Logo

Description automatically generated

ALY6015 Module 1 Project – R Practice

**Authored by: Prateek Singh**

Course: ALY6015 – Intermediate Analytics

Submission Date: 01/18/2022

**Instructor’s name: Jiyoung Yun**

# Introduction

In this assignment we are required to create two regression models out of the common dataset provided to us related to AMES Housing. We further will be involved to perform EDA and descriptive analysis of the dataset. Interpretation of results and correction of model to make it a best fit model is the ultimate objective of this exercise. This can help strategist or analyst improve the decision-making skills of an organization, understanding of the model and business process management.

Module Learning Objectives are to be able to perform correlation and regression analysis. This includes fitting of the regression model, interpreting that model and evaluating the efficiency of the model to finally select the best-fit model. While performing these, we will also deal with correcting the issues with the models with overfitting, linearity, multicollinearity & outliers.

# Analysis

Ames Housing data was given to analyze and build a regression model. The dataset has the information about the assessed property value of all the residential properties sold in Ames, IA during the period between 2006 and 2010. Upon importing the dataset, I performed the descriptive statistics to get the summary of the entire dataset. The dataset contains a total of 2930 observations (rows of data) and 82 variables (columns) of integer types and character types which can be categorized into 23 nominals, 23 ordinals, 14 discrete, and 20 continuous variables (and two additional observation identifiers). The str, summary and describe function of R were used to get these readings. The minimum, maximum, median and the quantile values of the integer/numerical variables and the length and class of categorical variables were obtained.

To get the idea of how the data available in the dataset is distributed and whether there are any data point outliers, I took couple of variables to test them using frequency/histogram plot and boxplot. Below is the graphical output from both the plots.

The histogram plot shows that the Sales Prices of the Ames housing properties is not normally distributed and is positively skewed. This can be verified by describe function too which shows the difference between mean (180796.1) and median (160000) of Sales Price.

Chart, histogram

Description automatically generated

Fig. 1: Histogram Plot for Sales Price of Ames Housing Property

The Boxplot for the ground floor living area (in sq. feet) for all the properties, illustrates that the mean (1499.69) and median (1442) is almost the same of this continuous variable. However, there are many outliers on the higher side of the boxplot and are not randomly distributed. This reduces the normality of the data distribution and affects the mean.

Chart

Description automatically generated

Fig. 2: Boxplot for Ground Floor Living Area

In the next steps, we were required to assign the missing values with the variable’s mean. To accomplish this task, we first need to get all the variables containing NA/NULL values, and then filter numerical variables to assign variable’s mean to the missing values. Post imputing missing values with the variable’s mean value, the data set is now ready to be analyzed by correlation matrix. However, we need to take care of one important factor that the correlation can only be performed between the numerical variables. So if the dataset has categorical or factor values as well, we need to first filter the numerical variables and list it as numeric matrix. This can be performed by either of following two methods: sapply and lapply function. Albeit the above function provide us with correlation matrix, but due to the presence of too many numeric/integer variables in the dataset, the output is not making much sense. To refine the output and produce some meaningful correlation between variables and Sales Price, I decided to take variables which can have an impact on the Sales of the property. The data frame, Ames\_Cor was created to subset the original DF and store the selected variables. This DF is then converted into correlation matrix to be able to generate a correlation matrix plot through it. The Plot clearly prints out the visual representation of the correlation between variables. Higher the correlation between variables, darker the color intensity is. I have also displayed the coefficient value inside the correlation shape and shade to support the result. The plot does have few negative correlations between the variables and is exhibited by the lighter shade of red. It is followed that the positive relations are represented by the shades of blue. Sales Price has its highest correlation of 0.8 with the overall material quality and finish of the house. It records its lowest correlation with overall condition rating of the house having coefficient value as -0.1. The variable with which the Sales Price correlation coefficient is close to 0.5, is Masonry Veneer Area. The coefficient value comes out to be 0.51.

Chart

Description automatically generated

Fig. 3: Correlation Matrix Plot

However, we just need continuous variables for the next part of the assignment. Masonry Veneer Area, Basement Area, Garage Area, Ground Floor Living Area, Sale Price, Linear feet of street connected to property and the Lot size are continuous variables and will be considered for further analysis. Now, Sales Price of the properties has its highest correlation of 0.71 with the Ground Floor Living Area variable, lowest correlation of 0.27 with Lot Area size variable and close to 0.50 with Masonry Veneer Area having 0.51 correlation coefficient. Using this information, I plotted three scatter plots to illustrate the difference in pattern when different correlation coefficients are considered. Highest Correlation graph clearly exemplifies the regression line (red) is almost like data trend line (blue) and nearly overlapping each other. This concludes the variable can be considered as a best fit for this regression analysis. On the contrary, when we look at the lowest correlation graph there is no sort of relationship between the regression line and the trend line. The trend line itself is a curve cubic line and doesn’t match with linear regression model. Comparatively, there was still some relationship amongst variables in around 50% correlation model with sales price. The regression line slope is relatively steeper than the trend line and it keeps going far away from the trend line with the increase in horizontal values.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Fig. 4: Highest, Lowest and 50% Correlation with Sales Price

In the further step, I have selected 5 variables out of 7 to fit the best regression model of continuous variables. I excluded the two variables having the least of correlation with Sales Price which were, Lot Frontage and the Lot Area. The summary of the regression model gives us the adjusted R2 as 0.6964 and the correlation coefficients of all the predictor variables.

Text

Description automatically generated

Fig. 5: Summary of Regression Model

The regression model equation is y′= a + bx where a is the y′intercept, and b is the slope of the line. The formula to present the above regression model in mathematical equation form: SalePrice = -42005.356 + (Overall.Cond)(3806.922) + (Mas.Vnr.Area)(63.605) + (Total.Bsmt.SF)(50.165) + (Gr.Liv.Area)(63.846) + (Garage.Area) 98.639. Also, the AIC and BIC values are calculated as 70980.52 and 71022.4, respectively.

Sale price is the response variable, and the remaining arguments are predictor variables.

Plotting the regression model, we get 4 graphs described ahead. To fit all the four graphs side by side we used mfrow parameter function and plotted the graph.

Plot A) Residual vs Fitted Plot for linearity: The red line shows the rough trend of the data. We can see the observation or data points are scattered randomly around the diagonal dotted line, however there is an obvious pattern observed with the trend line where it starts of the lowest around the diagonal line but moves further away going ahead. Trend line and the diagonal line are closer to each other where the data points are more, hence this concludes that there are few extreme outliers in the model and the data points are not linear.

Plot B) Normal Q-Q Plot for Normality: Here we check and want the data points to be present on to the diagonal line or very close to it. We can see that few data points at the left and the right of the plot are further away from the line which strengthened the assumption of possible outliers or very extreme/unusual observation. The model is concluded to be not normal in this case.

Plot C) Scale Location for Constant variance or Homoscedasticity: We do not wish to see an obvious patterns again in this kind of plots and data points should be randomly distributed. Nevertheless, we can see an increasing pattern for our model and thus this doesn’t satisfy homoscedasticity conditions.

Plot D) Residuals vs Leverage: This is used to identify unusual observations with the data. And like earlier plots this also supports the point of few extreme observations in our dataset.

We observed a common yield from all the graphs and three data points 2181, 2182 and 1499 are possible data points subjected to further investigation.

Chart, scatter chart

Description automatically generated

Fig. 6: Regression Model Plot

To accomplish the task of checking the multi-collinearity in the model VIF function was used and the result is displayed below. We can see that all the values are below 5 which means there is no multi-collinearity in the model, and we are good to proceed with this model. The highest value is recorded for Garage Area. If there was any multi-collinearity found in the model, we would need to drop those variables from the model. Additionally, there is a need of handling heavily correlated variables and partial least squares is one of the techniques to leverage.



Fig. 7: Multi-Collinearity Check

We did an outlier test to find possible outliers in our regression model. The extreme outliers should be removed to improve the model. Not to our surprise, three data points which resulted in extreme outlier conditions were 1499, 2181 and 2182. These are the same data points which were reflected in regression model plot above.

Text

Description automatically generated

Fig. 8: Outlier test output

Further to be sure I performed two more analysis, high leverage observation using hat plot and influential observations using Cook’s Model. 1499, 2181 and 2182 data points were the common amongst all the three tests, hence the decision of removing these row numbers from the data set is made.

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

Fig. 9: Outlier test Using Hat and Cook’s Model Plot

The regression model was run again with corrected model (post removing outliers) this time and we can clearly see it improved the model as the adjusted R2 value went up to 0.7514 from 0.6964.

Further I just want to see what is going to happen to the regression model if in the data frame only the predictor variables and response variable is modelled removing all other unnecessary variables. Hence, the data frame with 6 variables was made (1 response, 5 predictor). The adj R2 for this model comes out to b 0.7514 as observed earlier. Using the subset regression model, we tried to fit the best regression model. Ground Living area, Garage area and the total basement area comes out to be the best variables in the model. And when this best fit model was tested with these three best predictor variables, the adj R2 comes out to be 0.7323, which was less than the Best fit model with 5 predictor variables and more than the entire dataset adj R2. Hence, to conclude that I made the model worst while using only three best predictor variables (achieved via subset regression). This sometimes happen with the model and hence I will continue to use five predictor variables.

# Conclusion

1. We saw from the exercise that regression model is used to identify the best predictor variables related to the response variable to create the best fit model.
2. Correlation model presents the correlation coefficient between variables to determine how the two variables are affecting each other.
3. It is not always true that the transformation or subset regression analysis will improve the model, and is subject to different datasets and data point being analysed.

# Reference

1. Goyal, V. (2021, June 29). Calculate correlation matrix only for numeric columns in R. GeeksforGeeks. Retrieved January 22, 2022, from <https://www.geeksforgeeks.org/calculate-correlation-matrix-only-for-numeric-columns-in-r/>
2. Interpret the key results for Scatterplot. Minitab Express. (n.d.). Retrieved January 22, 2022, from <https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/scatterplot/interpret-the-results/key-results/>
3. Schork, J. (2021, October 15). Common main title for multiple plots in base R &amp; GGPLOT2 (2 examples). Statistics Globe. Retrieved January 22, 2022, from <https://statisticsglobe.com/common-main-title-for-multiple-plots-in-r>

# Appendix

# ALY6015 Module 1 R Practice: Singh Prateek ------------------------------------------------

#----------------- Author: Prateek Singh

#----------------- Submission Date: 18th Jan, 2022

#----------------- Tutor: Jiyoung Yun

# Step: Installing New Libraries ------------------------------------------------

install.packages("RColorBrewer")

install.packages("leaps")

# Step: Importing Libraries ------------------------------------------------

library(dplyr)

library(ggplot2)

library(corrplot)

library(RColorBrewer)

library(car)

library(psych)

library(ggpubr)

library(leaps)

library(MASS)

# Step: Load the Dataset and display summary ------------------------------------------------

Ames <- read.csv('AmesHousing.csv', header = TRUE)

Ames <- Ames %>% rename(Order = ï..Order, '1st.Flr.SF' = X1st.Flr.SF, '2nd.Flr.SF' = X2nd.Flr.SF)

str(Ames)

summary(Ames)

describe(Ames)

# Step: Checking normality of the variables and check if there are any outliers ------------------------------------------------

hist(Ames$SalePrice,

xlab = "Property Sales Price ($)", # Changing X-Axis Label

main = "Plot 1: Sales Price of AMES Housing Property", # Adding Title to the Scatter Plot

col.main = "Brown", # Changing Color of the Title

col.lab = "dark blue", # Color of the X and Y Axis labels

col= "cadetblue") # Color of the Histogram/Frequency Plot

axis(1,

col = "blue", # Axis line color

col.ticks = "green", # Ticks color

col.axis = "red") # Labels color

axis(2,

col = "blue", # Axis line color

col.ticks = "green", # Ticks color

col.axis = "red") # Labels color

describe(Ames$SalePrice)

ggboxplot(Ames$Gr.Liv.Area,

ylab = "Ground Living Area (Sq. Feet)",

xlab = FALSE,

color = "cyan4",

title = "Plot 2: Ground Living Area Boxplot",

bxp.errorbar = TRUE,

notch = TRUE,

fill = "#ff355e",

ggtheme = theme\_minimal())

describe(Ames$Gr.Liv.Area)

# Step: Imputing missing values with the variable's mean value ------------------------------------------------

colSums(is.na(Ames)) # To find NA values in dataset

Ames$Lot.Frontage[is.na(Ames$Lot.Frontage)] <- mean(Ames$Lot.Frontage,na.rm=TRUE)

Ames$BsmtFin.SF.1[is.na(Ames$BsmtFin.SF.1)] <- mean(Ames$BsmtFin.SF.1,na.rm=TRUE)

Ames$BsmtFin.SF.2[is.na(Ames$BsmtFin.SF.2)] <- mean(Ames$BsmtFin.SF.2,na.rm=TRUE)

Ames$Bsmt.Unf.SF[is.na(Ames$Bsmt.Unf.SF)] <- mean(Ames$Bsmt.Unf.SF,na.rm=TRUE)

Ames$Total.Bsmt.SF[is.na(Ames$Total.Bsmt.SF)] <- mean(Ames$Total.Bsmt.SF,na.rm=TRUE)

Ames$Bsmt.Full.Bath[is.na(Ames$Bsmt.Full.Bath)] <- mean(Ames$Bsmt.Full.Bath,na.rm=TRUE)

Ames$Bsmt.Half.Bath[is.na(Ames$Bsmt.Half.Bath)] <- mean(Ames$Bsmt.Half.Bath,na.rm=TRUE)

Ames$Garage.Yr.Blt[is.na(Ames$Garage.Yr.Blt)] <- mean(Ames$Garage.Yr.Blt,na.rm=TRUE)

Ames$Garage.Cars[is.na(Ames$Garage.Cars)] <- mean(Ames$Garage.Cars,na.rm=TRUE)

Ames$Garage.Area[is.na(Ames$Garage.Area)] <- mean(Ames$Garage.Area,na.rm=TRUE)

Ames$Mas.Vnr.Area[is.na(Ames$Mas.Vnr.Area)] <- mean(Ames$Mas.Vnr.Area,na.rm=TRUE)

# Step: Producing a correlation matrix of the numeric values ------------------------------------------------

options(max.print=1000000)

cor(Ames[sapply(Ames,is.numeric)])

cor(Ames[, unlist(lapply(Ames, is.numeric))])

# Creating a DF to have only required and effective numerical variables

Ames\_Cor <- data.frame(Ames$SalePrice,

Ames$Lot.Frontage,

Ames$Lot.Area,

Ames$Overall.Qual,

Ames$Overall.Cond,

Ames$Mas.Vnr.Area,

Ames$Total.Bsmt.SF,

Ames$Gr.Liv.Area,

Ames$Garage.Area)

Ames\_Cor <- cor(Ames\_Cor) # Converting DF to Correlation Matrix

# Step: Producing a Correlation Matrix Plot ------------------------------------------------

corrplot(Ames\_Cor,

order = "hclust",

hclust.method = "single",

method = "shade",

addCoef.col = "Dark orange",

title = "Plot 3: Correlation Matrix for AMES Housing with respect to Sales Price",

mar = c(1,1,2,1))

# Step: Plotting scatter plot for the continuous variable with the highest, lowest and close to 0.50 correlation with SalePrice ------------------------------------------------

# Highest Correlation with Sales Price

scatterplot(SalePrice ~ Gr.Liv.Area,

data = Ames,

axes = FALSE,

main = "Plot 4: Scatterplot for the Highest Correlation with Sales Price",

xlab = "Ground Floor Living Area (Sq. Feet)",

ylab = "Sales Price ($)",

col.lab = "dark blue",

regLine = list(method = lm, lty = 4, lwd = 2, col = "Red"),

grid = FALSE,

smooth = TRUE,

cex = 0.75,

cex.main = 1.25,

cex.lab = 1,

col.main = "Brown")

axis(1,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

axis(2,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

# Lowest Correlation with Sales Price

scatterplot(SalePrice ~ Lot.Area,

data = Ames,

axes = FALSE,

main = "Plot 5: Scatterplot for the Lowest Correlation with Sales Price",

xlab = "Lot Area Size (Sq. Feet)",

ylab = "Sales Price ($)",

col.lab = "dark blue",

regLine = list(method = lm, lty = 4, lwd = 2, col = "Red"),

grid = FALSE,

smooth = TRUE,

cex = 0.75,

cex.main = 1.25,

cex.lab = 1,

col.main = "Brown")

axis(1,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

axis(2,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

# ~ 50% Correlation with Sales Price

scatterplot(SalePrice ~ Mas.Vnr.Area,

data = Ames,

axes = FALSE,

main = "Plot 6: Scatterplot for around 50% Correlation with Sales Price",

xlab = "Masonry Veneer Area (Sq. Feet)",

ylab = "Sales Price ($)",

col.lab = "dark blue",

regLine = list(method = lm, lty = 4, lwd = 2, col = "Red"),

grid = FALSE,

smooth = TRUE,

cex = 0.75,

cex.main = 1.25,

cex.lab = 1,

col.main = "Brown")

axis(1,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

axis(2,

col = "blue",

col.ticks = "green",

col.axis = "red",

cex.axis = 0.75)

# Step: Fitting a regression model using continuous variable ------------------------------------------------

Ames\_Reg\_Model <- lm(formula = SalePrice ~ Overall.Cond +

Mas.Vnr.Area +

Total.Bsmt.SF +

Gr.Liv.Area +

Garage.Area, data = Ames)

summary(Ames\_Reg\_Model)

AIC(Ames\_Reg\_Model)

BIC(Ames\_Reg\_Model)

# Step: Plotting the regression model and interpreting the graphs produced ------------------------------------------------

par(mfrow = c(2,2))

plot(Ames\_Reg\_Model)

mtext("Plot 7: AMES Housing Regression Model", side=3, line = -2, outer=TRUE, col = "Brown", cex = 1.5)

dev.off()

# Step: Checking model for multi-collinearity ------------------------------------------------

vif(Ames\_Reg\_Model)

# Step: Checking model for outliers ------------------------------------------------

outlierTest(model = Ames\_Reg\_Model)

# High leverage observation

hat.plot <- function(Ames\_Reg\_Model) {

p <- length(coefficients(Ames\_Reg\_Model))

n <- length(fitted(Ames\_Reg\_Model))

plot(hatvalues(Ames\_Reg\_Model), main = "Hat Values Indexing")

abline(h = c(2,3)\*p/n, col = "Dark Orange", lty = 4)

identify(1:n, hatvalues(Ames\_Reg\_Model), names(hatvalues(Ames\_Reg\_Model)))

}

hat.plot(Ames\_Reg\_Model)

# Influential Observations

cutoff <- 4/(nrow(Ames) - length(Ames\_Reg\_Model$coefficients) - 2)

plot(Ames\_Reg\_Model, which = 4, cook.levels = cutoff)

abline(h = cutoff, lty = 2, col = "Dark Orange")

# Step: Removing outliers ------------------------------------------------

Ames\_Corrected <- Ames[-c(1499, 2181, 2182),]

Ames\_Corrected

# Running Regression Model post removing outliers

Ames\_Corrected\_Reg\_Model <- lm(formula = SalePrice ~ Overall.Cond +

Mas.Vnr.Area +

Total.Bsmt.SF +

Gr.Liv.Area +

Garage.Area, data = Ames\_Corrected)

summary(Ames\_Corrected\_Reg\_Model)

# Step: Using the all subsets regression method to identify the “best” model. ------------------------------------------------

Ames\_Corrected <- data.frame(Ames\_Corrected$SalePrice,

Ames\_Corrected$Overall.Cond,

Ames\_Corrected$Mas.Vnr.Area,

Ames\_Corrected$Total.Bsmt.SF,

Ames\_Corrected$Gr.Liv.Area,

Ames\_Corrected$Garage.Area)

colnames(Ames\_Corrected)[1] <- "SalePrice"

colnames(Ames\_Corrected)[2] <- "Overall.Cond"

colnames(Ames\_Corrected)[3] <- "Mas.Vnr.Area"

colnames(Ames\_Corrected)[4] <- "Total.Bsmt.SF"

colnames(Ames\_Corrected)[5] <- "Gr.Liv.Area"

colnames(Ames\_Corrected)[6] <- "Garage.Area"

Best\_Model <- regsubsets(SalePrice ~ ., data = Ames\_Corrected, nbest = 5)

summary(Best\_Model)

Best\_Model\_Summary <- summary(Best\_Model)

data.frame(

Adj.R2 = which.max(Best\_Model\_Summary$adjr2),

CP = which.min(Best\_Model\_Summary$cp),

BIC = which.min(Best\_Model\_Summary$bic)

)

Ames\_BestFit\_Reg\_Model <- lm(formula = SalePrice ~ Total.Bsmt.SF +

Gr.Liv.Area +

Garage.Area, data = Ames\_Corrected)

Ames\_BestFit\_Reg\_Model <- regsubsets(SalePrice ~ Total.Bsmt.SF +

Gr.Liv.Area +

Garage.Area, data = Ames\_Corrected, nbest = 5)

summary(Ames\_BestFit\_Reg\_Model)

Ames\_BestFit\_Reg\_Model\_Summary <- summary(Ames\_BestFit\_Reg\_Model)

data.frame(

Adj.R2 = which.max(Ames\_BestFit\_Reg\_Model\_Summary$adjr2),

CP = which.min(Ames\_BestFit\_Reg\_Model\_Summary$cp),

BIC = which.min(Ames\_BestFit\_Reg\_Model\_Summary$bic)

)