Topic 8a: Linear Regression and Correlation: Part 3

So far we have seen how to compare lines that we propose as models and we have seen that there is a value called the correlation coefficient, r, that tells us how good the best linear model is. What we need to see is how do we compute the equation for the best linear model and then how do we compute the value of the correlation coefficient. Let us start with some values.

gnrnd4(365340906, 6340400805, 2200003)

X:	23	3	16	11	21	19	23	13	20	17
Y:	38	16	30	33	42	41	44	32	47	26

Generate and verify the data in R.

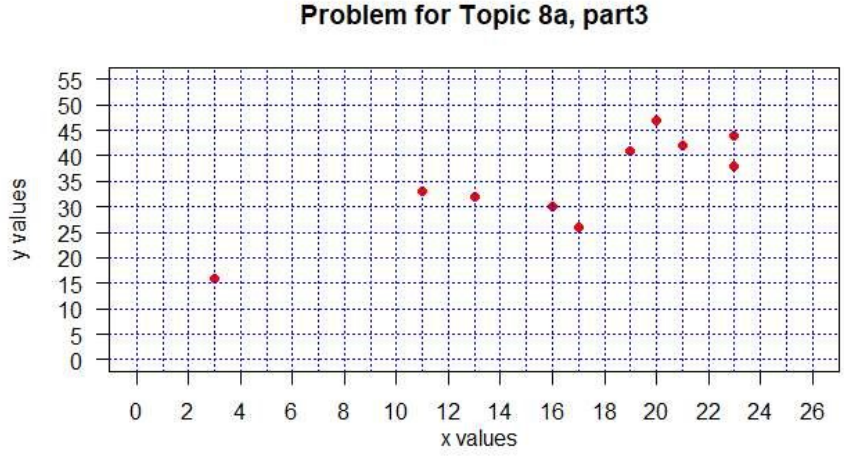
```
1  #
2  # script for topic 8a -- part 3
3  source("../gnrnd4.R")
4  gnrnd4(365340906, 6340400805, 2200003)
5  L1
6  L2
```

```
> source("../gnrnd4.R")
> gnrnd4(365340906, 6340400805, 2200003)
style= 6   size= 10    seed= 36534   num digits= 0   alt_sign=
 1
[1] "DONE "
> L1
 [1] 23  3 16 11 21 19 23 13 20 17
> L2
 [1] 38 16 30 33 42 41 44 32 47 26
```

It would be nice to get a plot of the values.

```
 8  #
 9  main_hold <- "Problem for Topic 8a, part3"
10  plot(L1,L2,main=main_hold,
11       xlim=c(0,26), ylim=c(0,55),
12       xaxp=c(0,26,13), yaxp=c(0,55,11), las=1,
13       pch=16, col="red", mar=c(2,0,0,0)+0.2,
14       ylab="y values", xlab="x values")
15  abline(h=seq(0,55,5), v=seq(0,26,1),
16         lty="dotted", col="blue")
```

```
> main_hold <- "Problem for Topic 8a, part3"
> plot(L1,L2,main=main_hold,
+      xlim=c(0,26), ylim=c(0,55),
+      xaxp=c(0,26,13), yaxp=c(0,55,11), las=1,
+      pch=16, col="red", mar=c(2,0,0,0)+0.2,
+      ylab="y values", xlab="x values")
> abline(h=seq(0,55,5), v=seq(0,26,1),
+        lty="dotted", col="blue")
>
```



Finding the linear regression equation means finding values for a and b in the linear equation y = a + bx, where a is the y-intercept and b is the slope. We can do this in R with the lm() function. However, we give the lm() function our two lists of values, L1 and L2, in a special way. The form is going to be lm(L2~L1). Notice that the y-list, L2, is given first and that the two lists are related by putting the "tilde" character, "~", between the lists.

```
16 # get the two values for our regression equation,
17 #  the intercept and the slope
18 lm( L2 ~ L1)
```

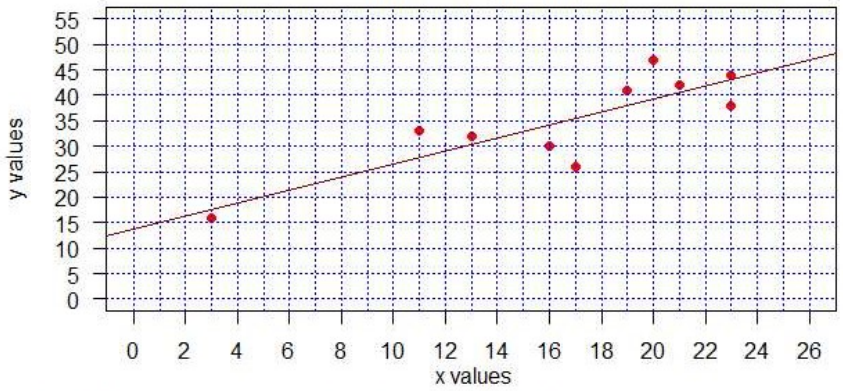
```
> # get the two values for our regression equation,
> #  the intercept and the slope
> lm( L2 ~ L1)

Call:
lm(formula = L2 ~ L1)

Coefficients:
(Intercept)           L1
      13.81         1.27
```

So, our linear regresson line is of the form y = 13.81 + 1.27x. We can add that to the graph.

```
19 #  From the output of that command we see that the
20 #  intercept is 13.81 and the slope is 1.27.  So
21 #  our regression equation is y = 13.81 + 1.27*x
22 #  We sill graph it on our plot
23 abline( 13.81, 1.27, col="darkred")
```

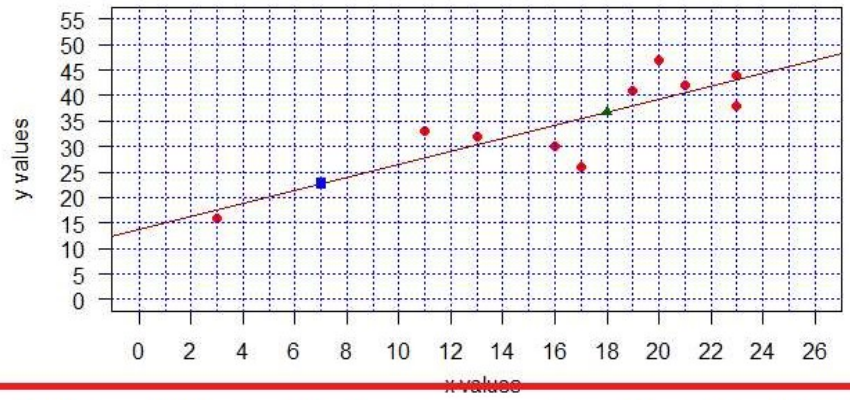
```
> #  From the output of that command we see that the
> #  intercept is 13.81 and the slope is 1.27.  So
> #  our regression equation is y = 13.81 + 1.27*x
> #  We sill graph it on our plot
> abline( 13.81, 1.27, col="darkred")
```

Problem for Topic 8a, part3




```
25 #  We can find the expected value when x=18
26 ev <- 13.81 + 1.27*18
27 ev
28 #  And then we can plot that point on the graph
29 #    with a green triangle
30 points( 18, ev, pch=17, col="darkgreen")
31 #  And we can find the expected value for x=7
32 ev <- 13.81 + 1.27*7
33 ev
34 #  And then we can plot that point on the graph
35 #  with a blue square
36 points( 7, ev, pch=15, col="blue")
```

Problem for Topic 8a, part3



To compute the correlation coefficient we use the cor() function and we just give it the two lists of values, as in cor(L1, L2).

```
#  Now find the correlation coefficient
cor( L1, L2 )
```

```
> #  Now find the correlation coefficient
> cor( L1, L2 )
[1] 0.8389308
```

Another part of doing linear regressions is to find the residual values. There is a residual value for each data point in our original observations (the original data). For each x value we have the original observed y value. The residual for that pair of values is the observed y minus the expected y, that is, the result of putting the x value into the regression equation.

```
#  One other issue is to find some residual values.
#  Here we do that the hard way, we will compute the
#  observed - expected value for a given x value.
#  Find the residual value when x=17.  The observed
#  value when x=17 is y=26.  We need to compute the
#  expected value and then subtract that from 26 to
#  get the residual value.
26 - ( 13.81 + 1.27*17)
```

```
> #  One other issue is to find some residual values.
> #  Here we do that the hard way, we will compute the
> #  observed - expected value for a given x value.
> #  Find the residual value when x=17.  The observed
> #  value when x=17 is y=26.  We need to compute the
> #  expected value and then subtract that from 26 to
> #  get the residual value.
> 26 - ( 13.81 + 1.27*17)
```

```
#  If we find all of the
#  residual values and
#  then get a scatter plot
#  of the x and residual
#  values, we want to see
#  the points all over the
#  scatter plot.
res_vals <- L2 -
       ( 13.81 + 1.27*L1)
plot( L1, res_vals,
      main="Residuals")
```