PROB STATS LAB ASSIGNMENT 3

COURSE: BMAT202P

SEMESTER: WINTER 2024-2025

SLOT: L33+L34

HARSHITH B

23BCE0559

BINOMIAL DISTRIBUTION

1)It is known that probability of an item produced by a certain machine will be defective is 0.05. If the produced items are sent to the market in packets of 20, then write down the R code to find the number of packets containing at least, exactly and at most 2 defectives Items in a consignment of 1000 packet.

Code:

```
> n=20

> p=0.05

> q=1-p

> N=1000

> k=2

> N1=round(N*(1-pbinom(k-1,n,p)))

> N1

[1] 264

> 264

[1] 264
```

```
> k=2
> N2=round(N*dbinom(k,n,p))
> N2
[1] 189
> 189
[1] 189
> k=2
> N3=round(N*pbinom(k,n,p))
> N3
```

Output:

Atleast 2 defective: 264

Exactly 2 defective: 189

Atmost 2 defective: 925

2) For a Binomial(7,1/4) random variable named X, i. Compute the probability of two success ii. Compute the Probablities for whole space iii. Display those probabilities in a table iv. Show the shape of this binomial Distribution.

Code:

```
>dbinom(2,7,1/4)

>dbinom(0:7,7,1/4)

>P=data.frame(0:7,dbinom(0:7,7,1/4))

>round(P,4)

>plot(0:7,dbinom(0:7,7,1/4),type="o")
```

Output:

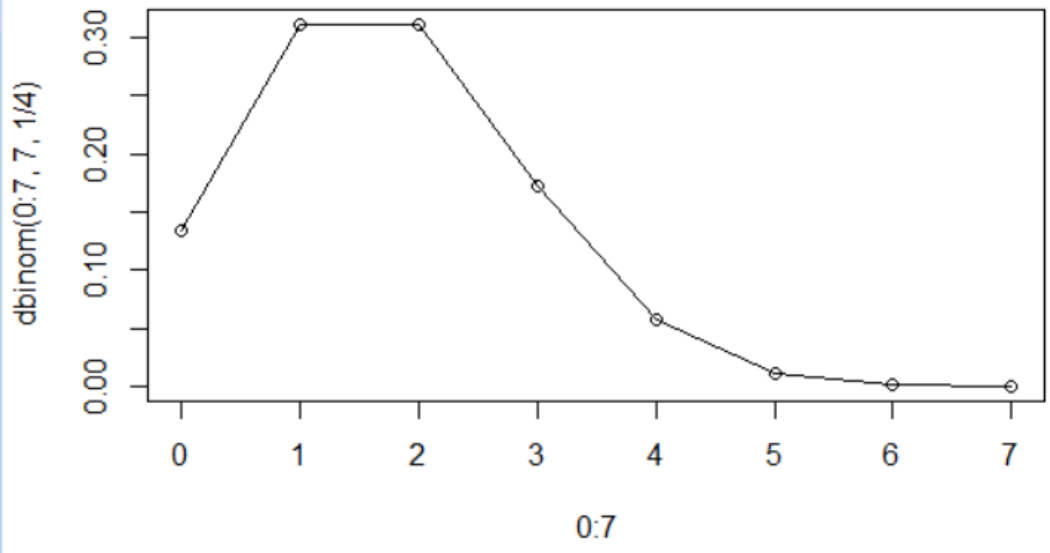
probability of two success: 0.3114624

probabilities for whole space: [1] 1.334839e-01 3.114624e-01 3.114624e-01 1.730347e-01 5.767822e-02 [6] 1.153564e-02 1.281738e-03 6.103516e-05

R Console

```
> dbinom(2,7,1/4)
[1] 0.3114624
> dbinom(0:7,7,1/4)
[1] 1.334839e-01 3.114624e-01 3.114624e-01 1.730347e-01 5.767822e-02
[6] 1.153564e-02 1.281738e-03 6.103516e-05
> P=data.frame(0:7,dbinom(0:7,7,1/4))
> round(P,4)
  X0.7 dbinom.0.7..7..1.4.
1    0                0.1335
2    1                0.3115
3    2                0.3115
4    3                0.1730
5    4                0.0577
6    5                0.0115
7    6                0.0013
8    7                0.0001
> plot(0:7,dbinom(0:7,7,1/4),type="o")
>
```

R Graphics: Device 2 (ACTIVE)



POISSON DISTRIBUTION

1)It is known that probability of an item produced by a certain machine will be defective is 0.05. If the produced items are sent to the market in packets of 20, then write down the R code to find the number of packets containing at least, exactly and at most 2 defectives Items in a consignment of 1000 packet.

```
>lam=1.5

>x=c(0,1,2)

>x=0

>P1=dpois(x,lam)

>P1

>x=1

>P2=ppois(x,lam)

>P2

>x=2

>P3=1-ppois(x,lam)

>P3
```

Output:

neither car is used: 0.2231302

at most one car is used: 0.5578254

some demand of car is not fulfilled: 0.1911532

2) Poisson distribution with parameter '2' 1. How to obtain a sequence from 0 to 10 2. Calculate P(0),P(1),...,P(10) when lambda =2 and Make the output prettier 3. Find P(x <= 6) 4. Sum all probabilities 5. Find P(Y>6) 6. Make a table of the first 11 Poisson probs and cumulative probs when # mu=2 and make the output prettier 7. Plot the probabilities Put some labels onthe axes and give the plot a title:

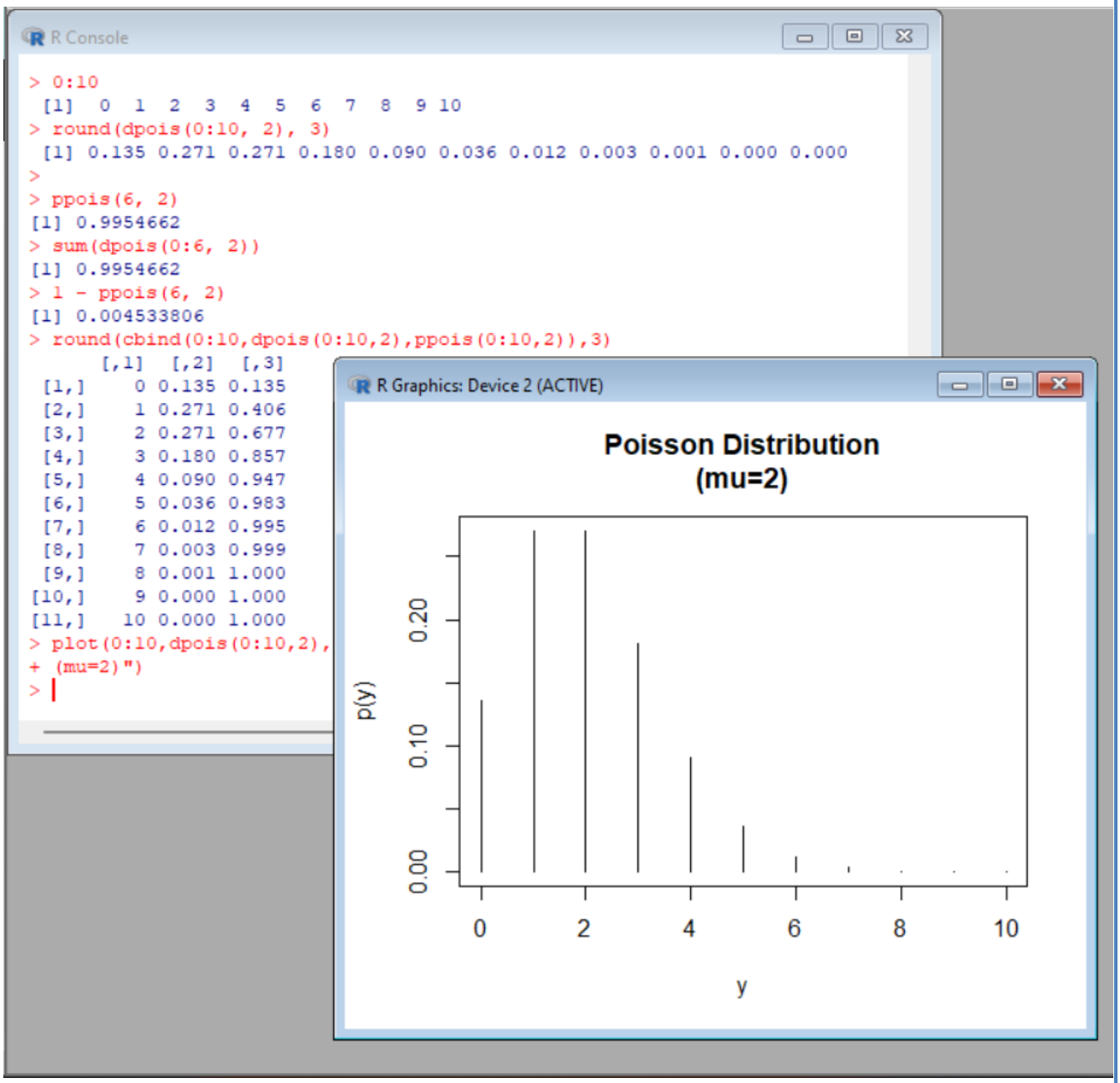
Code:

```
>0:10

[1] 0 1 2 3 4 5 6 7 8 9 10

>round(dpois(0:10,2),3)

[1] 0.135 0.271 0.271 0.180 0.090 0.036 0.012 0.003 0.001 0.000 0.000

>ppois(6,2)

[1] 0.9954662

>sum(dpois(0:6,2))

[1] 0.9954662

>1-ppois(6,2)

[1] 0.004533806

>round(cbind(0:10,dpois(0:10,2),ppois(0:10,2)),3)

>plot(0:10,dpois(0:10,2),type="h",xlab="y",ylab="p(y)",main="Poisson Distribution (mu=2)")
```

Output:

sum all probabillites: 0.9954662

P(Y>6): 0.004533806

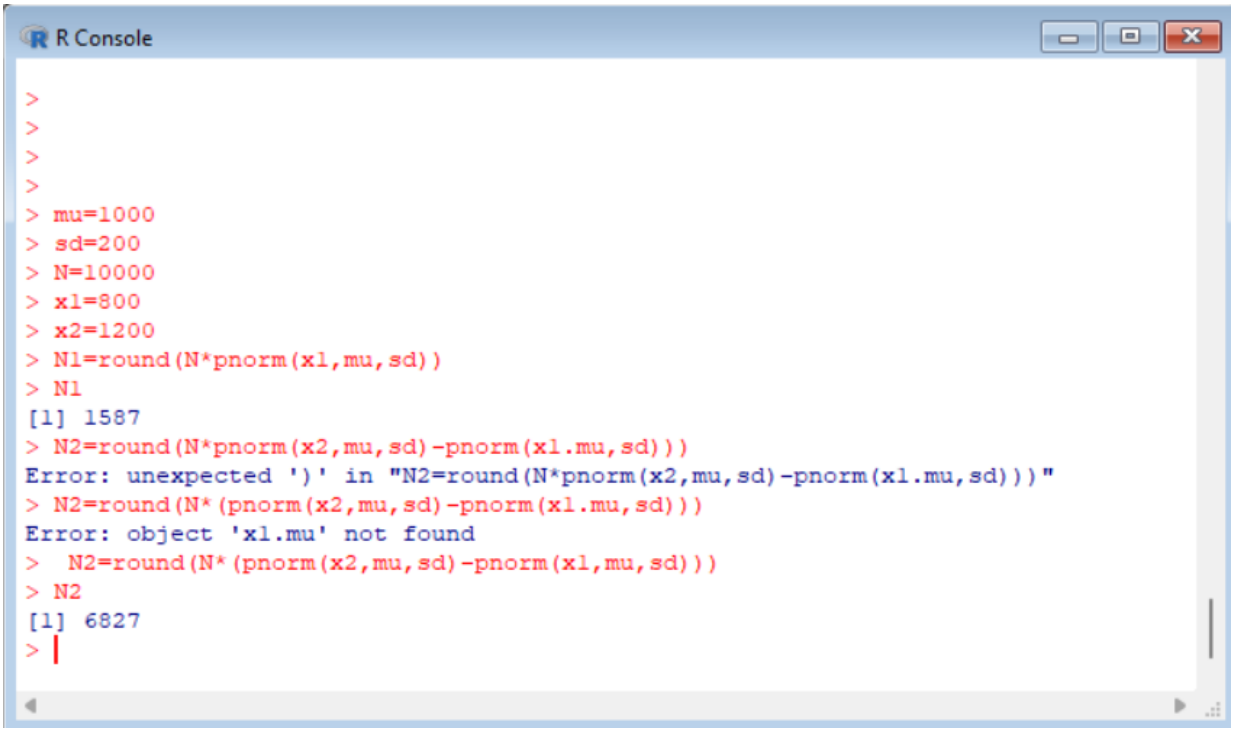


Normal distribution:

1)The local corporation authorities in a certain city install 10,000 electric lamps in the streets of the city with the assumption that the life of lamps is normally distributed. If these lamps have an average life of 1,000 burning hours with a standard deviation of 200 hours, then write down the R code to calculate the number of lamps might be expected to fail in the first 800 burning hours and also the number of lamps might be expected to fail between 800 and 1,200 burning hours.

Code:

```
>mu = 1000 #average lamp life

> sd = 200 #standard deviation

>N = 10000 #electric lamps

>x1 = 800 #first 800 hours

> x2 = 1200 #first 1200 hours

>

> # number of lamps fall in the first 800 burning hours

>x1 = 800

> N1=round(N*pnorm(x1,mu,sd))

> N1

[1] 1587

> # number of lamps fall in between 800 to 1200 burning hours

>x1 = 800

>x2 = 1200

> N2=round(N*(pnorm(x2,mu,sd)-pnorm(x1

> N2
[1] 6827
```



Output:

Failure(X>800):1587
Failure(800<X<1200):6827

2) In a test on 2000 Electric bulbs ,it was found that the life of particular make, was normally distributed with an average life of 2040 hours and S.D of 60 hours. Estimate the number of bulbs likely to burn for (i) More than 2150 hours (ii) Less than 1950 hours (iii) More than 1920 hours but less than 2160 hours (iv) More than 2150 hours

Code:

```
>(1 - pnorm(2150, mean=2040, sd=60))*2000

[1] 66.75302
```

(i) Less than 1950 hours

```
>(pnorm(1950, mean=2040, sd=60))*2000

[1] 133.6144
```

(ii) More than 1920 hours but less than 2160 hours

>(pnorm(2160, mean=2040, sd=60) - pnorm(1920, mean=2040,

sd=60))*2000

[1] 1908.999

Output:

The number of bulbs expected to burn for more than 2150 hours is 67

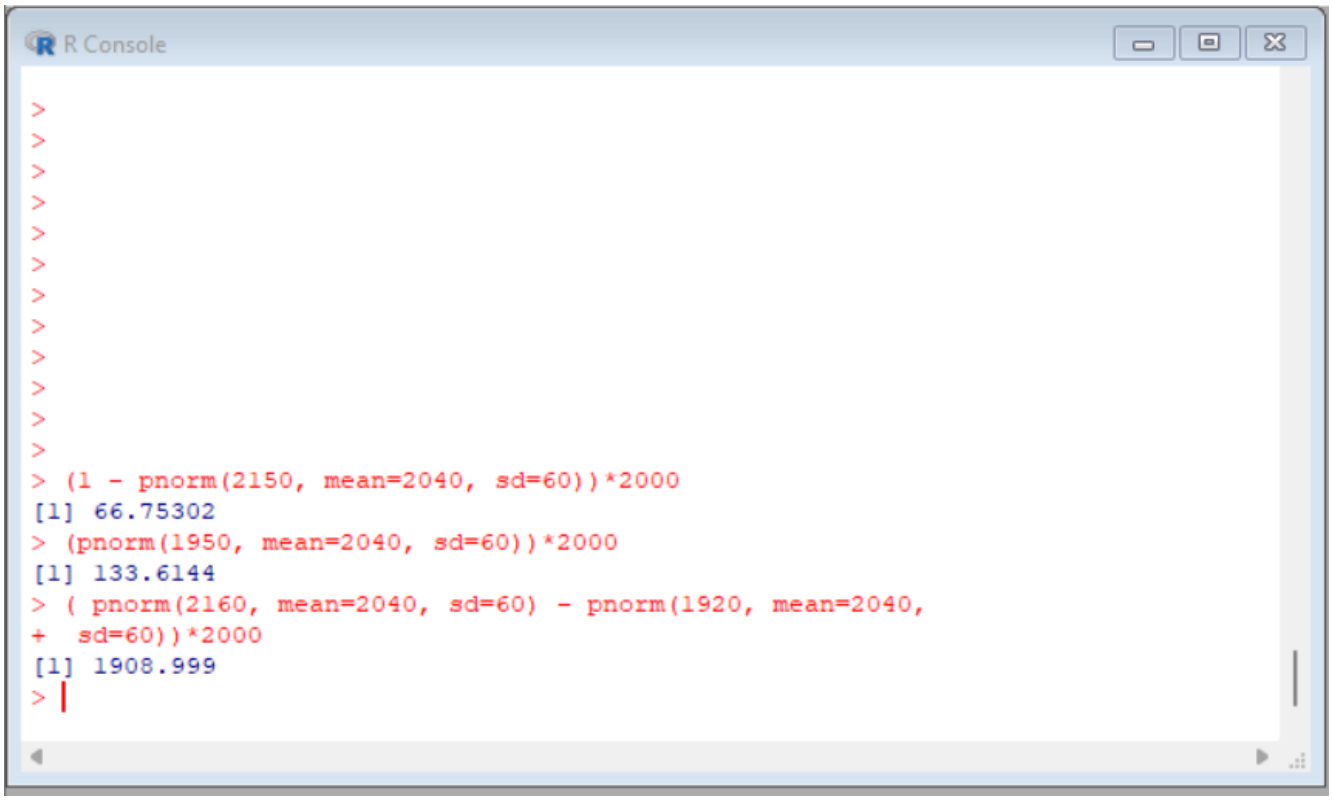
(approximately)

The number of bulbs expected to burn for less than 1950 hours is 134

(approximately)

The number of bulbs expected to burn more than 1920 hours but less

than 2160 is 1909 (approximately)



Problem : In a photographic process the developing times of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 seconds and a standard deviation 0.12 second. Find the probability that it will take

(i) Atleast 16.20 seconds to develop one of the prints;
(ii) atmost 16.35 seconds to develop one of the prints

Ans) Atleast 16.20 seconds to develop one of the prints;
Print developing time : $X \sim N(16.28, (0.12)^2)$

(Solution). (i) Required event : $[X \geq 16.20]$

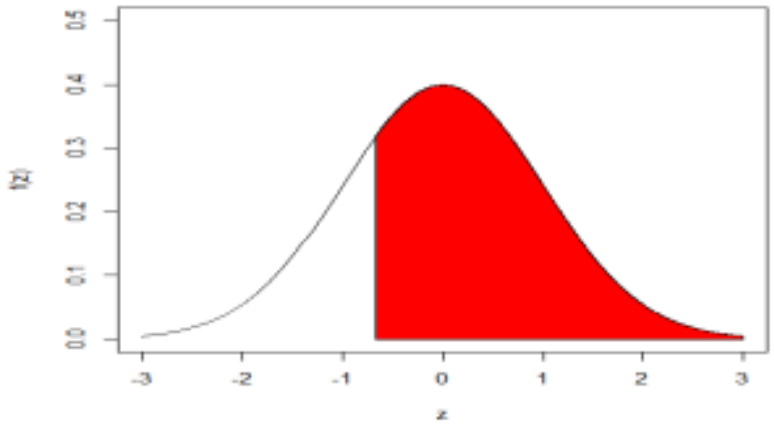
$$P[X \geq 16.20] = P\left[\frac{X - 16.28}{0.12} \geq \frac{16.20 - 16.28}{0.12}\right] = P[Z \geq -0.6667]$$

$$P[X \geq 16.20] = 0.7486(\text{Dotted area})$$

R Code:-

(i). $P[Z \geq -0.6667]$

```
> (1-pnorm (-0.6667))

[1] 0.7475181

> 1 - pnorm((-0.6667))
> plot.new()
> curve (dnorm, xlim = c(-3, 3), ylim = c(0, 0.5), xlab = "z", ylab="f(z)")
> z = -0.6667
> x = c(z, seq(z, 3, by=.001), 3)
> y = c(0, dnorm(seq(z, 3, by=.001)), 0)
> polygon(x, y, col="red")
```



(ii) Required event : [X ≤ 16.35]

$$P[X \leq 16.20] = P\left[Z \leq \frac{16.35 - 16.28}{0.12}\right]$$

$$= P[X \leq 0.5833] \text{(dotted area)}$$

$$= 0.5 + 0.2190 = 0.7190$$

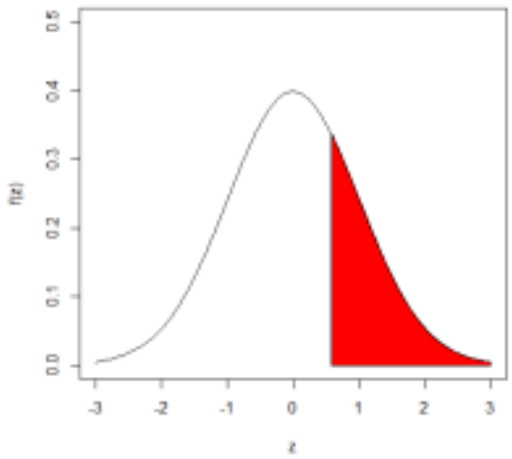
$$P[X \leq 16.35] = 0.7190$$

R code:-

```
> pnorm(0.5833)

[1] 0.7201543
```

```
>pnorm(0.5833)
> plot.new()
> curve (dnorm, xlim = c(-3, 3), ylim = c(0, 0.5), xlab = "z", ylab="f(z)")
> z = 0.5833
> x = c(z, seq(z, 3, by=.001), 3)
> y = c(0, dnorm(seq(z, 3, by=.001)), 0)
> polygon(x, y, col="red")
```

Output:-



Linear regression and Multiple Linear Regression

Regression Analysis

Description

Regression analysis can be defined as the process of developing a mathematical model that can be used to predict one variable by using another variable or variables. This section first covers the key concepts of two common approaches to data analysis: graphical data analysis and correlation analysis and then introduces the two main types of regression: linear regression and non-linear regression. The section also introduces a number of data transformations and explains how these can be used in regression analysis.

Learning Objectives

By the end of this section, you should be able to:

- Distinguish between a dependent variable and an independent variable and analyze data using graphical means.
- Examine possible relationships between two variables using graphical analysis and correlation analysis.
- Develop simple linear regression models and use them as a forecasting tool.
- Understand polynomial functions and use non-linear regression as a forecasting tool.
- Appreciate the importance of data transformations in regression modeling.

Assumptions

There are four principal assumptions that justify the use of linear regression models for inference or prediction:

(i) Linearity and Additivity of the Relationship Between Dependent and Independent Variables:

- The expected value of the dependent variable is a straight-line function of each independent variable, holding the others fixed.
- The slope of that line does not depend on the values of the other variables.
- The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) Statistical Independence of Errors:

- There should be no correlation between consecutive errors, especially in time series data.

(iii) Homoscedasticity (Constant Variance) of Errors:

Errors should have constant variance:

- Versus time (for time series data).
- Versus predictions (the variance of residuals should not change as the predicted values increase or decrease).
- Versus any independent variable (errors should not vary systematically across different values of independent variables).

(iv) Normality of the Error Distribution:

- The errors should be normally distributed.

If any of these assumptions is violated (e.g., nonlinear relationships, correlated errors, heteroscedasticity, or non-normality), then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be:

- Inefficient (producing less accurate estimates).
- Seriously biased or misleading (if the violations are severe).

Problem 1:

The following table shows the scores (X) of 10 students on Zoology test and scores (Y) on Bottony test. The maximum score in each test was 50. Obtain least square equation of line of regression of X on Y. If it is known that the score of a student in Botany is 28, Estimate his/her score in Zoology.

X	34	37	36	32	32	36	35	34	29	35
Y	37	37	34	34	33	40	39	37	36	35

R Code:

```
x=c(34,37,36,32,32,36,35,34,29,35)
y=c(37,37,34,34,33,40,39,37,36,35)
fit=lm(x~y)
fit
```

Output:

```
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
    18.9167       0.4167


[Execution complete with exit code 0]
```

The equation of the line of regression of X and Y is X=18.9167+0.4167Y. The required score of the student in Zoology is 30.58333

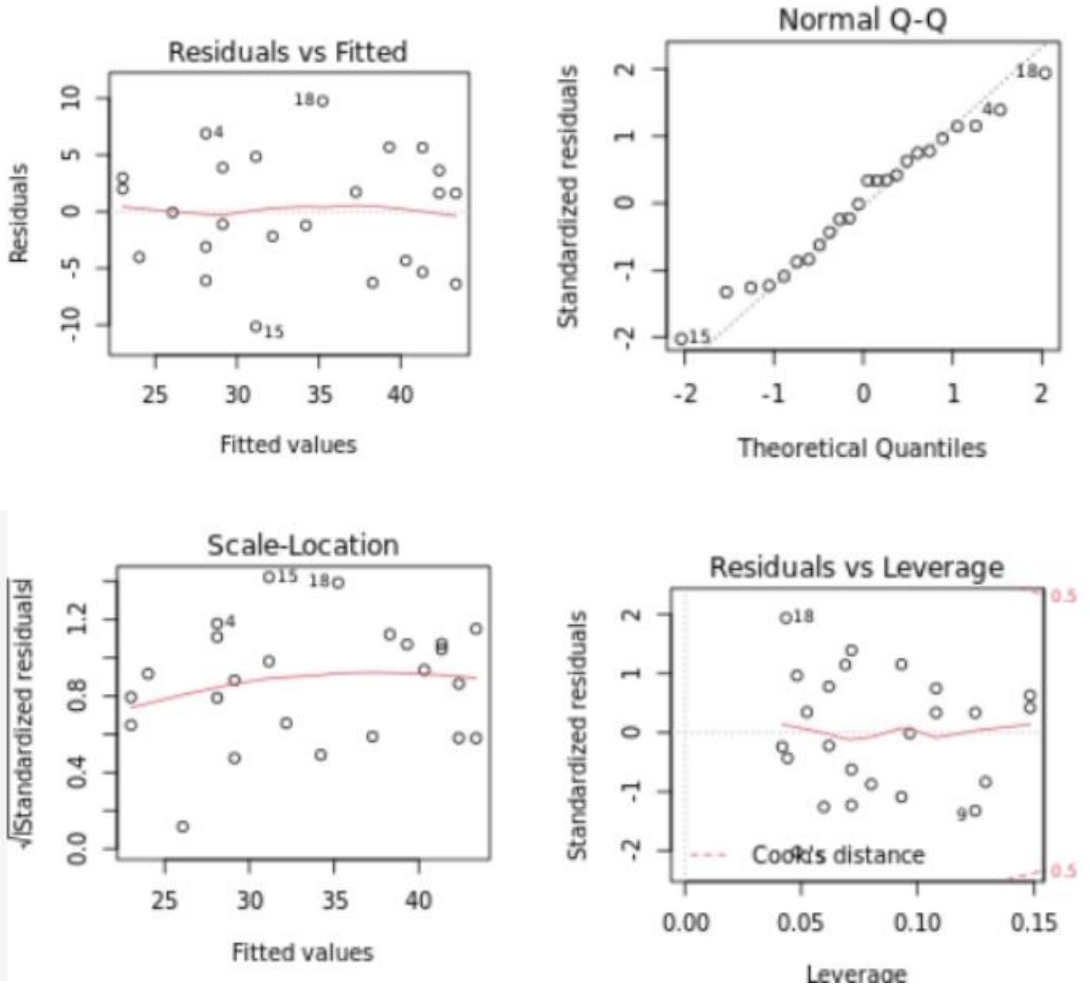
Problem 2 :- The following data pertain to the resistance in (ohms) and the failure times (minutes) of 24 overloaded resistors.

Resistance(x)	*43*	*29*	*44*	*33*	*33*	*47*	*34*	*31*	*48*
	34	*46*	*37*	*36*	*39*	*36*	*47*	*28*	*40*
	42	*33*	*46*	*28*	*48*	*45*			
Failure time(y)	*32*	*20*	*45*	*35*	*22*	*46*	*28*	*26*	*37*
	33	*47*	*30*	*36*	*33*	*21*	*44*	*26*	*45*
	39	*25*	*36*	*25*	*45*	*36*			

R Code:

```
x = c(43, 29, 44, 33, 33, 47, 34, 31, 48, 34, 46, 37, 36, 39, 36, 47, 28, 40, 42, 33, 46, 28, 48, 45

y = c(32, 20, 45, 35, 22, 46, 28, 26, 37, 33, 47, 30, 36, 33, 21, 44, 26, 45, 39, 25, 36, 25, 45, 36

fit = lm(y ~ x)
fit
par(mfrow=c(2,2));
plot(fit)
par(mforw=c(1,1));
```

Output:




```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -5.518        1.019
```

Problem 3 Problem 3: The sale of a Product in lakhs of rupees(Y) is expected to be influenced by two variables namely the advertising expenditure X1 (in'OOO Rs) and the number of sales persons(X2) in a region. Sample data on 8 Regions of a state has given the following results

Area	Y	X1	X2
1	110	30	11
2	80	40	10
3	70	20	7
4	120	50	15
5	150	60	19
6	90	40	12

7	70	20	8
8	120	60	14

Code :

```
Y=c(110,80,70,120,150,90,70,120)
X1=c(30,40,20,50,60,40,20,60)
X2=c(11,10,7,15,19,12,8,14)
input_data=data.frame(Y,X1,X2)
input_data
RegModel <- lm(Y~X1+X2, data=input_data)
RegModel
summary(RegModel)
```

Output :

```
    Y X1 X2
1 110 30 11
2  80 40 10
3  70 20  7
4 120 50 15
5 150 60 19
6  90 40 12
7  70 20  8
8 120 60 14

Call:
lm(formula = Y ~ X1 + X2, data = input_data)

Coefficients:
(Intercept)           X1           X2
    16.8314      -0.2442       7.8488
```

```
Call:
lm(formula = Y ~ X1 + X2, data = input_data)

Residuals:
      1       2       3       4       5       6       7       8
 14.157  -5.552   3.110  -2.355  -1.308 -11.250  -4.738   7.936

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.8314    11.8290   1.423   0.2140
X1           -0.2442     0.5375  -0.454   0.6687
X2            7.8488     2.1945   3.577   0.0159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.593 on 5 degrees of freedom
Multiple R-squared:  0.9191,    Adjusted R-squared:  0.8867
F-statistic:  28.4 on 2 and 5 DF,  p-value: 0.001862
```

Interpretation: Now the regression the regression model is Y X X = – + 16.834 0.2442* 1 7.8488* 2 Since R2 is 0.9593 and the ANOVA shows that the F-ratio is significant, this model can be taken as good-fit in explaining the sales in terms of the other two variables.