ANALYSIS OF IMPACT OF ADVERTISING MEDIAS YOUTUBE, FACEBOOK, NEWSPAPER ON SALES

**INTRODUCTION**

***Objective:*** Here we are taking a marketing dataset and we are interested in performing a multiple linear regression with three regressors i.e. advertising medias facebook, newspaper, youtube and analyze it by obtaining the matrix of scatter diagrams between the variables of interest and also find the matrix of coefficient of correlations. We are also interested in checking whether the regressors are independent to each other and fitting a multiple linear regression model and interpreting the estimated coefficients.We also want to test the significance of regression parameters using the t test and obtain the adjusted coefficient of determination and interpret the obtained results. We are also interested in obtainining the confidence interval of regression coefficients and further performing the residual analysis of the model.

***Data Description:*** Here in this practical we are considereing the marketing dataset which is a data frame containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales. The advertising experiment has been repeated 200 times.

Here, our response variable is sales and regressors are advertising budget of advertising medias youtube, newspaper, and facebook.

*#Loading the package datarium using library which consists of data we are interested in.*  
**library**(datarium)

## Warning: package 'datarium' was built under R version 4.0.3

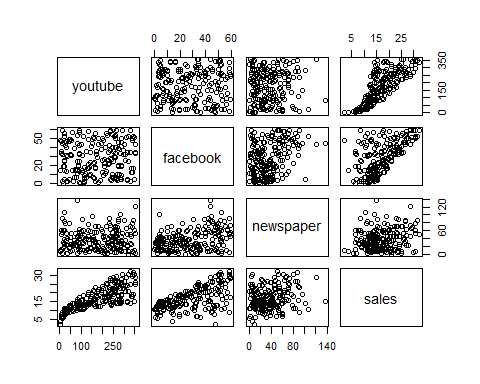
*#Since, we are interested in 'marketing' dataset, here we obtain the first few records of the dataset 'marketing'.*  
**head**(marketing)

## youtube facebook newspaper sales  
## 1 276.12 45.36 83.04 26.52  
## 2 53.40 47.16 54.12 12.48  
## 3 20.64 55.08 83.16 11.16  
## 4 181.80 49.56 70.20 22.20  
## 5 216.96 12.96 70.08 15.48  
## 6 10.44 58.68 90.00 8.64

**ANALYSIS**

*#We use attach function to access the variables present in the dataframe without calling the dataframe.*  
**attach**(marketing)

To Plot a matrix of scatter diagrams between the variables of interest and also find the matrix of coefficient of correlations and interpret it. And check if the regressors are independent to each other and Justifying our answer.  
  
*#Here we are plotting a matrix of scatter diagram between our variables of interest.*  
**pairs**(marketing[1**:**4])



From the above figure, we observe that there is a linear positive relationship between sales and youtube. We also observe that there is positive linear relationship between sales and facebook. There is a positive linear relationship between sales and newspaper. We also observe that there is no linear relationship between youtube and facebook. Also from the graph it is evident that there is no linear relationship between youtube and newspaper. But we observe that there is a positive linear relationship between facebook and newspaper.Since there is a positive linear relationship between two regressors(facebook and newspaper) hence we need to check the presence of multicollinearity.

Hence, we proceed for testing the significance of correlation.

*#Loading the package 'Hmisc'.*  
**library**(Hmisc)

## Warning: package 'Hmisc' was built under R version 4.0.3

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.0.3

## Loading required package: survival

## Warning: package 'survival' was built under R version 4.0.3

## Loading required package: Formula

## Warning: package 'Formula' was built under R version 4.0.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.3

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

*#Now we are obtaining the matrix of coefficient of correlation with rounded to two decimal points.*  
**rcorr**(**as.matrix**(marketing))

## youtube facebook newspaper sales  
## youtube 1.00 0.05 0.06 0.78  
## facebook 0.05 1.00 0.35 0.58  
## newspaper 0.06 0.35 1.00 0.23  
## sales 0.78 0.58 0.23 1.00  
##   
## n= 200   
##   
##   
## P  
## youtube facebook newspaper sales   
## youtube 0.4408 0.4256 0.0000  
## facebook 0.4408 0.0000 0.0000  
## newspaper 0.4256 0.0000 0.0011  
## sales 0.0000 0.0000 0.0011

From the above matrix, we observe that p value for newspaper and facebook is 0<0.05 therefore we reject the null hypothesis which implies that there is a significant relationship between newspaper and facebook.That is the two regressors newspaper and facebook are not independent of each other.

To fit a multiple linear regression model and interpret the estimated coefficients.

*#Fitting a multiple linear regression model to our dataset.*  
reg=**lm**(sales**~**youtube**+**facebook**+**newspaper,marketing)  
reg

##   
## Call:  
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)  
##   
## Coefficients:  
## (Intercept) youtube facebook newspaper   
## 3.526667 0.045765 0.188530 -0.001037

The fitted multiple linear regression model is,

sales = 3.526667 + 0.045765\*youtube + 0.188530\*facebook - 0.001037\*newspaper

From the fitted model we obtained the intercept as Beta0= 3.526667 and the regression coefficient for explantory variable advertising budget of youtube as Beta1=0.045765, the regression coefficient for explantory variable advertising budget of facebook as Beta2=0.188530 and the regression coefficient for explantory variable advertising budget of newspaper as Beta3=-0.001037.

To test the significance of regression parameters using the t test and interpret it and obtain the adjusted coefficient of determination and interpret it.

*#Getting the summary of the fitted regression model.*  
**summary**(reg)

##   
## Call:  
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.5932 -1.0690 0.2902 1.4272 3.3951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.526667 0.374290 9.422 <2e-16 \*\*\*  
## youtube 0.045765 0.001395 32.809 <2e-16 \*\*\*  
## facebook 0.188530 0.008611 21.893 <2e-16 \*\*\*  
## newspaper -0.001037 0.005871 -0.177 0.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.023 on 196 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956   
## F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

From the above summary, we observe that,

The calculated t-value for the regressor variable, advertising budget of youtube is 32.809 which is greater than the critical value 1.984 and hence we reject the null hypothesis and conclude that the regression coefficient of regressor variable, advertising budget of youtube is having a significant effect on our data.

The calculated t-value for the regressor variable advertising budget of facebook is 21.893 which is greater than the critical value 1.984 and hence we reject the null hypothesis and conclude that the regression coefficient of regressor variable, advertising budget of facebook is having a significant effect on our data.

The calculated t-value for the regressor variable advertising budget of newspaper is -0.177 which is smaller than the critical value 1.984 and hence we accept the null hypothesis and conclude that the regression coefficient of regressor variable, advertising budget of newspaper is having a not significant effect on our data.

We observe that the adjusted coefficient of determination is 0.8956.

*#To find the 95% confidence interval for the parameter.*  
confint(reg,level=0.95)

2.5 % 97.5 %

(Intercept) 2.78851474 4.26481975

marketing$youtube 0.04301371 0.04851558

marketing$facebook 0.17154745 0.20551259

marketing$newspaper -0.01261595 0.01054097

We observe that the confidence interval for the regression parameter of youtube is (0.04301371, 0.04851558), the confidence interval for the regression parameter of facebook is (0.17154745, 0.20551259) and the confidence interval for the regression parameter of newspaper is (-0.01261595, 0.01054097).

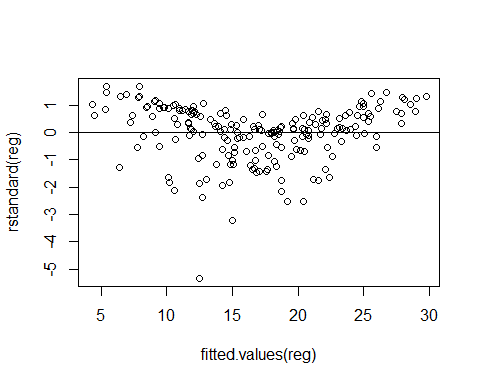
Residual Analysis

To check the validity of the following assumptions:

* The relationship between Y, X1, X2, …, Xp-1 is linear.
* Residuals has mean zero.
* Variance of residuals is constant.
* Residuals are uncorrelated.
* Residuals are normally distributed.

*#Plot of fitted values against residuals*

fitted\_values<-**fitted.values**(reg)  
**plot**(**fitted.values**(reg),**rstandard**(reg))  
**abline**(0,0)



The above plot of fitted values against residuals indicates the loss of linearity.  
  
*#Obtaining the standardized residuals*  
resid=**rstandard**(reg)  
resid

## 1 2 3 4 5 6   
## 0.947053927 -1.161041919 -1.820489933 0.539751949 -0.173316079 -3.208704720   
## 7 8 9 10 11 12   
## 0.041976938 0.641958204 0.646404591 -1.167476165 0.937628276 0.068750855   
## 13 14 15 16 17 18   
## -0.829759408 0.522187181 0.337328909 0.947716883 -0.200890561 0.706178146   
## 19 20 21 22 23 24   
## 0.804163171 0.258311560 -0.060194829 -1.342012197 -0.534278854 -0.623892469   
## 25 26 27 28 29 30   
## 0.927505559 -2.168146479 0.006256295 -0.687587187 -0.304980808 0.809602971   
## 31 32 33 34 35 36   
## -0.140251662 0.329963270 1.174471194 -0.880130614 1.153088140 -2.535381739   
## 37 38 39 40 41 42   
## 1.207652782 -0.554895703 0.114252168 0.628886367 0.132384588 -0.116738197   
## 43 44 45 46 47 48   
## -0.540358924 -0.635318884 -0.232151425 -0.155593935 1.031070189 0.887145002   
## 49 50 51 52 53 54   
## -0.875847231 0.915296995 -0.737517498 0.813139758 1.158245860 0.752563758   
## 55 56 57 58 59 60   
## -0.104473203 1.446309787 -1.814847849 0.253296247 1.141917164 0.159409996   
## 61 62 63 64 65 66   
## 1.413489147 0.786170291 -0.647344935 0.471835412 0.611308697 0.869076994   
## 67 68 69 70 71 72   
## 0.290600966 0.812636096 -0.045825579 0.713856429 0.310068057 1.056228279   
## 73 74 75 76 77 78   
## -0.939201445 0.655237802 -0.196589888 -1.929651879 1.455099428 0.231469119   
## 79 80 81 82 83 84   
## -2.112802404 0.790289863 0.211113050 -1.409255326 0.668391809 -0.492691009   
## 85 86 87 88 89 90   
## 0.549726896 0.011164799 0.239536865 0.243458398 0.715786544 -0.133268435   
## 91 92 93 94 95 96   
## 0.717581100 1.686295935 0.145745419 0.585711273 0.607071051 0.348975908   
## 97 98 99 100 101 102   
## -0.560451331 0.096993795 0.770760266 0.155090000 -1.309078051 0.343033731   
## 103 104 105 106 107 108   
## -1.706256309 -0.037063324 0.239712424 0.757566759 0.643162678 0.953321316   
## 109 110 111 112 113 114   
## 1.031056454 0.064254821 -0.818399847 0.395089231 0.131101587 -0.300745288   
## 115 116 117 118 119 120   
## -0.423072166 -0.191152820 0.131719293 1.695087413 0.201237061 -0.131693728   
## 121 122 123 124 125 126   
## 0.649114014 -0.502789986 -1.214482078 0.070077742 0.147856720 0.878200982   
## 127 128 129 130 131 132   
## -2.392061014 1.319147700 1.483099054 1.086726013 -5.336838368 -1.734346241   
## 133 134 135 136 137 138   
## -1.655277958 0.198766519 -0.624256420 -1.453423791 -1.177529475 -0.030778847   
## 139 140 141 142 143 144   
## -0.100633873 0.618659295 0.841688999 0.481867381 0.507085562 0.977565707   
## 145 146 147 148 149 150   
## 0.780463848 0.353882348 -1.255062810 1.285564299 -0.820877372 0.162928414   
## 151 152 153 154 155 156   
## -1.359451374 0.952505375 0.143121177 0.463109022 0.058702261 -1.266369879   
## 157 158 159 160 161 162   
## -0.050800581 0.051259868 -1.851751252 0.298079092 0.110794874 -0.154797854   
## 163 164 165 166 167 168   
## -0.027852508 0.388657361 0.496494269 -1.428973868 -1.696764638 -0.695535698   
## 169 170 171 172 173 174   
## -0.051535615 -1.768926662 0.600607247 0.084476356 -0.004604416 -0.161764665   
## 175 176 177 178 179 180   
## -1.344607324 1.333348427 0.131921254 -0.275908217 -2.533838316 0.128455641   
## 181 182 183 184 185 186   
## -0.052168787 -1.033934691 1.284422943 1.249566721 -0.560465294 1.079139042   
## 187 188 189 190 191 192   
## 0.363983038 0.133292737 -1.650932774 0.388283242 -1.019745535 0.882489489   
## 193 194 195 196 197 198   
## 0.861766581 0.670347650 0.482291641 1.337164319 0.917931162 0.008411539   
## 199 200   
## 1.044868687 -1.061710823

*#Obtaining the absolute values standardized residuals*  
**abs**(resid)

## 1 2 3 4 5 6   
## 0.947053927 1.161041919 1.820489933 0.539751949 0.173316079 3.208704720   
## 7 8 9 10 11 12   
## 0.041976938 0.641958204 0.646404591 1.167476165 0.937628276 0.068750855   
## 13 14 15 16 17 18   
## 0.829759408 0.522187181 0.337328909 0.947716883 0.200890561 0.706178146   
## 19 20 21 22 23 24   
## 0.804163171 0.258311560 0.060194829 1.342012197 0.534278854 0.623892469   
## 25 26 27 28 29 30   
## 0.927505559 2.168146479 0.006256295 0.687587187 0.304980808 0.809602971   
## 31 32 33 34 35 36   
## 0.140251662 0.329963270 1.174471194 0.880130614 1.153088140 2.535381739   
## 37 38 39 40 41 42   
## 1.207652782 0.554895703 0.114252168 0.628886367 0.132384588 0.116738197   
## 43 44 45 46 47 48   
## 0.540358924 0.635318884 0.232151425 0.155593935 1.031070189 0.887145002   
## 49 50 51 52 53 54   
## 0.875847231 0.915296995 0.737517498 0.813139758 1.158245860 0.752563758   
## 55 56 57 58 59 60   
## 0.104473203 1.446309787 1.814847849 0.253296247 1.141917164 0.159409996   
## 61 62 63 64 65 66   
## 1.413489147 0.786170291 0.647344935 0.471835412 0.611308697 0.869076994   
## 67 68 69 70 71 72   
## 0.290600966 0.812636096 0.045825579 0.713856429 0.310068057 1.056228279   
## 73 74 75 76 77 78   
## 0.939201445 0.655237802 0.196589888 1.929651879 1.455099428 0.231469119   
## 79 80 81 82 83 84   
## 2.112802404 0.790289863 0.211113050 1.409255326 0.668391809 0.492691009   
## 85 86 87 88 89 90   
## 0.549726896 0.011164799 0.239536865 0.243458398 0.715786544 0.133268435   
## 91 92 93 94 95 96   
## 0.717581100 1.686295935 0.145745419 0.585711273 0.607071051 0.348975908   
## 97 98 99 100 101 102   
## 0.560451331 0.096993795 0.770760266 0.155090000 1.309078051 0.343033731   
## 103 104 105 106 107 108   
## 1.706256309 0.037063324 0.239712424 0.757566759 0.643162678 0.953321316   
## 109 110 111 112 113 114   
## 1.031056454 0.064254821 0.818399847 0.395089231 0.131101587 0.300745288   
## 115 116 117 118 119 120   
## 0.423072166 0.191152820 0.131719293 1.695087413 0.201237061 0.131693728   
## 121 122 123 124 125 126   
## 0.649114014 0.502789986 1.214482078 0.070077742 0.147856720 0.878200982   
## 127 128 129 130 131 132   
## 2.392061014 1.319147700 1.483099054 1.086726013 5.336838368 1.734346241   
## 133 134 135 136 137 138   
## 1.655277958 0.198766519 0.624256420 1.453423791 1.177529475 0.030778847   
## 139 140 141 142 143 144   
## 0.100633873 0.618659295 0.841688999 0.481867381 0.507085562 0.977565707   
## 145 146 147 148 149 150   
## 0.780463848 0.353882348 1.255062810 1.285564299 0.820877372 0.162928414   
## 151 152 153 154 155 156   
## 1.359451374 0.952505375 0.143121177 0.463109022 0.058702261 1.266369879   
## 157 158 159 160 161 162   
## 0.050800581 0.051259868 1.851751252 0.298079092 0.110794874 0.154797854   
## 163 164 165 166 167 168   
## 0.027852508 0.388657361 0.496494269 1.428973868 1.696764638 0.695535698   
## 169 170 171 172 173 174   
## 0.051535615 1.768926662 0.600607247 0.084476356 0.004604416 0.161764665   
## 175 176 177 178 179 180   
## 1.344607324 1.333348427 0.131921254 0.275908217 2.533838316 0.128455641   
## 181 182 183 184 185 186   
## 0.052168787 1.033934691 1.284422943 1.249566721 0.560465294 1.079139042   
## 187 188 189 190 191 192   
## 0.363983038 0.133292737 1.650932774 0.388283242 1.019745535 0.882489489   
## 193 194 195 196 197 198   
## 0.861766581 0.670347650 0.482291641 1.337164319 0.917931162 0.008411539   
## 199 200   
## 1.044868687 1.061710823

Since one of the residual’s absolute value is greater than three hence there exist one outlier.  
  
*#We are obtaining the mean and variance of residuals.*  
**mean**(resid)

## [1] -0.001522272

**var**(resid)

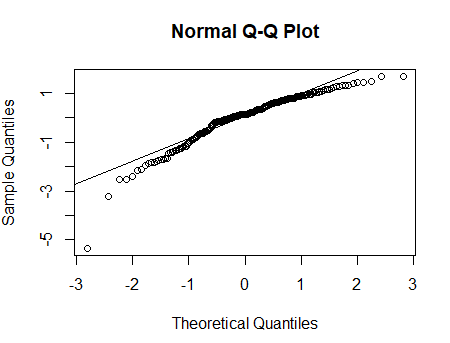
## [1] 1.01324

We also observe that mean of residuals is approximately 0 and variance is 1

*#Obtaining the Q-Q Plot.*

**qqnorm**(resid)

**qqline**(resid)



From the above Figure we observe that our plot is not a straight line it deviates hence we conclude that residuals are not normally distributed.  
  
  
*#To check if the residual series are normally distributed we perform shapiro test.*   
**shapiro.test**(resid)

##   
## Shapiro-Wilk normality test  
##   
## data: resid  
## W = 0.91662, p-value = 3.312e-09

Since p value is negligible hence we reject the null hypothesis and conclude that samples are not normally distributed.  
*#To examine the property of autocorrelation of residual sequence.*  
**Box.test**(resid)

##   
## Box-Pierce test  
##   
## data: resid  
## X-squared = 0.42548, df = 1, p-value = 0.5142

Since p value is 0.5142 which is greater that 0.05 hence we fail to reject the null hypothesis and conclude that the residuals from fitted model are uncorrelated.

**CONCLUSION**

1. From the above figure, we observe that there is a linear positive relationship between sales and youtube which means that as the advertising budget of youtube increases sales also increases. We also observe that there is positive linear relationship between sales and facebook which means that as the advertising budget of facebook increases sales also increases. There is a positive linear relationship between sales and newspaper which means that as the advertising budget of newspaper increases sales also increases. We also observe that there is no linear relationship between youtube and facebook. Also from the graph it is evident that there is no linear relationship between youtube and newspaper.

But we observe that there is a positive linear relationship between facebook and newspaper which means that as the advertising budget of facebook increases the advertising budget of newspaper also increases. Hence all the regressors are not independent.

Hence we proceed further to test the significance of correlation and from the matrix obtained, we observe that p value for newspaper and facebook is 0<0.05 therefore we reject the null hypothesis which implies that there is a significant relationship between newspaper and facebook.That is the two regressors newspaper and facebook are not independent of each other.

2. From the fitted model, we observe that the intercept Beta0=3.526667, which basically gives the average value of response variables when the regressor variables are zero therefore we conclude that when the advertising medias has not set any budget then the sales will be 3.526667 in thousands of dollars.

We also observe that the coefficient of youtube i.e. Beta1 is 0.045765 which means that when the advertising budget of advertising medias facebook and newspaper is kept constant then for one unit of change in advertising budget of advertising media youtube, sales changes 0.045765 times of advertising budget of youtube.

We also observe that the coefficient of facebook i.e. Beta2 is 0.188530 which means that when the advertising budget of advertising medias youtube and newspaper is kept constant then for one unit of change in advertising budget of advertising media facebook, sales changes 0.188530 times of advertising budget of youtube.

We also observe that the coefficient of newspaper i.e. Beta2 is -0.001037 which means that when the advertising budget of advertising medias facebook and youtube is kept constant then for one unit of change in advertising budget of advertising media newspaper, sales changes -0.001037 times of advertising budget of youtube.

3. From the summary of regression model we observe that the calculated t-value for regressor variable advertising budget of youtube is 32.809 which is greater than its critical value 1.984 hence we reject the null hypothesis which implies that advertising budget of youtube is statistically significant predictor of sales

we also observe that the calculated t-value for regressor variable advertising budget of facebook is 21.893 which is greater than its critical value 1.984 hence we reject the null hypothesis which implies that advertising budget of facebook is statistically significant predictor of sales

we also observe that the calculated t-value for regressor variable advertising budget of newspaper is -0.177 which is smaller than its critical value 1.984 hence we accept the null hypothesis which implies that advertising budget of newspaper is not statistically significant predictor of sales

4. Since the adjusted coefficient of determination is 0.8956, hence we conclude that 89.56% of total variation of sales is explained by the advertising budget of three advertising medias facebook, newspaper and youtube. Also since adjusted coefficient of determination is 0.8956>0.5 which means that our fitted regression model is of good quality.

Also since the p value is negligible for this dataset we can conclude that the performance of regression model is good.

5.From the confidence intervals we observe that the confidence interval for the regression parameter of youtube is (0.04301371, 0.04851558) that is we are 95% sure that the true value of regression parameter of youtube lies in the range (0.04301371, 0.04851558), the confidence interval for the regression parameter of facebook is (0.17154745, 0.20551259) that is we are 95% sure that the true value of regression parameter of youtube lies in the range (0.17154745, 0.20551259) and the confidence interval for the regression parameter of newspaper is (-0.01261595, 0.01054097) that is we are 95% sure that the true value of regression parameter of youtube lies in the range (-0.01261595, 0.01054097).

6. RESIDUAL ANALYSIS

From the plot of residuals sequence against fitted values it is evident that the linearity of the model is lost.

We also observe that only one residual value has absolute value greater than 3 hence there exists one outlier.

We also calculated that mean of the residual is approximately zero and variance is approximately one.

From the Q-Q Plot we observe that our plot is not a straight line it deviates hence we conclude that residuals are not normally distributed.

From shapirov wilks test we observe that Since p value is negligible therefore we reject the null hypothesis and conclude that samples are not normally distributed.

From Box test we observe that Since p value is 0.5142 which is greater that 0.05 hence we fail to reject the null hypothesis and conclude that the residuals from fitted model are uncorrelated.

Hence from the above residual analysis we observe that not all the assumptions for the models are valid hence the model is valid for all the assumptions except the assumption that residuals are normally distributed which is violated. Hence, the model is not a completely valid model.