NON LINEAR AND SIMPLE LINEAR REGRESSION MODEL TO A DATASET BASED ON ENZYMOLOGY EXPERIMENT

**INTRODUCTION**

***Objective:*** The dataset based on an enzymology experiment that represents the substrate concentration(S) and the observed velocity (v) is given to us and our objective is to,

* Fit a non linear regression model that relates velocity to concentration using Michaelis-Menten equation.
* Analyse and examine whether we can fit a simple linear regression model that relates velocity and substrate concentration by using a suitable transformation.

***Data Description:*** The dataset that we have considered in this assignment is based on an enzymology experiment that represents the substrate concentration and the observed velocity.

The variables in the dataset are,

* The ***substrate concentration*** is the ***independent variable denoted by S.***
* The ***observed velocity*** is the ***dependent variable denoted by V.***

**Entering the dataset under consideration and converting into the dataframe.**

*#Storing all the values of substrate concentration in the variable S.*  
S<-**c**(0,1,2,5,8,12,30,50)  
  
*#Storing all the values of observed velocity in the variable V.*  
V<-**c**(0.0,11.1,25.4,44.8,54.5,58.2,72.0,60.1)  
  
*#Converting the above data into a dataframe.*  
data=**data.frame**(S,V)  
data

## S V  
## 1 0 0.0  
## 2 1 11.1  
## 3 2 25.4  
## 4 5 44.8  
## 5 8 54.5  
## 6 12 58.2  
## 7 30 72.0  
## 8 50 60.1

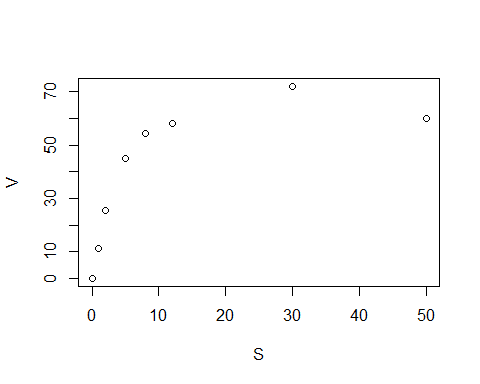
**ANALYSIS**

**Firstly, our objective is to fit a nonlinear regression model to the data using Michaelis-Menten equation.**

*#Attaching the dataset so that we can directly call the variables without calling the dataframe.*  
**attach**(data)

## The following objects are masked \_by\_ .GlobalEnv:  
##   
## S, V

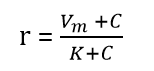
*#Obtaining the scatter plot for the above dataset.*  
**plot**(S,V)



**Interpretation:** From the figure 1 we observe that plot is not linear, hence we go for non linear reregression model. Also after a certaian point the graph has tendency to flatten therefore we proceed with Michaelis-Menten equation to fit the non linear model.

*#Loading the package 'stats' requred for fitting a non linear model.*  
**library**(stats)

Here, we make use of Michaelis-Menten equation i.e. given by the formula,



Where, Vm is the maximum rate in the system.

k is the concentration at which maximum rate is half.

r is the constant rate i.e. nothing but dependent variable.

c is the independent variable.

For fitting a nonlinear data in R using Michaelis-Menten equation the initial value of Vm and k should be specified,

*#We know that Vm is the maximum rate i.e. maximum observed velocity(with respect to this dataset) in the system therefore,*  
Vm=72.0  
  
*#Also, we know by the defination that k is the concentration at which maximum rate is half therefore,*   
*#Half of maximum rate is obtained as,*  
Vmh=Vm**/**2  
Vmh

## [1] 36

*#From the data table we observe that he concentration corresponding to Vmh=36 is,*  
k=5

*#Fitting a nonlinear model to the dataset.*  
nls\_model<-**nls**(V**~**vm**\***S**/**(k**+**S),start=**list**(k=5,vm=72.0))  
**summary**(nls\_model)

##   
## Formula: V ~ vm \* S/(k + S)  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## k 3.4372 0.8173 4.205 0.00565 \*\*   
## vm 73.2614 4.5825 15.987 3.8e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.106 on 6 degrees of freedom  
##   
## Number of iterations to convergence: 6   
## Achieved convergence tolerance: 6.206e-06

Hence, the the fitted nonlinear regression model is,

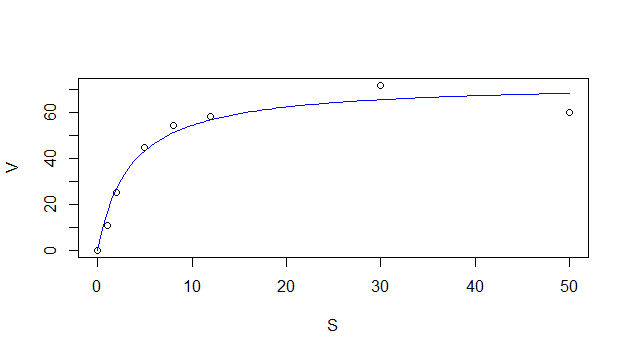
**V = vm \* S/(k + S)**

**i.e.**

**V = 73.2614\*S/(3.4372 + S)**

Therefore, the estimate of **Vm is 73.2614** and the estimate of **k is 3.4372.**

*#To plot the model to the plot.*  
x <- **seq**(**min**(S), **max**(S), length=100)  
y <- **predict**(nls\_model, **list**(S=x))  
**plot**(S,V)**points**(x, y, type='l', col='blue')



**Now our next objective is to analyse and examine whether we can fit a simple linear regression model that relates velocity and substrate concentration by using a suitable transformation.**

From figure 1 we observe that the plot is not linear hence we cannot fit a simple linear regression model to the dataset.

Further we try to fit simple linear regression model to the data and check where our results lead us to.

We try to fit a simple linear regression model to the dataset.

*#Fitting a simple linear regression model to the dataset.*  
reg=**lm**(V**~**S,data = data)  
**summary**(reg)

##   
## Call:  
## lm(formula = V ~ S, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.244 -17.162 4.351 15.771 19.245   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.2442 9.2299 2.952 0.0256 \*  
## S 1.0014 0.4328 2.314 0.0600 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.21 on 6 degrees of freedom  
## Multiple R-squared: 0.4715, Adjusted R-squared: 0.3834   
## F-statistic: 5.352 on 1 and 6 DF, p-value: 0.05998

**Interpretation:** From the above summary we obtain the following polynomial regression model,

**Y = B0+(B1*x) = 27.2442+(1.0014\**x)**

**i.e.**

**velocity = B0+(B1\*concentration)**

**velocity = 27.2442+(1.0014\* concentration)**

From the above model we observe that,

The intercept is 27.2442 which means that when the concentration is 0 the velocity is 27.2442.

B1=1.0014 indicates that for one unit of change in concentration the velocity will change by 1.0014 amount.

Since, coefficient of determination is 0.3834 which is lesser than 0.5, hence it indicates that our fitted regression model is of bad quality.

Now, we try to validate the following assumtions regarding the fitted model,

*#1. To check if the he relationship beetween y and x is linear.*

**Interpretation:** From the figure 1 we observe that the relationship between y and x is not linear hence we observe that the assumption about the linearity is violated.

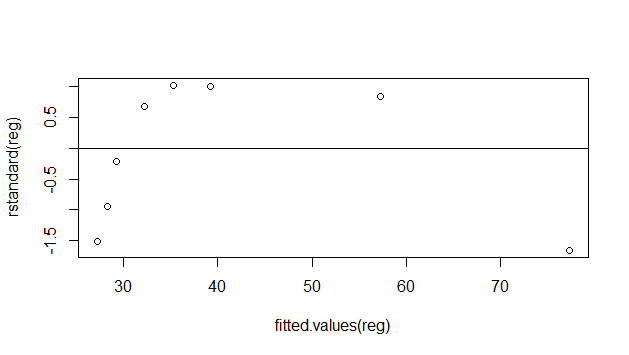
*#2. To check if the errors have zero mean.*  
residuals=**resid**(reg)  
**mean**(residuals)

## [1] 2.220446e-16

**Interpretation:** Hence we observe that the errors has mean 2.220446e-16 which is almost equal to zero hence the assumption of error mean to be zero is validated.

*#3. Assumption of homoscedasticity, i.e. the errors have constant variance.*  
*#Plot of fitted values against residuals.*  
fitted\_values<-**fitted.values**(reg)  
**plot**(**fitted.values**(reg),**rstandard**(reg))

**abline**(0,0)





**Interpretation:** From figure 3 i.e. the plot of fitted values against residuals we observe that all the points are not evenly spread around the horizontal line which means the errors have non-constant variance.

Since the residuals has non constant variance we try to stabilize the variance and check.

*#Removing the zeroes from the dataset*  
data\_new=data[**-**1,]  
data\_new

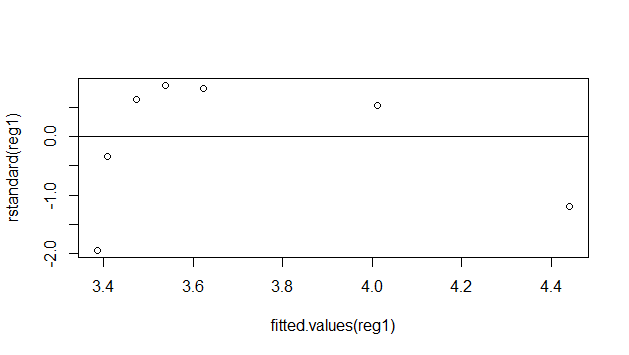
## S V  
## 2 1 11.1  
## 3 2 25.4  
## 4 5 44.8  
## 5 8 54.5  
## 6 12 58.2  
## 7 30 72.0  
## 8 50 60.1

*#Stabilizing the variance*  
  
*#Step 1 - Transforming the response variable.*  
y=**log**(data\_new**$**V)  
  
*#Step 2 - Fit and validate a polynomial regression model of order 2 in the transformed variable.*  
reg1=**lm**(y**~**data\_new**$**S,data=data\_new)  
**summary**(reg1)

##   
## Call:  
## lm(formula = y ~ data\_new$S, data = data\_new)  
##   
## Residuals:  
## 2 3 4 5 6 7 8   
## -0.9792 -0.1730 0.3299 0.4613 0.4408 0.2659 -0.3456   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.36463 0.29890 11.257 9.67e-05 \*\*\*  
## data\_new$S 0.02154 0.01311 1.643 0.161   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5822 on 5 degrees of freedom  
## Multiple R-squared: 0.3505, Adjusted R-squared: 0.2207   
## F-statistic: 2.699 on 1 and 5 DF, p-value: 0.1613

*#Now we check the assumption of homoscedasticity again.*  
  
*#Plot of fitted values against residuals.*  
fitted\_values<-**fitted.values**(reg1)  
**plot**(**fitted.values**(reg1),**rstandard**(reg1))

**abline**(0,0)





**Interpretation:** Again from the above plot we observe that the residuals follow some non random pattern i.e. the residual points are not randomly distributed in the plot hence we conclude that the residuals have non constant variance.

Hence, it is examined from the coefficient of determination and suitable transformation that we cannot fit a simple linear regression to this data.

**CONCLUSION**

From the above analysis it is observed that,

1. The fitted nonlinear regression model to the data is,

V = vm \* S/(k + S)

i.e.

V = 73.2614\*S/(3.4372 + S)

and the estimate of Vm is 73.2614 and the estimate of k is 3.4372.

1. On fitting a simple linear regression to the data we also observe that the coefficient of determination is very low i.e. 0.3834 which indicates that our model is of bad quality.

Also on doing the residual analysis we observe that the residuals follow a non random pattern and the assumptions about errors are violated. Hence we conclude that we cannot fit a simple linear regression that relates velocity and substrate concentration.