STAT S 520
HOMEWORK 12
RAMPRASAD BOMMAGANTY(rbommaga)

Problem 1:

Solution: Yes, there is a simpler explanation for the mentioned phenomenon. This can be attributed to the regression effect, where it is stated that experimental units with extreme X quantiles will have less extreme Y quantiles. In this case, if we consider the “tries” as the experimental units, then a very good performance on the first try will most probably be followed by a performance that is not as good as the first one. Though it seems like praise might be a factor in decrease in performance, it really is not. It is just the regression effect coming into play.

The chapter 15 deals with regression effect and simple linear regression. Since the given claim that praise reduces performance is a phenomenon attributed to regression effect, it is related to the chapter and is worth discussing.

Problem 2: Exercise 4:

Solution:

a. We know that 5’10’’ = 70 inches.
We need to estimate the proportion of all brothers who are at least 5’ 10’’.

If we consider X as the random variable for heights of sisters and Y as the random variable for heights of brothers, then we need to estimate:

$P(Y \geq 70) = 1 - P(Y \leq 70)$

The R code computation is as follows:

```
> 1 -pnorm(70,69,sqrt(7.4))
[1] 0.356583
```

Therefore, the proportion of all brothers who are at least 5’10’’ = 0.356583

b. Given that Carol’s height is 5’1’’= 61 inches.

E(Y|X=61) = $\bar{y} + r\frac{s_y}{s_x}(61 - \bar{x})$

E(Y|X=61) = 69 + 0.55 * (sqrt(7.4)/sqrt(6.6)) * (61 - 64) = 67.25286

Therefore, her brother’s height is 67.25.

c. We know that

Y| X = x ~ $Normal(\mu_y + \rho\frac{\sigma_y}{\sigma x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$

If we take Y = 70 and X = 61, then the distribution we get is as Normal(67.25,5.09).

Now, P(Y ≥ 70) = 1 - P(Y ≤ 70)

The R code for computation is as follows:

```
> 1 - pnorm(70,67.25,sqrt(5.09))
[1] 0.111438
```

Therefore, the proportion of brothers who are at least 5’ 10’’ is 0.11.

Problem 3 : Exercise 8:

Solution:

a. The parameters for n = 33 are as follows: (75,64,10^2,12^2,0.5)

Since Jill suggests to take her first test score, we can take X = x and estimate the Y values that is obtained.

We know that,

Y| X = x ~ $Normal(\mu_y + \rho\frac{\sigma_y}{\sigma x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$

Here, Y| X = 80 ~ Normal(66.73, 9)

Therefore, P(Y ≥ 80| X = 80) = 1 - pnorm(80,66.73,sqrt(9)) = 4.859481e-06

Since, the probability that Jill would score more than 80 on the second test is very low, the Professor should not allocate her score as 80, based on her request.

We could compute Y | X =80 and predict her score instead.

The answer is the same as the one we got above: 66.73.

Thus, I would recommend the Professor to allocate a score of 66.73 to Jill.

b.
From the problem, we know that ρ = 0.5 and not zero. Hence, we cannot follow Jack’s suggestion that his score be allocated 85, one standard deviation above $\bar{x}$.

We try to fit another line where the slope = 0.5 * 10/12 = 0.41

The intercept is 75 - (0.41 * 64) = 48.76

Therefore the predicted value is : 48.76 + 0.41*76 = 79.92

Problem 4:

Solution:

a. Let us take X is the random variable for the number of wins in year 1 and Y is the random variable for number of wins for year 2.

Using the binorm.estimate function for the data containing X and Y, we get the following output:

```
> binorm.estimate(all_wins)
[1]  34.9745597  32.3052838 152.3032731 142.9497563   0.3176933
```

By taking the sum of the total wins for both years, we get the following summary:

```
> summary(sum_of_wins)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   53.50   67.00   67.28   84.00  109.00
```

We need to estimate P(Z $\geq$ 84.5) = 1 - P(Z $\leq$ 84.5)

```
> 1 - pnorm(84.5,67.28,19.72)
[1] 0.1912702
```

Thus, the probability that a randomly selected team will at least 84.5 games across both years is 0.191.

b. Given that X = 95 and if we assume Y = 84.5, then the resulting distribution is of the form

Y| X = x ~ $Normal(\mu_y + \rho \frac{\sigma_y}{\sigma x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$

```
> binorm.estimate(all_wins)
[1]  34.9745597  32.3052838 152.3032731 142.9497563   0.3176933
```

Y | X = 95 ~ Normal(50.32, 129.20)

P(Y ≥ 84.5) = 1 - P(Y ≤ 84.5)

The computation is as follows:

```
baseball_data <- read.table("/home/ramprasad/Downloads/baseball-wins.txt")
year1_wins <- baseball_data$V3
year2_wins <- baseball_data$V5
all_wins <- cbind(year1_wins,year2_wins)
binorm.scatter(all_wins)
summary(all_wins)
comp <- binorm.estimate(all_wins)
year1_wins <- as.numeric(year1_wins)
year2_wins <- as.numeric(year2_wins)
sum_of_wins <- year1_wins + year2_wins
summary(sum_of_wins)
sd(sum_of_wins)
mean = comp[2]+(comp[5]*(sqrt(comp[4])/sqrt(comp[3]))*(95-comp[1]))
variance = (1-(comp[5]^2))*(comp[4])
```

```
> 1-pnorm(84.5,mean,sqrt(variance))
[1] 0.001467875
```

Therefore, the probability that a team that won 95 games first season wins at least 84.5 games next season is 0.0014.

c. In this case, X = 75.

Y | X = 75 ~ Normal (44.32, 129.20)

P(Y ≥ 84.5) = 1 - P(Y ≤ 84.5)

The computation is as follows:

```
> 1 -pnorm(75,44.32,sqrt(129.20))
[1] 0.003476031
```

Therefore, the probability that a team that won 75 games first season wins at least 84.5 games the next season is 0.0034.

Problem 5:

a. The R code for setting up the distribution is as follows:

```
exam_anxiety <- read.table("/home/ramprasad/Downloads/examanxiety.txt")
exam <- exam_anxiety$V3
anxiety <- exam_anxiety$V4
exam <- as.numeric(exam)
anxiety <- as.numeric(anxiety)
data <- cbind(exam,anxiety)
comp <- binorm.estimate(data)
```

Slope = -0.73

Intercept = 111.24

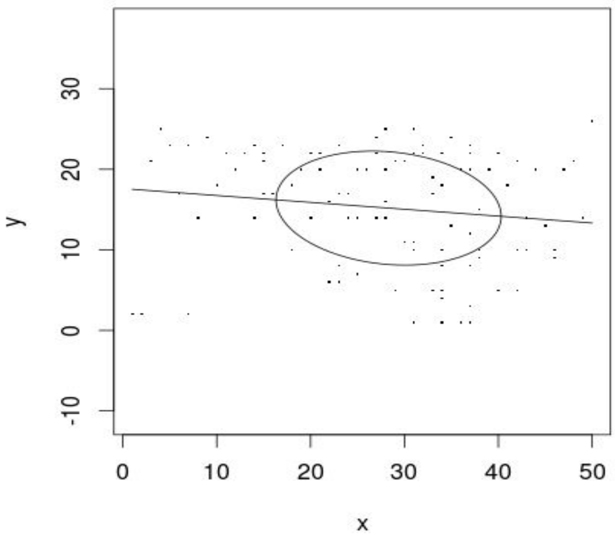
The regression line equation is: 111.24 - (0.73 * anxiety)

b.

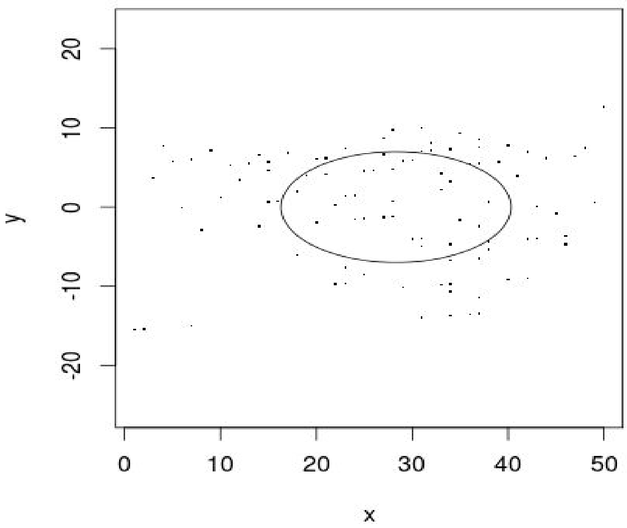
```
binorm.regress(data)
x <- binorm.resid(data)
binorm.scatter(x)
qqnorm(data)
```

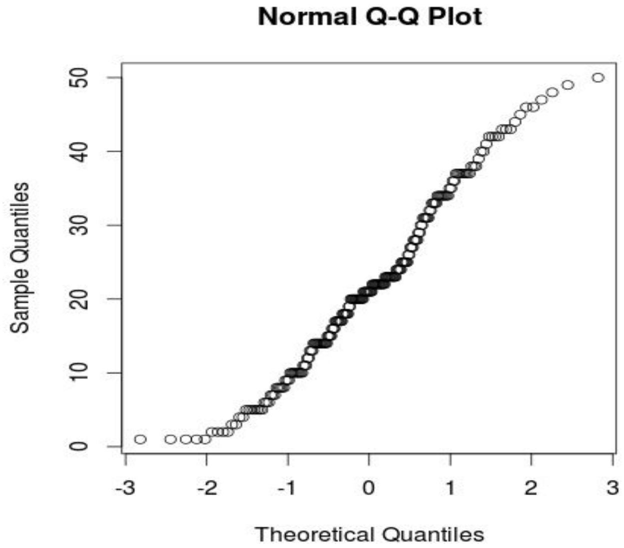
The plots are as follows:

Regression Line



Scatter Diagram

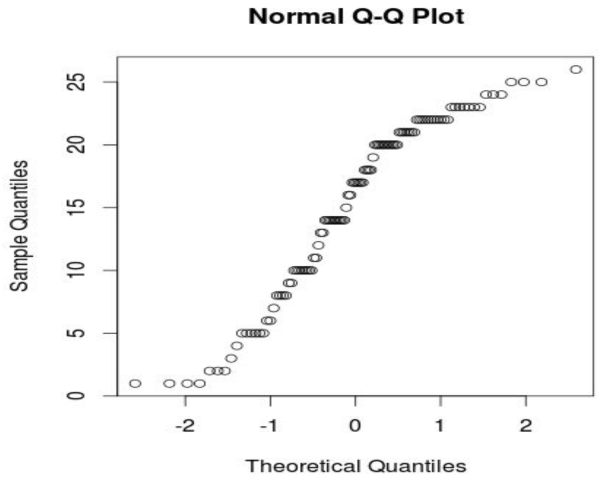


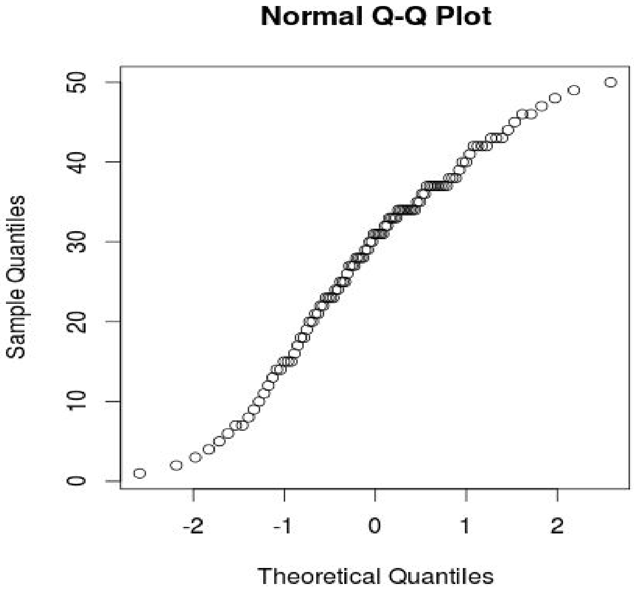


i. The scatter-plot and correlation coefficient reveal a linearity.
ii. Independence is also a possible assumption, since scores of one student do not affect the scores of other students.
iii. Homoscedasticity does not seem plausible.
iv. The qqnorm plot looks almost like a straight line except for the edges, hence normality is possible.

c.

```
exam <- as.numeric(exam)
anxiety <- as.numeric(anxiety)
qqnorm(exam)
```





From the above qq plots we can say that the random variables are from a normal distribution. Hence, we can use the bivariate normal.

Problem 6:

Solution:

The R code for computation of the difference is as follows:

```
data<-read.table("/home/ramprasad/Downloads/testscores.txt", header=TRUE)
intermediate_Data <- cbind(data$first.test[1:20],data$second.test[1:20])
est<- binorm.estimate(cbind(data$first.test[1:20],data$second.test[1:20]))
Matrix <- matrix(nrow=20)
averageValue <- (est[2]+(est[5]*(sqrt(est[4])/sqrt(est[3]))*(myData[,1]-est[1])))
average<-mean(averageValue-intermediate_Data[1:20,1])
average
```

```
 [1] -2.15
```

The above value is the predicted value in the difference that will occur between first test and second test and it is similar to what was actually observed by the Board when the results were in.

This is a classic case of regression effect, where people who performed well in the first test were bound to perform in a less effective manner in the second test. Hence, the class did perform as per expectations and the Board need not be unhappy with the teacher of Class A.