STAT S 670 – Exploratory Data Analysis – Homework #4

Ganesh Nagarajan
gnagaraj@indiana.edu

Solutions

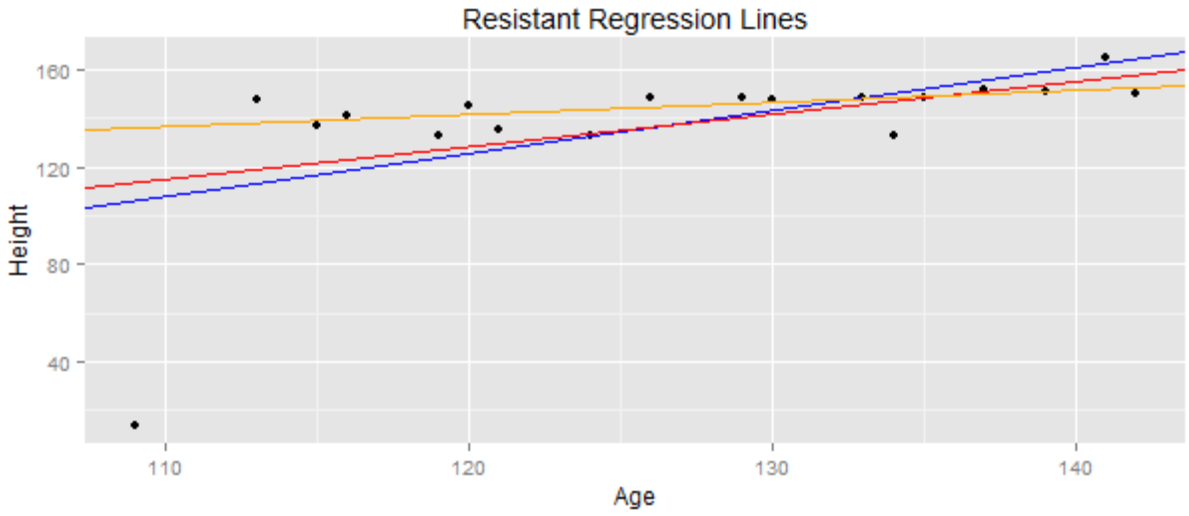
1. Line equations

a) "Resistant Regression y = 0.506976760317702 x+ 80.8151141174279"

b) "Least Squares y = 1.75866327779347 x -85.3926812890269"

c) "Barlett Line y = 1.34220588235294 x -32.5720016339868"

The Actual plot with the these regression types are as follows,



Legend: Blue : Linear Model, Red : Barlett, Orange : RR

2. Given that the points are (-2,-2),(-1,-1),(1,1),(2,2),(u,v)

a) Considering resistance lines, y=x for (u,v)

From UREDA, Pg 132, the slope of the resistant line can be given as follows $B_0=\frac{Y_R-Y_L}{X_R-X_L}$

Intercept a = $\frac{1}{3}[(Y_L - b_0X_L) + (Y_M - b_0X_M) + (Y_R - b_0X_R)]$

Since y=x, as in the question b=1 and a=0

So we have $(Y_R - Y_L) = (X_R - X_L)$ #Eq 2.1

So from the previous question, intercept a=0;

Hence, $X_R + X_L + X_M = Y_R + Y_L + Y_M$

Case 1:

Assume (u,v) is in the right group. From above equation 2.1,

Hence (-2,-2),(-1,-1) belongs to single group, (1,1) belongs to middle group, (2,2) and (u,v) belongs to third group.

$Y_R = \frac{(2+V)}{2} X_L = \frac{(2+u)}{2}$

$X_R = \frac{-3}{2}\ Y_R = \frac{-3}{2}$

Substituting on #Eq 2.1, we get u=v
Hence the observation is, wen (u,v) is in the right third of the three resistant line and u=v, there will be no impact to the resistant line.

Case 2:
Assume (u,v) is in the left group. Using #Eq 2.1
$X_L = \frac{(u-2)}{2}\ Y_L = \frac{(v-2)}{2}$
$Y_R = \frac{3}{2}\ X_R = \frac{3}{2}$
Substituting above equation, we get u=v.
Hence as same as above, when (u,v) is in the left of the third, and when u=v, there will be no impact to the resistant line.

Case 3:
Consider now, (u,v) is in the middle of the third group.
$Y_R = \frac{3}{2}\ X_R = \frac{3}{2}$
$Y_L = -\frac{3}{2}\ X_L = -\frac{3}{2}$
Even now as seen in above cases, with Eq 2.1, u=v.
Hence when u=v, this point does not impact the resistance line.

b) For least squares regression, given y=x find all possible positions of (u,v)
slope b = $\frac{\sum(\mathrm{xi}-\bar{x})(\mathrm{yi}-\bar{y})}{\sum(\mathrm{xi}-\bar{x})2}$

Given that slope =1, as y=mx+c, c=0 and m=1, $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})^2$

$\bar{x} = \frac{u}{5}, \bar{y} = \frac{v}{5}$

So substituting this to the mean square equation, Consider the points (-2,-2),(-1,-1),(1,1),(2,2),(u,v)

$$\left(-2-\frac{u}{5}\right)\left(-2-\frac{v}{5}\right)+\left(-1-\frac{u}{5}\right)\left(-1-\frac{v}{5}\right)+\left(1-\frac{u}{5}\right)\left(1-\frac{v}{5}\right)+\left(2-\frac{u}{5}\right)\left(2-\frac{v}{5}\right)+\left(u-\frac{u}{5}\right)\left(v-\frac{v}{5}\right)$$
$$=\left(2-\frac{u}{5}\right)^2+\left(1-\frac{u}{5}\right)^2+\left(-2-\frac{u}{5}\right)^2+\left(-1-\frac{u}{5}\right)^2+\left(u-\frac{u}{5}\right)^2$$

Hence, it becomes apparent that u=v, Hence the inference is if u=v, then y=x.

c) is there any point where y=-x, b=-1 which reflects the

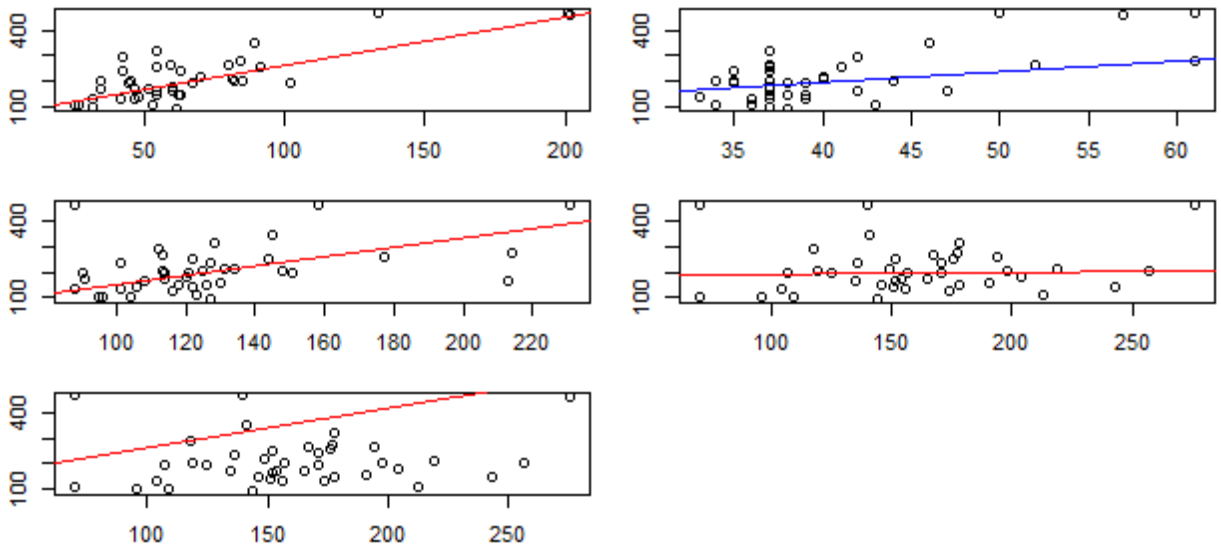
$$\left(-2-\frac{u}{5}\right)\left(-2-\frac{v}{5}\right)+\left(-1-\frac{u}{5}\right)\left(-1-\frac{v}{5}\right)+\left(1-\frac{u}{5}\right)\left(1-\frac{v}{5}\right)+\left(2-\frac{u}{5}\right)\left(2-\frac{v}{5}\right)+\left(u-\frac{u}{5}\right)\left(v-\frac{v}{5}\right)=$$
$$-\left\{\left(2-\frac{u}{5}\right)^2+\left(1-\frac{u}{5}\right)^2+\left(-2-\frac{u}{5}\right)^2+\left(-1-\frac{u}{5}\right)^2+\left(u-\frac{u}{5}\right)^2\right\}$$

This reduces to $U^2+UV+25 = 0$.

Since y=-x, u=-v, we get $U^2-U^2+25 \neq 0$, Thus is a contradiction.
Thus there is no point which can have y=-x

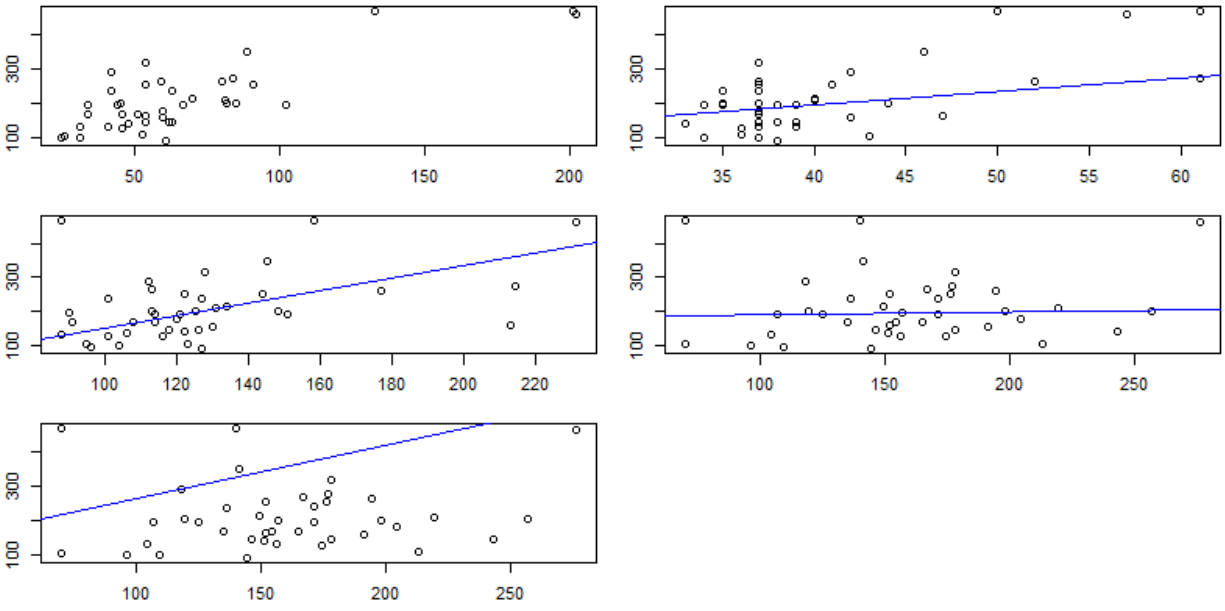
d) Least Absolute residual is given by $\sum(y_i - x_i) = 0$
Considering the points (-2,-2),(-1,-1),(1,1),(2,2),(u,v)
Hence adding all the points, u=v, hence equation y=x

3. a) Using RR program for plotting, following is the plot of all other variables vs y.



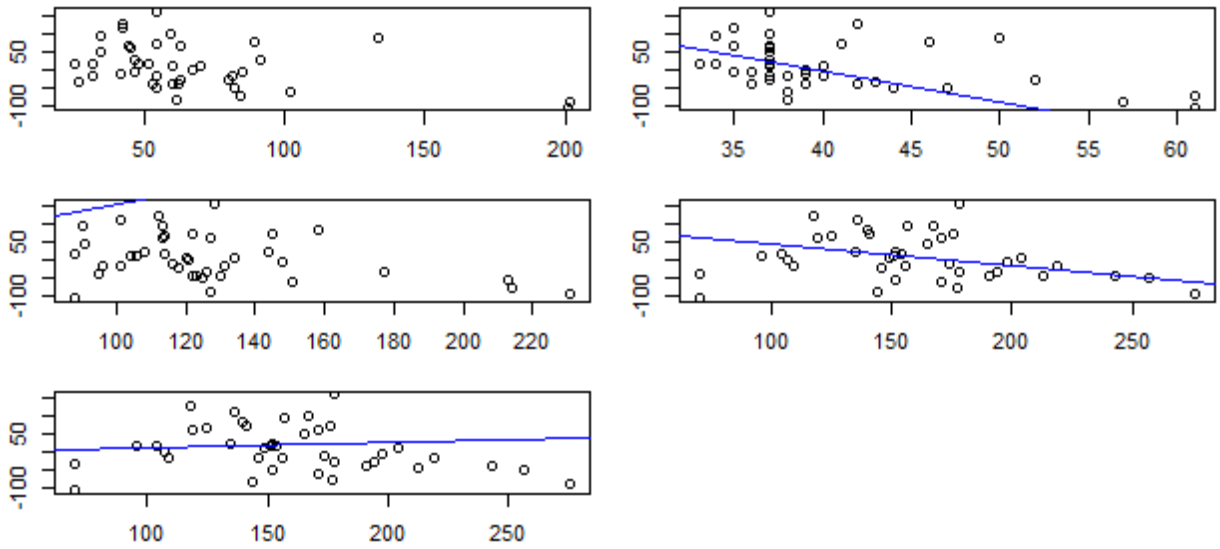
It can be clearly seen that X1, X2 and X3 have an almost linear relationship with y.

Hence the process should be to remove the effect of X1, : y1<-y-(a$a+a$b*X1)



Still in for X2, there is dependency.

```
rr<-run.rrline(X2,y1)
y12<-y1-(rr$a+rr$b*X2)
```



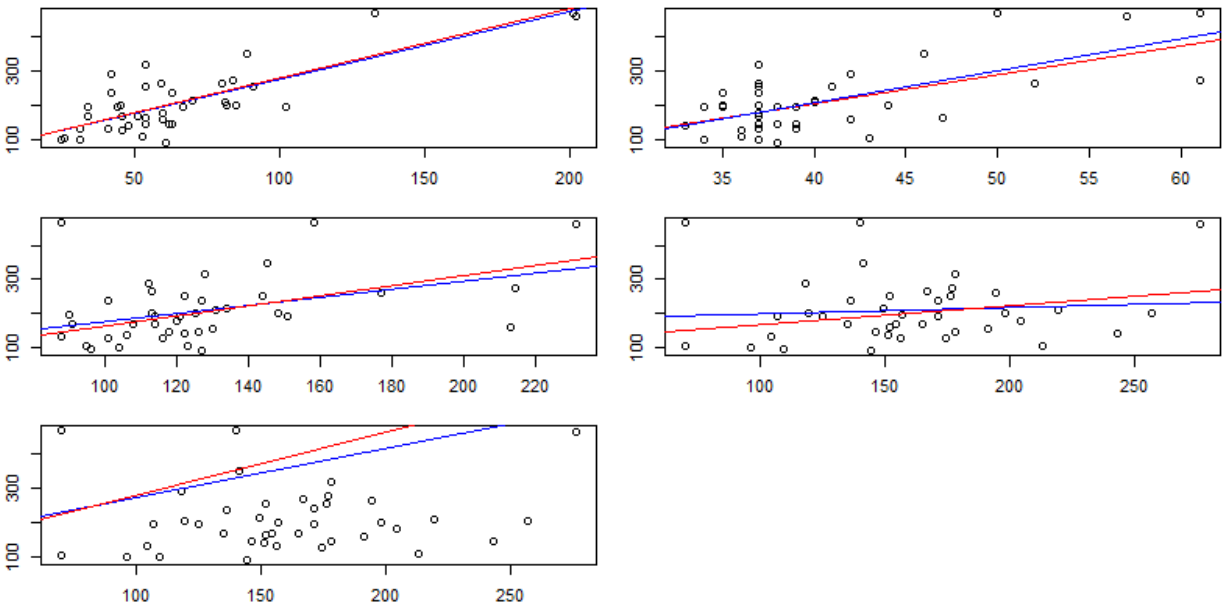
Thus now it can be seen that there are no visible dependency on any X and y.

Fit a line, fit1<-lm(y12~X3)

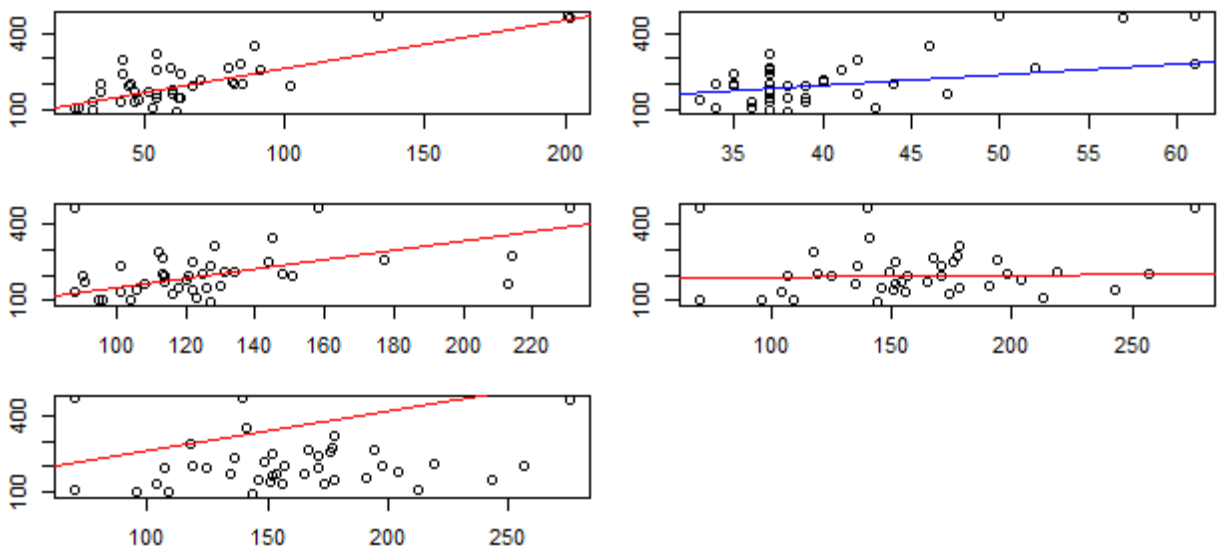
This gives slope = 0.46 and Intercepts=-51.0, hence line is of form y=0.46x-51.0

b) When talking about resistant regression, with continual iterations, the influence of even heavy outliers can be reduced.

c) Least Squares analysis



The above series of plots gives the effect of the outlier in the Linear Model. The blue line refers to the dataset without the outliers. The red line refers to the line with outliers. It can be seen that the line clearly deviates from the actual expected line.



The above equation is the Resistant regression line. It can be clearly seen that these lines overlap and the outliers doesn't have much impact on the RR line.