**Lab 9**

Name: Jesto Peter

Title: Robust Regression

Date: 28/01/2023

Class: 2MSTAT

* **Objective**
* Select any two datasets and fit a **robust regression model**.
* To prepare a report that consists of an introduction, analysis, and conclusion.
* **Procedure**
* Dataset 1: Steam

#Loading the Package  
library(robustbase)

library(olsrr)

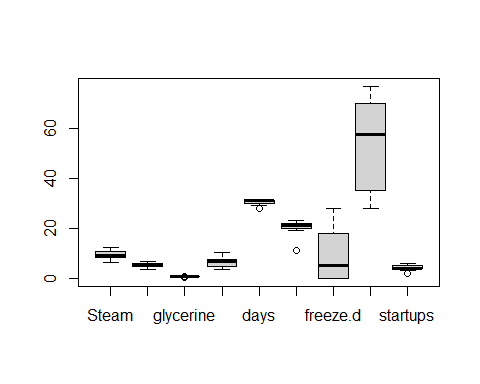
*#Using the Steam Datset(Dataset 1)*

data=steamUse  
attach(data)  
data

## Steam fattyAcid glycerine wind days op.days freeze.d temperature startups  
## 1 10.98 5.20 0.61 7.4 31 20 22 35.3 4  
## 2 11.13 5.12 0.64 8.0 29 20 25 29.7 5  
## 3 12.51 6.19 0.78 7.4 31 23 17 30.8 4  
## 4 8.40 3.89 0.49 7.5 30 20 22 58.8 4  
## 5 9.27 6.28 0.84 5.5 31 21 0 61.4 5  
## 6 8.73 5.76 0.74 8.9 30 22 0 71.3

**Data Interpreation**: Here we have our Dependent Variable as steam which denotes the monthly usage of steam per months, and we have other 8 variables(regressors) in our data set.

#Checking for Outliers using Boxplot  
boxplot(data)

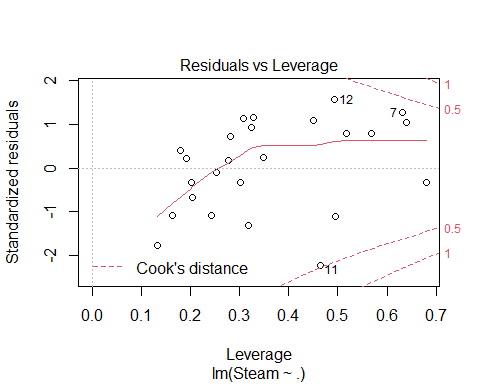


#Finding the exact values of the Outliers  
boxplot(data,plot=FALSE)$out

## [1] 0.42 0.45 0.95 28.00 11.00 11.00 2.00

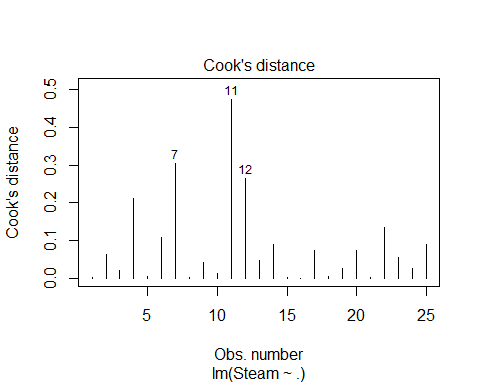
#Interpretation: Here we see that there are 7 leverage points,since these are more than 3 hence we cant delete them.

#Checking Cooks Distance and Highly Influential points  
  
#Cooks Distance  
mlr=lm(Steam~.,data)  
plot(mlr,which=5)



**#Intepretation**: We see that the value 11 is almost near the dottedline and hence maybe affecting the data in general linear modelling.

**#Highly Influential Points**  
plot(mlr,which=4)



**#Interpretation**: We see that the points 7, 11 and 12 are highly influential and from previous outliers plot we know that 11 is an outlier. Soif we remove this outlier our model will be affected on a big scale.

#Forward Selection(Tofind significant variables)  
  
ols\_step\_forward\_aic(mlr,details = FALSE)

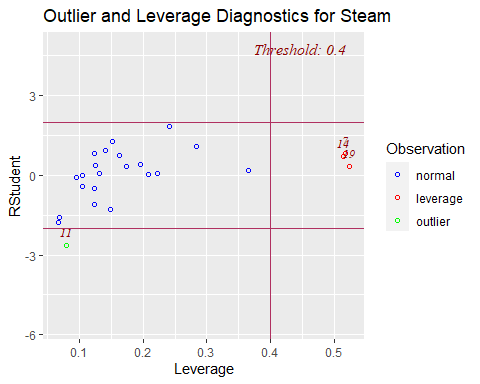
##   
## Selection Summary   
## -----------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -----------------------------------------------------------------  
## temperature 69.043 45.592 18.223 0.71444 0.70202   
## fattyAcid 53.214 54.884 8.931 0.86004 0.84732   
## op.days 51.456 56.131 7.685 0.87958 0.86238   
## days 50.406 57.014 6.802 0.89341 0.87209   
## -----------------------------------------------------------------

#**CONCLUSION:** We get the significant variables as temperature + fattyAcid + op.days + days

#Multiple Linear Regression model after forward selection  
mlr1=lm(Steam~temperature + fattyAcid + op.days + days,data)  
summary(mlr1)

##   
## Call:  
## lm(formula = Steam ~ temperature + fattyAcid + op.days + days,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.30055 -0.23729 0.07946 0.34148 0.88383   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.098781 5.350204 0.018 0.9855   
## temperature -0.075582 0.007197 -10.502 1.38e-09 \*\*\*  
## fattyAcid 0.297672 0.235842 1.262 0.2214   
## op.days 0.142301 0.060350 2.358 0.0287 \*   
## days 0.288734 0.179222 1.611 0.1228   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5832 on 20 degrees of freedom  
## Multiple R-squared: 0.8934, Adjusted R-squared: 0.8721   
## F-statistic: 41.91 on 4 and 20 DF, p-value: 1.88e-09

#Plot for checking outliers and residuals  
ols\_plot\_resid\_lev(mlr1)



#**Conclusion :** The outlier is 11 and the rest of the points in red are the leverage points

#Robust Regression  
rob1=lmrob(Steam~temperature + fattyAcid + op.days + days,data)  
summary(rob1)

##   
## Call:  
## lmrob(formula = Steam ~ temperature + fattyAcid + op.days + days, data = data)  
## \--> method = "MM"  
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2277 -0.2022 0.1209 0.3049 3.0403   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.46344 3.38569 0.728 0.475   
## temperature -0.08320 0.00641 -12.980 3.35e-11 \*\*\*  
## fattyAcid 0.26923 0.17857 1.508 0.147   
## op.days 0.38925 0.07869 4.947 7.77e-05 \*\*\*  
## days 0.05922 0.14636 0.405 0.690   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Robust residual standard error: 0.4602   
## Multiple R-squared: 0.8952, Adjusted R-squared: 0.8743   
## Convergence in 15 IRWLS iterations  
##   
## Robustness weights:   
## 2 observations c(7,19) are outliers with |weight| = 0 ( < 0.004);   
## one weight is ~= 1. The remaining 22 ones are summarized as  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4567 0.9248 0.9732 0.9041 0.9886 0.9989   
## Algorithmic parameters:   
## tuning.chi bb tuning.psi refine.tol   
## 1.548e+00 5.000e-01 4.685e+00 1.000e-07   
## rel.tol scale.tol solve.tol eps.outlier   
## 1.000e-07 1.000e-10 1.000e-07 4.000e-03   
## eps.x warn.limit.reject warn.limit.meanrw   
## 1.395e-10 5.000e-01 5.000e-01   
## nResample max.it best.r.s k.fast.s k.max   
## 500 50 2 1 200   
## maxit.scale trace.lev mts compute.rd fast.s.large.n   
## 200 0 1000 0 2000   
## psi subsampling cov   
## "bisquare" "nonsingular" ".vcov.avar1"   
## compute.outlier.stats   
## "SM"   
## seed : int(0)

**#Inter**: We see that after 15 iterations the consecutive values of regression coefficients are same and hence converge and we get the optimal regression coefficients as shown in the table of summary

#Checking which is a better model  
  
summary(mlr1)$sigma

## [1] 0.5831847

summary(rob1)$sigma

## [1] 0.4601965

**#Conclusion:** We see that the Residual Standard Error is lower for the Robust Regression Model**(0.4601965),** than the mlr **(0.5831847).**  
  
#Hence Robust Regression gives a better model.

* **Dataset 2: Toxicity**

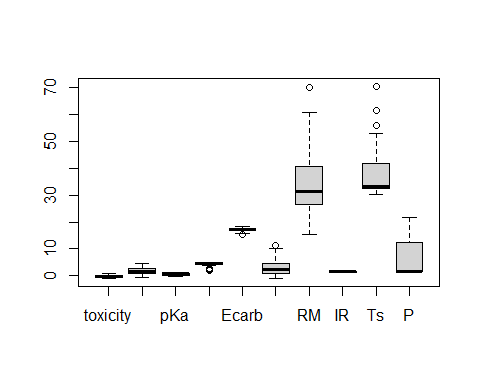
data(toxicity, package="robustbase")  
attach(toxicity)

data2=toxicity  
data2

## toxicity logKow pKa ELUMO Ecarb Emet RM IR Ts P  
## 1 -0.15 1.68 1.0000 4.81 17.8635 1.4838 31.36 1.425 31.3 12.430  
## 2 -0.33 0.94 0.9800 4.68 16.9491 0.0000 22.10 1.408 30.4 8.760  
## 3 -0.34 1.16 0.9600 4.86 17.1806 0.2778 26.73 1.418 30.9 10.590  
## 4 0.03 2.75 1.0000 4.83 18.4794 3.5836 40.63 1.435 31.8 16.100  
## 5 -0.57 0.79 0.9700 4.80 16.8022 1.0232 22.14 1.411 32.5 8.770  
## 6 0.08 2.64 1.0100 4.90 18.3937 3.7145 40.63 1.435 31.8

**Data Interpretation:** Here we have our Dependent Variable as “toxicity” which denotes toxicity of carboxylic acids, and we have other 9 variables(regressors) in our data set, which are attributes fro carboxylic acids

#Checking for Outliers using Boxplot  
boxplot(data2)

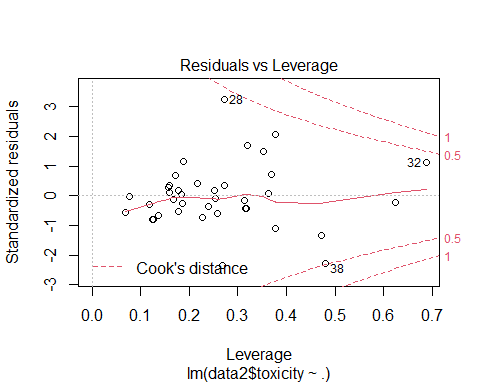


#Finding the exact values of the Outliers  
boxplot(data2,plot=FALSE)$out

## [1] 2.7700 1.8600 2.6700 1.8600 2.3900 15.4004 11.2898 70.0300 56.1000  
## [10] 61.6000 70.5000

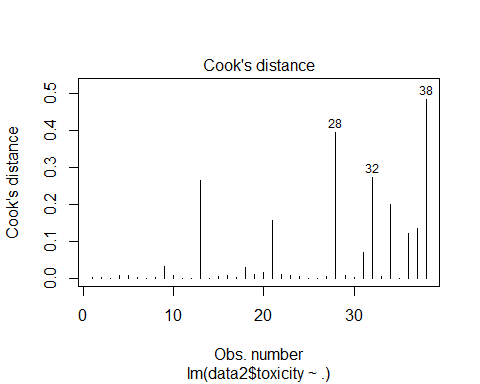
**#Interpretation:** Here we see that there are 11 leverage points,since these are more than 3 hence we cant delete them.

#Checking Cooks Distance and Highly Influential points  
  
#Cooks Distance  
mlr2=lm(data2$toxicity~.,data2)  
plot(mlr2,which=5)



**#Intepretation:** We see that the value 38 is almost near the dottedline and hence maybe affecting the data in general linear modelling using ols.

**#Highly Influential Points**  
plot(mlr2,which=4)



**#Interpretation**: We see that the points 28, 32 and 38 are highly influential and from previous outliers plot we know that 11 is an outlier. So if we remove this outlier our model will be affected on a big scale.  
#We saw that 38 was a influential leverage point using cooks distance. So we cant remove it.

#**Forward Selection(Tofind significant variables)**  
  
ols\_step\_forward\_aic(mlr2,details = FALSE)

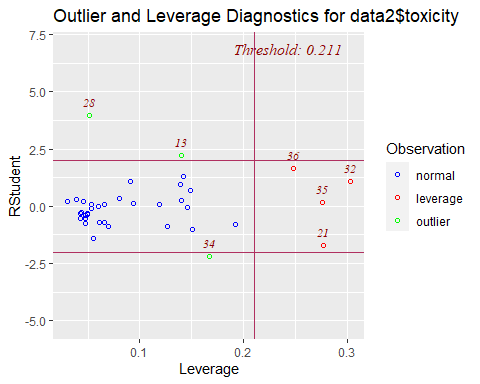
##   
## Selection Summary   
## ---------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ---------------------------------------------------------------  
## logKow 1.762 4.177 1.990 0.67733 0.66837   
## ELUMO -11.842 4.848 1.320 0.78600 0.77377   
## RM -19.764 5.151 1.017 0.83517 0.82063   
## ---------------------------------------------------------------

**#CONCLUSION**: We get the significant variables as logKow + ELUMO + RM

#Multiple Linear Regression model  
mlr3=lm(data2$toxicity~logKow + ELUMO + RM,data2)  
summary(mlr3)

##   
## Call:  
## lm(formula = data2$toxicity ~ logKow + ELUMO + RM, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.32946 -0.11379 -0.00531 0.04656 0.55854   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.428985 0.181029 2.370 0.0236 \*   
## logKow 0.390551 0.039164 9.972 1.25e-11 \*\*\*  
## ELUMO -0.182345 0.036243 -5.031 1.57e-05 \*\*\*  
## RM -0.012943 0.004064 -3.185 0.0031 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1729 on 34 degrees of freedom  
## Multiple R-squared: 0.8352, Adjusted R-squared: 0.8206   
## F-statistic: 57.42 on 3 and 34 DF, p-value: 2.138e-13

**#Plot for checking outliers and residuals**  
ols\_plot\_resid\_lev(mlr3)



#**Conclusion :** The outliers are 28 ,13 and 34 and the rest of the points in red are the leverage points

**#Robust Regression**  
rob2=lmrob(data2$toxicity~logKow + ELUMO + RM,data2)  
summary(rob2)

##   
## Call:  
## lmrob(formula = data2$toxicity ~ logKow + ELUMO + RM, data = data2)  
## \--> method = "MM"  
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.396022 -0.073732 -0.006431 0.075998 0.599970   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.542666 0.168623 3.218 0.00283 \*\*   
## logKow 0.398679 0.045247 8.811 2.69e-10 \*\*\*  
## ELUMO -0.222918 0.041634 -5.354 5.95e-06 \*\*\*  
## RM -0.011806 0.005341 -2.211 0.03389 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Robust residual standard error: 0.1148   
## Multiple R-squared: 0.8971, Adjusted R-squared: 0.888   
## Convergence in 34 IRWLS iterations  
##   
## Robustness weights:   
## observation 28 is an outlier with |weight| = 0 ( < 0.0026);   
## 2 weights are ~= 1. The remaining 35 ones are summarized as  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2094 0.8462 0.9572 0.8784 0.9805 0.9983   
## Algorithmic parameters:   
## tuning.chi bb tuning.psi refine.tol   
## 1.548e+00 5.000e-01 4.685e+00 1.000e-07   
## rel.tol scale.tol solve.tol eps.outlier   
## 1.000e-07 1.000e-10 1.000e-07 2.632e-03   
## eps.x warn.limit.reject warn.limit.meanrw   
## 1.274e-10 5.000e-01 5.000e-01   
## nResample max.it best.r.s k.fast.s k.max   
## 500 50 2 1 200   
## maxit.scale trace.lev mts compute.rd fast.s.large.n   
## 200 0 1000 0 2000   
## psi subsampling cov   
## "bisquare" "nonsingular" ".vcov.avar1"   
## compute.outlier.stats   
## "SM"   
## seed : int(0)

**#Inter:** We see that after 34 iterations the consecutive values of regression coefficients are same and hence converge and we get the optimal regression coefficients as shown in the table of summary.

**#Checking which is a better model**  
  
summary(mlr3)$sigma

## [1] 0.1729155

summary(rob2)$sigma

## [1] 0.1147799

**#Conclusion:** We see that the Residual Standard Error is lower for the Robust Regression Model(0.1729155), than the mlr (0.1147799).  
  
#Hence Robust Regression gives a better model.

**Conclusion**

**Dataset1 :** We see in the Outlier and Leverage Diagnostic Plot, that in the OLS model the point 11 is an outlier and the rest of the points in red are the leverage points. Hence Robust Regression can be applied

**Dataset2 :** We see in the Outlier and Leverage Diagnostic Plot, that in the OLS model the points 28,34 and 13 are outliers and rest 4 points are leverage points. Hence the model can be better obtained using Robust Regression.

We obtained the residual standard error of the Robust Model lower than that in the OLS model in both datasets. Hence conclude that it was a better fit model.