**II. PRELIMINNARY MULTI-LINEAR REGRESION MODEL ANALYSIS:**

The general form of fitted regression line is given as:

Yi = β0 + βXi1 + β2Xi2 + β3Xi3 + β4Xi4 + εi

Where,

Yi = House Price in US dollars (Response Variable)

Xi1 = Total Living Area in sq. feet (Predictor Variable 1)

Xi2 = Living Area Above Base in sq. feet (Predictor Variable 2)

Xi3 = Basement Area in sq. feet (Predictor Variable 3)

Xi4 = Grade (Predictor Variable 4)

εi = Random Error

2.1 Analysis of variance and estimation of parameters using SAS:

A screenshot of a social media post

Description automatically generated

After analyzing the SAS output with the house price as a response and 4 predictors mentioned above Preliminary MLR model is given as:

 **Ŷ (House price in USD) = -484438 + 215.02003\*(Total Living Area in ft2) - 97.62612\*(Living Area Above Basement in ft2) + 90.85933\*(Basement Area in ft2) + 72886\*(Grade) + εi**

**2.2 MODEL ASSUMPTIONS:**

1. **The Model form is Reasonable:** This first assumption is made by plotting residuals Vs predictor plot for all the four predictors. There is not curvature in any of the graphs below hence we can say that our MLR model is reasonable for further analysis.

A screenshot of a social media post

Description automatically generated

To support this assumption further matrix scatter graphs is plotted in SAS. It can be seen that there is high correlation between response variable and predictor variables.

A screenshot of a cell phone

Description automatically generated

1. **The constant variance in residual:** This is concluded on the preliminary basis with the fact that there is no funnel shape present in the residuals vs predicted value plot. We are considering that there are no outliers as per assumption 5. The graph of residuals vs predicted valur is shown below,

A close up of a map

Description automatically generated

To make sure that the variance is constant modified levene test is conducted. The result of the test is in the next section of this report.

1. **Residuals are normally distributed:** Normal probability plot for preliminary model follows straight line pattern except few outliers residual which can be ignored. It gives preliminary impression that the residuals are normally distributed. The normality test is done to confirm this assumption.

A close up of a map

Description automatically generated

1. **The residuals are not correlated:** (INSERT TIME PLOT)
2. **There no outlier in the model:** Few outliers are present. But it is not significantly influencing the model. The outliers are neglected for the analysis purpose. We are assuming that no outliers are present.
3. **No strong association between the predictors:** The predictors are not associated with each other that means change in any one predictor doesn’t really affect the other predictors. Hence this assumption is valid too.

**2.6 ANOVA:**

A screenshot of a cell phone

Description automatically generated

**SSR:** SSR is Sum of Squares for Regression or regression sum of squares. From the ANOVA table above, the value of SSR is 6.229293E13 with 4 dof.

page13image26535488

**SSE:** SSE is the sum of squares for error or error sum of squares. From the ANOVA table above, the value of SSE is 5.279064E13 with 4 dof.

**SSTo:** Total sum of squares is the addition of SSE and SSR. Here SSTo is 1.150836E14

**MSE:** Mean sum of errors is 53109292111.

**MSR:** Mean sum for regression is 1.557323E13.

page13image26733056

**Coefficient of Determination And F- Value:**  
The coefficient of determination is the R square value which measures how well the regression model fits in the data. The range lies between 0 to 1 where 1 explains the 100% variability explained by predictors for the expected response. R square is calculated as,

R2= MSR/MSE.

In our model the R2 value is 0.5413. There we can conclude that 54.13% of the total variability in house price is explained by the 4 predictors.

**Adjusted R2:**

From the ANOVA table adj R2 is 0.5394. This value is useful to understand the significance of the predictors taken into consideration. The difference between the R2 and adjusted R2 indicates the significance of predictor.

**2.6.4 significance of predictors:** The significance of individual parameter can be decided by comparing their p values with significance level which is 0.1 (90% confidence). From the parameter estimate table given below the p-value is 0.0001 for each parameter which is far lesser than 0.1. Therefore, we can say that each predictor is significant to analyze this MLR model.

A screenshot of a cell phone

Description automatically generated

**2.6.5 Significance test:**

To check the significance of the model we will conduct this test. Comparing the P-value of the model with the level of significance which is 0.1.

Hypothesis Test:

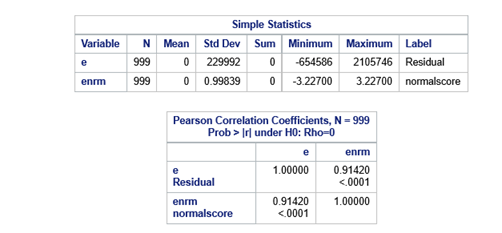
H0: Yi = ẞ0 + Ԑi (Reduced model)

H1: Yi = ẞo + ẞ1xi1 + ẞ2xi2 + ẞ3xi3 + ẞ4xi4 + ẞ5xi5 + Ԑi (Full model)

Decision statement: If p< α, reject H0

Conclusion: From ANOVA table, the value of p is 0.0001 which is lesser than the significance level of 0.1. Hence, we can conclude that our model is significant. Here we are 90% confident that our model is significant for analysis.

**NORMALITY TEST:**



1. **Hypothesis:** H0:Normality is OK.

H1:Normality is violated.

# Ø Test statistics:

From the SAS output, we found out the ρ-value =0.97645(from table Pearson Correlation Coefficient) Assuming α=0.05, n=40.

Cut Off (α, n) =C (0.10,999) = 0.989(from table B.4),

Ρ<C.

1. **Decision**: Reject H0.

we reject H0, we conclude that Normality is Not satisfied.

**OUTLIER TEST**

Based on our scatter plots we can assume that our data has a linear relationship. We observe many outliers but we actually don’t know what outliers they are. The outliers can be Influencer or leverage. Outliers are just the extreme observations found in the model and they are detected by the **studentized residual**. There are many types of studentized residuals such as regular and deleted studentized etc.

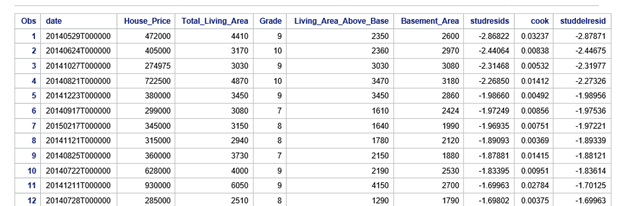
This plot shows how they take residuals and they divide with standard deviation and depicts who much standard deviation away.



Anything that is in red are away from 3 standard deviation and can be stated as outlier. This will be more comfortable when you see from the table.

**INFLUENCE POINTS**

Influence points are any just observation which has a significant impact on regression coefficient. All outliers may or may not be a influence points. These can be calculated by **cooks distance**. The cut off is given by d>=4/n=0.004. Anything above this cut off values are called influence points.



**LEVERAGE POINTS**

Leverage points are the observations which are far away from the line but these points does not have any significant impact on regression coefficients. But the leverage points affect other statistics such as R^2 and standard error of coefficients.

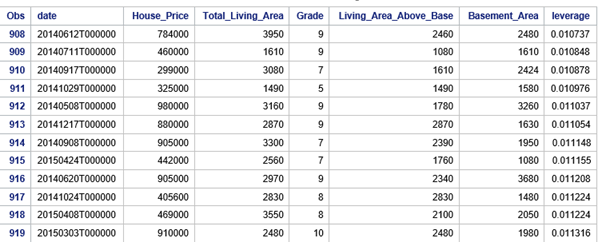
Cutt of is 2(K+1)/n

Kàno of predictors

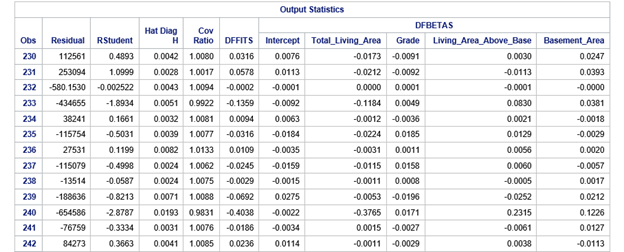
nàno of observations

2(4+1)/999=0.01001001.

Anything above this cut off value is said to be leverage points

 .

**OTHER WAYS**



**DFFITS**

The leverage can also be calculated with other ways by using DFFITS value and its cut off points.

Cut of point is 2(p/n)^0.5=0.126554. Anything above this value is not a influential points.

Where pàno of predictors

nàno of observations

**DFBETAS**

The cut-off value is calculated by (2/√𝑛); (Nà No of observations), and the value is 0.063277. This value is compared with the table value and if the cut-off value is higher, then it is considered as an influential outlier.

**SECTION 3**

**Exploration of Interaction Terms**

**3.1** Explore interactions using partial regression plots.

Discuss the addition of possibly useful interaction terms.

Check correlations involving the added interaction terms before and after standardization.

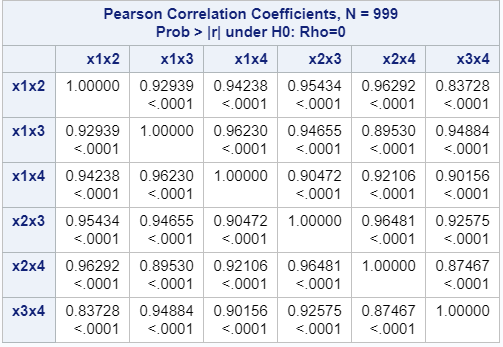
**3.1 Partial Interaction Plots**



Fig. X – Residual vs. Interaction Plots

Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model. Interaction terms are obtained by taking the cross product of two independent variables. The interactions used for this analysis were evaluated before and after standardization. Standardizing the variables is obtained by centering the mean to zero and sealing the variance to 1. Sealing is important for numerical stability and permits direct comparisons for LSEs. It can be inferred from the plots above that the interaction terms most likely would not add any significant impact on the regression model. This is because a linear relationship between the residuals and the interaction terms does not exist. Therefore, it is not useful to add any of these interaction terms to the regression model.

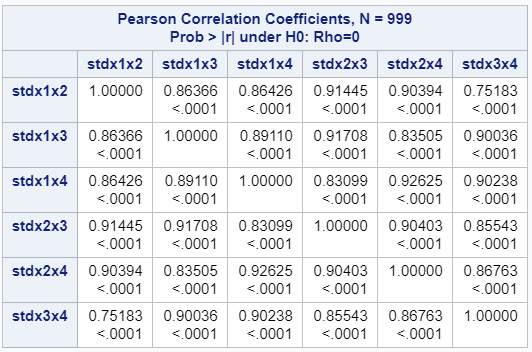
**Correlation Matrix before Standardization**



**Fig. X** - Correlation Matrix before Standardization

Fig. X shows the correlation matrix before standardization. It can be seen that all of the values are avove the threshold, 0.7. This gives an indication that there might be serious multicollinearity issues. For this reason, these terms were standardized and shown in Fig. X.

**Correlation Matrix after Standardization**



**Fig. X** - Correlation Matrix after Standardization

**SECTION 4**

**Model Search**

Table X: Best Subsets Regression Output

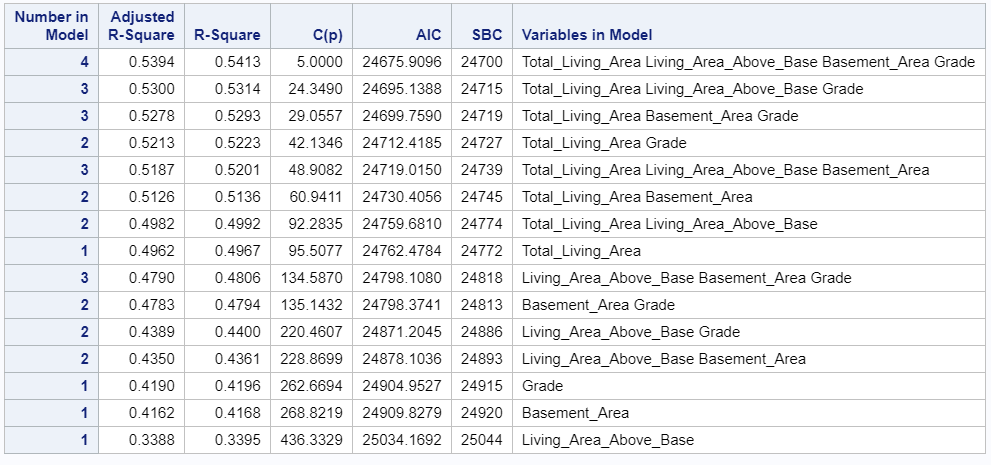


Table X – Stepwise Regression

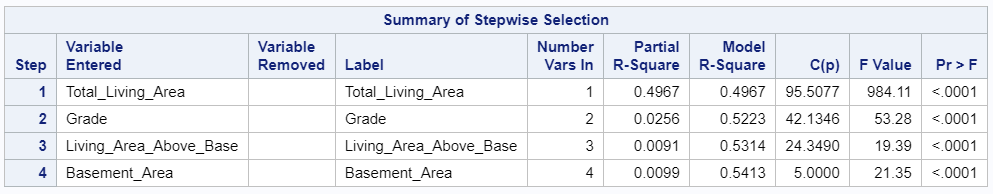


Table X: Backward Deletion Regression Model

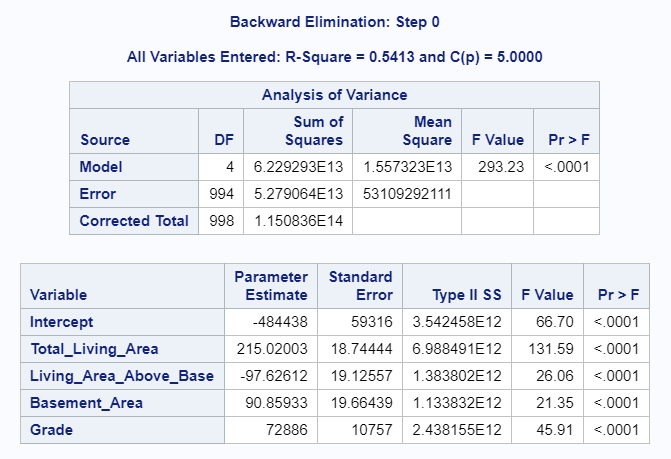


Table X – VIF Values of Model 1

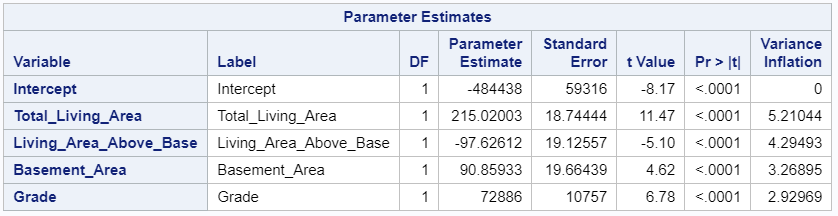


Table X – Correlation Matrix of Model 1

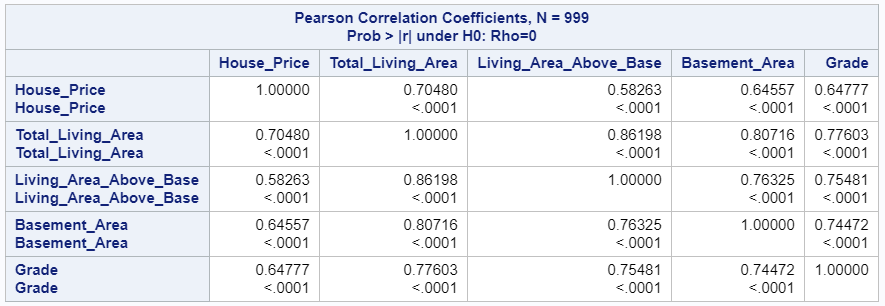


Table X – VIF Values of Model 2

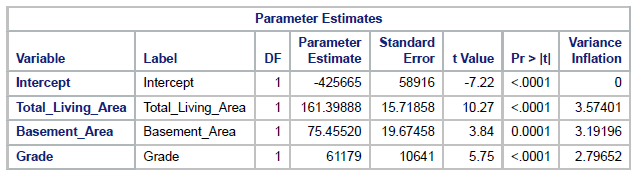
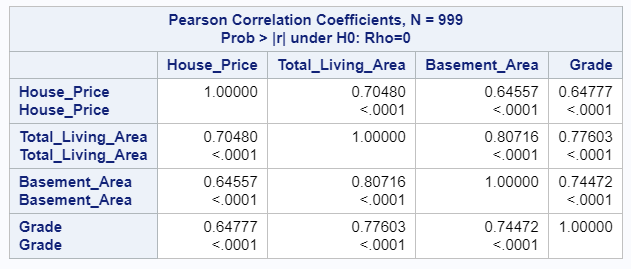


Table X – Correlation Matrix of Model 2



**Best Model 1** : house price = -484438 +215.02\*(total living area) - 97.63\*(living area above base) + 90.86\*(basement area) + 72886\*(grade)

**Best Model 2** = house price = -425665 +161.399\*(total living area) + 75.46\*(basement area) + 61179\*(grade)

**Section 5**

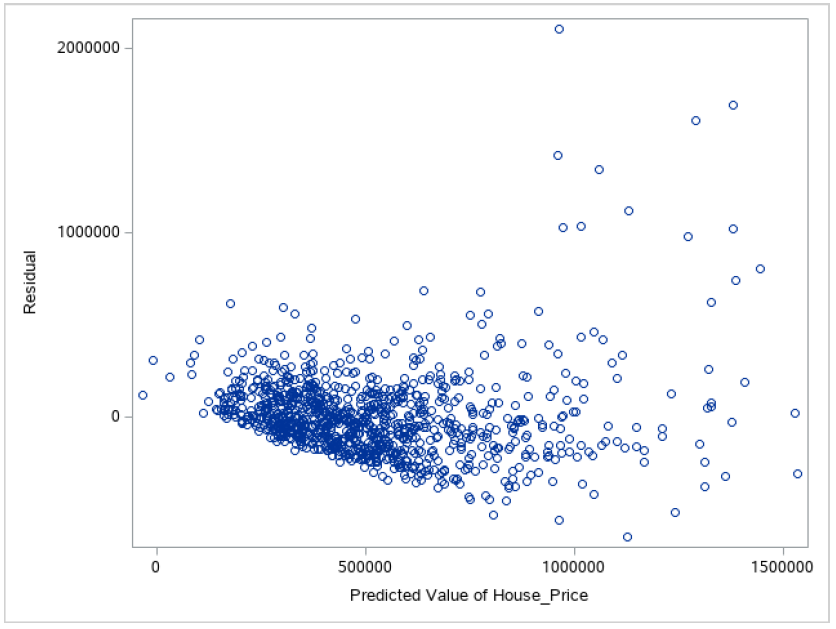
**Model Selection**

Model 1 – Assumption Verification

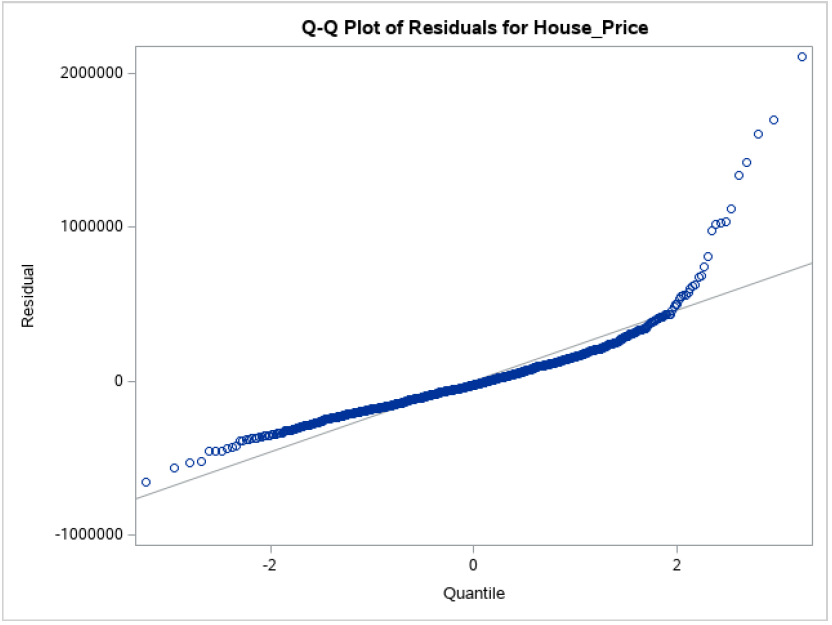
1. Model Form (no curvature)



1. Check for Constant Variance



1. Check for Normally Distributed Data

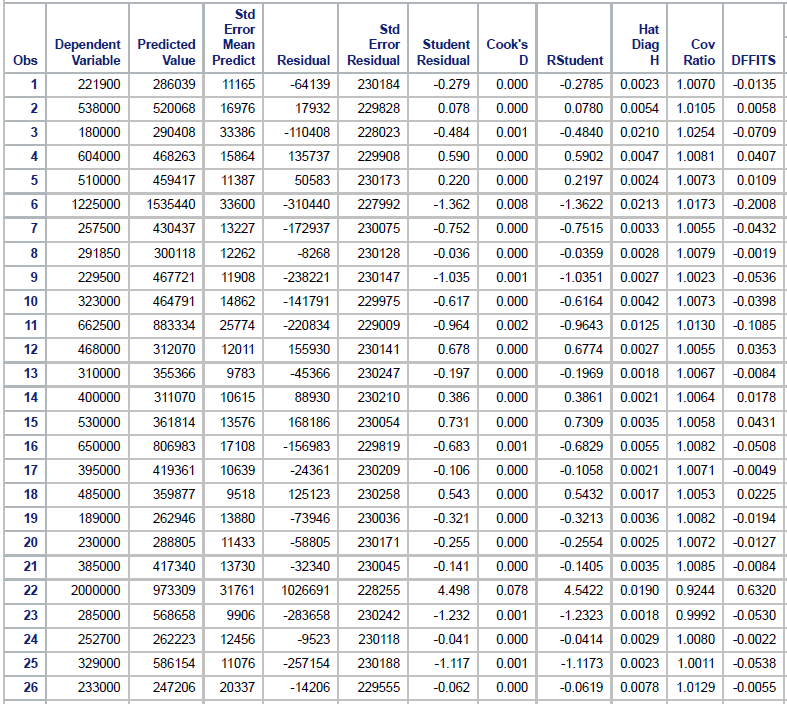


1. Check for Uncorrelated Errors

(INSERT TIME PLOT)

Check Diagnostics:

1. Check for Outliers
2. Check for Influencers



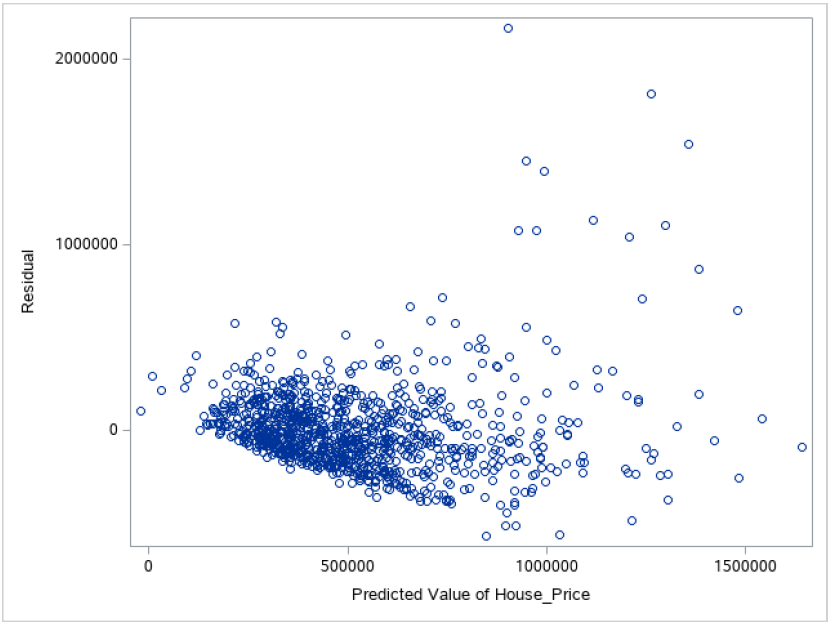
1. Check for Multicollinearity (done previously)

Model 2 – Assumption Verification

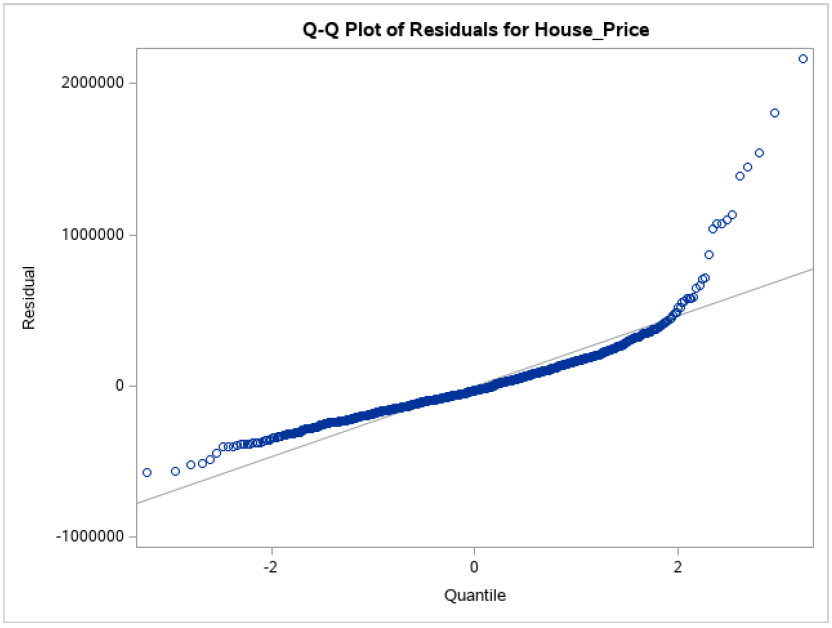
1. Model Form (no curvature)



1. Check for Constant Variance



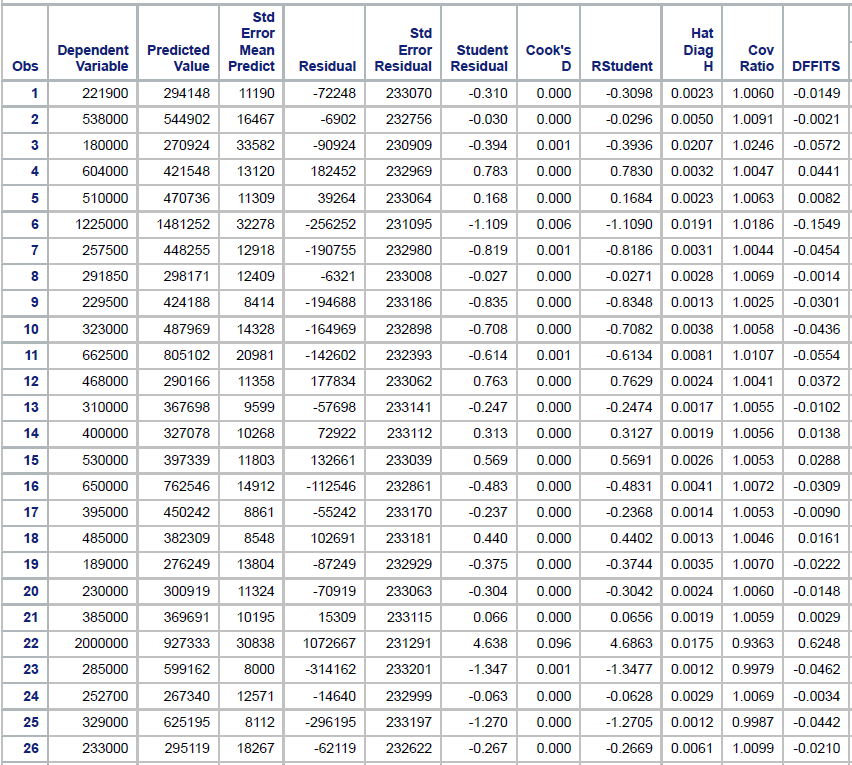
1. Check for Normally Distributed Data



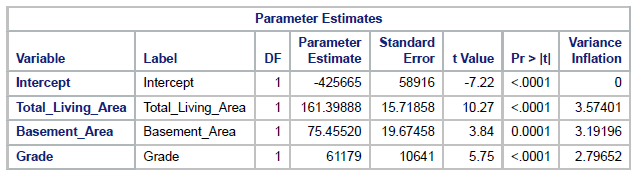
1. Check for Uncorrelated Errors (previously done)

Check Diagnostics

1. Check for Outliers
2. Check for Influencers



1. Check for Multicollinearity



**Section 6**

VI. Final Multiple Linear Regression Model

Present and interpret the meaning of your final model.

Discuss the fit of the model and interpret inferences (explained variability, joint C.I. for the parameters; C.I., C.B., and P.I. calculated at one xh of interest).

**6.1 Final Multiple Linear Regression Model**

**Present and interpret the meaning of your final model. (What theory to add here?)**

**Discuss the fit of the model:**

The Fitted line for this model is given below.

**ŷ i = -484438 + 215.02003xi1-97.62612xi2 – 90.85933xi3 +** **72886xi4**

We will use hypothesis test to verify if our model is significant;

H0 = β0 + β1xi1+Ԑi (Reduced model)

H1= β0 + β1 xi1 + β2 xi2 + β3 xi3+ β4 xi4 + Ԑi (Full Model)

Here we are 95% confident that model is significant. Significance value α = 0.1 and from the SAS output the p-value for each predictor is 0.0001 which is lesser than the 0.1. As p< α, we reject the null hypothesis. With this hypothesis test we conclude that our model is significant.

ANOVA and parameter estimates for final selected model:

A screenshot of a cell phone

Description automatically generated

**Explained Variability:**

The variability in the response variable corresponding to the changes in the predictors is explained by the R2 value which is 0.5413. This value is called as coefficient of determination. Here from the output we can say that variability by each predictor is explained at 54.13%.

**Bonferroni Joint Confidence Interval:**

We have 5 parameters and 4 predictor variables, hence taking g=4.

two-sided interval: bk ± B\*S{bk}

where, B is the Bonferroni Coefficient given as, B = t (1 – (α/2g); n-p).

Hence, B = t (1 – 0.1/2\*4; 999-5) = t (0.9875;994) = 2.2448

CI for β1= b1 ± B\*s{b1} = 215.02003 ± (2.2448\*18.74444) = (172.942511, 257.097549)

CI for β2= b2 ± B\*s{b2} = -97.62612 ± (2.2448\*19.12557) = (-140.5592, -54.69304)

CI for β3= b3 ± B\*s{b3} = 90.85933 ± (2.2448\*19.66439) = (46.7167073, 135.001953)

CI for β4= b4 ± B\*s{b4} = 72886 ± (2.2448\*10757) = (48738.6864, 97033.3136)

Conclusion: We are 95% confident that β1 lies in (172.942511, 257.097549), β2 lies in between (-140.5592, -54.69304), β3 lies in (46.7167073, 135.001953), β4 lies in between (48738.6864, 97033.3136) simultaneously.

**Confidence Interval at New Xh:**

Taking new predictor values as follows:

Total Living Area= 2200 sq. feet, Living Area Above Basement= 1000 sq. feet, Basement Area=1500 sq. feet, Grade= 8 for 90% CI.

Here, XhT = (1 2200 1000 1500 8) at α=0.1  
**ŷ**h = -484438+215.02003\*(2200)-97.62612\*(1000)-90.8593\*(1500) +72886\*(8) = $337778.996

The standard error of predicted variable is calculated as follows;

S{**ŷ**h}= √MSE (xhT(XTX)−1xh ) = ?????????????????????

Where, MSE = 53109292111

t(1-α/2; n-p) = t(1-0.1/2;999-5) = t (0.95; 994)= 1.646

95% Confidence Interval for new response is given as:

**ŷ**h ± t(1 − α⁄2 ; n − p)\*S{ **ŷ**h } =21471.4971± 1.646\*??????? = 21471.4971±????????=?????????

Hence, we get the C.I for new Xh is (??????,??????). (units)

**Confidence Band**:

Formula for confidence band is **ŷ**h ± W \* S{**ŷ**h} we can calculate the confidence band of mean response. W is the working-hotelling coefficient is given as,

W2= p\*F (1-α; p, n–p)

Where, p = 5 n =40 α = 0.1; From the table we get F (1-0.1, 5, 994) = 1.85

Hence, W2=p\*F (1-α, p, n – p) = 5\*1.85=9.25

W=3.04138127  
By substituting the values in the formula **ŷ**h ± W \* S{**ŷ**h}, we got the confidence band region of (???????, ????????).

Conclusion: We can conclude that the confidence band for mean response of house price lies between ??????? and ???????. (units)

**Prediction Interval:**

prediction interval for the response at new Xh is given as, **ŷ**h ± t (1-α/2; n-p) \* S{pred}

s{pred} = √MSE + S{Yh}2 = √?????? + ?????? = ???????

step itself by substituting the known values we get S2{pred} =???????

S{pred} =???????

by substituting above values in the PI formula, we get the prediction interval for house price response as (???????,???????)

Conclusion: Hence we can conclude that we are 90% confident that the mean response variable house price at the considered predictor variables Total Living Area= 2200 sq. feet, Living Area Above Basement= 1000 sq. feet, Basement Area=1500 sq. feet, Grade= 8 will lie between ????????? and ?????????. (include units)

**Section 7**

**7.1 Final Discussion**