**Problem 3 – All Greens Franchise**

**Problem Statement:** We have the **All greens franchise** dataset in which we need to find how the independent variables X2,X3,X4,X5,X6 affect the dependent variable X1(Sales).

The following variables are used in the data (each observation represents one store):

X1 = annual net sales/$1000

X2 = number sq. ft./1000

X3 = inventory/$1000

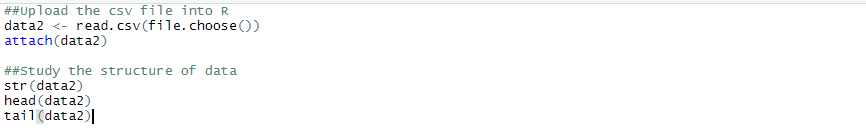
X4 = amount spent on advertising/$1000

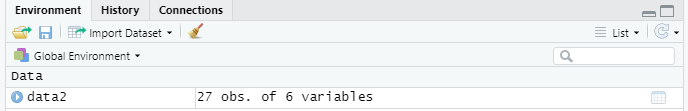
X5 = size of sales district/1000 families

X6 = number of competing stores in district

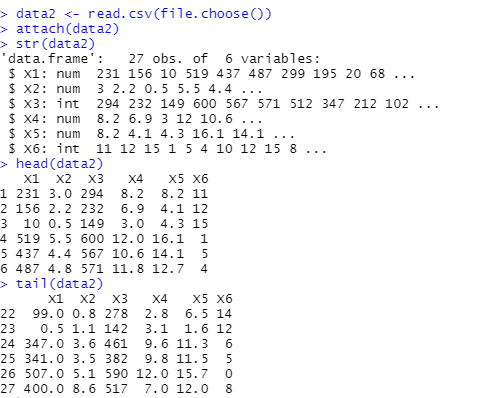
**Objective**: Our objective here is to analyse how different factors affect the sales at All Greens Franchise. Some inter-factor relations will be analysed as well on for more detail.

Let us explore the data in R.





The data set has 27 observations of 6 variables. The structure of the data reveals them to be numeric and integers



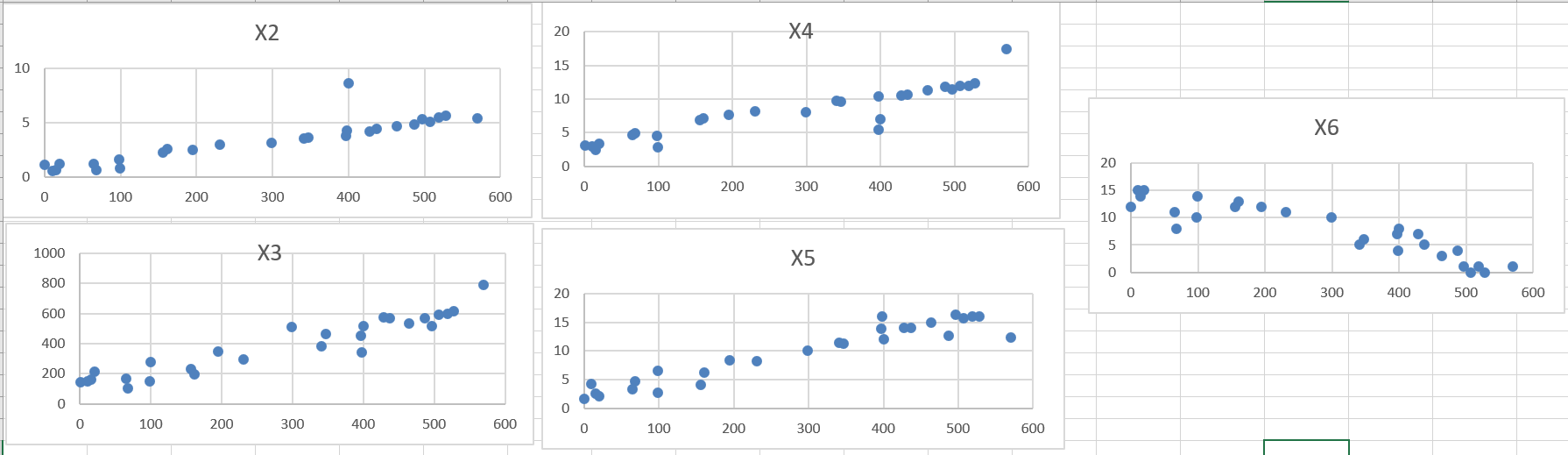
To find how the variables X2, X3, X4, X5, and X6 affect the performance of X1, we can do a regression analysis.

Before going ahead with the regression let us check if the dataset satisfies all the regression Analysis assumptions:

**Assumptions:**

1.**Linear relationship**: There should be linear relationship between independent and the dependent variable.(X1,X2),(X1,X3),(X1,X4),(X1,X5),(X1,X6)

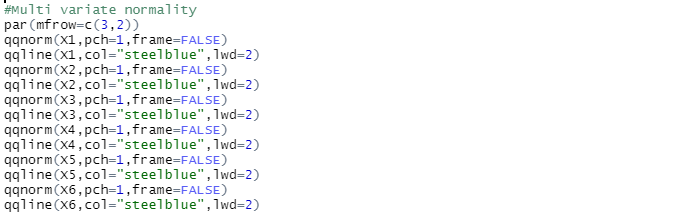
Let us check the same by scatterplot:

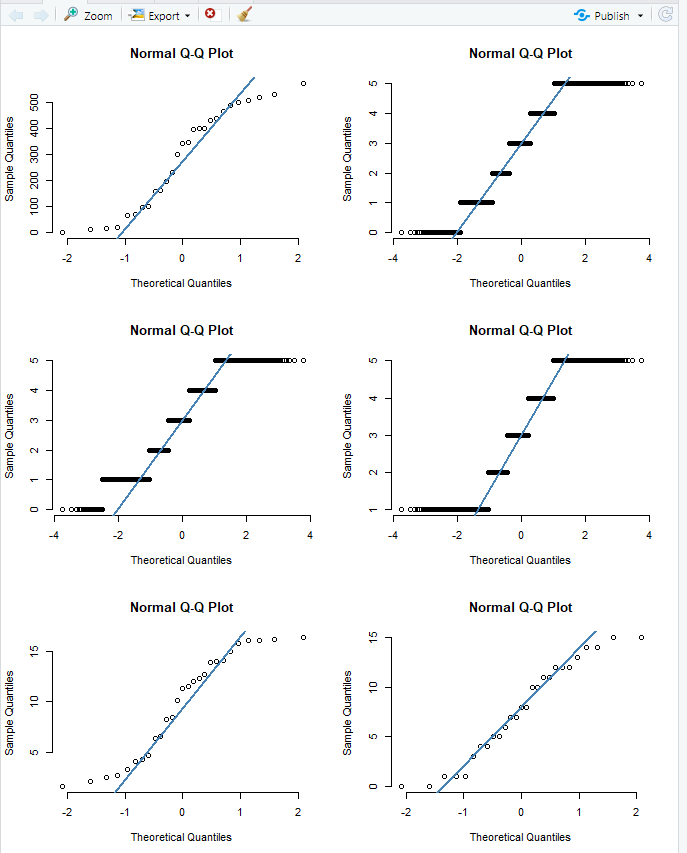


We can see linearity between the independent and the dependent variable here.

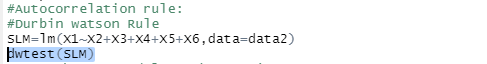
**2. Multivariate normality**: All the variables need to be normal

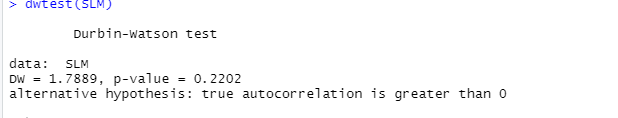
Let us check the normality through the qqplot:





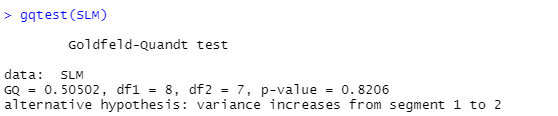
3. **Autocorrelation rule**: Let us check this with the Durbin Watson Rule in which the null hypothesis states that residuals are not linearly auto-correlated





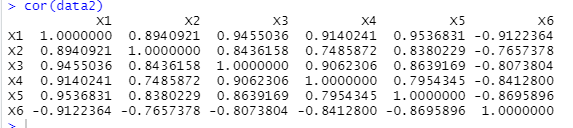
There is no autocorrelation. Since calculated DW value (1.7) lies between 1.5 and 2.5, it shows that there is no auto-correlation

**4. Homoscedasticity:** We need to check if the residuals are equal across the regression line. We can perform Goldfield-Quant Test:

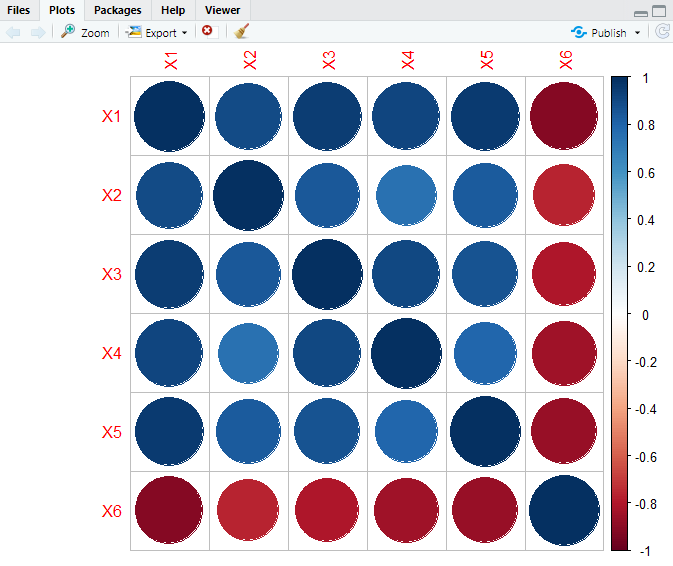


Since calculated P>0.05 which means we accept null hypothesis, that there is homoscedasticity.

**5.Multicollinearity:** We need to check if there is no correlation between the independent variables



Let us see the visual representation of this correlation through corrplot.



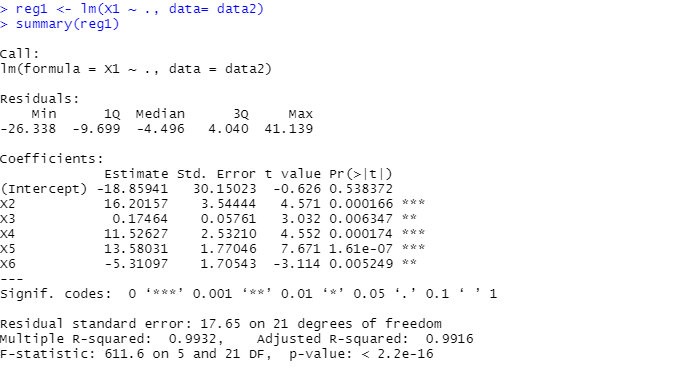
From both the calculation and the plot we can see that that there is multi collinearity between the independent variables and the dependent variable (sales)

Let us actually find which particular variable is creating this multi collinearity by calculating the Variance Inflation Factor (VIF).For this we should first have the regression model in place.

Performing the regression model using all the variables.

**Regression Model 1**: Let us formulate the hypothesis as below:

* **Null hypothesis (Ho):** Net sales (X1) is not dependent on independent variables: X2, X3, X4, X5, or X6.
* **Alternate hypothesis (H1):** Net sales (X1) is dependent on independent variables.

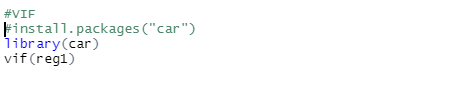


**Observation:**

We can see from the model that Multiple R-squared value is 0.9932 / Adjusted R squared value is 0.9916. This indicates that there is a strong relation between the sales and other variables, explaining little more than 99% of variation in sales.

The p value of all the independent variables is also significantly less than significance level(0.05), So, we can reject the null hypothesis and say- Net sales (X1) is dependent on independent variables.

But, we also need to take care of the multi collinearity through VIF (Variance Inflation Factor) as suggested above. .Generally, if the VIF value is above 10, the variable is said to have multi -collinearity





From the above VIF values we could see that the variable X3(Inventory) is causing the multi collinearity as the value is >10.

**So, how can we actually deal with this multi collinearity?**

We can:

* Perform the centering of data by deducting the mean of the variable from each score
* Remove highly correlated predictors from the model- Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared.
* Use methods like component analysis, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

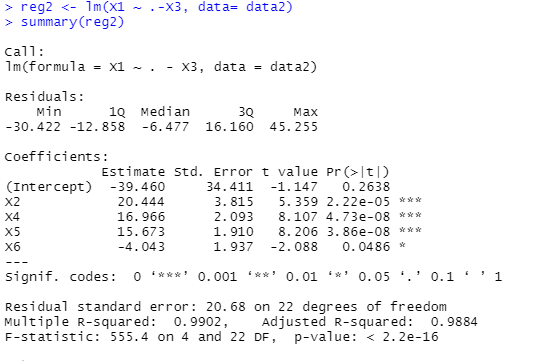
Simplest way among the methods listed above is to remove the independent variable with high VIF value (X3) and again perform the regression analysis.

Regression after excluding redundant variable (X3):

**Regression Model 2:**

Let us formulate the hypothesis as below:

* **Null hypothesis (Ho):** Net sales (X1) is not dependent on independent variables: X2, X4, X5, or X6.
* **Alternate hypothesis (H1):** Net sales (X1) is dependent on independent variables.



**Observation**: From the above regression model we can see that Multiple R-squared: 0.99 and adjusted R-squared: 0.9884 value is pretty large which states that these variables (X2,X4,X5,X6) together explain more than 98% of the model. The pvalue of the overall model and each variable is also pretty low. So we would reject the null hypothesis and state that net sales is highly dependent on the independent variables.

Let us re-run the VIF to check the multi collinearity of this model.



From the above VIF result we can say that there is no effect of multi collinearity after the removal of X3 variable.

**Interpretation:**

Let us interpret the data on the basis of the 2 regression models which we have executed.

**Regression model 1:**

The regression equation for the model1 is as follows:

Y=-18.85+16.20X2+0.17X3+11.52X4+13.58X5-5.31X6

-Let us see how the dependent variables individually effect the sales variable

X2 = Area is directly correlated and explains about 16.20 percent of the sales data

X3 = Inventory explains about 0.17 percent of the sales data

X4 = amount spent on advertising has around 11.52 percent of sales

X5 = size of sales district also explains around 13.58 percent of sales generation

X6 = number of competing stores in district, explains the sales variable little more than 5 percent negatively.

The above equation represents the same information as reflected in the linear plot. All the variables have positive correlation except X6, which is true because the competition would actually reduce the store sales. However, we also notice that the Inventory variable which was causing the multi collinearity has the least effect on the sales.

This model says that all the independent variables together explain more than 99% of the dependent variable.

**Regression model 2:**

The regression equation for the model2 is as follows:

Y=-39.460+20.44X2+16.96X4+15.67X5-4.04X6

In the regression model 2 we could see that even after dropping the single variable the r square .9902/adjusted r square .9884 is still quite high .Which means that the dependent variables could explain about 99 % of the sales data.

-Let us see how the above independent variables individually effect the sales variable

X2 = Area co-efficient has increased and now explains 20.44 percent of the sales data

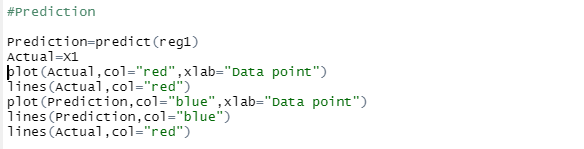
X4 = amount spent on advertising: Coefficient has also increased to 16.96 percent of sales

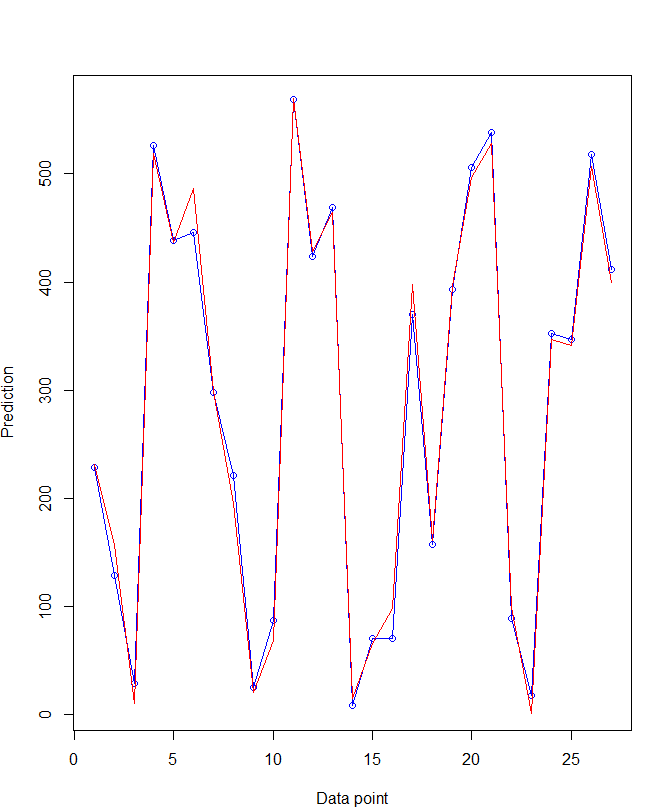
X5 = size of sales district also explains around 15.67 percent as the number of the families increase and population density would lead to sales generation

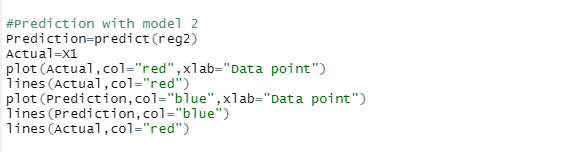
X6 = number of competing stores in district; explains the sales variable little more than 4 percent negatively .Which is rightly explained as the competition would increase with the increase in the neighbouring stores.

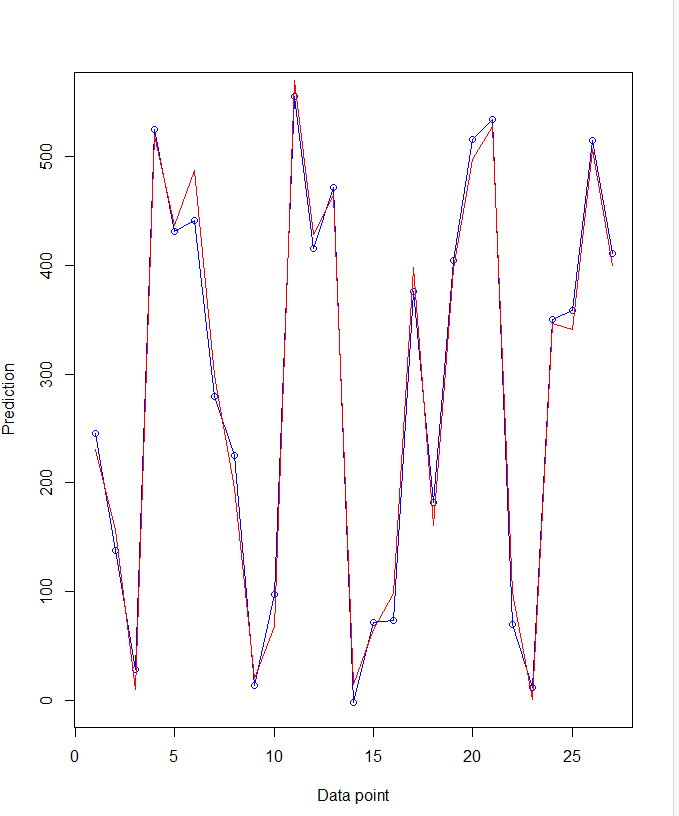
**Conclusion**: While performing the regression analysis ,we had the problem of the multi- collinearity which we chose to resolve by eliminating the redundant variable, as the contribution of that variable in the overall model was only 0.17 %.So, we had the regression analysis run again after dropping the inventory variable (this variable was causing the multi-collinearity-VIF)..Even after dropping the variable we could see that our model still holds great because the dependent variables were able to explain around 99% of our sales data.

So, we can say multicollinearity affects the coefficients and the p value, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit. So let us also check the backtrack model here.









**Conclusion**: With effect of each variable explained in the models we can also see that the prediction of our both models explains the goodness of fit. Hence, our model hold good explaining the relation between all the independent and the dependent variable.