Advertising

Gokul Shanth Raveendran

2/17/2020

Data Analysis on Advertising Dataset

Loading the libraries

rm(list=ls())  
library(rio)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.2

library(corrplot)

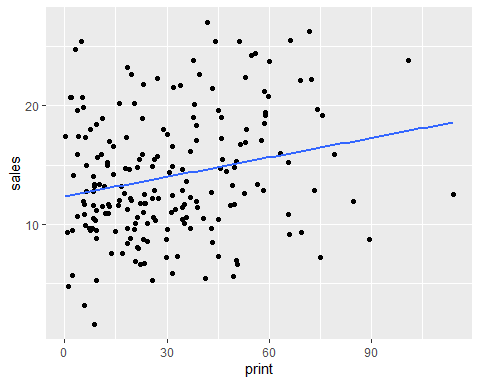
## corrplot 0.84 loaded

library(car)

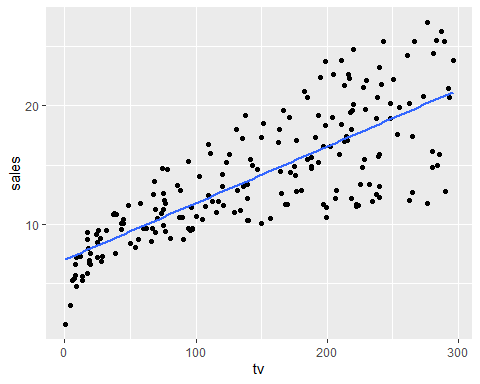
## Loading required package: carData

Importing the dataset and checking scatter plots for the three columns in the dataset.

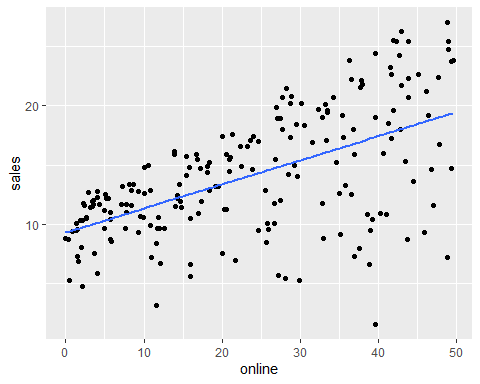
advertising=import("Advertising.csv")  
ggplot(data = advertising, aes(x = print, y = sales)) +   
 geom\_point(color='black') +  
 geom\_smooth(method = "lm", se = FALSE)



ggplot(data = advertising, aes(x = tv, y = sales)) +   
 geom\_point(color='black') +  
 geom\_smooth(method = "lm", se = FALSE)



ggplot(data = advertising, aes(x = online, y = sales)) +   
 geom\_point(color='black') +  
 geom\_smooth(method = "lm", se = FALSE)



We can see a linear relationship between tv spending and sales, online and sales but not so much between print and sales

tv.out = lm(sales~tv, data = advertising)  
summary(tv.out)

##   
## Call:  
## lm(formula = sales ~ tv, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.3860 -1.9545 -0.1913 2.0671 7.2124   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.032594 0.457843 15.36 <2e-16 \*\*\*  
## tv 0.047537 0.002691 17.67 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.259 on 198 degrees of freedom  
## Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099   
## F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16

online.out = lm(sales~online, data = advertising)  
summary(online.out)

##   
## Call:  
## lm(formula = sales ~ online, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.7305 -2.1324 0.7707 2.7775 8.1810   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.31164 0.56290 16.542 <2e-16 \*\*\*  
## online 0.20250 0.02041 9.921 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.275 on 198 degrees of freedom  
## Multiple R-squared: 0.332, Adjusted R-squared: 0.3287   
## F-statistic: 98.42 on 1 and 198 DF, p-value: < 2.2e-16

print.out = lm(sales~print, data = advertising)  
summary(print.out)

##   
## Call:  
## lm(formula = sales ~ print, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2272 -3.3873 -0.8392 3.5059 12.7751   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.35141 0.62142 19.88 < 2e-16 \*\*\*  
## print 0.05469 0.01658 3.30 0.00115 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.092 on 198 degrees of freedom  
## Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733   
## F-statistic: 10.89 on 1 and 198 DF, p-value: 0.001148

Since all the models have a significant p value, we can reject the null hypothesis and we can say that there is a relationship between sales and the independent variables. We can use the R square to assess the accuracy of the model. R squared explains variablity of in the data, higher the R square the more closer are the data to the fitted regression line.

Since the R squared values are low, we can combine them to see if they have a better explanation.

print.out = lm(sales~print+online+tv, data = advertising)  
summary(print.out)

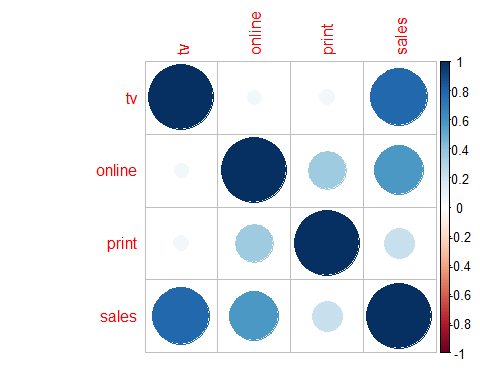
##   
## Call:  
## lm(formula = sales ~ print + online + tv, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8277 -0.8908 0.2418 1.1893 2.8292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.938889 0.311908 9.422 <2e-16 \*\*\*  
## print -0.001037 0.005871 -0.177 0.86   
## online 0.188530 0.008611 21.893 <2e-16 \*\*\*  
## tv 0.045765 0.001395 32.809 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.686 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

This gives us an improved R square of almost 90%. However you’ll notice that in a simple linear regression between print and sales was statistically significant however in multiple it is giving us a negative beta coefficient suggesting that it causes a decrease in sales? To build a better model lets take a look at the correlation matrix

cor(advertising)

## tv online print sales  
## tv 1.00000000 0.05480866 0.05664787 0.7822244  
## online 0.05480866 1.00000000 0.35410375 0.5762226  
## print 0.05664787 0.35410375 1.00000000 0.2282990  
## sales 0.78222442 0.57622257 0.22829903 1.0000000

corrplot(cor(advertising), method="circle")



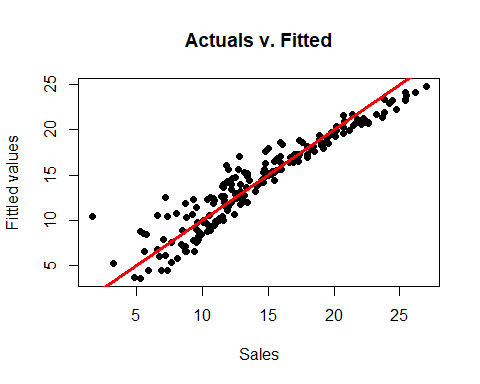
We can see that the correlation between tv and sales, online and sales are 0.7 and 0.57 but correlation between online and print is higher than between print and sales. What we can interpret by this is that there has been a tendency to spend more in print in markets where we spend more on online advertising. To put things in perspective, even though print doesn’t increase sales it, gets recognition for the impact of online spending on sales when we do a simple linear regression. Therefore removing it will not alter the regression significantly in predicting sales.

onlinetv = lm(sales~online+tv, data = advertising)  
summary(onlinetv)

##   
## Call:  
## lm(formula = sales ~ online + tv, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7977 -0.8752 0.2422 1.1708 2.8328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.92110 0.29449 9.919 <2e-16 \*\*\*  
## online 0.18799 0.00804 23.382 <2e-16 \*\*\*  
## tv 0.04575 0.00139 32.909 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.681 on 197 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962   
## F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

As we can see from the above summary there is not really that much difference in the R square. The p values indicate that tv and online are related to sales but there is no proof that print is associated to sales in the presence of tv and online. According to the above model the beta coefficients for online and tv are independent of each other, in the sense that increase in spending for online will only increase sales only by the beta coefficient for online. The above can be interpreted as for every $1000 increase in advertising, sale through online increase by 187 units and by tv 457 units.

plot(advertising$sales,onlinetv$fitted.values,  
 xlab="Sales", ylab="Fittled values",  
 pch=19,main="Actuals v. Fitted")  
abline(0,1,col="red",lwd=3)



Since we remove print because we saw that it doesn’t significantly affect the regression model because there was no proof that print is associated to sales in the presence of tv and online, we can check the synergy/interaction in markets for both tv and online spending because if we split a fixed budget for tv and online equally it, will increase/decrease the overall sales compared to just allocating to one of the two. We can check for interaction between online and tv and how both together will have its effect on the regression.

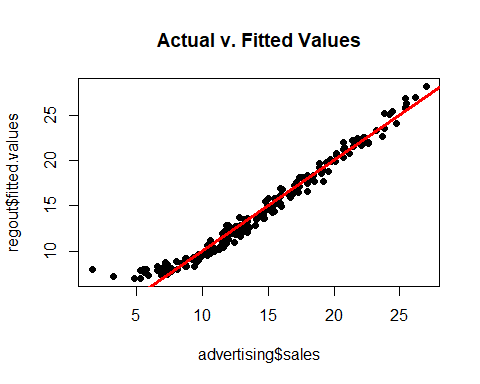
regout = lm(sales~online+tv+online:tv, data = advertising)  
summary(regout)

##   
## Call:  
## lm(formula = sales ~ online + tv + online:tv, data = advertising)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3366 -0.4028 0.1831 0.5948 1.5246   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 \*\*\*  
## online 2.886e-02 8.905e-03 3.241 0.0014 \*\*   
## tv 1.910e-02 1.504e-03 12.699 <2e-16 \*\*\*  
## online:tv 1.086e-03 5.242e-05 20.727 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9435 on 196 degrees of freedom  
## Multiple R-squared: 0.9678, Adjusted R-squared: 0.9673   
## F-statistic: 1963 on 3 and 196 DF, p-value: < 2.2e-16

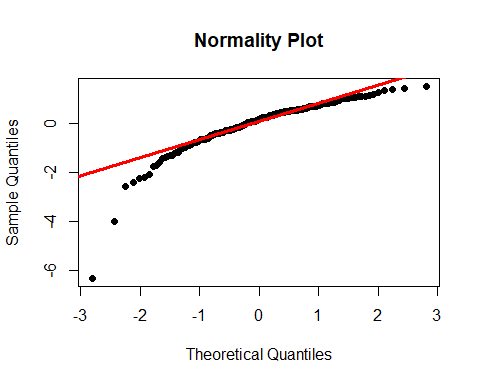
After adding the interaction term we can see a significant increase in the R squared value and statistical significance of the interaction term. The results from the summary suggest that this model is superior to the other models. It explains that, the interaction term explains most of the variability in sales. Thus we can say that the model isn’t just additive. We can interpret the interaction term beta coefficent as increase in effectiveness of the online advertisiment for a one unit increase in tv advertisement and vice versa.

In order for the model to be usable, it needs to conform to certain assumptions. Let’s check for Linearity, Normality and Equality of Variance.

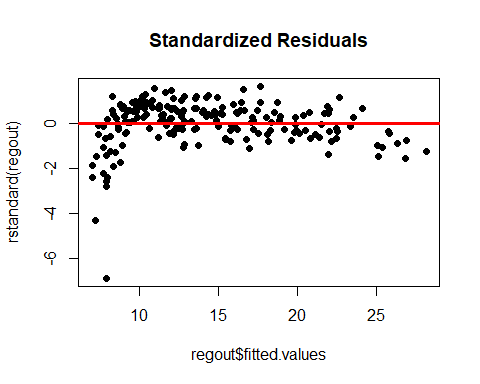
#Linearity  
plot(advertising$sales,regout$fitted.values,pch=19,main="Actual v. Fitted Values")  
abline(0,1,col="red",lwd=3)



qqnorm(regout$residuals,pch=19,main="Normality Plot")  
qqline(regout$residuals,col="red",lwd=3)



#Equality of Variances  
plot(regout$fitted.values,rstandard(regout),pch=19,  
 main="Standardized Residuals")  
abline(0,0,col="red",lwd=3)



sqrt(mean((regout$residuals)^2))

## [1] 0.9340326

From plot 1, fitted vs sales, we can see that it is linear. From plot 2, we can see that the residuals are almost normal From plot 3, we can see that it is almost homoscedastic.

##Questions: What statistical method will you use for your analysis and why?

For my analysis I will use a linear regression. Firstly, the scatter plots suggest a linear increase between sales and the other independent variables. After modeling and testing the model for regression assumptions, we can see that it satisfies linearity, normality and homoscedasity. Thus I think regression is an appropriate model.

How do you know that your model is good enough? With an R square of 96 and an RMSE of 0.93 I think the model is good enough, because 96% of variance is explained. The model accuracy can be explained by the high R squared value

What do the “standard errors” of beta coefficients mean? The standard error of a coefficient, tells us how much sampling variation there is if we resample and re estimate the coefficient.

How did the model compute “degrees of freedom”? Degrees of freedom is the number of independent pieces of information that went into calculating the estimate. It is computed by subractng the total number of observations minus the number of parameters used. In our case we have 3 beta coefficients and the Intercept, therefore the degrees of freedom is 196.

What is the best measure of model fit in this problem? The best measure of model fit is R square, Root mean squared deviation and the F test.

What are “residuals” in your model and why are they important? The residuals in the model are the difference between the actual value vs the value on the regression line. They are important because, residuals need to follow the assumptions of linearity, normality and equality of variance to validate our model. Residuals help us see outliers, if the relationship is linear or non-linear and if we have overlooked groups of observations.