Multiple Linear Regression and Least Squares

Why is it needed in this case?

When dealing with defects in a business it is easy to quantify what percentage of your product is defective, it is not as easy to pinpoint where those defects come from. When various inputs are required to make a certain output it is harder to quantify how dependent a product is from one input. This is where multiple linear regression plays a role in helping continuous improvement specialists eliminate defects in various products.

How does it work?

Simple Linear Regression works by plotting the data in a scatterplot and then running the linear regression line through the data points. **Y = a + bX + e** is the equation that is used in graphing the line. Y is the value of the output. A is the estimated Y intercept. 'b' is the correlation from -1 to 1 which signifies the relationship form input to output. e is an error term representing the unexplained or residual variance.

How does it relate to Linear Algebra?

When computing this line we use the least squares method. When using the least squares method there are techniques in linear algebra to find a, b, x, and e.

We have our model equation Y=a+b1x1+b2x2….bnxn

where each one of these symbolizes a vector or a matrix which turns into

$$Y= xb + e$$

Y is then a vector of the observables,

x being the vector of correlations between 0-1

b being the vector of independent variables

e being the vector of residuals or the distance from our observables to our estimated values. (These could be negative)

I will test it on a specific application about a salesforce:

Our data:

A Sales Manager is analyzing the performance of the sales force, which exhibits wide variation between individuals. It appears that the more experienced salespeople generate consistently higher sales. The Sales Manager wonders if sales performance

can be predicted from experience and training. Although this is a simple linear regression question it will prove useful in our pedagogical approach to understanding multiple linear regression

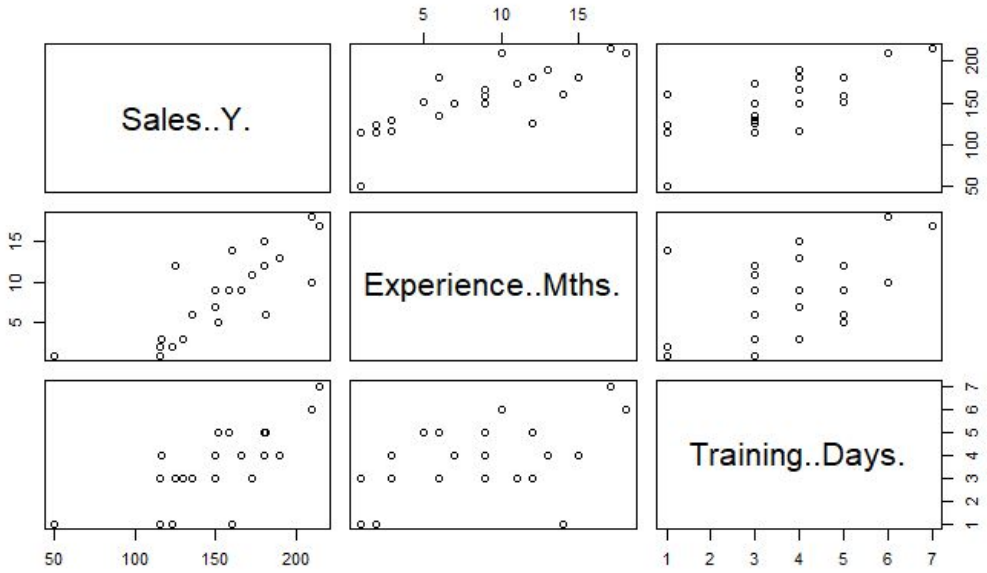
In understanding this data we are going to run a few functions in R and then explain are findings.

Our R code:sales<-read.csv("file:///C:/Users/Hayden.DESKTOP-B22L498/Documents/R/sales_effectiveness_data.csv")

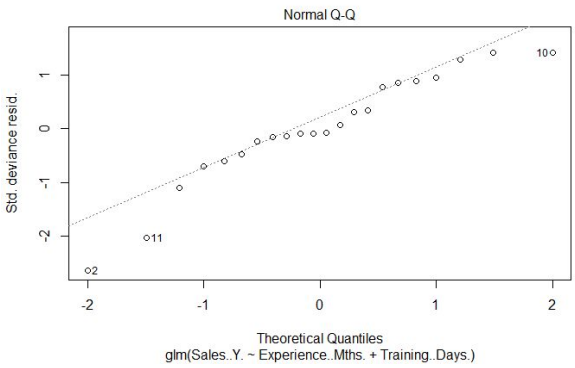
View(sales) #Seeing how are data was integrated into R

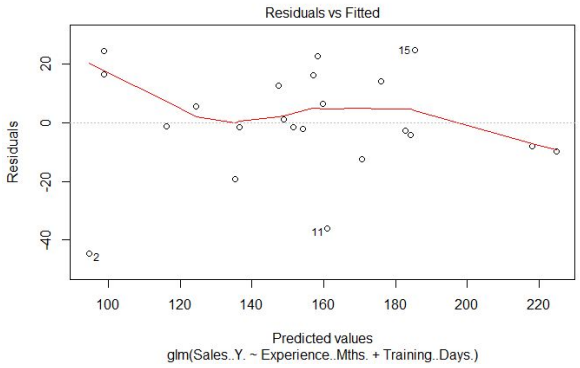
attach(sales)

plot(sales) #Checking if our data is normally distributed.



plot(res2)#This checks whether or not our residuals are homoskedastic and normally distributed.





res2 = glm(Sales..Y. ~ Experience..Mths. + Training..Days.) # This is the actual function we used to see what variables have the most significant impact.

summary(res2)

at the alpha = .05 we see that both of our categories have a significant effect on the total sales.

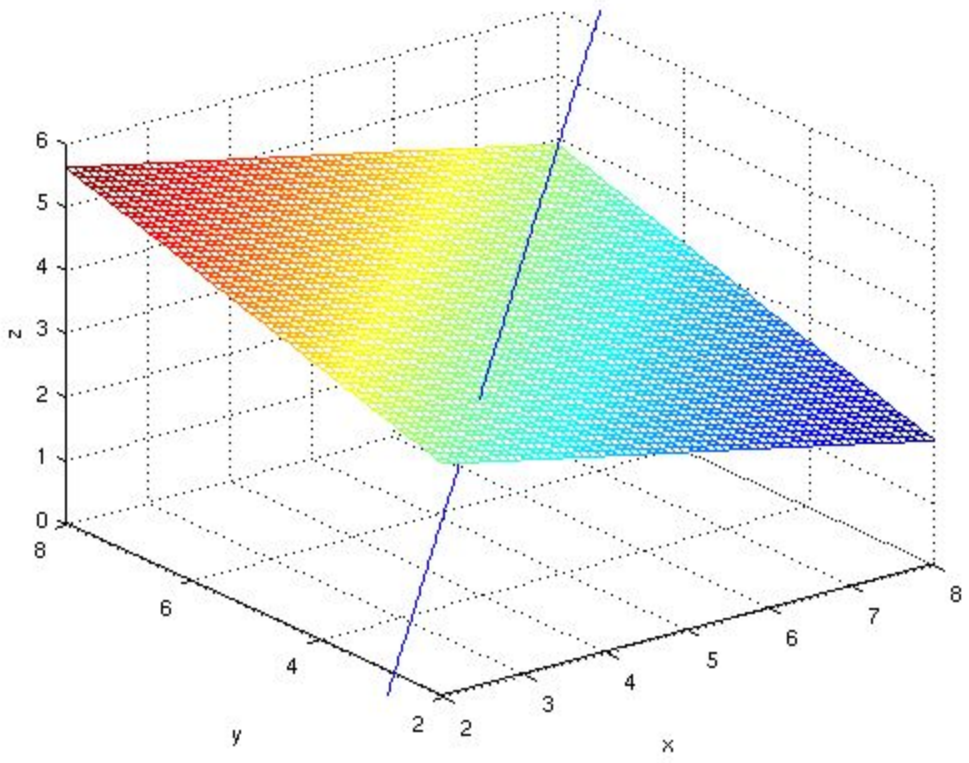
Assumptions

What we first need to do is make sure that we check our assumptions. We do this by plotting the data and making sure that our residuals are normally distributed; this checks out. What we also want to know is if our data is homoskedastic. This is measuring the variance of our residuals, this also checks out. Additionally, we would like to know is if our data is identically independently distributed. We will trust that our experiment was done correctly.

Results

	-Estimate	-Std. Error	t value	-Pr(>|t|)
(Intercept)	**79.7183**	**9.9627**	**8.002**	**1.67e-07 *****
Experience..Mths.	**4.0657**	**0.9226**	**4.407**	**0.000303 *****
Training..Days.	**10.8400**	**2.8337**	**3.825**	**0.001142 ****

From our data we see that the null hypothesis is rejected on both cases saying that both independent variables have a significant impact on the dependent sales variable. What we also see is that the Estimate which estimates our slope of the variable. Because we have multiple variables it is harder to graph in 2 dimensions. What we have right now is something like this: Where the least squares line is actually our least squares plane.



Sales (Y)	Experience (Mths)	Training (Days)
116	3	4
50	1	1
150	7	4
158	9	5
115	1	3
166	9	4
180	15	4
150	9	3
215	17	7
123	2	1
125	12	3
210	18	6
115	2	1
135	6	3
210	10	6
180	12	5
190	13	4
152	5	5
173	11	3
130	3	3
160	14	1
181	6	5