea neighborhood;

### Stepwise

/\*trainTest is data set with log transform Sale Price and influential observation removed\*/

/\*finding the best subset of variables using stepwise\*/

proc glmselect data=trainTest plots=ALL;

class MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition;

model lSalePrice=MSSubClass LotFrontage LotArea OverallQual

OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1

BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF

LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces

GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF

EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal

MoSold YrSold MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition/selection=stepwise;

/\*creating model with the best selected variables\*/

proc glm data=trainTest plots=all;

class Mszoning Street LotConfig

Neighborhood Condition1 BldgType Exterior1st

MasVnrType BsmtExposure Heating CentralAir

KitchenQual Functional GarageQual GarageCond

PoolQC SaleCondition ;

model lSalePrice=LotFrontage lLotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1

TotalBsmtSF GrLivArea BsmtFullBath FullBath HalfBath

KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF

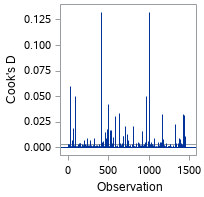
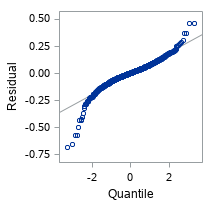
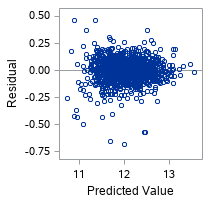
EnclosedPorch ScreenPorch PoolArea Mszoning Street LotConfig

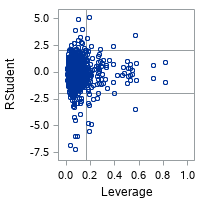
Neighborhood Condition1 BldgType Exterior1st

MasVnrType BsmtExposure Heating CentralAir

KitchenQual Functional GarageQual GarageCond

PoolQC SaleCondition ;





### Forward

/\* trainTest is dataset\*/

/\*Finding best subset of varialbes\*/

/\*\*/

proc glmselect data=trainTest plots=ALL;

class MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition;

model lSalePrice=MSSubClass LotFrontage LotArea OverallQual

OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1

BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF

LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces

GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF

EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal

MoSold YrSold MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition/selection=forward;

/\*dummy coding variables to use proc reg\*/

/\*Removing Houses with id 411 and 1000 because they are influential\*/

proc glmmod data=trainTest outdesign=GLMDesign outparm=GLMParm NOPRINT;

where Id not in (411,1000);

class Mszoning LotConfig

Neighborhood Condition1 Exterior1st

BsmtExposure HeatingQC CentralAir

KitchenQual Functional GarageQual

PoolQC SaleCondition ;

model lSalePrice=MSSubclass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1

TotalBsmtSF GrLivArea BsmtFullBath FullBath HalfBath

Fireplaces GarageCars WoodDeckSF OpenPorchSF

EnclosedPorch ScreenPorch Mszoning LotConfig

Neighborhood Condition1 Exterior1st

BsmtExposure HeatingQC CentralAir

KitchenQual Functional GarageQual

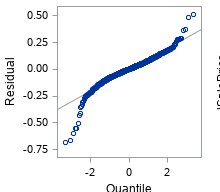
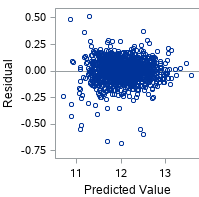
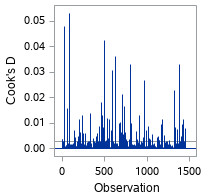
PoolQC SaleCondition;

/\*creating regression model from dummy coded varialbes\*/

proc reg data=GLMDesign;

DummyVars: model lSalePrice = COL2-COL118

/influence;



### Backward

\*Select best subset of variables from all;

/\*cleanTrainTranInf is the dataset after transformation and influence observation removal\*/

proc glmselect data=cleanTrainTranInf plots=ALL;

class MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition;

model lSalePrice=MSSubClass LotFrontage LotArea OverallQual

OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1

BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF

LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces

GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF

EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal

MoSold YrSold MSZoning Street Alley

LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1

Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond

Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive PoolQC Fence

MiscFeature SaleType SaleCondition/selection=backward;

/\*Dummy code the best categorical variables\*/

proc glmmod data=trainTest outdesign=GLMDesign outparm=GLMParm NOPRINT;

where Id ne 411;

class Mszoning Foundation

CentralAir

KitchenQual Functional

SaleCondition ;

model lSalePrice= OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2

BsmtUnfSF X1stFlrSF X2ndFlrSF LowQualFinSF

Fireplaces GarageCars WoodDeckSF ScreenPorch Mszoning Foundation

CentralAir

KitchenQual Functional

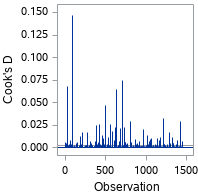
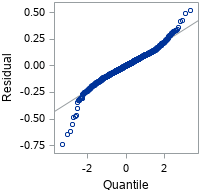
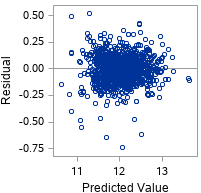
SaleCondition;

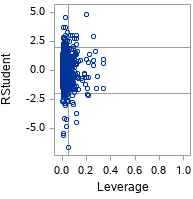
/\*Build model using the dummy coded categorical variables and best numerical variables\*/

proc reg data=GLMDesign;

DummyVars: model lSalePrice = COL2-COL45

/influence;





model lSalePrice=num5 num17 num28 num13 num6

num7 fac9 num10 fac43 fac24 fac12 fac31 num4 ;