Multiple Regression

By Saba Akram

1/19/2021

# Input and prepare Cereals data.  
cereal <- read.csv("D:/M.Sc in Banking and Financial Analytics/Sem 3/Data Analytic and Machine learning/Data Mining and Predictive Analysis/Data sets/cereals.CSV",stringsAsFactors = TRUE,header = TRUE)  
head(cereal)

## Name Manuf Type Calories Protein Fat Sodium Fiber Carbo  
## 1 100%\_Bran N C 70 4 1 130 10.0 5.0  
## 2 100%\_Natural\_Bran Q C 120 3 5 15 2.0 8.0  
## 3 All-Bran K C 70 4 1 260 9.0 7.0  
## 4 All-Bran\_with\_Extra\_Fiber K C 50 4 0 140 14.0 8.0  
## 5 Almond\_Delight R C 110 2 2 200 1.0 14.0  
## 6 Apple\_Cinnamon\_Cheerios G C 110 2 2 180 1.5 10.5  
## Sugars Potass Vitamins Shelf Weight Cups Rating Cold Nabisco Quaker  
## 1 6 280 25 3 1 0.33 68.40297 1 1 0  
## 2 8 135 0 3 1 1.00 33.98368 1 0 1  
## 3 5 320 25 3 1 0.33 59.42551 1 0 0  
## 4 0 330 25 3 1 0.50 93.70491 1 0 0  
## 5 8 NA 25 3 1 0.75 34.38484 1 0 0  
## 6 10 70 25 1 1 0.75 29.50954 1 0 0  
## Kelloggs GeneralMills Ralston AHFP  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 0 0 1 0  
## 6 0 1 0 0

which(is.na(cereal$Sugars))

## [1] 58

## Here we observe that record 58 is not available. Hence we will omit record 58 from the cereal data. Therefore we have :-  
  
cereal<-cereal[-58]  
  
## Now after removing the row with the missing value we will now create the dataframe of the "Ratings", "Sugars" and "Fibre" with "Ratings" being dependent variable in applying the multiple Regession. Therefore we have :-  
  
dat<-data.frame(Ratings=cereal$Rating, Sugars=cereal$Sugars, Fibre=cereal$Fiber)  
  
## Data frame has already been created with three variables i.e. Rating which will be dependent variable and Sugars and Fiber which is dependent variable..  
  
head(dat)

## Ratings Sugars Fibre  
## 1 68.40297 6 10.0  
## 2 33.98368 8 2.0  
## 3 59.42551 5 9.0  
## 4 93.70491 0 14.0  
## 5 34.38484 8 1.0  
## 6 29.50954 10 1.5

## Here we will test the relationship among the varibles using the multiple Regression ahead.

# Three Variable Scatter plot  
## We will now try to observe the pattern among variables by using scatterplot.  
  
library(scatterplot3d)  
##rg<-colorRampPalette(c("red","green"))(76)  
##sp<-scatterplot3d(z=sort(cereal$Rating),y=cereal$Sugars,x=cereal$Fiber,color = rg,pch = 16,xlab = "Fiber",ylab = "Sugars",zlab = "Ratings",main = "3D Scatterplot")

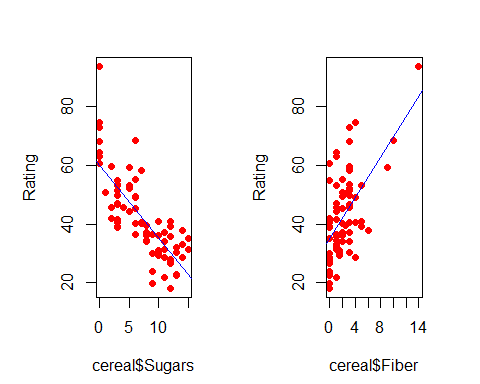
# Individual Variable Scatter plots of Ratings vs. Sugar and Fiber.  
## Now we will check the relationship between the Ratings vs. Sugars and the fiber individually in order to obtain the individual relationship between them by rnning the simple linear regression.  
  
par(mfrow=c(1,2),mar=c(4.5,4,3,3),oma=c(0,1,0,0))  
lm91<-lm(Rating~Sugars,data = cereal)  
lm92<-lm(Rating~Fiber,data = cereal)  
lm91

##   
## Call:  
## lm(formula = Rating ~ Sugars, data = cereal)  
##   
## Coefficients:  
## (Intercept) Sugars   
## 59.853 -2.461

## Here by running the linear regression analysis, we observe that its intercept is 59.853 and its slope is -2.461. The slope of -2.461 indicates that the increase in 1 unit of sugar amount will lead to the decrease in the rating amount by a unit of 2.461.  
  
lm92

##   
## Call:  
## lm(formula = Rating ~ Fiber, data = cereal)  
##   
## Coefficients:  
## (Intercept) Fiber   
## 35.257 3.443

## Here by running the simple linear regression Analysis, we observe that we have 35.257 as the intercept and slope as 3.443. The slope of 3.443 says that there is the direct relationship between the Rating of the cereal and the fiber content. According to the analysis we get that, as we increase the fiber content of 1 unit, there is an increase of 3.443 units of ratings for that cereal.  
  
## Now lets plot these to observe the relation over the graphs.  
  
plot(cereal$Rating~cereal$Sugars,pch=16,col="red",ylab = "Rating")  
abline(lm91,col="blue") ### To show the line.  
plot(cereal$Rating~cereal$Fiber,pch=16,col="red",ylab = "Rating")  
abline(lm92,col="blue")



## After running scatter plots for the two individual variables with the dependent variable i.e. Ratings we observe the same as we observed in the regression model, i.e. Sugar is directly related whereas Fiber is indirectly related.  
  
## Now reset the plot area.  
  
par(mfrow=c(1,1))

## Now Coming for Multiple Regression  
  
mreg1<-lm(Rating~Sugars+Fiber,data = cereal)  
summary(mreg1)

##   
## Call:  
## lm(formula = Rating ~ Sugars + Fiber, data = cereal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.159 -4.415 -1.151 2.584 16.732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52.1742 1.5556 33.541 < 2e-16 \*\*\*  
## Sugars -2.2436 0.1632 -13.750 < 2e-16 \*\*\*  
## Fiber 2.8665 0.2979 9.623 1.28e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.127 on 73 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.8164, Adjusted R-squared: 0.8114   
## F-statistic: 162.3 on 2 and 73 DF, p-value: < 2.2e-16

ma1<-anova(mreg1)  
ma1

## Analysis of Variance Table  
##   
## Response: Rating  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Sugars 1 8711.9 8711.9 232.045 < 2.2e-16 \*\*\*  
## Fiber 1 3476.6 3476.6 92.601 1.276e-14 \*\*\*  
## Residuals 73 2740.7 37.5   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## After running the multiple Regression with two predictor variables i.e. Sugars and Fiber, and a target variable and taking their summary we get that the estimated nutritional ratings equals 52.1742 minus 2.2436 times the grams of sugars plus 2.8665 times the grams of fibers. We here note that the coefficient of sugar is negative indicating an inverse relationship between sugars and Ratings while the coefficient of fiber is positive indicating the direct relationship between the Ratings and Fibers.  
  
## The interpretations of the slope coefficients b1 and b2 are slightly different than for the simple linear regression case. For example, to interpret b1=-2.2436, we say that "The estimated decrease in nutritional Ratings for a unit increase in sugar content is 2.2436 points, when fiber content is held constant."  
## Similarly we interpret b2=2.8665 as follows, "The estimated increase in nutritional Ratings for a unit increase in fiber content is 2.8665 points, when fiber content is held constant"  
  
## These results concur with the plots which we did earlier.  
  
## Now coming to the P value. By observing the P values for each of the variables both the target variable and the two predictor variables, we have all the P values close to 0. The P values close to 0 indicates to reject the null hypothesis i.e. b1 and b2 are 0. Hence it shows that some relationship exists between target variable and the two predictor variables.  
  
## Also we have high F statistics of 162.3 on 2 and 73 degrees of freedom indicating the significance of the model.  
  
## Now talking about R-Square also called Coefficient of Determination, which is simply SSR/SST. R-Square can be interpreted as the proportion of variability in the target variable that is accounted for by its linear relationship with the set of predictor variables. After analyzing the summary we observe that the value of R-Square is 0.8164 i.e. 81.64% which means that 81.64% of the variability in the nutritional rating is accountes for by the linear relationship (the Plane) between Rating and the set of predictors, sugar content and fiber content.  
  
## Now when we look at the Residual Standard Error i.e. s, we observe that its value is 6.127 rating points. Therefore our estimation of the nutritional ratings of the cereals, based on sugar and the fiber content, is typically in error by 6.127 points.  
  
  
## Overall we have a significant Regression Model with low P value, high F Statistics, high R-square and low Residual Standard Error.

# Calculating the Confidence Intervals.  
## Confidence Intervals for Beta Coefficients.  
  
confint(mreg1,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 49.074025 55.274460  
## Sugars -2.568736 -1.918369  
## Fiber 2.272853 3.460221

## Mathematically, the formula for calculating the confidence interval of b (beta) is b(i)+-(t(n-m-1))(sb(i)), where t(n-m-1) is based on (n-m-1) degrees of freedom, and sb(i)represents the standard error of the ith coefficient estimate.  
## After constructing 95% confidence interval for the true value of the coefficient of sugar content whose point estimate is b1=-2.2436, we get -2.568736 and -1.918369 as the lower bound and the upper bound respectively. We say that we are 95% confident that the value for the coefficient of sugar lies between -2.568736 and -1.918369.  
## Similarly for the Fiber, we have 2.272853 and 3.460221 as the confidence intervals and for the intercepts, we have confidence intervals of 49.074025 and 55.274460.  
  
## Constructing the confidence interval at y=y\*.  
  
predict(mreg1,newdata = data.frame(Sugars=5,Fiber=5),interval = c("confidence"))

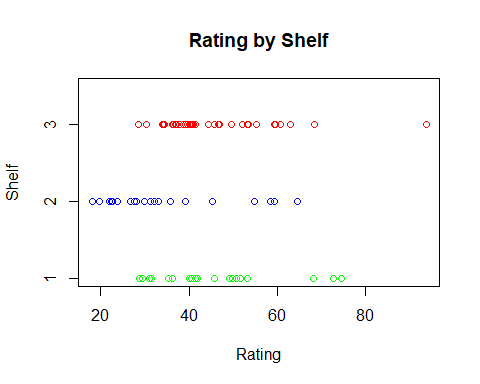
## fit lwr upr  
## 1 55.28916 53.06209 57.51623

## By Constructing the confidence interval we get 53.06209 and 57.274460 as the lower interval and upper interval respectively.  
  
## Calculating the Prediction Interval for the given value of y  
  
predict(mreg1,newdata = data.frame(Sugars=5,Fiber=5),interval = c("prediction"))

## fit lwr upr  
## 1 55.28916 42.876 67.70233

## By Constructing the confidence interval we get 42.876 and 67.70233 as the lower interval and upper interval respectively.

#Dotplot of Rating by Shelf.  
  
## We will create the indicator variables to understand it to the further level.  
  
cereal$Shelf1<-ifelse(cereal$Shelf==1,1,0)  
cereal$Shelf2<-ifelse(cereal$Shelf==2,1,0)  
stripchart(Rating~Shelf,data = cereal,method="stack",pch=1,col=c("green","blue","red"),main="Rating by Shelf",offset=0.5,ylab="Shelf")



## After observing the abve stripchart, we expect the coefficient of shelf2 (slope) to be negative, because shelf2 cereals have a lower mean rarings compared to shelf3. Similarly we can also expect coefficient of shelf 1 to be essentially negligible but slightly positive(slpe) because the mean rating of Shelf 1 is slightly greater compared t shelf 3.  
  
## Also this is an indipendent stripchart and doesnot take into account the presence of other variables, such as sugar content and fiber content.  
  
## In regression a categrical variables with k categories must be transformed into k-1 indicator variables.

# Regression using shelf effect.  
  
## Now we will run the Regression Analysis including all the Shelf Effect also.  
mreg2<-lm(Rating~Shelf1+Shelf2,data = cereal)  
summary(mreg2)

##   
## Call:  
## lm(formula = Rating ~ Shelf1 + Shelf2, data = cereal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.403 -8.749 -4.205 5.447 48.485   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.2200 2.2325 20.256 < 2e-16 \*\*\*  
## Shelf1 0.9254 3.7356 0.248 0.80503   
## Shelf2 -10.2472 3.6780 -2.786 0.00677 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.39 on 74 degrees of freedom  
## Multiple R-squared: 0.1147, Adjusted R-squared: 0.09075   
## F-statistic: 4.793 on 2 and 74 DF, p-value: 0.01103

anova(mreg2)

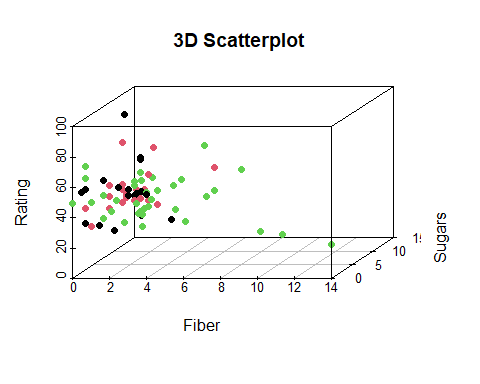
## Analysis of Variance Table  
##   
## Response: Rating  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Shelf1 1 327.1 327.14 1.8234 0.181030   
## Shelf2 1 1392.7 1392.70 7.7623 0.006773 \*\*  
## Residuals 74 13277.0 179.42   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Here we note that for the shelf 2 dummy variable, its coefficient is -10.247, which will be equal to the difference in the mean Nutritional Ratings between cereals on shelf 2 and shelf 3. Similarly the coefficients for the shelf 1 dummy variable is 0.9254 which will be equal to the difference in the mean ratings between cereals on shelf 1 and shelf 3.  
  
## Also we have lower R-Square and low F-Statistics.  
  
## Regressing the model using Sugar, Fiber, Shelf1, Shelf2.  
  
mreg3<-lm(Rating~Sugars+Fiber+Shelf1+Shelf2,data = cereal)  
summary(mreg3)

##   
## Call:  
## lm(formula = Rating ~ Sugars + Fiber + Shelf1 + Shelf2, data = cereal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.7512 -4.3085 -0.6918 2.9774 17.4020   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.5245 1.8512 27.293 < 2e-16 \*\*\*  
## Sugars -2.3183 0.1729 -13.409 < 2e-16 \*\*\*  
## Fiber 3.1314 0.3186 9.827 7.08e-15 \*\*\*  
## Shelf1 2.1011 1.7948 1.171 0.2457   
## Shelf2 3.9154 1.8646 2.100 0.0393 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.021 on 71 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.8276, Adjusted R-squared: 0.8179   
## F-statistic: 85.21 on 4 and 71 DF, p-value: < 2.2e-16

## After analyzing the Multiple Regression Summary,we get that cereals on each shelf are modeled as following the exact same slope in the sugars dimension(-2.3183) and the exact same slope in the fiber dimension(3.1314), which reveals the three parallel planes. The only difference lies in the value of the y-intercept for the cereals on the three shelves.

# 3D Scatter plots with groups.  
library(scatterplot3d)  
sp<-scatterplot3d(z=sort(cereal$Rating),y=cereal$Sugars,x=cereal$Fiber,color=cereal$Shelf,pch=16,xlab="Fiber",ylab="Sugars",zlab="Rating",main="3D Scatterplot")



# Sequential Sum of Squares (SS)  
mreg4.1<-lm(Rating~Sugars+Fiber+Shelf1+Shelf2,data = cereal)  
anova(mreg4.1)

## Analysis of Variance Table  
##   
## Response: Rating  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Sugars 1 8711.9 8711.9 240.3194 < 2.2e-16 \*\*\*  
## Fiber 1 3476.6 3476.6 95.9034 8.184e-15 \*\*\*  
## Shelf1 1 7.0 7.0 0.1935 0.66137   
## Shelf2 1 159.9 159.9 4.4096 0.03929 \*   
## Residuals 71 2573.9 36.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## We already know that the SSR(Sum of Square of Regression) represents the proportion of the variability in the target variable that is explained by the linear relationship of the target variable with the set of the predictor variables. The \*\*Sequential Sum of Squares\*\* partition the SSR into unique portions of the SSR that are explained by the particular predictors, given any earlier predictors. Thus the values of the \*\*Sequential Sum of Squares\*\* depend on the order that the variable are entered into the model.  
  
## Now after running ANOVA we get that the Sum of Squares for Sugar is 8711.9, and represents the variability in Nutritional Rating that is explained by the linear relationship between Rating and Sugar Content. Also the first Seqential Sum of Squares is exactly the value for SSR from the Simple Linear Regression of Nutritional Rating on Sugar Content.  
  
## The Second Sequential Sum of Squares, for fiber content, equals 3476.6. This represents the amount of unique additional variability in nutritional Rating that is explained by the linear relationship of Rating vs Fiber Content, given that the variability explained by Sugars has already been extracted.  
  
## The third Sequential Sm of Squares, for shelf 1 is 7. This represents the amount of unique additional variability in nutritional Rating that is accounts for by location on Shelf 1 (Compared to the reference class Shelf 3), given that the variability accountd for by Sugars and fiber has already been separated out.## This tiny value for the sequential sum of squares for shelf 1 seems to be of less importance.  
  
mreg4.2<-lm(Rating~Shelf1+Shelf2+Sugars+Fiber,data = cereal)  
anova(mreg4.2)

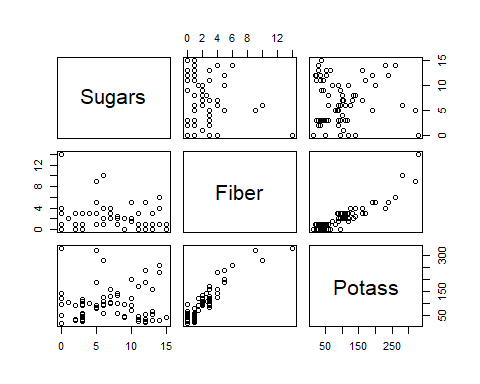
## Analysis of Variance Table  
##   
## Response: Rating  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Shelf1 1 282.7 282.7 7.7989 0.006713 \*\*   
## Shelf2 1 1392.7 1392.7 38.4178 3.340e-08 \*\*\*  
## Sugars 1 7179.0 7179.0 198.0328 < 2.2e-16 \*\*\*  
## Fiber 1 3501.0 3501.0 96.5764 7.084e-15 \*\*\*  
## Residuals 71 2573.9 36.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Now after changing the orders of the variables in the regression model we observe that the values of Sequential Sum of Squares also changed.

# Multicollinearity   
## Lets check the multicollinearity among the three predictor variables i.e. Fiber, Sugar and Shelf 2.  
  
datam<-matrix(c(cereal$Fiber,cereal$Sugars,cereal$Shelf2),ncol = 3)  
colnames(datam)<-c("Fiber","Sugars","Shelf2")  
cor(datam)

## Fiber Sugars Shelf2  
## Fiber 1.0000000 NA -0.3225486  
## Sugars NA 1 NA  
## Shelf2 -0.3225486 NA 1.0000000

## In data set with severe multicollinaerity, it is possible for the F-test for the overall regression to be to be significant, while none of the t-tests for the individual predictors are significant.  
  
pairs(~Sugars+Fiber+Potass,data = cereal)



##VIF  
## The Value of Variance Inflation Factor or VIF warn us of the presence of multicollinearity in our Regression models.   
  
## VIF=1/(1-R^2)  
## R^2 represents the R-Square value obtained by regressing one predictor variable on the other predictor variables.  
  
## When R^2=0(No correlation at all), then we have VIF=1. This is the minimum value of VIF. However as th degree of correlation between one predictor variable and other predictor variable increases, R^2 will also increase. The maximum value can reach infinity as the R^2 reaches 1.  
  
  
mreg5<-lm(Rating~Sugars+Fiber+Shelf2+Potass,data = cereal)  
library(car)

## Loading required package: carData

vif(mreg5)

## Sugars Fiber Shelf2 Potass   
## 1.460974 6.951804 1.417057 7.156593

##Here we observe that the Variance Inflation Factor(VIF) of Sugar and Shelf2 are 1.460974 and 1.417057 respectively, quite close to 1, which is reasonably fine. The variables are not approximately correlated while the Variance Inflation Factor(VIF) of Fiber and Potass are quite high, 6.951804 and 7.156593 respectively showing the evidence of multicollinearity.  
  
## A rough rule off thumb for interpreting the value of the VIF is to consider VIF>=5 to be an indicator of moderate multicollinearity, and to consider VIF>=10 to be an indicator of severe multicollinearity. A VIF=5 corresponds to R-square =0.80 and VIF=10 corresponds to R-square=.90.

# Now we will take the example of Gas Mileage Dataset in order to do the further examples.  
  
#auto\_gas <- read.csv("C:/Users/sabaa/Downloads/archive/auto-mpg.csv",stringsAsFactors = TRUE)  
#gas<-auto\_gas[-c(33,126,330,336,354,374),]  
#head(gas)  
#gas$"lnMPG"<-log(gas$mpg)  
#gas$"lnHP"<-log(gas$horsepower)  
#gas1<-gas[,c(7,2,8,5,6)]  
#names(gas1)  
#pairs(gas1[,1]~gas1[,2]+gas1[,3]+gas1[,4]+gas1[,5],labels=names(gas1),cex.labels=1)

##### Now we will see the application of variable selection methods. To assist the data analysts in determing which variables should be included in a multiple regression model, several different variable selection methods have been developed. These are forward selection, backward selection, stepwise selection and best subsets. These variable selection methods are essentially algorithms to help construct the model with the optimal set of predictors.

```