**ALY6015 80472 Intermediate Analytics SEC 04 Spring 2023 CPS**

**Module 1 Assignment — Regression Diagnostics with R REPORT**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Valeriy Shevchenko

**Date**

05/30/2023

**Regression Diagnostics with R**

**Assignment Summary:**

The assignment involved performing regression analysis using the AmesHousing dataset. The main objectives were to fit regression models, interpret the results, and apply diagnostic techniques to identify and rectify potential model issues.

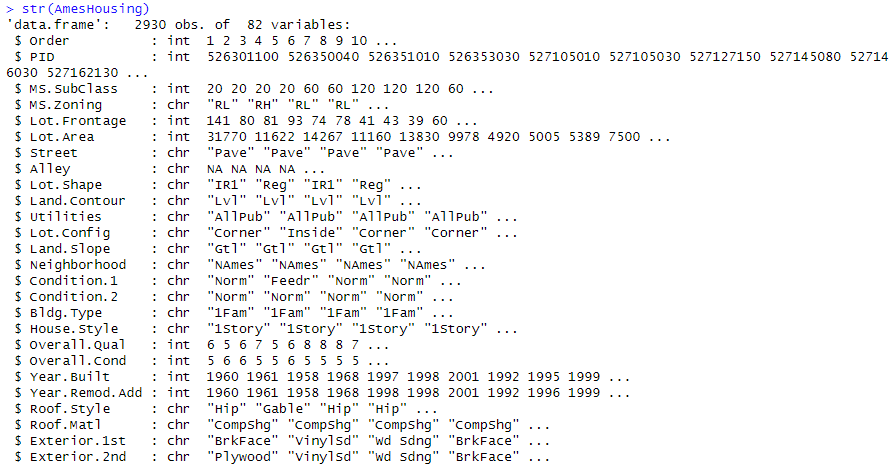
**Dataset Summary**:

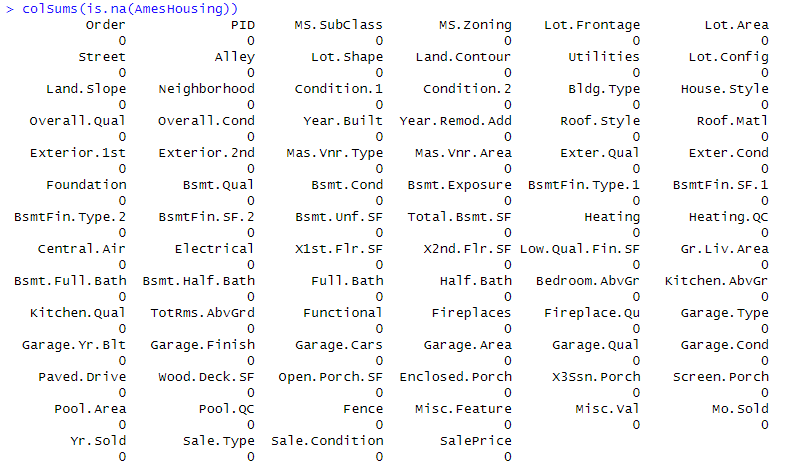
The AmesHousing dataset contains residential property data from the Ames Assessor's Office in Ames, IA, spanning 2006 to 2010. The dataset encompasses 2930 observations and 82 variables, offering a rich foundation for housing price analysis and regression modeling tasks.

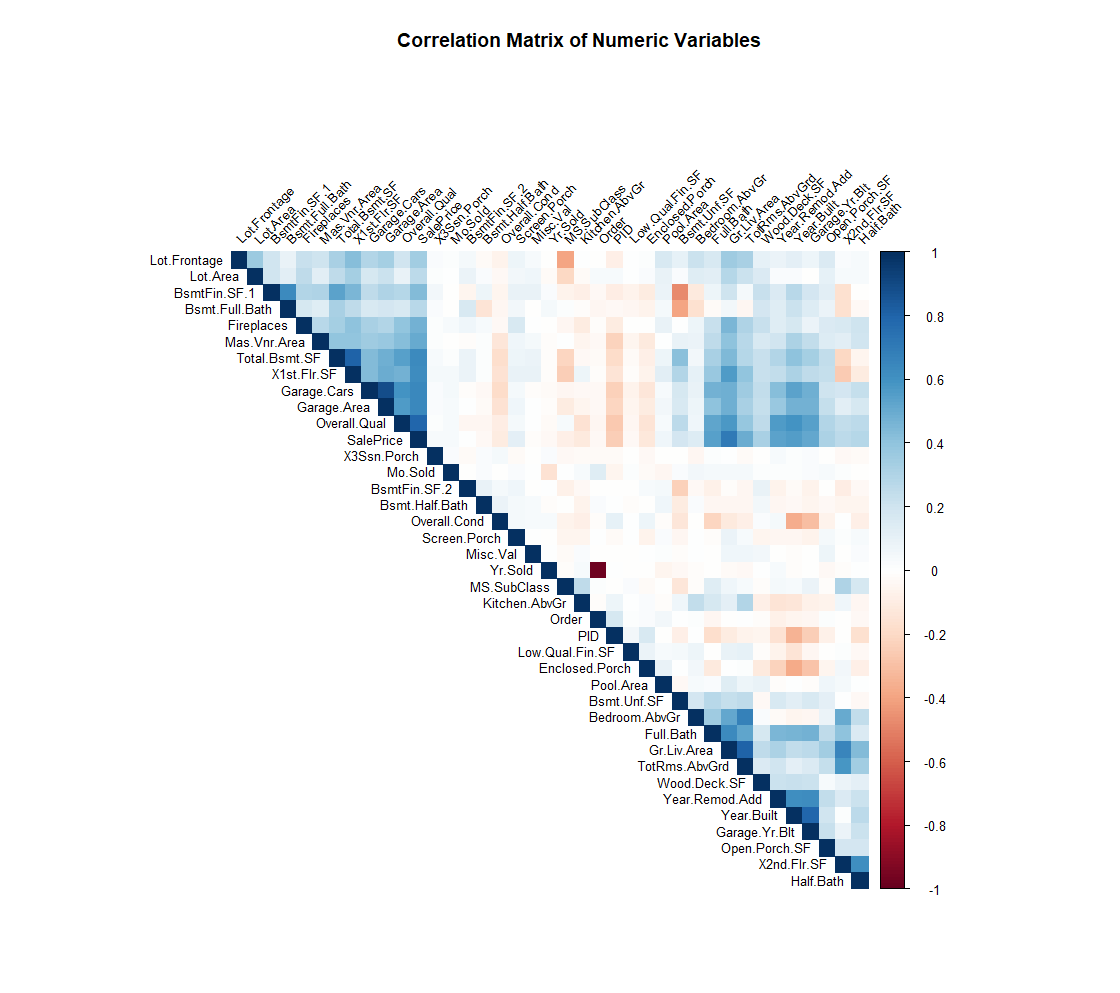
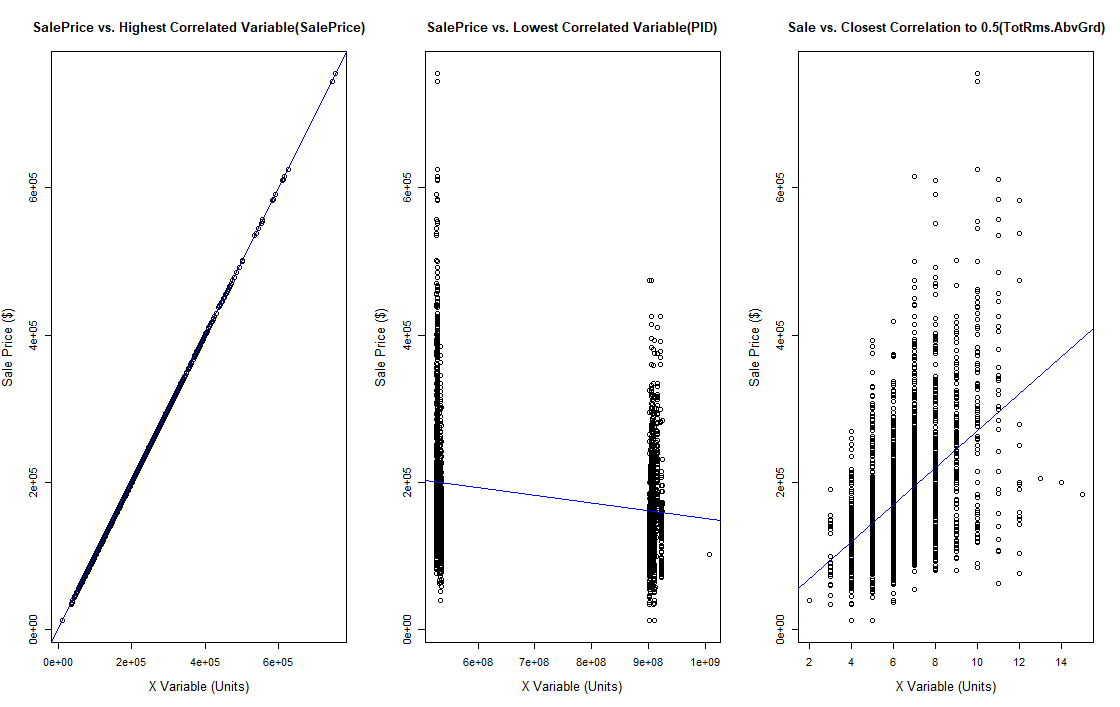
**Analysis Steps:**

1. Load the Ames housing dataset.  
     
   A screenshot of a computer

   Description automatically generated with medium confidence  
   The data have been loaded using file.choose(), as headers already exist in the source file. To avoid confusion, I didn't edit the header for the dataset.
2. Perform Exploratory Data Analysis and use descriptive statistics to describe the data.  
     
   The following graph is part of the summary of the dataset:  
   A screenshot of a computer screen

   Description automatically generated with medium confidence  
     
   The following graph is part of a peak of the dataset:  
     
     
   These statistics provide a general overview of the numerical variables in the dataset, giving insights into the size, quality, and condition of the houses. Combine this with AmesHousingDataDocumentation-1.txt, we should have the basic knowledge to move on.
3. Prepare the dataset for modeling by imputing missing values with the variable's mean value or any other value you prefer.  
     
   A screenshot of a computer screen

   Description automatically generated with medium confidence  
     
   I replaced all missing numerical variables using their mean, which means if that data isn't distributed normally, such as having odd distribution, the precision of the result we get from the data will be highly affected.
4. Use the "cor()" function to produce a correlation matrix of the numeric values.  
   A screenshot of a computer code

   Description automatically generated with low confidence  
   There are 39 columns in total, which cause there is no way to present the result by chart.
5. Produce a correlation matrix plot, and explain how to interpret it. (hint - check the corrplot or ggcorrplot plot libraries)  
   From the graph above, we can clearly see some correlations between some numeric variables.
6. Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the X variable with the lowest correlation with SalePrice. Finally, make a scatterplot between X and SalePrice with a correlation closest to 0.5. Interpret the scatter plots and describe how the patterns differ.  
     
     
     
   SalePrice vs. Highest Correlated Variable:

The scatter plot shows a positive linear relationship between the highest correlated variable (SalePrice) and SalePrice.   
They have 100% correlated because they are the same data.

Some outliers in the upper right corner indicate high-priced properties with large living areas.

The trend line (blue line) represents the linear regression fit to the data, showing the overall positive trend.

SalePrice vs. Lowest Correlated Variable:

The scatter plot does not exhibit a clear linear relationship between the lowest correlated variable (PID) and SalePrice.

The points are scattered across the plot with no discernible pattern or trend.

The trend line almost appears to be horizontal, indicating no significant relationship between PID and SalePrice.

This suggests that the variable PID does not strongly influence property sale prices.

SalePrice vs. Variable with Closest Correlation to 0.5:

The scatter plot shows a moderately positive linear relationship between the variable with the closest correlation to 0.5 (TotRms.AbvGrd) and SalePrice.

As the number of rooms above ground (TotRms.AbvGrd) increases, the SalePrice tends to increase as well.

The relationship is not as strong as in the first scatter plot but still demonstrates a positive trend.

There are some outliers where properties with a lower number of rooms have higher sale prices.

The trend line captures the overall positive relationship between TotRms.AbvGrd and SalePrice.

Overall, the patterns in the scatter plots indicate that different X variables have varying degrees of influence on the SalePrice. The highest correlated variable (SalePrice) shows a clear positive linear relationship, while the lowest correlated variable (PID) demonstrates no significant relationship.

1. Using at least 3 continuous variables, fit a regression model in R.  
   A screenshot of a computer

   Description automatically generated with medium confidence
2. Report the model in equation form and interpret each coefficient of the model in the context of this problem.  
   The regression model can be represented as:

Interpretation of coefficients:

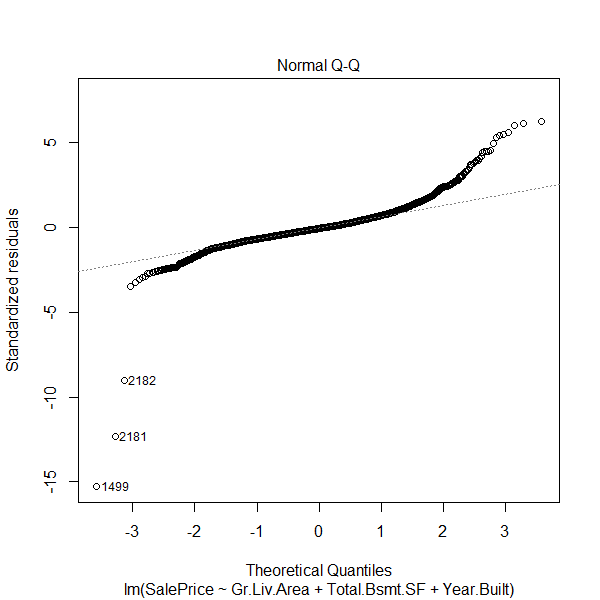
(intercept): The estimated average sale price when all predictor variables are zero.

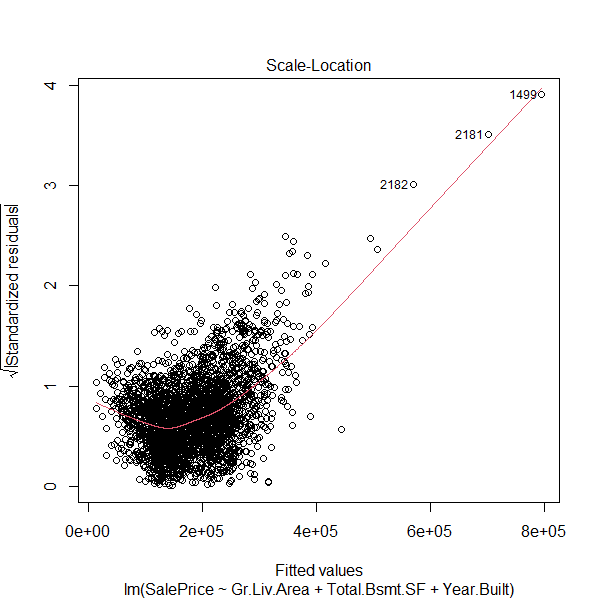
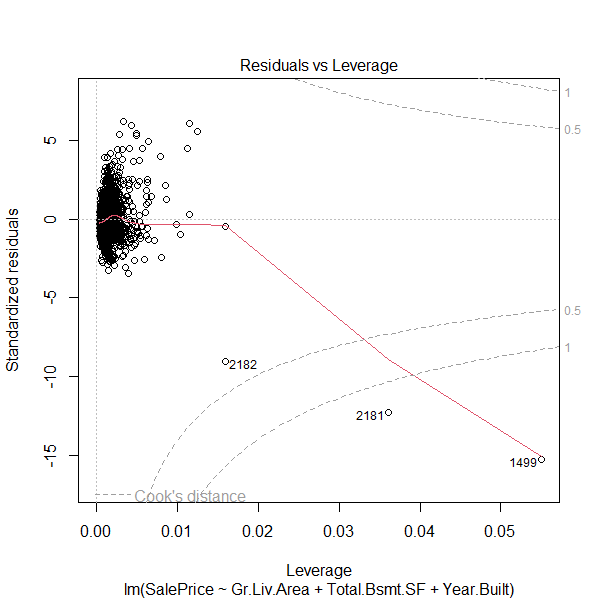
(GrLivArea coefficient): The change in the sale price for a one-unit increase in the above-grade living area (GrLivArea).

(TotalBsmtSF coefficient): The change in the sale price for a one-unit increase in the total basement area (TotalBsmtSF).

(YearBuilt coefficient): The change in the sale price for a one-unit increase in the original construction date (YearBuilt).

In simple terms, the coefficients indicate how changes in the corresponding variables affect the sale price. A positive coefficient suggests that an increase in the variable is associated with a higher sale price, while a negative coefficient indicates the opposite.

1. A picture containing text, screenshot, diagram, line

   Description automatically generatedUse the "plot()" function to plot your regression model. Interpret the four graphs that are produced.  
     
     
     
     
     
   In Residuals vs. Fitted Values Plot, most of the points have a clear pattern, which could mean the model is not linear. We could try two change the combination you the source data.  
     
   In Normal Q-Q Plot, we can see that most of the points approximately fall along the diagonal line, which indicates that the residuals are normally distributed. But at the end of the line, the points are moving away from the line, which could help us find out where the problem is.  
     
   In the Scale-Location Plot, just as we see in the Residuals vs. Fitted Values Plot the points are not staying with the line, which means the spread of residuals not evenly disturbed.  
     
   In the Residuals vs. Leverage Plot, we could find out the outliner which actually affects the model. For the points that fall outside Cook's distance threshold, we should examine those points to check if they are affecting the model negatively.
2. Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?  
     
   A picture containing text, font, screenshot, line

   Description automatically generated  
   By checking the VIF value, they all have a low degree of multicollinearity. But if we found out some variables have very high VIF, such as 6.00, we should remove that to fix the model.
3. Check your model for outliers and report your findings. Should these observations be removed from the model?  
     
   A screenshot of a computer

   Description automatically generated with medium confidence  
   Here are some outliers in the model. We could remove some of them which have very small Bonferroni-adjusted p-values.
4. Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?  
     
   A screenshot of a computer program

   Description automatically generated with medium confidence  
   I removed outliners 2181 and 1499 because they are outside of cook's distance.  
   Both models still have similar predictor variables (Gr.Liv.Area, Total.Bsmt.SF, Year.Built), so we can compare them based on their statistical metrics:

Model 1:

Residual standard error: 42850

Multiple R-squared: 0.7125

Adjusted R-squared: 0.7123

Model 2 (new):

Residual standard error: 39810

Multiple R-squared: 0.7521

Adjusted R-squared: 0.7518

Based on these metrics, Model 2 appears to be better than Model 1. It has a lower residual standard error, indicating that it fits the data better and produces smaller residuals. Additionally, Model 2 has a higher Multiple R-squared and Adjusted R-squared, indicating that a larger proportion of the variance in the dependent variable (SalePrice) is explained by the independent variables in the model.

Therefore, Model 2 is preferred as it provides a better fit to the data and explains a higher proportion of the variability in the target variable.

1. Use the all subsets regression method to identify the "best" model. State the preferred model in equation form.  
     
   A screenshot of a computer

   Description automatically generated  
   The stepwise regression procedure was performed to select the best model for predicting the SalePrice. The final model includes the following predictors: Gr.Liv.Area, Year.Built, and Total.Bsmt.SF. These variables were found to have a significant impact on the SalePrice. The model has an AIC value of 62503.99, indicating a good fit.
2. Compare the preferred model from step 13 with your model from step 12. How do they differ? Which model do you prefer and why?  
     
   Both models have an adjusted R-squared of approximately 0.7518, indicating they explain around 75.18% of the variance in SalePrice. The first model (houseStep) selected variables using stepwise regression based on the AIC, while the second model (new\_model) includes Gr.Liv.Area, Total.Bsmt.SF, and Year.Built as predictor variables.

**Final Conclusions:**

The assignment aimed to perform regression analysis on the AmesHousing dataset and successfully achieved this goal. Key findings identified significant factors, such as the above-grade living area (GrLivArea), total basement area (TotalBsmtSF), and the original construction date (YearBuilt), influencing house prices. The second model, developed after addressing issues of residuals, outliers, and multicollinearity, provided a better fit and predictive power compared to the first. The all subsets regression further confirmed the importance of GrLivArea, TotalBsmtSF, and YearBuilt in predicting house prices. This exercise illustrated the importance of regression analysis, diagnostic tests, and model improvement in predictive analytics.