COMPARISON OF OLS AND ROBUST REGRESSION MODEL FOR THE DATA (WITH OUTLIERS) ON MAGAZINE ADVERTISING

***INTRODUCTION***

***Robust* Regression:**

In [robust statistics](https://en.wikipedia.org/wiki/Robust_statistics), robust regression is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) designed to overcome some limitations of traditional [parametric](https://en.wikipedia.org/wiki/Parametric_statistics) and [non-parametric methods](https://en.wikipedia.org/wiki/Non-parametric_statistics). Regression analysis seeks to find the relationship between one or more [independent variables](https://en.wikipedia.org/wiki/Dependent_and_independent_variables#Statistics) and a [dependent variable](https://en.wikipedia.org/wiki/Dependent_and_independent_variables#Statistics). Certain widely used methods of regression, such as [ordinary least squares](https://en.wikipedia.org/wiki/Ordinary_least_squares), have favorable properties if their underlying assumptions are true, but can give misleading results if those assumptions are not true; thus ordinary least squares is said to be not [robust](https://en.wikipedia.org/wiki/Robust_statistics) to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process.

In particular, [least squares](https://en.wikipedia.org/wiki/Least_squares) estimates for [regression models](https://en.wikipedia.org/wiki/Regression_model) are highly sensitive to [outliers](https://en.wikipedia.org/wiki/Outliers). While there is no precise definition of an outlier, outliers are observations that do not follow the pattern of the other observations. This is not normally a problem if the outlier is simply an extreme observation drawn from the tail of a normal distribution, but if the outlier results from non-normal measurement error or some other violation of standard ordinary least squares assumptions, then it compromises the validity of the regression results if a non-robust regression technique is used. Thus, to overcome this problem we make use of robust regression.

***Objective:***

Here our objective is to,

1. Bulid a regression model ln R on ln P using least square method of estimation.
2. Build a robust fit for the full data using M-estimation and regression model of ln R on ln P by removing the 4 outliers and compare the results.

***Data Description:***

The dataset consists of 41 records of the details of magazine published in 1986 and the advertising pages (in hundreds) and advertising revenue (in million of dollars) of each magazine.

The variables we are going to build our model on are,

* y = ***log R*** is the ***dependent variable***, where R is advertising revenue (in millions of dollars).
* x = ***log P*** is the ***independent variable***, where P is advertising pages (in hundreds).

*#Importing the Salary dataset from current working directory.*  
**library**(readxl)  
Magazine <- **read\_excel**("Magazine.xlsx")  
  
*#Obtaining the first few records of the dataset.*  
**head**(Magazine)

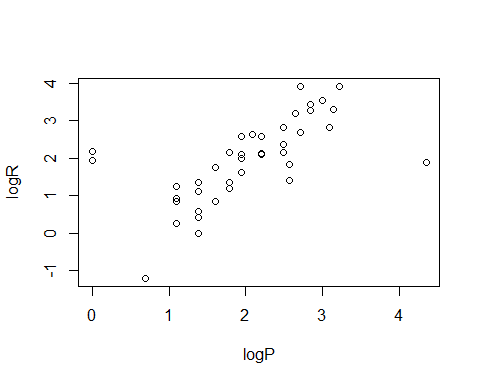
## # A tibble: 6 x 3  
## Magazine P R  
## <chr> <dbl> <dbl>  
## 1 Cosmopolitan 25 50   
## 2 Redbook 15 49.7  
## 3 Glamour 20 34   
## 4 Southern Living 17 30.7  
## 5 Vogue 23 27   
## 6 Sunset 17 26.3

*#Since we want to build a regression model of Log P on Log R.*  
logP=**log**(Magazine**$**P)*#taking log of P*  
logR=**log**(Magazine**$**R)*#taking log of R*  
  
*#Forming a new dataframe with new values.*  
Magazine1=**data.frame**(Magazine**$**Magazine,logP,logR)  
  
*#Obtaining the first few records of the dataset.*  
**head**(Magazine1)

## Magazine.Magazine logP logR  
## 1 Cosmopolitan 3.218876 3.912023  
## 2 Redbook 2.708050 3.906005  
## 3 Glamour 2.995732 3.526361  
## 4 Southern Living 2.833213 3.424263  
## 5 Vogue 3.135494 3.295837  
## 6 Sunset 2.833213 3.269569

**ANALYSIS**

*#Obtaining the scatter plot of log R and log P.*  
**plot**(logP,logR)

From the Figure 1 we observe thatlog P and logR are positively related that means as lgP increases logR will also increase. But we also observe that there exists outliers in the dataset.

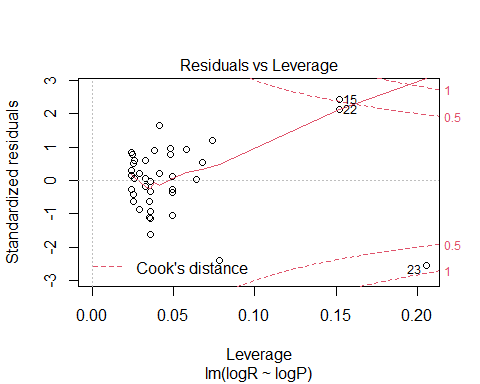
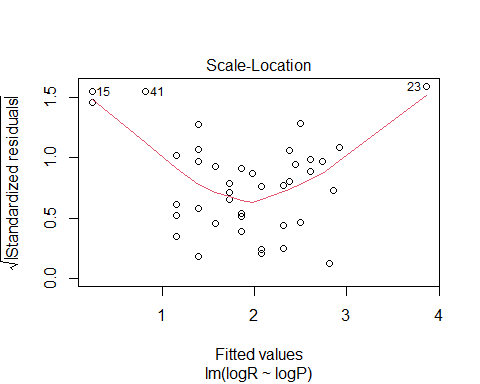
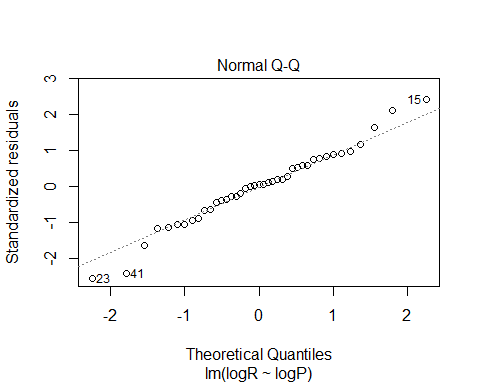
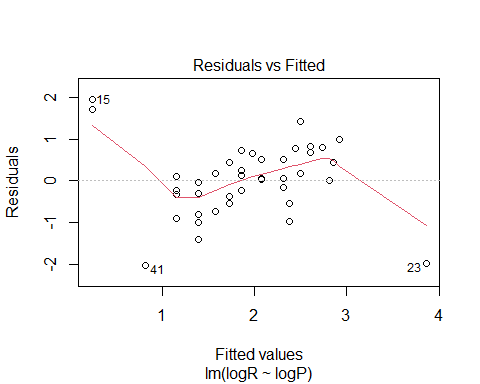
1. Regression Model of log R on log P using least squares.

*#Fitting a simple linear regression using least square method.*  
reg=**lm**(logR**~**logP,Magazine1)  
  
*#Obtaining the summary of the regression model.*  
**summary**(reg)

##   
## Call:  
## lm(formula = logR ~ logP, data = Magazine1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.01534 -0.53524 0.04836 0.50718 1.94245   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.2323 0.3398 0.684 0.498   
## logP 0.8354 0.1571 5.318 4.57e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8719 on 39 degrees of freedom  
## Multiple R-squared: 0.4203, Adjusted R-squared: 0.4055   
## F-statistic: 28.28 on 1 and 39 DF, p-value: 4.575e-06

From the above summary we observe that since there exists outliers in the dataset weight of the estimate is increased and also the estimate associated logP is not significant hence we cannot proceed with least square method of estimation in the presence of outliers.

*#Obtaining the residual versus fitted plot of the fitted regression model.*  
**plot**(reg)



On observing Figures above, we see that there exists outliers that are creating problems to the fitted model. We observe that 15th,22nd,23rd,41st observations are creating problems. Hence proceed on fitting the robust regression model using M Estimation.

1. Robust regression using M ESTIMATION method of estimating parameter.

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

## Warning: package 'MASS' was built under R version 4.0.4

*#Fitting a linear model by robust regression using an M estimator.*  
rr.bisquare<-**rlm**(logR**~**logP,Magazine,psi=psi.bisquare)*#using bisquare method we assign less weightage to outliers.*  
**summary**(rr.bisquare)

##   
## Call: rlm(formula = logR ~ logP, data = Magazine, psi = psi.bisquare)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5460 -0.3412 0.1370 0.4087 3.3571   
##   
## Coefficients:  
## Value Std. Error t value  
## (Intercept) -1.1823 0.2311 -5.1163  
## logP 1.5229 0.1068 14.2561  
##   
## Residual standard error: 0.5425 on 39 degrees of freedom

*#Extracting weights*  
bweight<-**data.frame**(Magazine=Magazine**$**Magazine,resid=rr.bisquare**$**resid,weight=rr.bisquare**$**w)  
  
*#Arranging the extracted weights in ascending order.*  
bweightasc<-bweight[**order**(rr.bisquare**$**w),]  
bweightasc

## Magazine resid weight  
## 15 Modern Bride 3.35705137 0.0000000  
## 22 Town and Country 3.12820980 0.0000000  
## 23 True Story -3.54600513 0.0000000  
## 27 Yankee -1.31298881 0.5375559  
## 41 Soap Opera Digest -1.07729782 0.6730130  
## 2 Redbook 0.96409452 0.7329487  
## 40 High Times -0.92894968 0.7507188  
## 24 Brides -0.89942649 0.7652703  
## 12 Life Magazine 0.79900339 0.8121247  
## 30 New Woman 0.76193711 0.8283487  
## 8 New York Magazine -0.69787273 0.8549335  
## 11 Psychology Today 0.64009425 0.8771765  
## 16 Parents 0.61687251 0.8856607  
## 37 Saturday Evening Post -0.52348457 0.9169765  
## 25 Book Digest Magazine 0.48907301 0.9273178  
## 17 Architectural Digest -0.46200902 0.9350245  
## 35 Self -0.43587579 0.9420594  
## 28 Playgirl 0.43202688 0.9430490  
## 33 Mother Earth News 0.42546488 0.9447419  
## 13 Smithsonian 0.40866087 0.9489624  
## 7 House and Garden 0.36590834 0.9589748  
## 31 Ms. -0.35252805 0.9619053  
## 34 1001 Decorating Ideas 0.34208327 0.9640978  
## 36 Decorating & Craft Ideas -0.34116301 0.9642996  
## 19 Apartment Life 0.32292071 0.9679742  
## 5 Vogue -0.29704705 0.9728788  
## 4 Southern Living 0.29173572 0.9738203  
## 10 Mademoiselle -0.26088889 0.9790471  
## 14 Rolling Stone -0.24122118 0.9820736  
## 38 McCall's Needlework and Craft -0.22846159 0.9839113  
## 39 Weight Watchers -0.22846159 0.9839113  
## 9 House Beautiful 0.21333354 0.9859566  
## 21 Gourmet 0.20666091 0.9868189  
## 1 Cosmopolitan 0.19215353 0.9885981  
## 29 Saturdat Review -0.18547396 0.9893826  
## 32 Cuisine 0.16966261 0.9911062  
## 26 W -0.15197290 0.9928661  
## 3 Glamour 0.14632629 0.9933794  
## 6 Sunset 0.13704201 0.9941916  
## 20 Bon Appetit -0.05981721 0.9988939  
## 18 Harper's Bazaar -0.04769585 0.9992969

Hence it can be observed from the above results that the huber’s method has assigned least weightage to the outliers so that they donot effect the fitted model.

**Our robust regression model is,**

**logR = -1.1823 + 1.5229\*logP**

1. Fiiting a regression model using the method of least square after deleting the four outliers.

*#Removing the observations which are creating problems.*  
Magazine2=Magazine1[**c**(**-**15,**-**22,**-**23,**-**41),]  
Magazine2

## Magazine.Magazine logP logR  
## 1 Cosmopolitan 3.218876 3.9120230  
## 2 Redbook 2.708050 3.9060049  
## 3 Glamour 2.995732 3.5263605  
## 4 Southern Living 2.833213 3.4242627  
## 5 Vogue 3.135494 3.2958369  
## 6 Sunset 2.833213 3.2695689  
## 7 House and Garden 2.639057 3.2027464  
## 8 New York Magazine 3.091042 2.8273136  
## 9 House Beautiful 2.484907 2.8154087  
## 10 Mademoiselle 2.708050 2.6810215  
## 11 Psychology Today 2.079442 2.6246686  
## 12 Life Magazine 1.945910 2.5802168  
## 13 Smithsonian 2.197225 2.5726122  
## 14 Rolling Stone 2.484907 2.3608540  
## 16 Parents 1.791759 2.1633230  
## 17 Architectural Digest 2.484907 2.1400662  
## 18 Harper's Bazaar 2.197225 2.1162555  
## 19 Apartment Life 1.945910 2.1041342  
## 20 Bon Appetit 2.197225 2.1041342  
## 21 Gourmet 1.945910 1.9878743  
## 24 Brides 2.564949 1.8245493  
## 25 Book Digest Magazine 1.609438 1.7578579  
## 26 W 1.945910 1.6292405  
## 27 Yankee 2.564949 1.4109870  
## 28 Playgirl 1.386294 1.3609766  
## 29 Saturdat Review 1.791759 1.3609766  
## 30 New Woman 1.098612 1.2527630  
## 31 Ms. 1.791759 1.1939225  
## 32 Cuisine 1.386294 1.0986123  
## 33 Mother Earth News 1.098612 0.9162907  
## 34 1001 Decorating Ideas 1.098612 0.8329091  
## 35 Self 1.609438 0.8329091  
## 36 Decorating & Craft Ideas 1.386294 0.5877867  
## 37 Saturday Evening Post 1.386294 0.4054651  
## 38 McCall's Needlework and Craft 1.098612 0.2623643  
## 39 Weight Watchers 1.098612 0.2623643  
## 40 High Times 1.386294 0.0000000

*#Fitting a multiple linear regression model to our dataset.*  
reg\_ols=**lm**(logR**~**logP,Magazine2)  
reg\_ols

##   
## Call:  
## lm(formula = logR ~ logP, data = Magazine2)  
##   
## Coefficients:  
## (Intercept) logP   
## -0.9941 1.4351

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

##   
## Call:  
## lm(formula = logR ~ logP, data = Magazine2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2759 -0.3202 0.1032 0.3523 1.0137   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9941 0.2845 -3.494 0.00131 \*\*   
## logP 1.4351 0.1318 10.891 8.67e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5185 on 35 degrees of freedom  
## Multiple R-squared: 0.7722, Adjusted R-squared: 0.7657   
## F-statistic: 118.6 on 1 and 35 DF, p-value: 8.667e-13

**The regression model by the method of least square on deleting the outliers from the original dataset is,**

**logR = -0.9941 + 1.4351\*logP**

**CONCLUSION:**

We observe from the above analysis that on fitting the regression model for the data with outliers using least square method does not lead to accurate. In that case, due to the higher weights of outliers the error gets reduced which might leads to the wrong prediction result. Also the presence of outliers in the dataset causes the regression coefficient to be insignificant also the weight assigned to them is more.Since the least square method gives more weight to outlying observation.

Thus, We proceed to robust regression and we observed that it assigns less weight to the outlying observation and give us the better estimate.

The **regression model obtained by robust regression** is,

**logR = -1.1823 + 1.5229\*logP**

and further we proceed to compare it with method of least square on removing the outliers,

The **regression model by the method of least square** **on deleting the outliers** from the original dataset is,

**logR = -0.9941 + 1.4351\*logP**

Thus, on comparing both the results we observe that the method of M-Estimation with the presence of outliers and the method of least square on deleting the outliers gives the same result.