PW
Simple Regression

Work by:

Alaa Eddine Cherif
Moataz Hellal

GL3

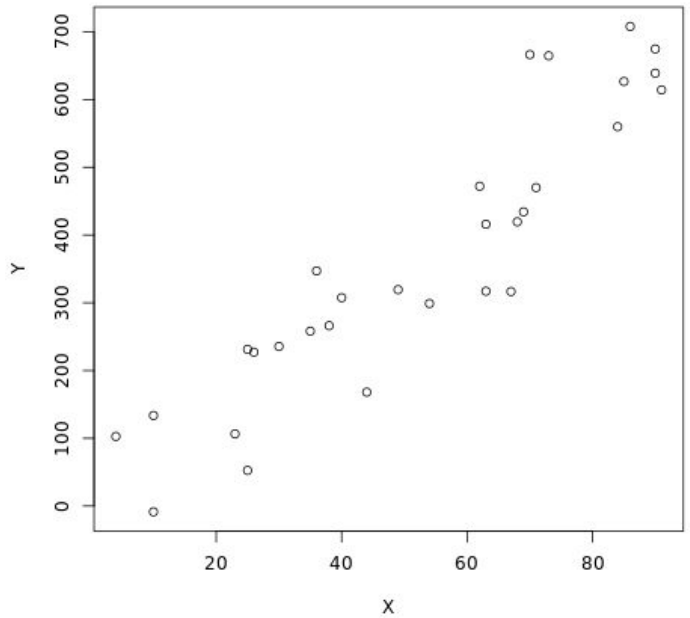
1. Define in R the variable `x=1:100` and then the variables `X=sample(x,30,replace=T)` and `Y=7*X+3+rnorm(30,0,100)`.

```
x=1:100
X=sample(x,30,replace=T)
Y=7*x+3+rnorm(30,0,100)
```

Output:

x	int [1:100] 1 2 3 4 5 6 7 8 9 10 ...
X	int [1:30] 2 34 8 93 49 32 85 22 35 81 ...
Y	num [1:30] -125 195 112 776 324 ...

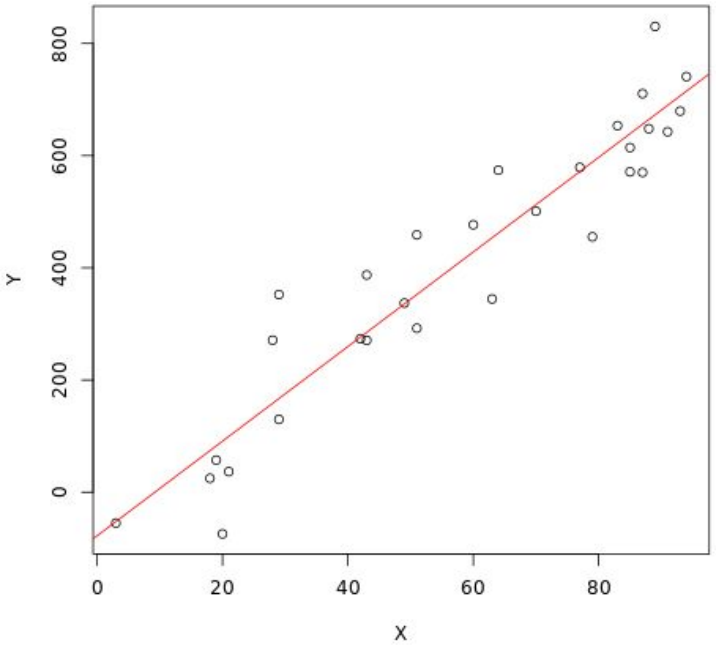
Plot the scattergram of `X` and `Y`, use the command `plot()`, save your graphic as a pdf file under the name scatterXY.



3. Apply the command `lm(Y ~ X)` and save the results in `regXY`.

4. Use the command `abline(regXY$coefficients)` and plot the regression line and save it under the name LS.pdf.

```
x=1:100
X=sample(x,30,replace=T)
Y=7*X+3+rnorm(30,0,100)
regXY<-lm(Y~X)
plot(X,Y)
abline(regXY$coefficients,col="red")
```



5. Use the command **anova(regXY)** and save the result in a variable **anovXY**.

```
Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
X          1 1089711 1089711  132.27 4.03e-12 ***
Residuals 28  230673    8238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

6. Explain how you can compute by hand all outputs of the command `anova(regXY)` following the notations below (write down the R' commands):

```
> y_mean=mean(Y)
> fss= sum((regXY$coefficients[2]*X+regXY$coefficients[1]-y_mean)^2)
> options
function (...)
.Internal(options(...))
<bytecode: 0x0000016b96cdc720>
<environment: namespace:base>
> options(digits=16)
> fss
[1] 1713962.371182605
> rss=sum((Y-(regXY$coefficients[2]*X+regXY$coefficients[1]))^2)
> sfss=fss^0.5
> srss=rss^0.5
> f_value=(fss/rss)*(30-2)
> pf(f_value,1,30-2,lower.tail = FALSE)
[1] 1.910997520163141e-13
```

f_value	171.002630010822
fss	1713962.3711826
offset	32.9398137454011
rss	280644.492953564
sfss	1309.18385690575
srss	529.758900778046
ssx	898.96091954023
sx	29.9826769908931
sy	262.258549982986
T	0.0632713014180791
x	int [1:100] 1 2 3 4 5 6 7 8 9 10 ..
X	int [1:30] 2 34 8 93 49 32 85 22 3
x_mean	59.2666666666667
Y	num [1:30] -125 195 112 776 324 ...
y_mean	428.580814907823

7. By using the command `confint()` give a confidence interval of the regression coefficients with a level of confidence of 5%.
use the notations: `CI_slope` and `CI_Interp`.

```
df=confint(regXY,level=0.05)
rownames(df)=c("CI_slope","CI_Interp")
df
```

```
                47.5 %    52.5 %
CI_slope    38.483944 42.888623
CI_Interp    6.125651  6.201152
```

8. Explain how you can obtain the same confidence interval by hand but using "R" !

```
> T=qt(1-((1-0.05)/2),30-2)
> beta=regXY$coefficients[2]
> alpha=regXY$coefficients[1]
> sx=sd(X)
> sy=sd(Y)
> offset=T*(sy/sx)
> beta_max=beta+offset
> beta_min=beta-offset
> x_mean=mean(X)
> ssx=var(X)
> offset=T*sy*(((1/30)+((x_mean)^2)/ssx)^0.5)
> alpha_min=alpha-offset
> alpha_max=alpha+offset
```

alpha	Named num -52
alpha_max	Named num -19
alpha_min	Named num -84.9
beta	Named num 8.11
beta_max	Named num 8.66
beta_min	Named num 7.55

9. Use the command `predict()` to predict the value of Y for $X = 500$, `Y*500`.

```
newdata_1<-data.frame(X=c(Y*500))
newdata_2<-data.frame(X=c(500))
predict(regXY,newdata_1)
```

```
1       2       3       4       5       6       7
148086.80 1444794.87  643363.91  554750.19  255422.95 2513801.63 1774629.91
      8       9      10      11      12      13      14
393227.63 1453164.61  104187.50 2702676.88 1822875.70 2211075.29  807488.19
     15      16      17      18      19      20      21
347692.91 -372361.80  577052.66 1657010.05  148434.69   44434.68 1219032.01
     22      23      24      25      26      27      28
2767321.29 2122408.17 1693797.55 2362920.51 2083896.77 2715375.62 3359280.68
     29      30
1755261.14 1651172.72
```

```
newdata_2<-data.frame(X=c(500))
predict(regXY,newdata_2)
```

```
       1
3613.307
```