

Kaggle Project



December 4, 2019

From: Billy Nayden, Muchigi Kimari, Vijayasrikanth Kanit & Paritosh Rai

**Introduction:** We are a team of data scientists from Southern Methodist University seeking to create models of housing data to answer potential questions from home buyers.

**Data Description:** This is a set of 1459 observations from 79 different data points regarding Ames housing, compiled by Dean De Cock for data science education.

**Analysis Question 1:**

**Restatement of Problem:** Century 21 Ames (a real estate company) in Ames Iowa only sells houses in the NAmes, Edwards and BrkSide neighborhoods. This analysis is to estimate the Sale Price of the house in relation to square footage of living area our client Century 21 Ames.

We identified four data points as outliers (**Table 1.3**) and removed from original data. The following is fit model based on the regression line with slope and intercept table (**Table 1.2**).

**Build and fit the model:** Predicted Sales Price = β0 + β1GeLivArea + β2 Neighborhood Edwards + β3 Neighborhood NAmes+ β4 GrLivArea\*Neighborhood Edwards+ β5 GrLivArea\*Neighborhood NAmes

**Fit model after removing outliers** (**Table 1.2**)**:** Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*Neighborhood Edwards + 60354.2\*Neighborhood NAmes – 17GrLivArea\*Neighborhood Edwards – 37.6\*GrLivArea\*Neighborhood NAmes

**Checking Assumptions:**

**Residual Spread (constant Variance):** (**Figure 1.1**)

* Even scatter with modified data
* No trends and no curvature
* Random and equal spread of residual data is indication of constant variance.
* There were few visible outliers that were removed in modified data.

The spread of the responses around the residual line is the same at all levels of the data.

**Influential point analysis (Cook’s D and Leverage):** (**Figure 1.2**) We identified four outliers based on scatter plot, residual plot and Cook’s D plot. Removed the outliers from the original data set and created new data set as modified data set for analysis.

**Assumptions:**

Linearity: (**Figure 1.3** and **Figure 1.4**) Identified few outliers in original data set. Based on multiple graphs, identified four outliers (**Table 1.3**) and removed them. The spread of the responses around the residual line improved, and is the same at all levels of the explanatory variable (Living Area). There is enough evidence from Q-Q plot and scatter plot of modified data to demonstrate the linearity of data. As the square footage of living area is increases, sale price also increases (**Figure 1.4**).

Normality: (**Figure 1.5**) There is enough evidence to suggest that the original data follows a normal distribution, as per the Q-Q Plot and histogram (**Figure 1.5**). Also, the data set is big enough to utilize Central Limit Theorem. The data becomes more linear and closer to trend line after removing the outliers in the scatter plot. Also, the right tail is reduced in histogram, and more aligned with normal curve.

Equal standard deviations: (**Figure 1.1**) The residual graph shows equal spread of data along the zero (0) line. This is enough evidence to suggest that the data has equal standard deviation. Four identified outliers, removed from the original data set, improved the standard errors for most coefficients.

Independence: We will assume independence, although not much is known about how these samples were chosen.

Outliers: (**Figure 1.1, Figure 1.2, Figure 1.3, Figure 1.4 and Figure 1.5**) Based on the residual plot, scatter plot, Cook’s D and Q-Q plot, we identified four outliers (**Table 1.3**). They were removed fromm the original data set, and we created new data set to be used for the analysis.

Based on assumption being met, we will proceed to make inferences on existing data.

**Comparing Competing Models:**

Model => SalePrice = GrLivArea TotalBsmtSF

Competing Model => SalePrice = GrLivArea TotalBsmtSF

**Adj R2  and CV Press:**

|  |  |  |
| --- | --- | --- |
| Model | Adj R2 | CV Press |
| Model (Outliers removed) | 44.90 % (Ref Table 1.4) | 2.6533E11 |
| Competing (Outliers removed) | 52.06% (Ref Table 1.5) | 2.35125E11 |

Model for Neighborhood Edwards:

(Neighborhood Edwards = 1 and Neighborhood NAmes = 0)

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*Neighborhood Edwards + 60354.2\*Neighborhood NAmes – 17GrLivArea\*Neighborhood Edwards – 37.6\*GrLivArea\*Neighborhood NAmes

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*(1) + 60354.2\*(0) – 17GrLivArea\*(1) – 37.6\*GrLivArea\*(0)

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 – 17GrLivArea

**Predicted Sales Price Edwards = 37100.42 + 70.16GrLivArea**

Model for Neighborhood NAmes:

(Neighborhood Edwards = 0 and Neighborhood NAmes = 1)

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*Neighborhood Edwards + 60354.2\*Neighborhood NAmes – 17GrLivArea\*Neighborhood Edwards – 37.6\*GrLivArea\*Neighborhood NAmes

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*(0) + 60354.2\*(1) – 17GrLivArea\*(0) – 37.6\*GrLivArea\*(1)

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 60354.2– 37.6GrLivArea

**Predicted Sales Price for NAmes = 80325.71 + 49.56\*GrLivArea**

Model for Neighborhood BrkSide:

(Neighborhood Edwards = 0 and Neighborhood NAmes = 0)

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*Neighborhood Edwards + 60354.2\*Neighborhood NAmes – 17GrLivArea\*Neighborhood Edwards – 37.6\*GrLivArea\*Neighborhood NAmes

Predicted Sales Price = 19971.51 + 87.16GrLivArea + 17128.91 \*(0) + 60354.2\*(0) – 17GrLivArea\*(0) – 37.6\*GrLivArea\*(0)

Predicted Sales Price = 19971.51 + 87.16GrLivArea

**Predicted Sales Price for BrkSide= 19971.51 + 87.16GrLivArea**

| **Parameter** | **Estimate** | **95% Conf Limits** | | **Interpretation** |
| --- | --- | --- | --- | --- |
| **Intercept** | 19971.51 | -1039.27 | 40982.30 | With zero living area predicted mean sale price will be ~$20,000, likely the cost of land. We are 95% confident the true mean value is between -$1K and $41K. |
| **GrLivArea** | 87.16 | 70.52 | 103.80 | With every increase in square foot of living area, predicted sales price increases by ~$87. We are 95% confident the true mean increase in sales price per square foot is between $70 and $103. |
| **Neighborhood Edwards** | 17128.91 | -10704.48 | 44962.30 | The predicted premium to buy homes in the Edwards neighborhood holding other variables constant is ~$17K. We are 95% confident the true sale price premium is between -$10K and $45K. |
| **Neighborhood NAmes** | 60354.20 | 36640.01 | 84068.38 | The predicted premium to buy homes in NAmes neighborhood holding other variables constant is ~$60K. We are 95% confident that the true sale price premium is between -$36K and $84K. |
| **Neighborhood BrkSide** | 0.0 | . | . | Using BrkSide as reference Neighborhood. |
| **GrLivArea\*Neighborho Edwards** | -17.00 | -38.74 | 4.73 | Predicted mean sale price in Edward neighborhood decreases by $17 per square foot holding other variables constant. We are 95% confident the true change in sale Price per square foot living area is between -$38 and $4. |
| **GrLivArea\*Neighborho NAmes** | -37.60 | -56.09 | -19.11 | Predicted mean sale price in NAmes neighborhood decreases by $37.60 per square foot holding other variables constant. We are 95% confident the true mean decrease is between -$56.07 and -$19.11. |
| **GrLivArea\*Neighborho BrkSide** | 0.0 | . | . | Using BrkSide as reference Neighborhood. |

**Conclusion:** The model with removed outliers represents the proportion of the variance for sales price that is explained by neighborhood and living area in a regression model is only 44% (Adjusted R2).

However, the competing model with basement area explained 52% of variance in sales price in a regression model.

We should consider building a stronger model that includes more than living area and basement area to get better predictions.

**Analysis Question 2:**

**Restatement of Problem:** Our client wants us to use the variables available to us in the data to build a model predicting sale price based on relevant variables.

**Checking Assumptions:**

**Residual Spread (constant Variance):** (**Figure 2.1**)

* Even scatter with modified data
* No trends and no curvature
* Random and equal spread of residual data is indication of constant variance.
* There were few visible outliers that were removed in modified data.

The spread of the responses around the residual line is the same at all levels of the data.

**Influential point analysis (Cook’s D and Leverage):** (**Figure 2.2, 2.3, 2.4**) We identified outliers based on scatter plot, residual plot and Cook’s D plot. Removed the outliers from the original data set and created new data set as modified data set for analysis.

**Assumptions:**

Linearity: (**Figure 2.2** and **Figure 2.3**) Identified few outliers in original data set. Based on multiple graphs, identified four outliers and removed them. The spread of the responses around the residual line improved, and is the same at all levels of the explanatory variable (Living Area). There is enough evidence from Q-Q plot and scatter plot of modified data to demonstrate the linearity of data.

Normality: (**Figure 2.2, 2.3, 2.4**) Judging from scatter plot, q-q plot, and histogram of residuals, there isn’t enough evidence against normality. Also, the data set is big enough to utilize Central Limit Theorem.

Equal standard deviations: (**Figure 2.2**) The residual graph shows equal spread of data along the zero (0) line. This is enough evidence to suggest that the data has equal standard deviation. Independence: We will assume independence, although not much is known about how these samples were chosen.

Outliers: Based on the left cooks D plot on the left, we Identified a few outliers. We later deleted them from the original data, and we created a new data set to be used for the analysis. (**Figure 2.4**)

Based on assumption being met, we will proceed to make inferences on existing data.

**Model Selection:**

We identified 13 data points as outliers via iterative testing and removed them from our model, in order to ensure higher accuracy.

**Forward selection:** We utilized the quantitative variables in the data set to produce a forward selection model with twelve parameters and an adjusted r-squared of 0.86.(**Table 2.1**)

**Backward selection:** We utilized the quantitative variables in the data set to produce a backward selection model with twelve parameters and an adjusted r-squared of 0.86. (**Table 2.2**)

**Stepwise section:** We utilized the quantitative variables in the data set to produce a stepwise selection model with twelve parameters and an adjusted r-squared of 0.86. (**Table 2.3**)

**Custom model:** We fine tuned the model using the twelve quantitative variables that came from the initial model investigations. We then used ANOVAs to determine which categorical variables had a significant effect on sales price. Utilizing the 40 categorical variables that had significant differences in mean sales price between levels, we created scales based on the median sale price of each and incorporated them into the model. The result is a model with 19 parameters and an adjusted r-squared of 0.88. (**Table 2.4**)

**Custom Model:**

*Sale Price* = -855,372 -91.46(*MSSubClass*) +0.69(*LotArea*) +11,169(*OverallQual*) +4,568.31(*OverallCond*) +261.57(*YearBuilt*) +23.69(*BsmtFinSF1*) +18.16(*TotalBsmtSF*) +62.49(*GrLivArea*) -9644.24(*BedroomAbvGr*) -21,889(*KitchenAbvGr*) +3,571.02(*TotRmsAbvGrd*) +29.82(*GarageArea*) +36,983(*LotConfig*) +45,763(*RoofStyle*) +12,482(*ExterQual*) +16,149(*KitchenQual*) +47,432(*Functional*) +26,058(*GarageFinish*) +25,345(*SaleCondition*)

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Estimate** | **Description** |
| Intercept | -855,372 | A house with no land, square footage or features would be worth -$855,372. |
| MSSubClass | -91.46 | Every increase in class corresponds to a decrease of $91.46 in value. |
| LotArea | 0.69 | Every increase in lot area corresponds to an increase of $0.69 in value. |
| OverallQual | 11,169 | Every increase in overall quality corresponds to an increase of $11,169 in value. |
| OverallCond | 4,568.31 | Every increase in overall condition corresponds to an increase of $4,568.31 in value. |
| YearBuilt | 261.57 | Every increase in year built corresponds to an increase of $261.57 in value. |
| BsmtFinSF1 | 23.69 | Every increase in finished basement square footage corresponds to an increase of  $23.69 in value. |
| TotalBsmtSF | 18.16 | Every increase in basement square footage corresponds to an increase of  $18.16 in value. |
| GrLivArea | 62.49 | Every increase in living area square footage corresponds to an increase of  $62.49 in value. |
| BedroomAbvGr | -9,644.24 | Every increase in bedrooms corresponds to an decrease of $9,644.24 in value. This offsets the total rooms parameter. |
| KitchenAbvGr | -21,889 | Every increase in bedrooms corresponds to an decrease of $21,889 in value. This  offsets the total rooms parameter. |
| TotRmsAbvGr | 3,571.02 | Every increase in total rooms corresponds to an increase of $3,571.02 in value. |
| GarageArea | 29.82 | Every increase in garage square footage corresponds to an increase of $29.82 in  value. |
| LotConfig | 36,893 | Every increase in lot configuration corresponds to an increase of $36,893 in  value. The levels are based off our custom scale of sale price medians. |
| RoofStyle | 45,763 | Every increase in roof style corresponds to an increase of $45,763 in value. The  levels are based off our custom scale of sale price medians. |
| ExterQual | 12,482 | Every increase in exterior quality corresponds to an increase of $12,482 in value.  The levels are based off our custom scale of sale price medians. |
| KitchenQual | 16,149 | Every increase in kitchen quality corresponds to an increase of $16,149 in value. The levels are based off our custom scale of sale price medians. |
| Functional | 47,432 | Every increase in functionality corresponds to an increase of $47,432 in value. The  levels are based off our custom scale of sale price medians. |
| GarageFinish | 26,058 | Every increase in garage interior finish to an increase of $26,058 in value. The levels  are based off our custom scale of sale price medians. |
| SaleCondition | 25,345 | Every increase in sale condition corresponds to an increase of $25,345 in value. The  levels are based off our custom scale of sale price medians. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** (Fig 2.7) |
| Forward (Table 2.1) | 0.86 | 1.333179E12 | 0.60337 |
| Backward (Table 2.2) | 0.86 | 1.354469E12 | 0.60048 |
| Stepwise (Table 2.3) | 0.86 | 1.353647E12 | 0.60238 |
| CUSTOM (Table 2.4) | 0.88 | 8.21749E11 | 0.21619 |

**Conclusion**

We ultimately chose the custom model because it had the highest adjusted r-squared and the lowest CV PRESS and Kaggle score. We believe this improvement is due to the addition of relevant categorical variables. Utilizing this model will be helpful for our client, as it can explain 88% of the variance in sale price across all neighborhoods in Ames, Iowa.

**Appendix**

**Code to plot Original data**

**PROC** **IMPORT** OUT= train

DATAFILE= "C:\Paritosh\SMU\6371 Statistical Foundation for Data Science\Project\train.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**run**;

/\* Print table

Proc print data = train;

run; \*/

**proc** **glm** data=WORK.train plots=all;

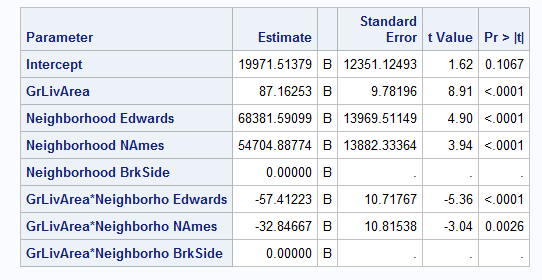
Where Neighborhood = "BrkSide" | Neighborhood = "NAmes" | Neighborhood = "Edwards";

class Neighborhood (REF="BrkSide");

model SalePrice = GrLivArea | Neighborhood/solution clm cli;

**run**;

Table: 1.1: Regression Line with Slope and Intercept table (Original Data)



Code to remove outliers

**proc** **glm** data=WORK.train plots=all;

Where Neighborhood = "BrkSide" | Neighborhood = "NAmes" & Id ne **643** | Neighborhood = "Edwards" & Id ne **524** & Id ne **725** & Id ne **1299** ;

class Neighborhood (REF="BrkSide");

model SalePrice = GrLivArea | Neighborhood/solution clm cli;

**run**;

Table 1.2: Regression Line with Slope and Intercept table (Mod data after removing four Outliers):

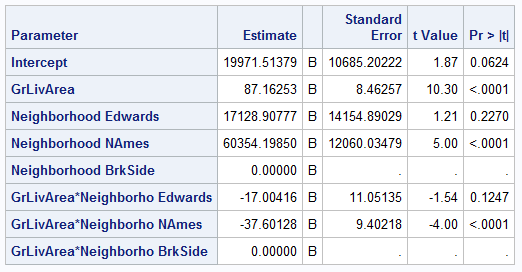


Table 1.3: List of data points removed:



Table 1.4 CV Press: Model with removed Outliers

/\* Calculating CV Press Model with Outlier Removed\*/

**proc** **glmselect** data = train plots (stepaxis = number)=(criterionpanel ASEPLOT);

Where Neighborhood = "BrkSide" | Neighborhood = "NAmes" & Id ne **643** | Neighborhood = "Edwards" & Id ne **524** & Id ne **725** & Id ne **1299** ;

model SalePrice = GrLivArea

/ selection = stepwise (select = cv choose = cv stop = cv) CVDETAILS;

**run**;

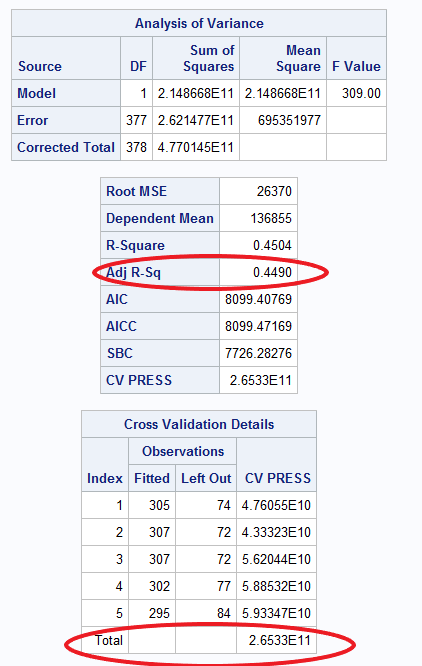


Table 1.5 CV Press: Competing Model with removed Outliers

/\* Calculating CV Press Competing Model with Outlier Removed\*/

**proc** **glmselect** data = train plots (stepaxis = number)=(criterionpanel ASEPLOT);

Where Neighborhood = "BrkSide" | Neighborhood = "NAmes" & Id ne **643** | Neighborhood = "Edwards" & Id ne **524** & Id ne **725** & Id ne **1299** ;

model SalePrice = GrLivArea TotalBsmtSF

/ selection = stepwise (select = cv choose = cv stop = cv) CVDETAILS;

**run**;

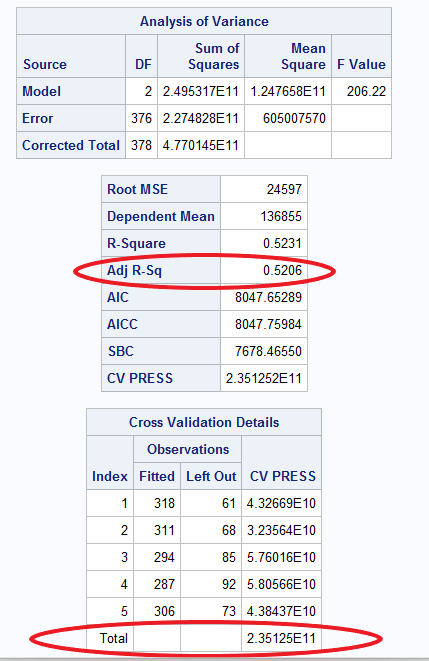


Fig 1.1: **Residual Spread (constant Variance):**

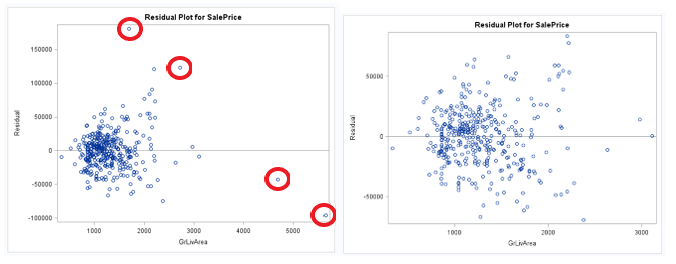


Fig 1.2: **Cook’s D:**

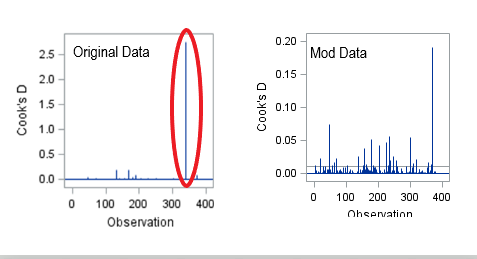
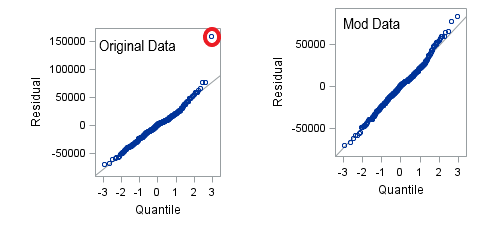
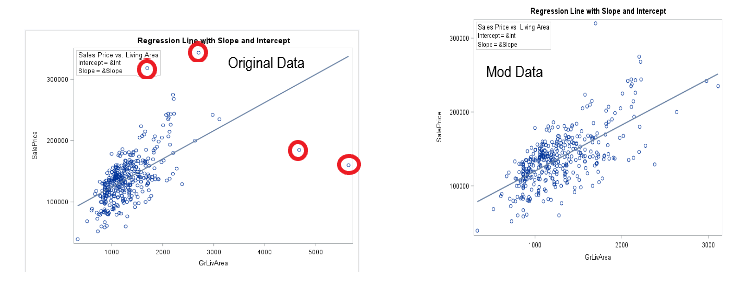


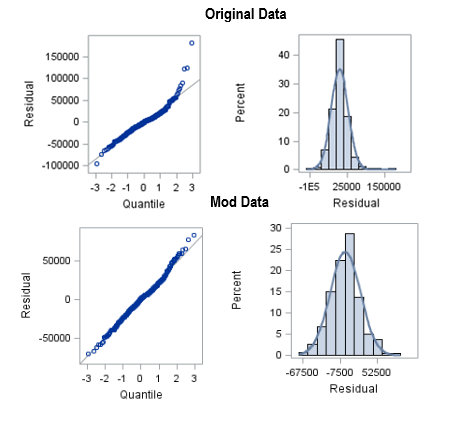
Fig 1.3: **Q-Q Plot for Linearity**



**Fig 1.4 Scatter Plot (Linearity)**



**Fig 1.5 : Q-Q Plot and Histogram with normality curve (Linearity)**



**Analysis 2:**

Fig 2.0Code to Import data SAS after cleaning up the headers. Header of fields should start with alphabet.

/\* Import data to SAS. Adjusted data field to start with alphabet vs. numeric \*/

**PROC** **IMPORT** OUT= train1

DATAFILE= "C:\Paritosh\SMU\6371 Statistical Foundation for Data Science\Project\train 2.0.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**run**;

/\* print train \*/

**proc** **print** data= Work.train1;

**run**;

/\* Remove Outliers from Data\*/

**data** train;

set Work.train1;

if \_n\_= **1299** then delete;

if \_n\_= **497** then delete;

if \_n\_= **739** then delete;

if \_n\_= **314** then delete;

if \_n\_= **334** then delete;

if \_n\_= **250** then delete;

if \_n\_= **693** then delete;

if \_n\_= **524** then delete;

**run**;

**Plot data**

**proc** **reg** data = train\_1 plots= all;

model Saleprice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearBuilt BedroomAbvGr OverallCond LotArea TotRmsAbvGrd WoodDeckSF ScreenPorch BsmtFullBath/vif;

**run**;

**Table 2.1:**

Code to run **Forward Selection Method** (Removed the outliers)

/\* Run code to for forward Predication Model\*/

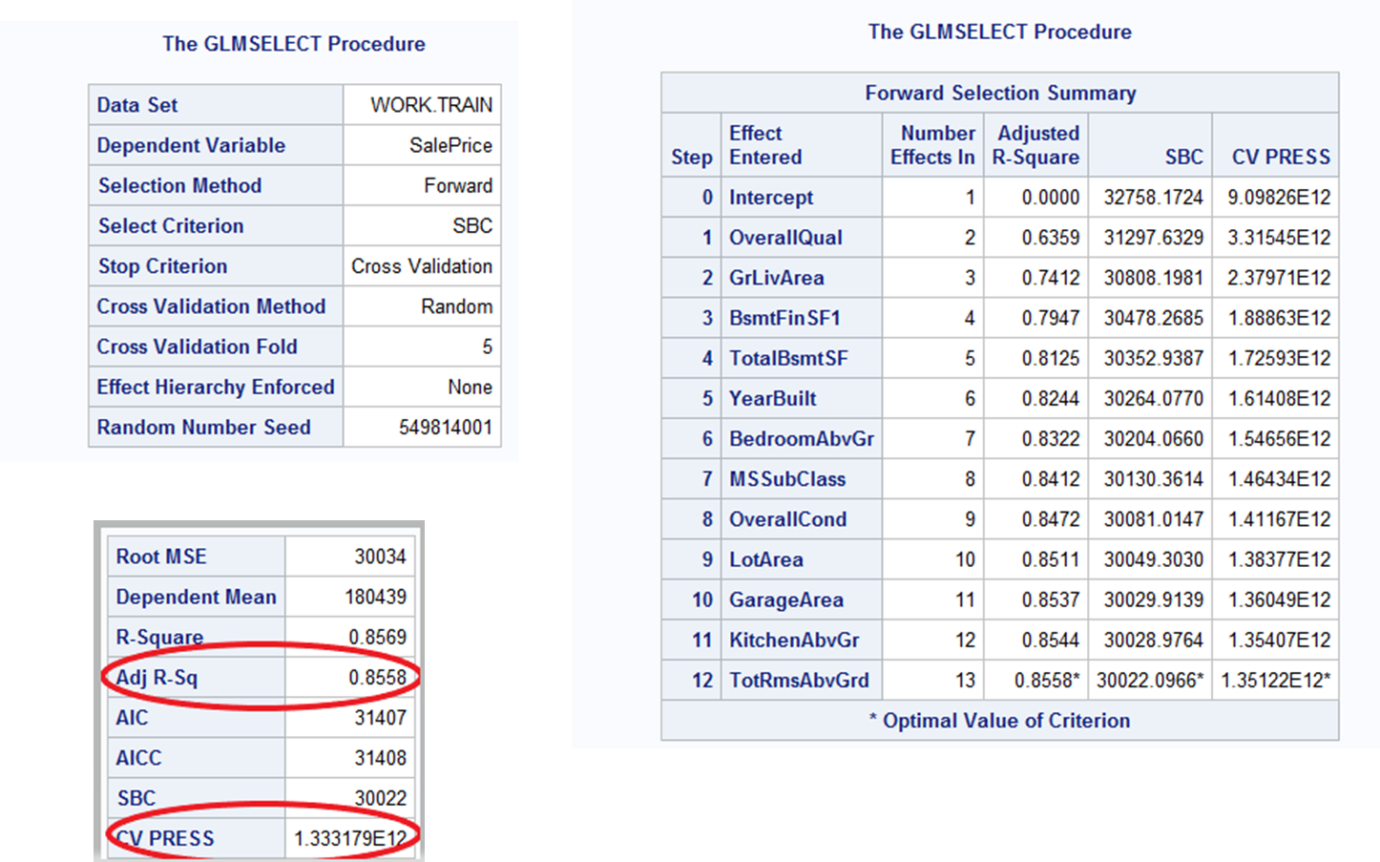
proc glmselect data = train;

model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold FstFlrSF SndFlrSF TSsnPorch

/ selection = forward (stop = cv) cvmethod = random(5) stats= adjrsq;

run;

Out Put of Forward Predicative Model code



**Table 2.2**

Code to run **Backward Selection Method** (Removed the outliers)

/\* Run code to for Backward Prediction Model\*/

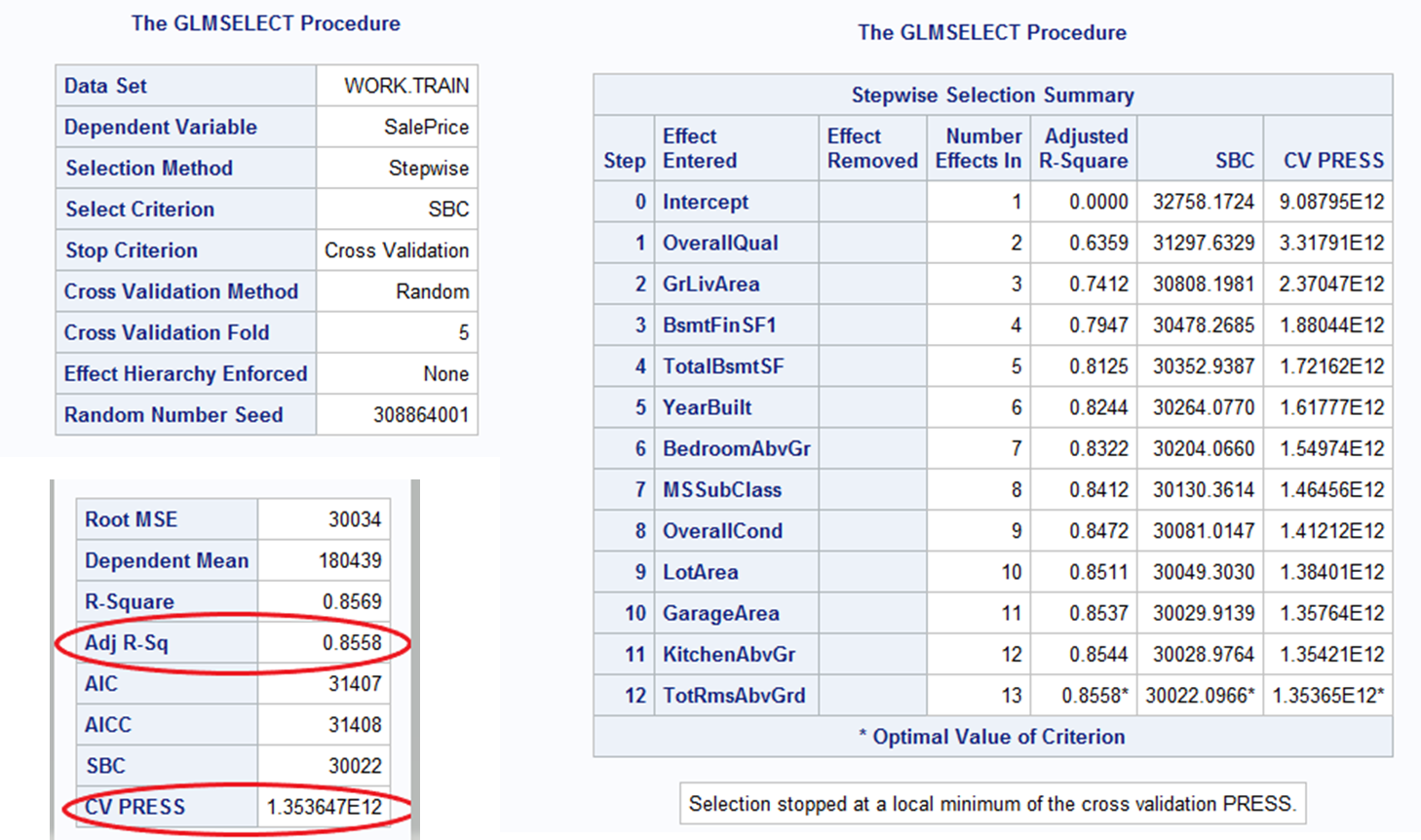
**proc** **glmselect** data = train;

model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold FstFlrSF SndFlrSF TSsnPorch

/ selection = backward (stop = cv) cvmethod = random(**5**) stats= adjrsq;

**run**;

**Out Put of Backward Selection Method**



**Table 2.3**

Code to run **Stepwise selection method** (Removed the outliers)

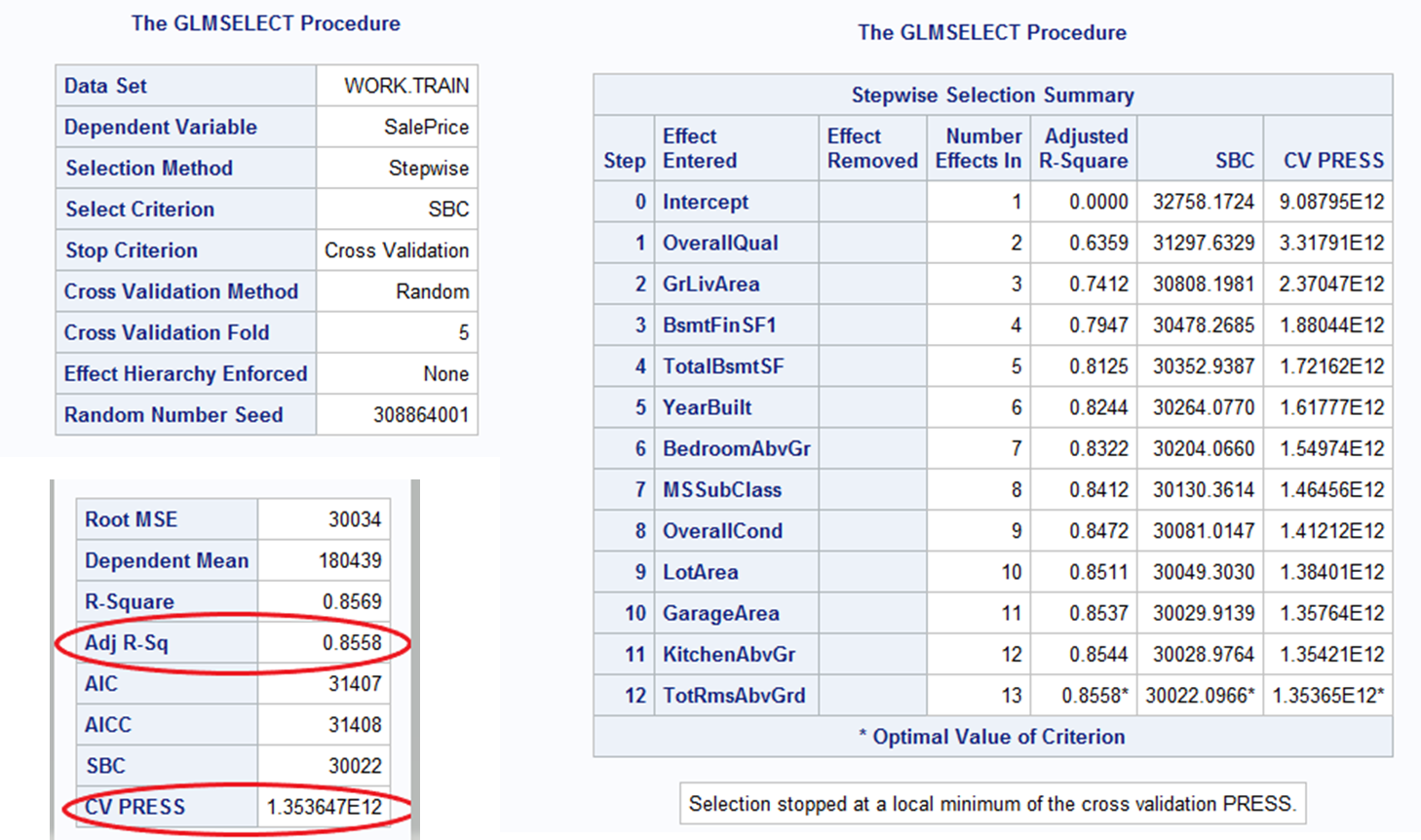
**proc** **glmselect** data = train;

model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold FstFlrSF SndFlrSF TSsnPorch

/ selection = stepwise (stop = cv) cvmethod = random(**5**) stats= adjrsq;

**run**;

**Output of Stepwise Selection Method**

****

**Table 2.4**

Code to run **stepwise selection method** for Custom Model (Removed the outliers)

/\* Run step model with quantification\*/

**proc** **glmselect** data =train;

model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 TotalBsmtSF GrLivArea BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageArea ScreenPorch LotShapeQ LotConfigQ NeighborhoodQ Condition1Q HouseStyleQ RoofStyleQ RoofMatlQ Exterior1stQ Exterior2ndQ MasVnrTypeQ ExterQualQ ExterCondQ FoundationQ HeatingQ HeatingQCQ KitchenQualQ FunctionalQ FireplaceQuQ GarageTypeQ GarageFinishQ GarageQualQ GarageCondQ PavedDriveQ PoolQCQ MiscFeatureQ SaleTypeQ SaleConditionQ

/ selection = stepwise (stop = cv) cvmethod = random(**5**) stats= adjrsq;

**run**;

**Output of Stepwise Selection Method**

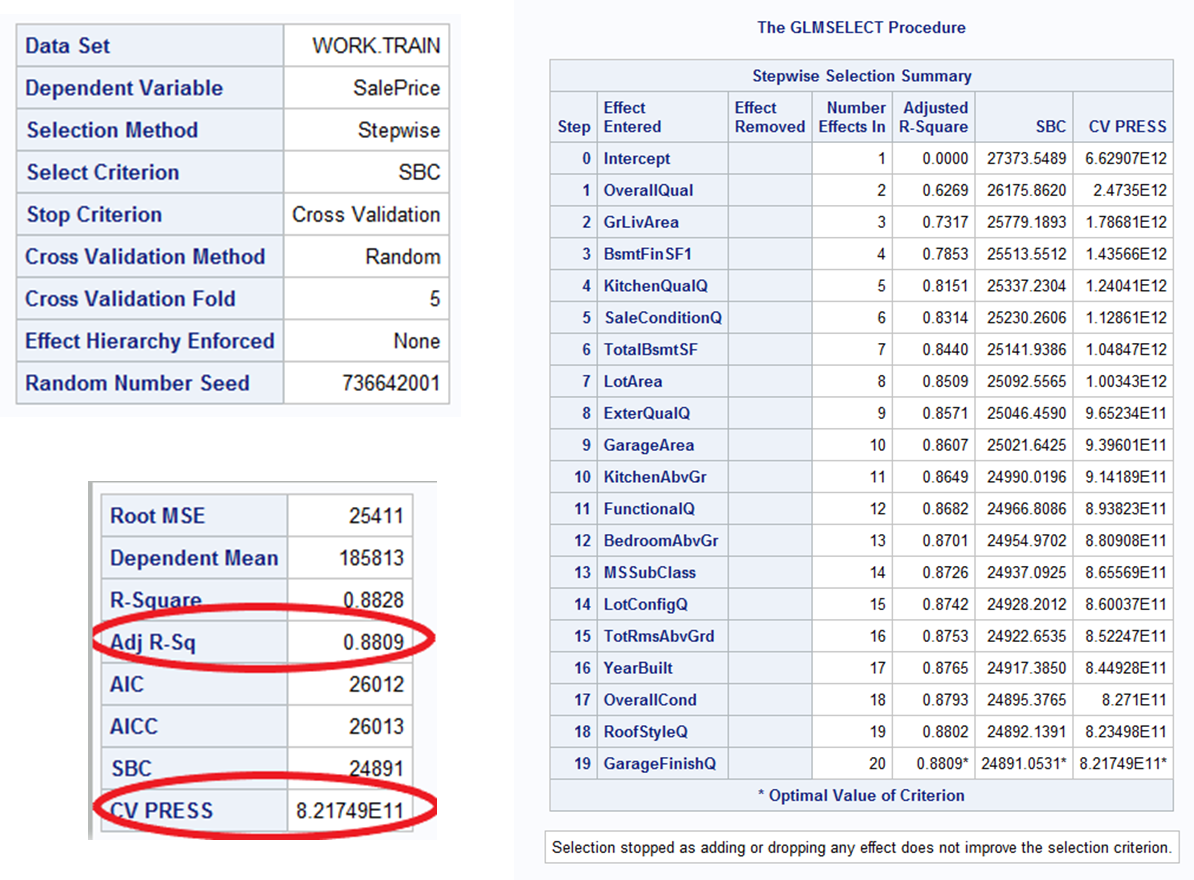
****

Fig 2.1: Assumption Graphs

Code:

/\* Plot data \*/

**proc** **reg** data = train plots= all;

model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 TotalBsmtSF GrLivArea BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageArea ScreenPorch LotShapeQ LotConfigQ NeighborhoodQ Condition1Q HouseStyleQ RoofStyleQ RoofMatlQ Exterior1stQ Exterior2ndQ MasVnrTypeQ ExterQualQ ExterCondQ FoundationQ HeatingQ HeatingQCQ KitchenQualQ FunctionalQ FireplaceQuQ GarageTypeQ GarageFinishQ GarageQualQ GarageCondQ PavedDriveQ PoolQCQ MiscFeatureQ SaleTypeQ SaleConditionQ

/vif;

**run**;

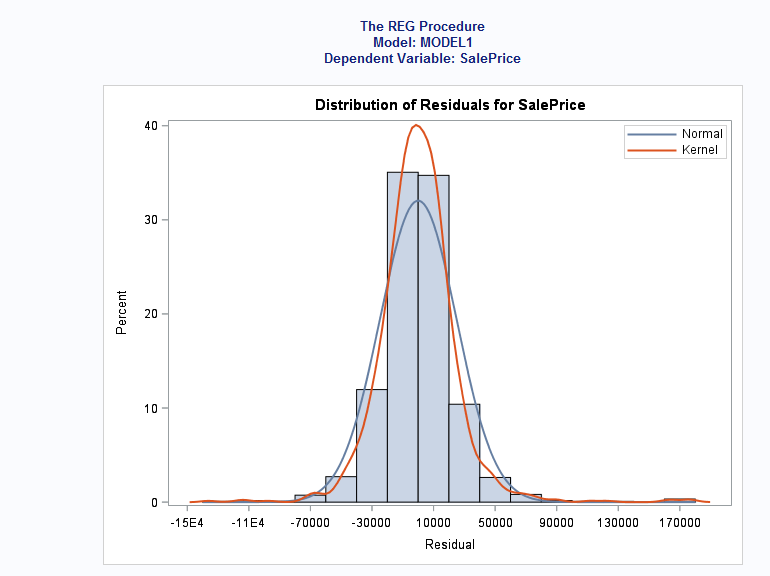


Fig 2.2:

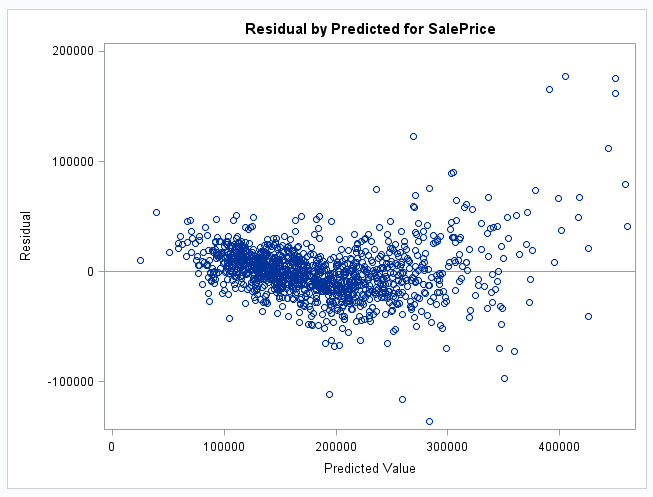


Fig 2.3:

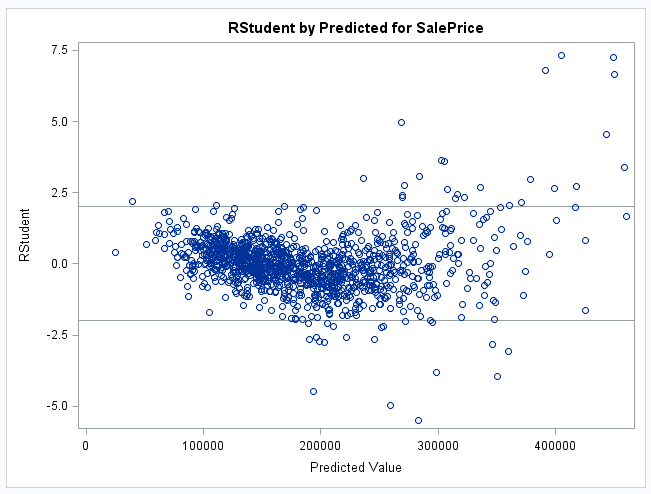


Fig 2.4

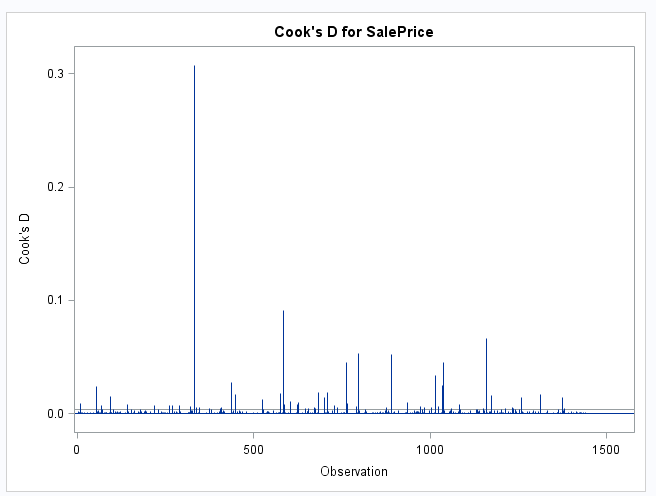
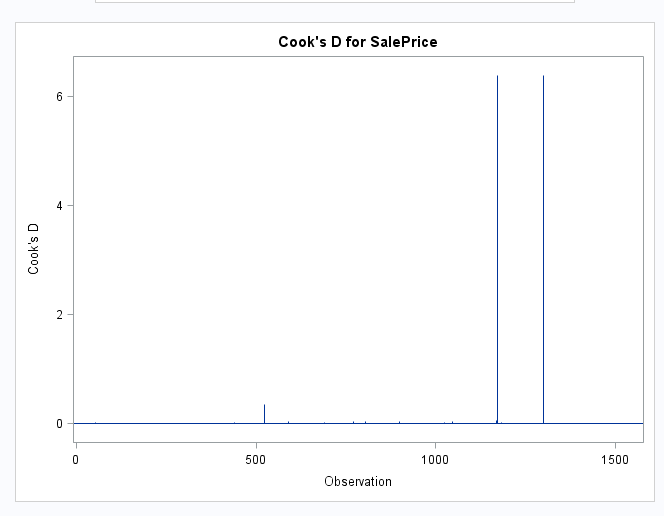


Fig 2.5

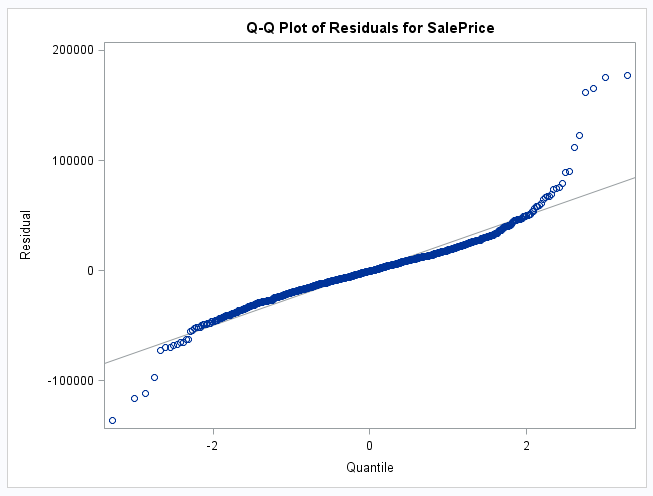


Fig 2.6

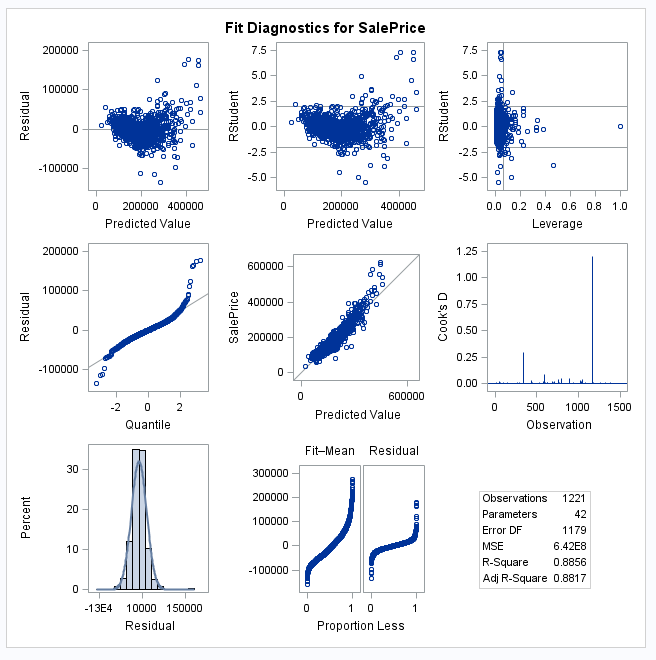


Fig 2.7 Kaggle score

