Problem:

The following table gives the amount of time required by the route driver in selling soft drinks:

Delivery time Minutes (Y)	Distance Feet (X)
16.68	560
.	.
.	.
.	.
.	.
10.75	150

(i) Fit a linear regression model of Yon X.

(ii) Draw a scatter plot of X and Y. Is there any unusual observation present in the data?

(iii) Repeat (i) and (ii), if we change the 9^{th} data of Y from 79.24 to 65.24?

(iv) Repeat (i) and (ii), if we change the 9^{th} data of Y from 79.24 to 65.24, the 22^{nd} data of Y from 52.32 to 35.32 and the 22^{nd} data of X from 810 to 610?

(v) Find the influential observation, high leverage point and outlier (if any) from the above data.

(vi) Comment on your findings.

Answer:

The linear regression model of Y on X is Y=………

The fitted regression line is $\hat{Y} = ……………$

We know that $e_i = Y_i - \hat{Y}$

We also know that $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Standardized residuals: $d_i = \frac{|e_i|}{\sqrt{MS_{RES}}}$; where $MS_{RES} = \frac{SS_{\mathrm{Re}s}}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$

Studentized residuals: $r_i = \frac{|e_i|}{\sqrt{MS_{RES}(1 - w_{ii})}}$

Cook's distance: $CD_i = \frac{h_{ii}}{P(1-h_{ii})} r_i^2$

Table 1:

SL	X	e_i	h_{ii}	d_i	r_i	CD_i
		-8.788	0.0472	0.6288	0.6442	0.0102

Cut-off point: $w_{ii} > \frac{2p}{n}$, p = no. of parameter. Then the corresponding observation will be high leverage point.

Cut-off point: d_i or r_i > 3, Then the corresponding value will be outlier.

Cut-off point: $CD_i > 1$, Then the corresponding observation will be influential.

Comment:

```
y<-c( 16.68, 11.50, 12.03, 14.88, 13.75, 18.11,  8.00, 17.83, 79.24, 21.50, 40.33, 21.00, 13.50,
19.75, 24.00 ,29.00, 15.35, 19.00,  9.50, 35.10, 17.90, 52.32, 18.75 ,19.83, 10.75)

x<-c(560,  220,  340,   80,  150,  330,  110 , 210, 1460,  605,  688,  215,  255 , 462,  448,  776,
200,  132 ,  36,  770,  140,  810 , 450,  635,  150)

##(i)
model<-lm(y~x);model
##(ii)
plot(x,y)
abline(model)

#######(iii)
###High leverage value
d=(x-mean(x))^2/(sum((x-mean(x))^2));d
n=length(x);n
hii=(1/n)+d
hii

######
h<-hat(x);h

###Cut-off Point#####
p=2
CP=2*(p/n);CP ### Twice the mean rule

############Identification###########
CP>hii
which(hii>CP)

#####outlier
r<-model$residuals;r
msr<-sum(r^2)/(n-p);msr
ar<-abs(r);ar
di<-ar/sqrt(msr);di  ##Standardized residuals
out<-di[di>3];out

#or
ri<-ar/sqrt(msr*(1-h));ri  ##Studentized residuals
out<-ri[ri>3];out
```

```
#### Influential Observation#####
cdi<-(h*ri^2)/(p*(1-h));cdi
IO<-cdi[cdi>1];IO

###
cooks.distance(model)

####
y.hat=4.96116+.04257*x
y.hat
ei=y-y.hat;ei
sl.<-c(1:25)
data2=data.frame(sl.,y,x,y.hat,ei,hii,di,ri,cdi);data2

View(data2)
```