Assignment #2
Nitin Gaonkar

Introduction:

The purpose of this assignment is to build regression models for the home sale price(simple linear and multi regression models) and to find the best regression model to predict the sale price.

Results:

1. Correlation procedure:

From the below correlation procedure results we can see that the variable MasVnrArea is correlated approximately 0.5 with saleprice, below table also provides the measure of variability and helps to understand the data in a better way.

The CORR Procedure

2 Variables:	MasVnrArea SalePrice

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
MasVnrArea	2907	101.89680	179.11261	296214	0	1600
SalePrice	2930	180796	79887	529732456	12789	755000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	MasVnrArea	SalePrice
MasVnrArea	1.00000 2907	0.50828 <.0001 2907
SalePrice	0.50828 <.0001 2907	1.00000 2930

x

Simple regression model:

Below out gives us the number of observation read, no of the observation used in the procedure, also provides number of observation that are missing the values.

Below is the model in equation form:

saleprice= 157303 + 226.47763 * MasVnrArea

Increase of one unit of MasVnrArea the sale price goes up by 226.If the MasVnrarea is zero then the sale price of the house is 157303.
The **P value** in the anova suggests that the regression is significant and there is some linear relationship with the dependent and the independent variables.

R-square: Here the R square value is 0.2584, this value implies that how close the data are to the fitted regression line, it is the percentage of the response variable variation that is explained by a linear model. Higher the value of r-squared the better model fits our data.
Here its model/corrected total = 4.78E/1.85E= 0.25
Co-eff var is nothing but the root mse/dependent mean, By looking at the values of the P values in the Parameter estimates we can see that the estimates are significant.

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2907
Number of Observations with Missing Values	23

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.781879E12	4.781879E12	1011.96	<.0001
Error	2905	1.372718E13	4725361826		
Corrected Total	2906	1.850905E13			

Root MSE	68741	R-Square	0.2584
Dependent Mean	180380	Adj R-Sq	0.2581
Coeff Var	38.10908		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	157303	1466.89502	107.24	<.0001
MasVnrArea	1	226.47763	7.11940	31.81	<.0001

Below is the scatter plot where we have the maxvaarea the independent variable (x) and the sale price (y) dependent variable and the points are the actual values and the line in between is the regression line. This regression line minimizes the sum of the squared errors. Now in the fig(4) ,we can see that the residuals or the errors are plotted against zero, its more of the regression from fig(3) is flattened in the fig(4) and the errors are plotted. If you notice carefully we can see the actual values which are plotted in the fig(3) are the same points which are plotted as residual in the fig(4)

Fig 3

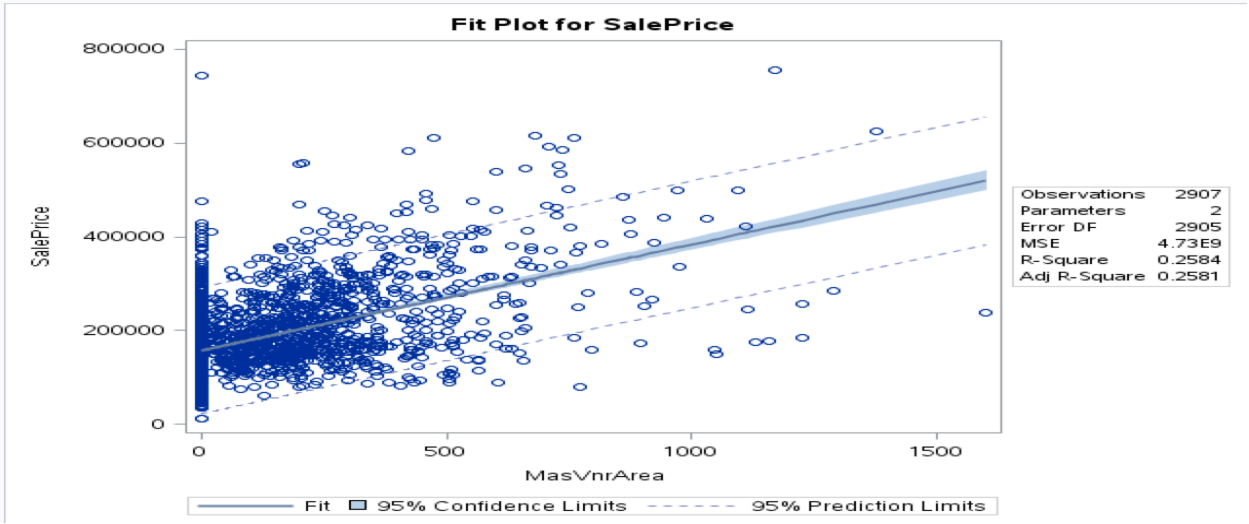
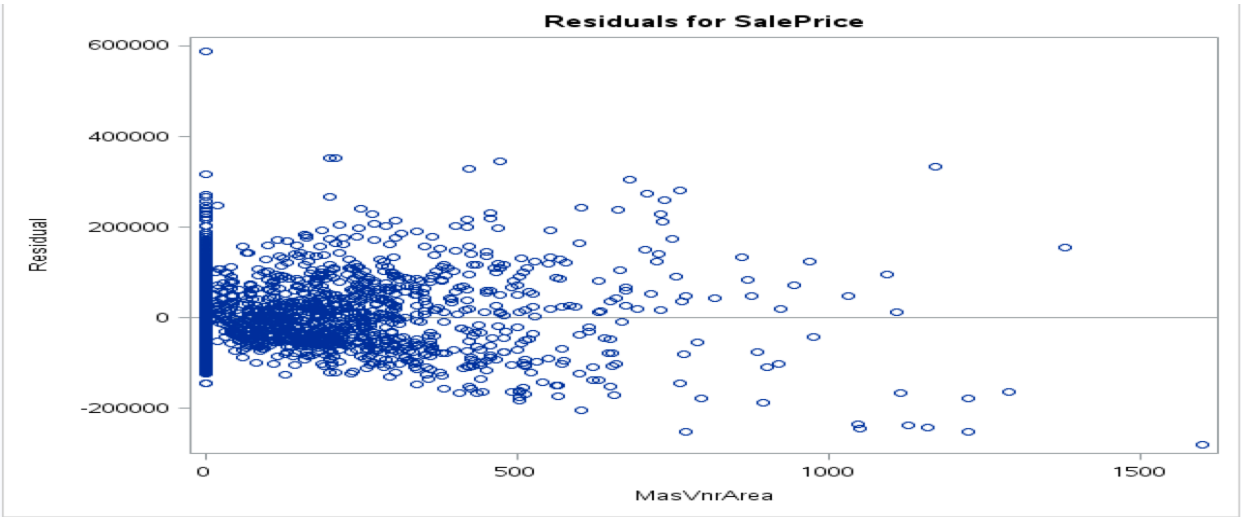


Fig4:



Below is the qq plot(fig 5) for this model, we can see that the plot start off pretty nicely along the line, as we go at the end we can see that the graph deviates from the straight line indicating that there could be few outliers there and also we can observe the last point which is a outlier and definitely a potential problem, overall this plot looks ok. In the Fig6 we have histogram of the residual and we can see that the histogram is close to the curve put by SAS.

Fig5

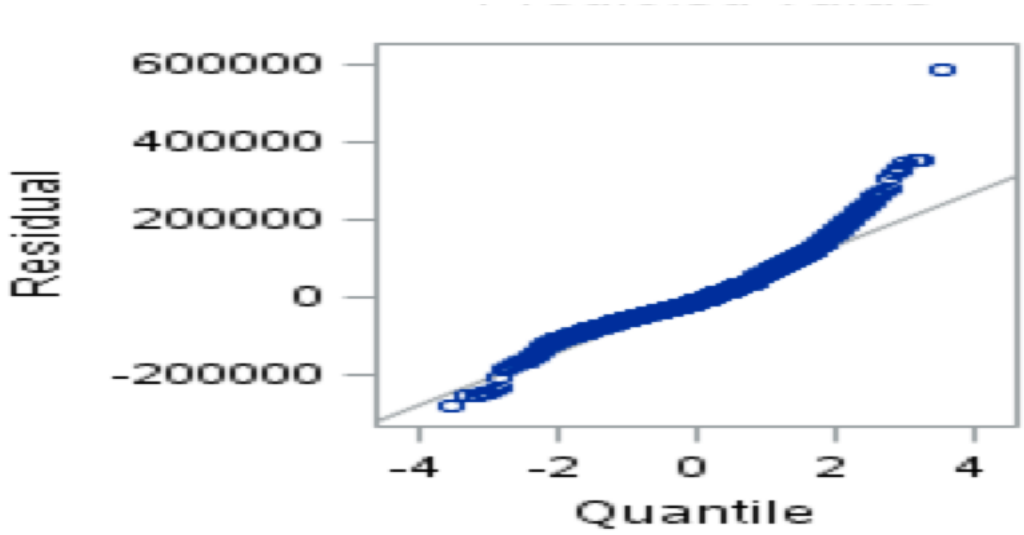
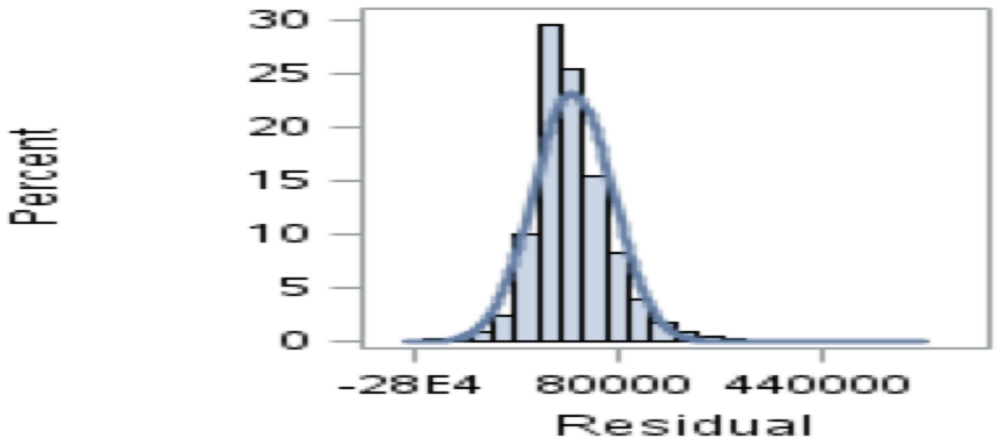
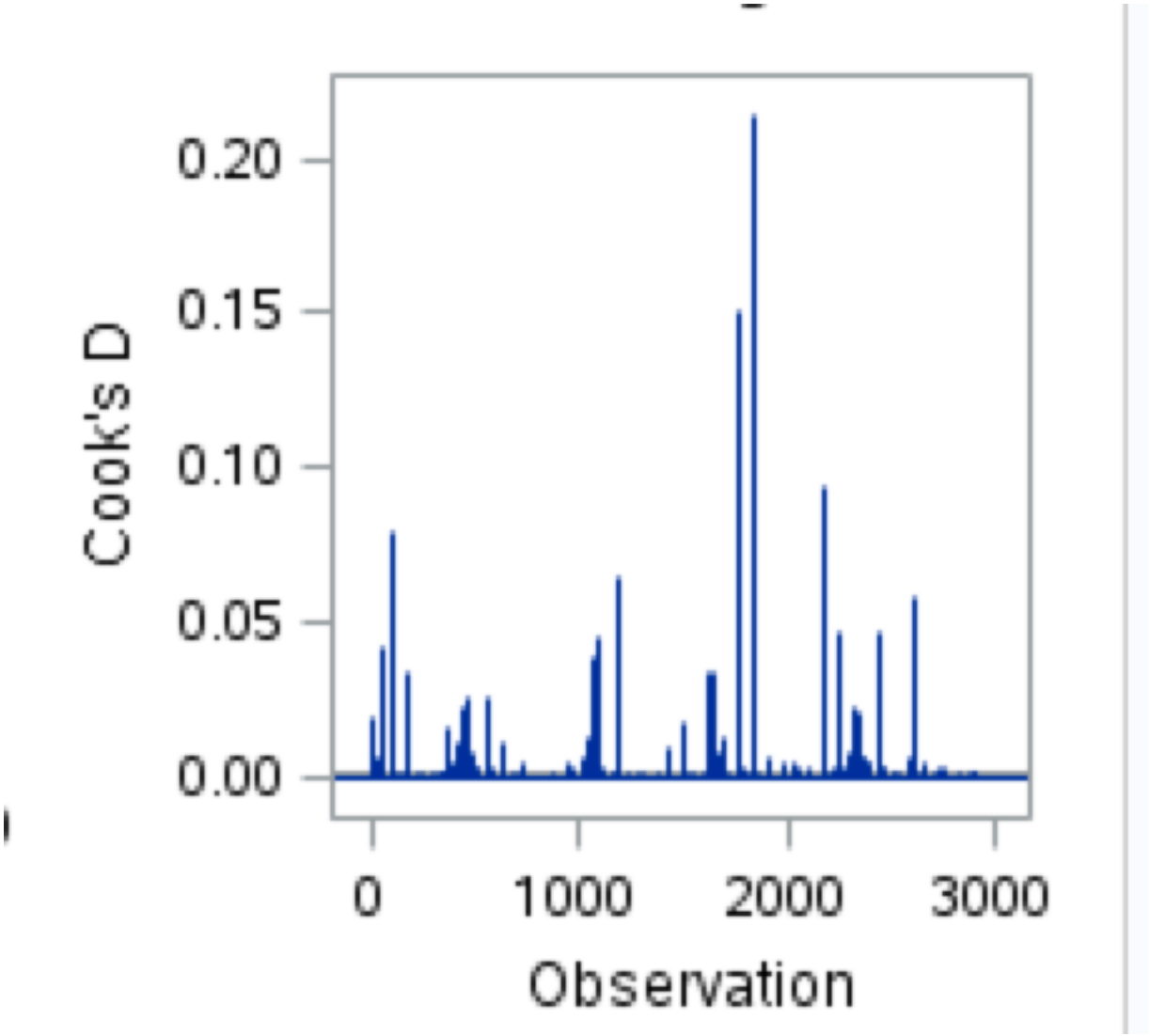


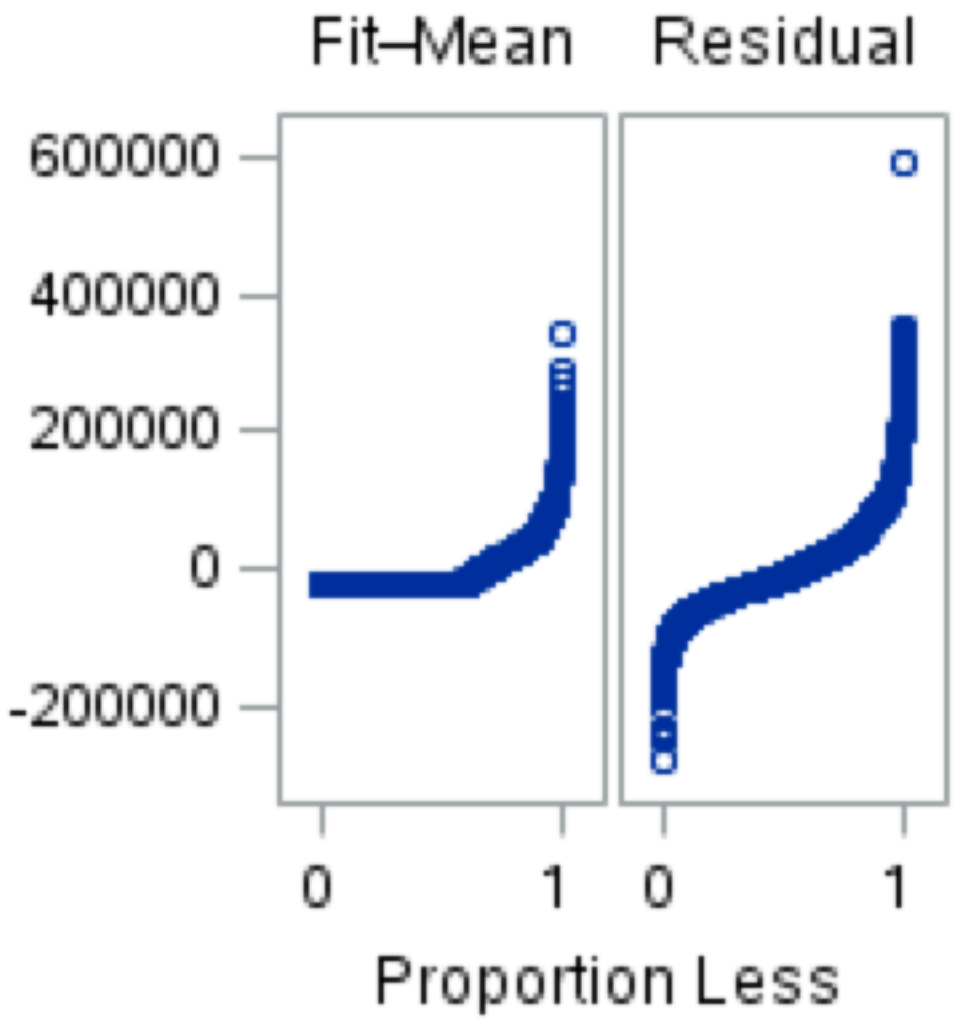
Fig6



Cooks d plot helps us to identify the influential observations, we could see that in our below graph there are two observations near close to 2000 have a larger value of D than others, so we need to have a closer look at these observation in order to decide whether these are influential or not



In the below plot we have to compare the the spread of the fit to the spread of the residuals, since the left side of the plot is taller than the right so we can conclude that the spread of the residual is less than the spread of the fitted value.



2

I used below continues variables for the simple linear regression

Grlivarea, GarageCars, GarageArea, TotalBsmtSF, FIrstFlrSf, MasVnrArea

Grliv area is the best fit as the R-square is high and also the fit-mean and the residual plot looks good as the fit mean is greater than the residual, there are couple of outliers in the cook's plot. Even the QQ plot looks good and we could see few outliers at the end of the tail and there are few outliers on the scatter plot and but most of the values fall within prediction limits.

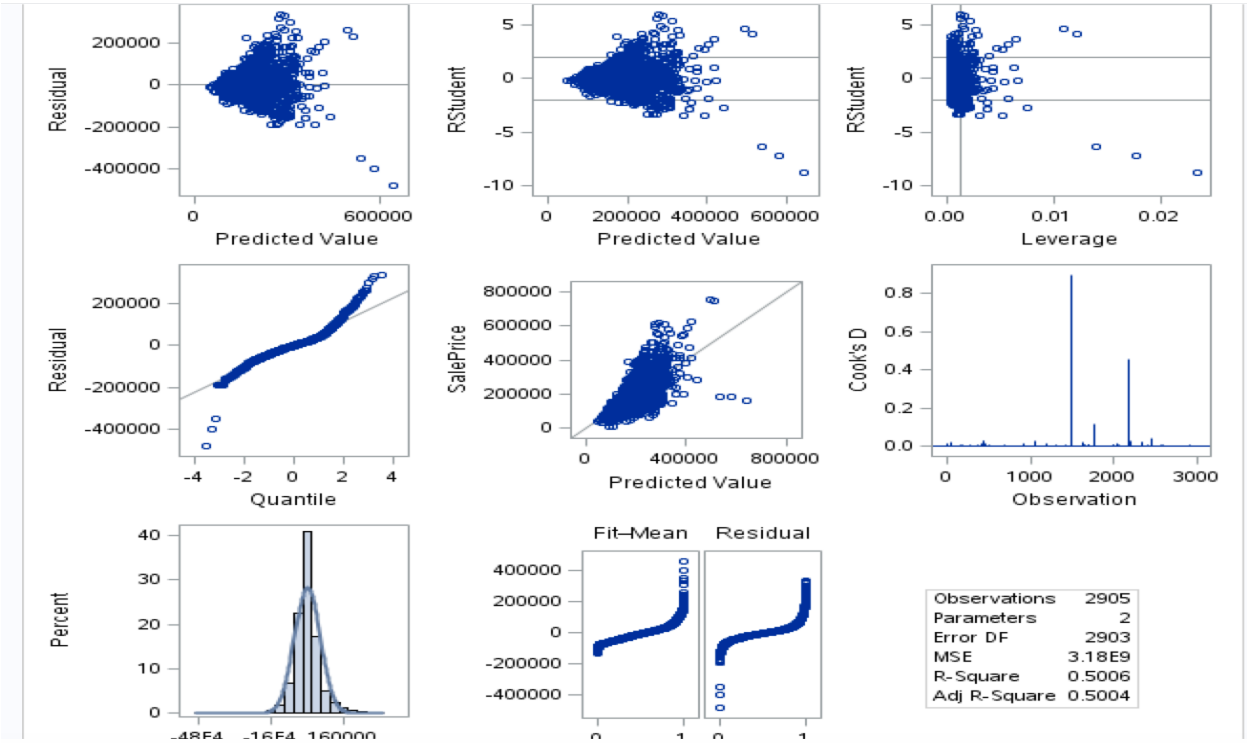
The equation of the model:

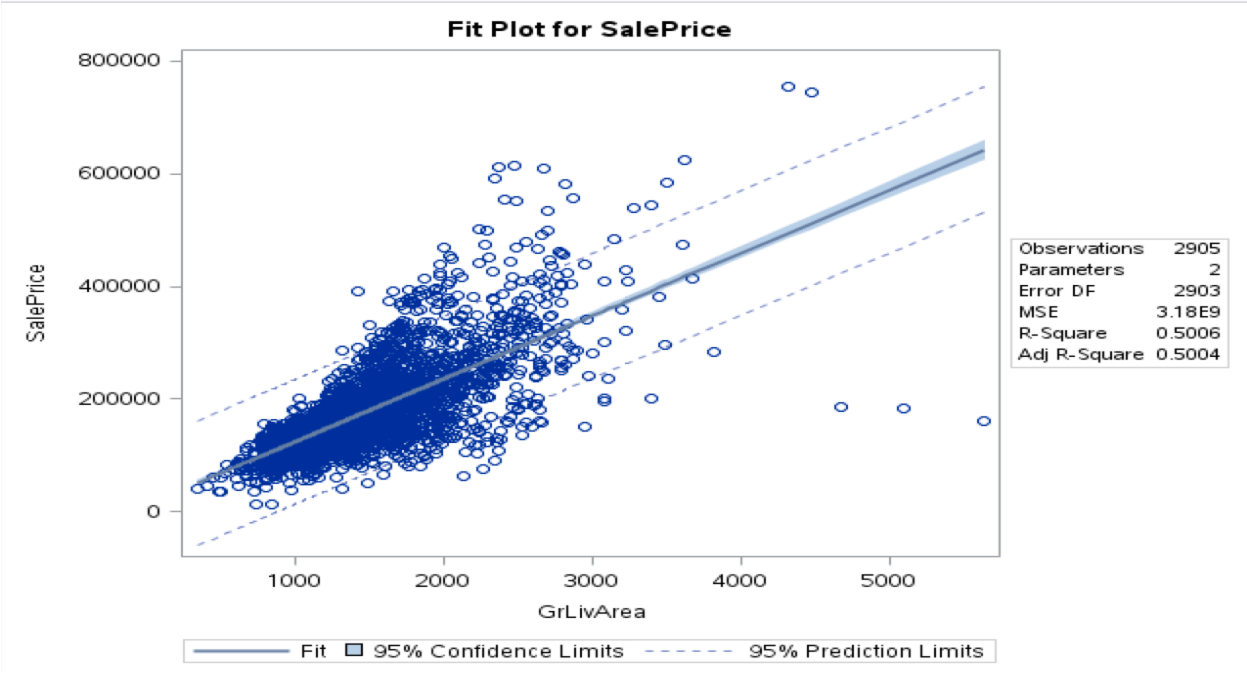
Saleprice= 13290 + 111.69400 Grlivarea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
R-Square Selection Method

Number of Observations Read	2930
Number of Observations Used	2905
Number of Observations with Missing Values	25

Number in Model	R-Square	Variables in Model
1	0.5006	GrLivArea
1	0.4185	GarageCars
1	0.4086	GarageArea
1	0.4002	TotalBsmtSF
1	0.3885	FirstFlrSF
1	0.2582	MasVnrArea


Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Fit–Mean
Residual
Observations 2905
Parameters 2
Error DF 2903
MSE 3.18E9
R-Square 0.5006
Adj R-Square 0.5004


Fit Plot for SalePrice
SalePrice
GrLivArea
Observations 2905
Parameters 2
Error DF 2903
MSE 3.18E9
R-Square 0.5006
Adj R-Square 0.5004
Fit
95% Confidence Limits
95% Prediction Limits

3. For this regression I have picked the below categorical variable:

below is the output of the reg proc:

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.20966E12	4.20966E12	851.07	<.0001
Error	2928	1.448288E13	4946337816		
Corrected Total	2929	1.869254E13			

Root MSE	70330	R-Square	0.2252
Dependent Mean	180796	Adj R-Sq	0.2249
Coeff Var	38.90030		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	145729	1770.04460	82.33	<.0001
Fireplaces	1	58512	2005.67334	29.17	<.0001

The model equation is:

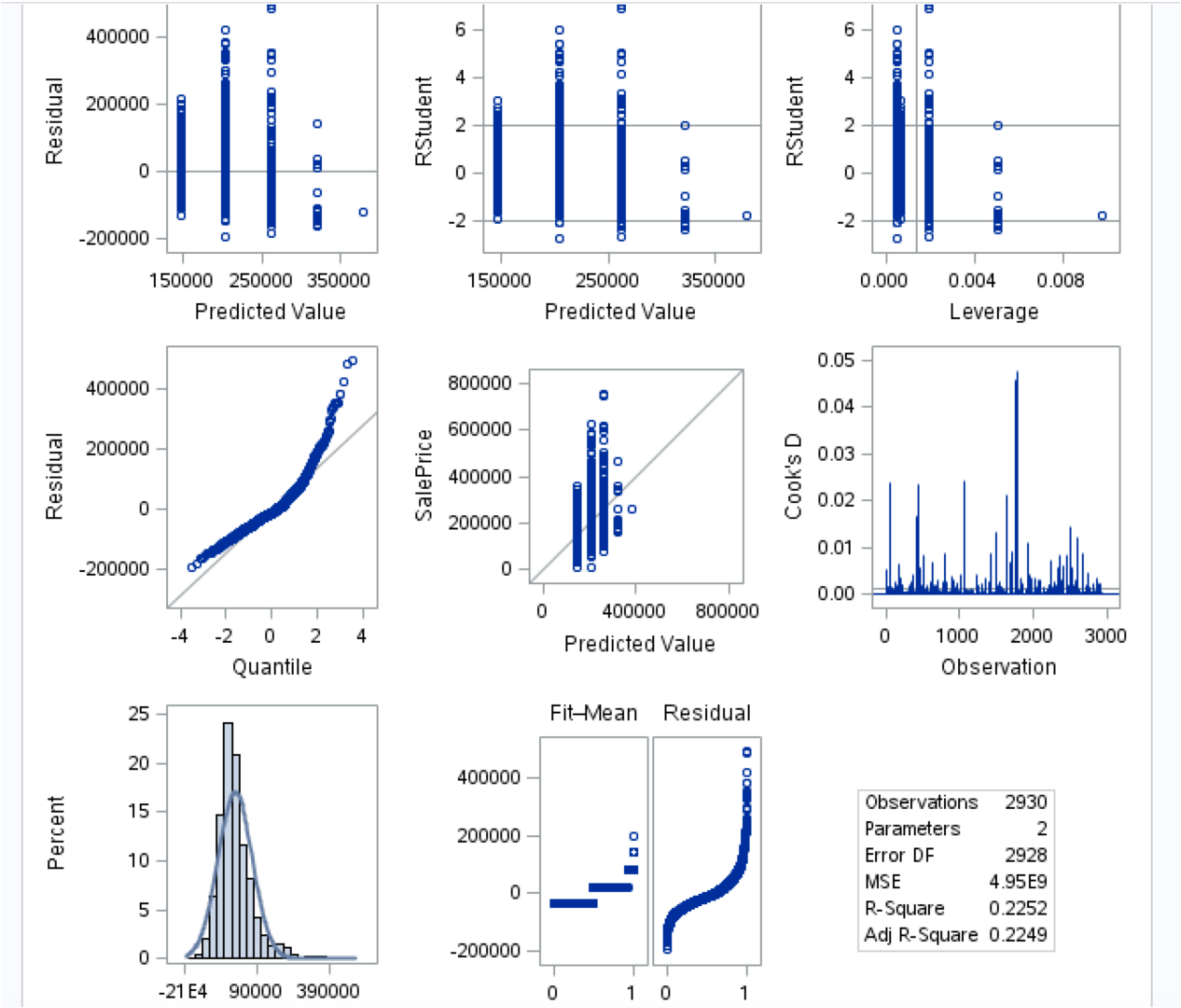
Saleprice=145729 + 58512 fireplaces

Increase of one unit of fireplaces the sale price goes up by 58512.If the fireplaces is zero then the sale price of the house is 145729.
The **P value** in the anova suggests that the regression is significant and there is some linear relationship with the dependent and the independent variables.

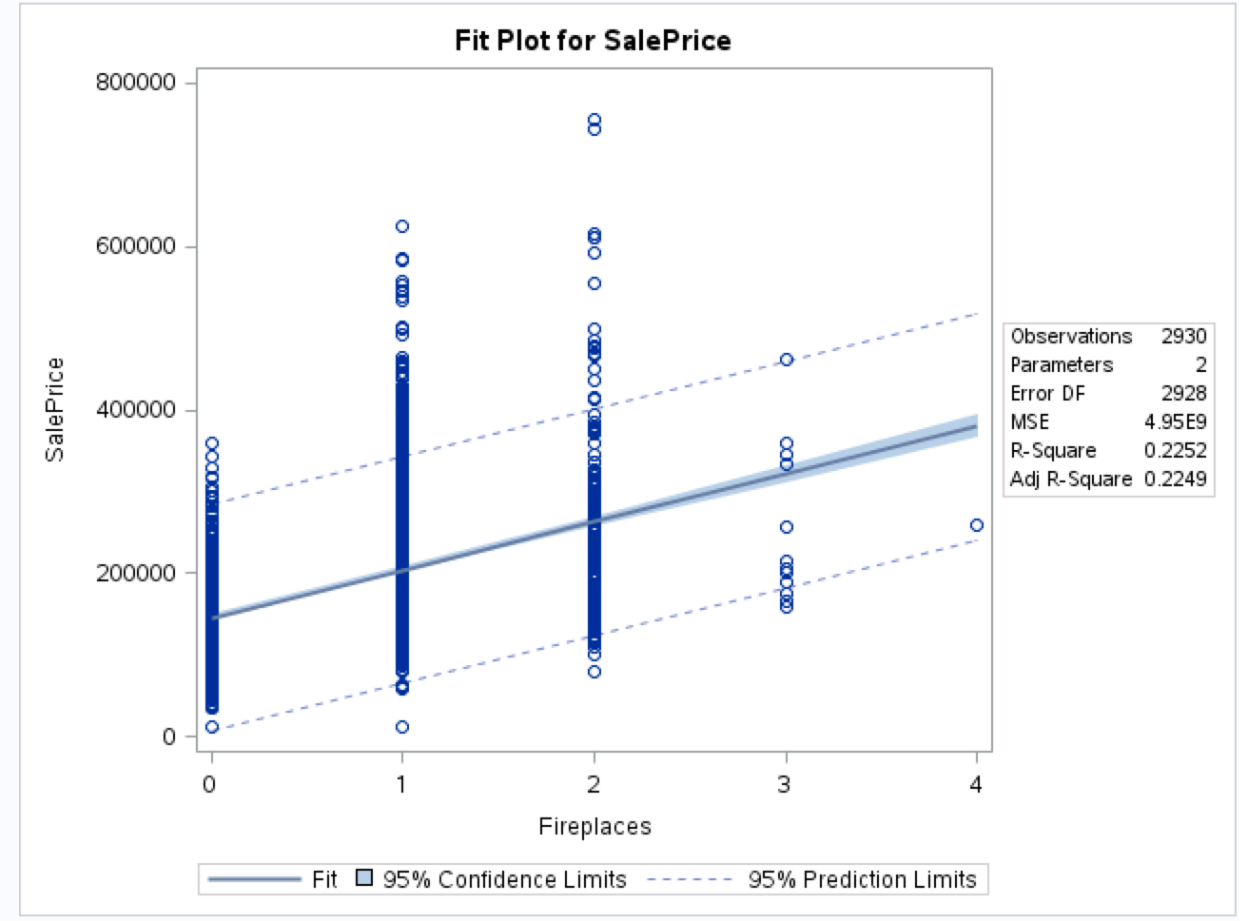
R-square: Here the R square value is 0.2252, this value implies that how close the data are to the fitted regression line, it is the percentage of the response variable variation that is explained by a linear model. Higher the value of r-squared the better model fits our data.
Here its model/corrected total = 4.209E/1.86E= 0.2252
Co-eff var is nothing but the root mse/dependent mean, By looking at the values of the P values in the Parameters estimates we can see that the estimates are significant.

By observing the below graphs, we can see that in the residual graph data points are scattered and we can see lot of outliers
In the Cook's D we can see outliers around observation 2000 also few outliers between 0 to 1000.

Also In fit mean and residual plot we can see the residual plot is higher than the fir mean and also the Rsquare value is only 0.2252, thus doesn't look like a great fit.



From the below scatter plot we can say that most of the data points are between 0 to 2, meaning most of the houses have either 1 or 2 fireplaces and very few houses with 3 fireplaces, we can also see lot of outliers outside the prediction limits. By looking at the scatter plot it looks like the the prediction model goes through the mean of the Y.



The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2930	180796.06	79886.69	12789.00	755000.00

4.

Below are the comparisons of all the models:

a. Below are the R-square values for the three models:

Model 1: 0.25

Model2: 0.500

Model3: 0.2552

Here we can see that model2 has the higher value of r –square.

b. By checking the p value, we can see that all the p values for all the models are significant.

c. By comparing the scatter plots of all the three models, I can say that the distribution of the model 2 Is the better than the other two models.

d. BY comparing the residual and the cook's d plots of all the three models respectively I think model 2 is the best fit.

By checking all the above parameters, I think model is the best fit.

5

Multiple regression model:

Model equation is:

Saleprice= 26547 + 118.54695 MasVnrArea + 94.60302 GrLivArea

Forward Selection: Step 2

Variable MasVnrArea Entered: R-Square = 0.5599 and C(p) = 3.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.036256E13	5.181279E12	1846.98	<.0001
Error	2904	8.146497E12	2805267561		
Corrected Total	2906	1.850905E13			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	26547	3141.93715	2.002617E11	71.39	<.0001
MasVnrArea	118.54695	5.99550	1.096743E12	390.96	<.0001
GrLivArea	94.60302	2.12104	5.580679E12	1989.36	<.0001

Bounds on condition number: 1.1946, 4.7784

All variables have been entered into the model.

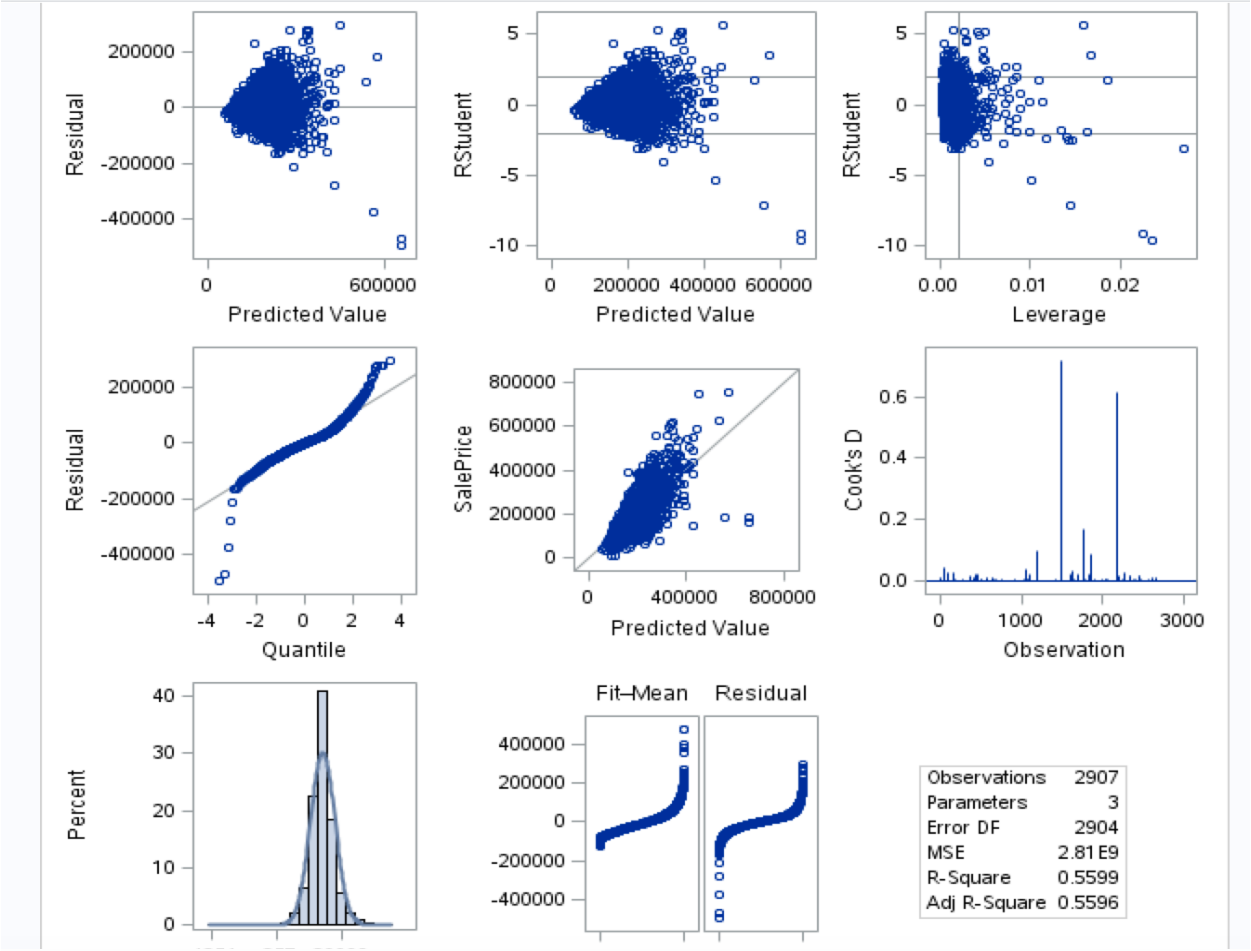
Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea	1	0.5006	0.5006	391.958	2912.09	<.0001
2	MasVnrArea	2	0.0593	0.5599	3.0000	390.96	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
R-Square Selection Method

Number of Observations Read	2930
Number of Observations Used	2907
Number of Observations with Missing Values	23

Number in Model	R-Square	Variables in Model
2	0.5599	MasVnrArea GrLivArea



- The residual plots looks ok, but there are few data points which are away from the cluster and may require some investigation, in the cook's D couple of observation are outliers, the histogram looks ok as the points are around the SAS curve, the fit mean curve looks better since its greater than the residual and the R square is higher in this model 0.5599, the qq plots also look better except the few data points. Yes, this model looks better than the simple regression as the value of R square is high and all the graphs looks better, this model is looks like a better fit than the simple regression.

6

I added the variable BsmtFinSF2 into my model since this had the least correlation with the Y. After running the model I did not see much changes in the model after adding the new variable, the r square value did not change much and also the residual plots and the other plots did not change much.
I don't think more predictor variables means a better fit, the main criteria for comparing models are: rsquare values, residual plots, cook's D observations and the scatter plots.

Bounds on condition number: 1.1945, 4.7779

Forward Selection: Step 3

Variable BsmtFinSF2 Entered: R-Square = 0.5602 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.036248E13	3.454159E12	1232.01	<.0001
Error	2902	8.136296E12	2803685698		
Corrected Total	2905	1.849877E13			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	25997	3159.53176	1.898187E11	67.70	<.0001
MasVnrArea	118.65010	5.99412	1.098536E12	391.82	<.0001
GrLivArea	94.62084	2.12098	5.579961E12	1990.22	<.0001
BsmtFinSF2	10.46253	5.78953	9156235321	3.27	0.0708

Bounds on condition number: 1.1946, 10.169

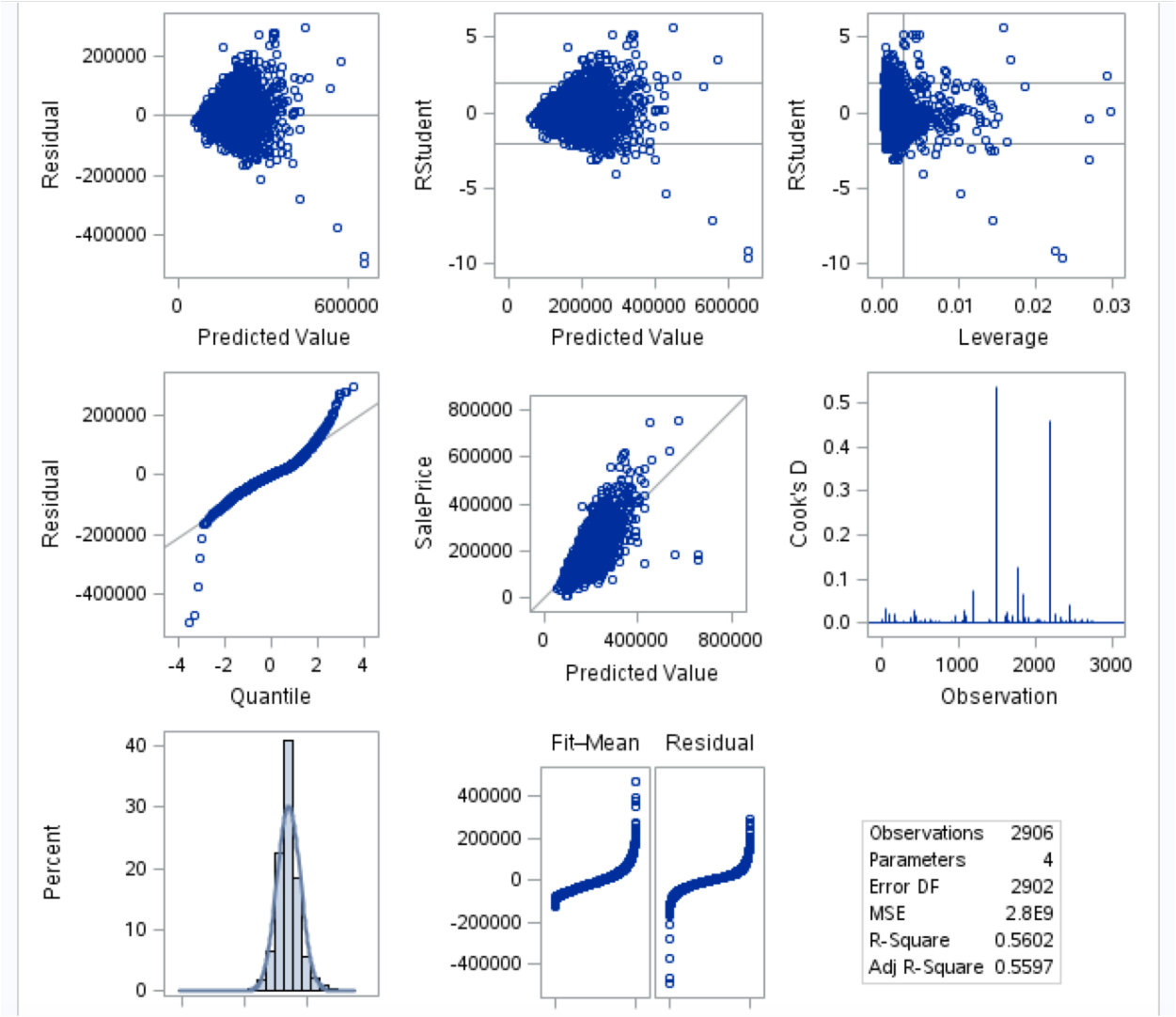
All variables have been entered into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea	1	0.5004	0.5004	394.400	2908.60	<.0001
2	MasVnrArea	2	0.0593	0.5597	5.2658	390.83	<.0001
3	BsmtFinSF2	3	0.0005	0.5602	4.0000	3.27	0.0708

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
R-Square Selection Method

Number of Observations Read	2930
Number of Observations Used	2906
Number of Observations with Missing Values	24

Number in Model	R-Square	Variables in Model
2	0.5597	MasVnrArea GrLivArea
2	0.5008	GrLivArea BsmtFinSF2
2	0.2585	MasVnrArea BsmtFinSF2
3	0.5602	MasVnrArea GrLivArea BsmtFinSF2



Conclusion:

Overall in this assignment I have tried to fit simple and multi regression and tried to find the best regression model to predict Y. I think if overdispersion seems to be an issue then it would make me think that the model is not appropriately specified. I think the next step in modelling would be identifying the outliers and investigating on those, we need to check the impact of the outliers to see whether they really affect the model or not, also we need to test the assumptions and also resolve any kind of data issues and finally interpret the results.

Code:

Paste your code in at the end.

```
libname mydata "/scs/wtm926/" access=readonly;

proc datasets library=mydata;

run;

 quit;


data my_assign;


set mydata.ames_housing_data;

proc contents data=my_assign;

run;
```

Question 1:

Finding co-relation for all the numeric variables with rank

```
proc corr data=my_assign rank;

var _numeric_ ;

with saleprice ;

run;
```

```
###### simple linear regression#########

proc corr data=my_assign;

var MasVnrArea saleprice;

run;


ods graphics on;

proc reg data=my_assign;

model saleprice=MasVnrArea;

run;



ods graphics on;

proc reg data=my_assign;

model saleprice = grlivarea GarageCars GarageArea TotalBsmtSF FirstFlrSF MasVnrArea/

selection=rsquare start=1 stop=1;


ods graphics on;

proc reg data=my_assign;

model saleprice=Fireplaces;

run;


ods graphics on;
```

```
proc reg data=my_assign;

model saleprice = MasVnrArea grlivarea/

selection=forward;
```