Statistical Structures in Data

Assignment 1

Anurag Shukla (22BM6JP08)

Variable Dictionary for R-Script used:

Name	Description	Name	Description
X_dis	Distribution of “dis”	skew_dis	Skewness of “dis”
Y_dis	Distribution of “dis”	kurt_dis	Kurtosis of “dis”
X_dis_mean	Mean of “dis”	fd_dis	Frequency Distribution of “dis”
Y_dis_mean	Mean of “dis”	dis_norm	Normal fitted on “dis”
X_dis_median	Median of “dis”	fd_dis_norm	Frequency Distribution of Normal fitted “dis”
Y_dis_median	Median of “dis”	dis_fit	KS Statistic on “dis”
X_dis_mode	Mode of “dis”	regress	Linear regression model
Y_dis_mode	Mode of “dis”	fstat	F-statistc
range_dis	Range of “dis”	rsquared	R-Squared value
iqr_dis	IQR of “dis”	cooks_dist	Cooke’s Distance
sd_dis	Standard Deviation of “dis”	influential_points	Influential Points

Note: “dis” can take poisson, geo, exp and gamma for various distributions.

CSV files for data:

X_poisson_table	Frequency table for Poisson Distribution
X_geo_table	Frequency table for Geometric Distribution
Y_exp_table	Frequency table for Exponential Distribution
Y_gamma_table	Frequency table for Gamma Distribution

poi_norm_table	Frequency table for fitted Normal on Poisson Distribution
gamma_norm_table	Frequency table for fitted Normal on Gamma Distribution
anova	ANOVA table for linear regression model
influential_points	List of influential points along with their Cooke's Distance

Data Generated as follows (with seed value 08):

1. X_poisson: Discrete Poisson Distribution (lamba = 3)
2. X_geo: Discrete Geometric Distribution (p = 0.2)
3. Y_exp: Continuous Exponential Distribution (lamda = 3)
4. Y_gamma: Continuous Gamma Distribution (gamma = 5)

Solution to question 1 :-

(i) Frequency Distribution of the generated data is as follows:

(a) Discrete Poisson (X) -

Value (X_i)	**Frequency (f_i)**
0	53
1	153
2	226
3	219
4	159
5	101
6	56
7	23
8	8
9	2
Total	1000

(Table generated from R (X_poisson_table.csv) available in the folder)

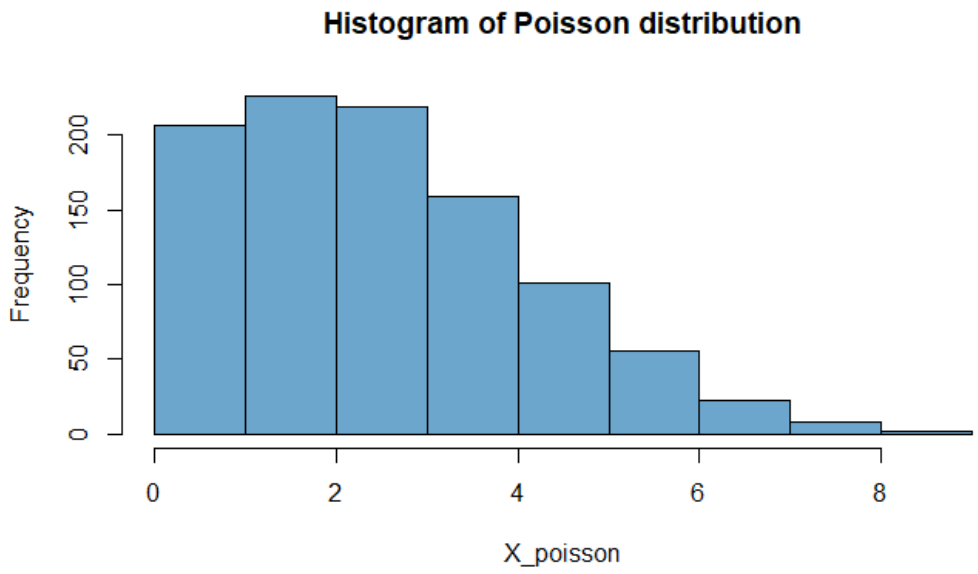
(b) Continuous Gamma (Y) -

Interval (Y_i)	**Frequency (f_i)**
[0,2)	63
[2,4)	305
[4,6)	319
[6,8)	199
[8,10)	81
[10,12)	24
[12,14)	7
[14,16)	1
[16,18)	1
Total	1000

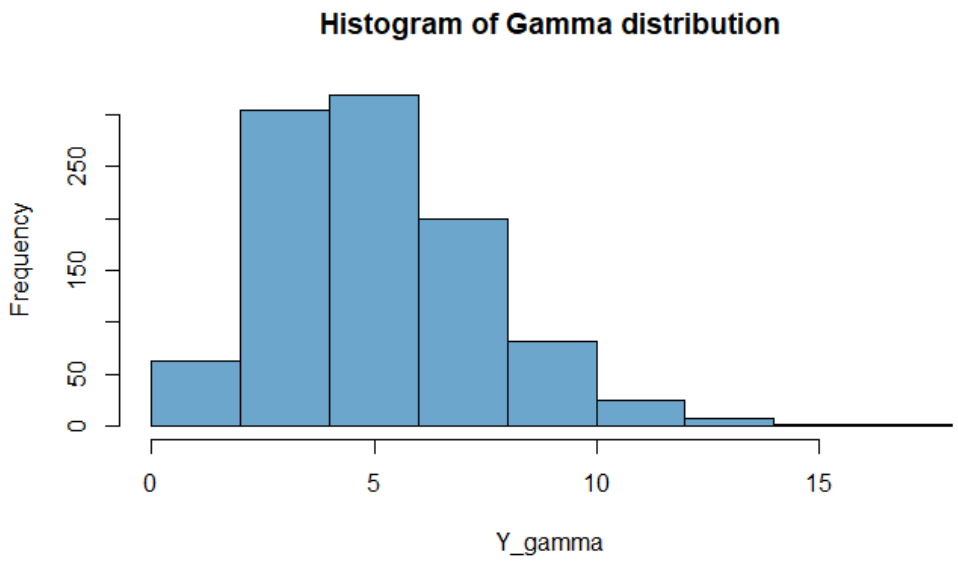
(Table generated from R (Y_gamma_table.csv) available in the folder)

(ii) Histograms of generated Poisson and Gamma Distribution by default (in-built) method in R:

(a) Discrete Poisson Distribution -



(b) Continuous Gamma Distribution –



The default method used by R to decide bin width can be broken down into two steps-

(a) First R calls the ***nclass.Sturges*** function which uses **Sturges' formula** based on bin sizing on range of data given by **ceil($\log_2$ (length(x)) + 1)**. This only finds the recommended value of bin width.

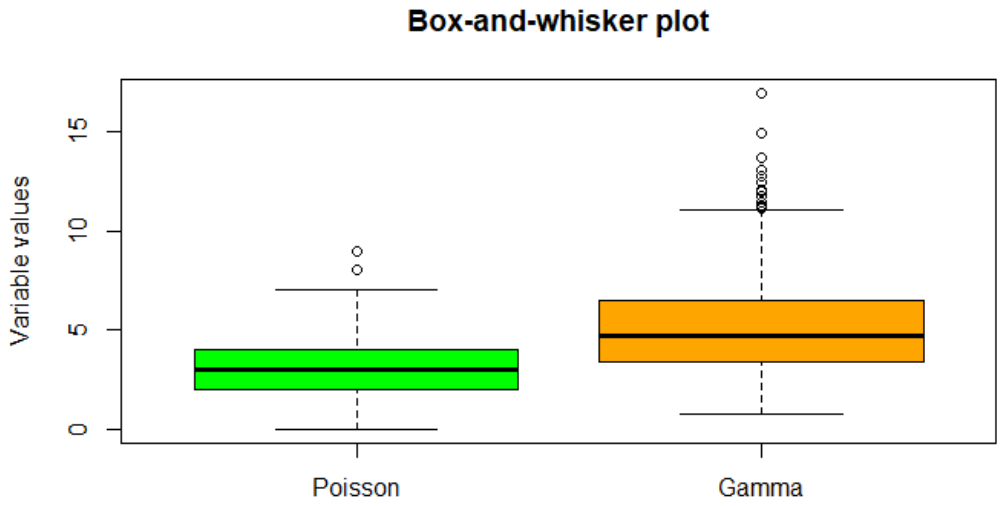
(b) Second step is passing data and the recommended value to ***pretty*** function which computes a sequence of n + 1 equally spaced 'round' values to cover range of data. The values are chosen so that they are 1,2 or 5 times a power of 10.

(iii) Measures of central tendency, dispersion, skewness and kurtosis:

Statistic	Poisson (X)	Gamma (Y)
Mean	2.982	5.056
Median	3	4.723
Mode	2	4.06
Range	[0 , 9]	[0.796 , 16.942]
Interquartile Range	2	3.095
Standard Deviation	1.739	2.316

Skewness	0.507	0.835
Kurtosis	2.950	4.026

(iv) Box-and-Whisker Plot for generated Poisson and Gamma Distribution:



(v) Following can be concluded from above results:

(a) The mean of both distribution (2.982 and 5.056) is **very close to theoretical values** (3 and 5) as our sample size is large enough (1000).

(b) Interquartile range of Gamma distribution (3.095) is **higher than** that of Poisson distribution (2). This is evident from box plot thickness also.

(c) Both the data are **positively (right) skewed data**. Skewness is more for Gamma distribution compared to Poisson distribution that is, Poisson distribution is more symmetric than Gamma distribution (for generated data).

(d) Both the data are **Leptokurtic** that is have kurtosis value more than 1. Comparing both values, Gamma distribution is more peaked (higher kurtosis) than Poisson distribution (for generated data).

(e) From Box-and-Whisker plot we observe that Gamma distribution has **many more outliers** compared to Poisson distribution.

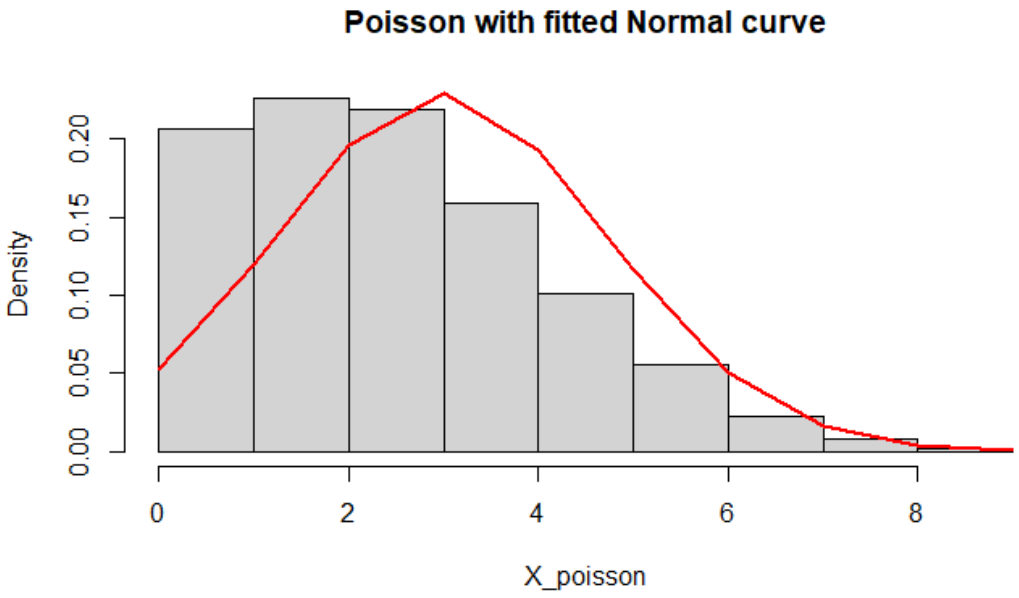
Solution to question 2 :-

(a) Observed and expected frequency table:

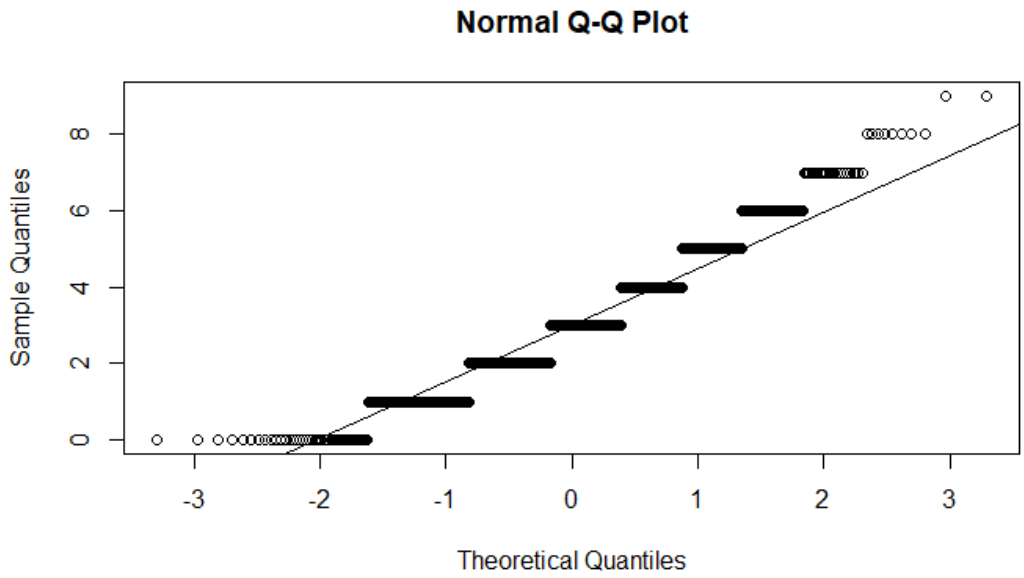
(i) For Discrete Poisson Distribution-

Discrete Variable	Poisson Frequency	Normal Frequency
0	53	57
1	153	130
2	226	190
3	219	212
4	159	198
5	101	104
6	56	59
7	23	17
8	8	3
9	2	0
Total	1000	1000

(Table generated from R (poi_norm_table.csv) available in the folder. Normal continuous approximated to nearest integer to get discrete values)



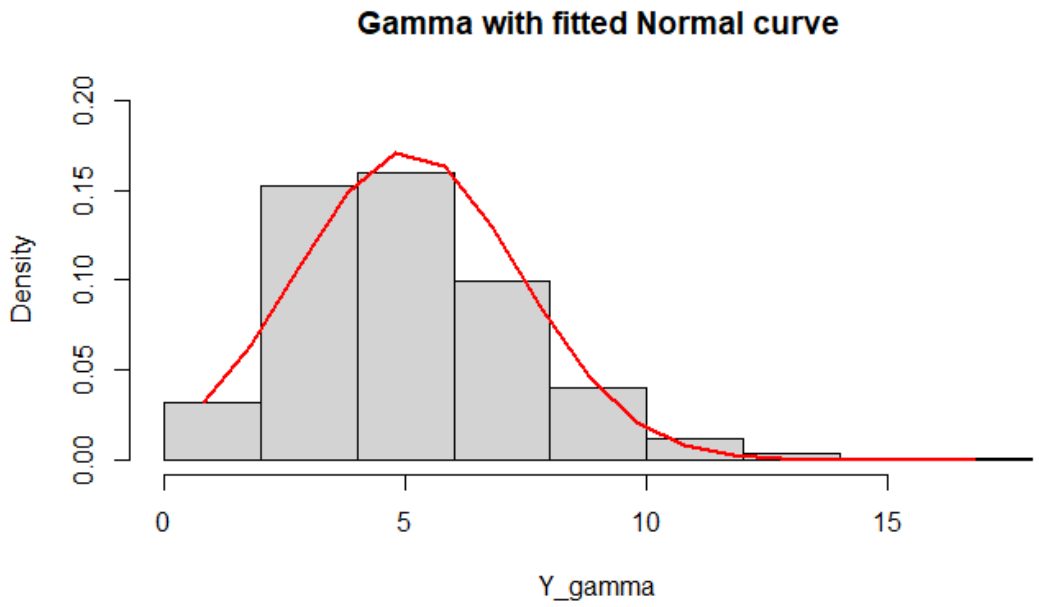
Histogram of Poisson with fitted Normal curve (red)



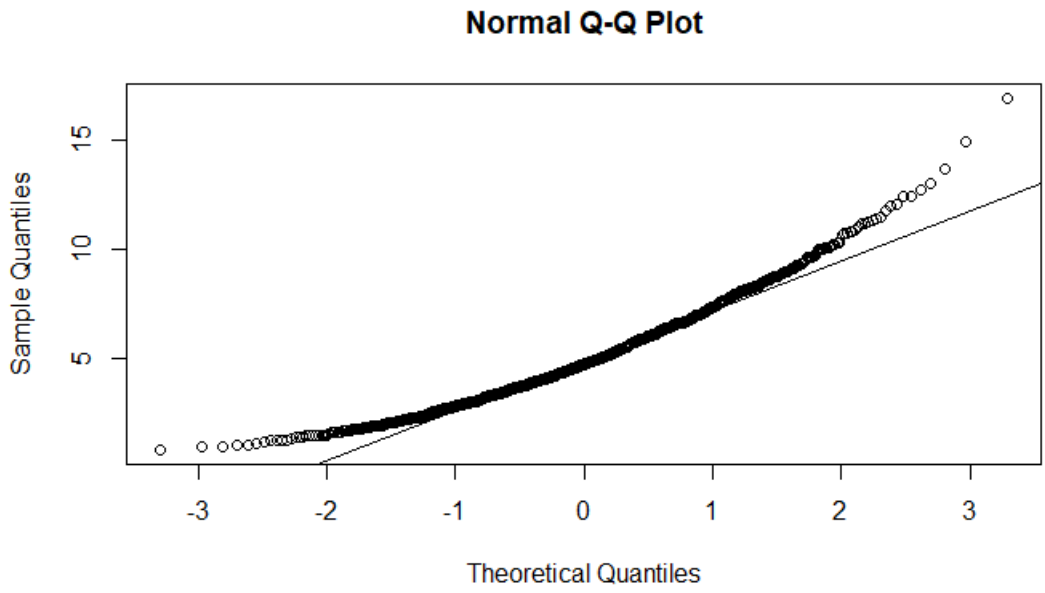
(ii) For Continuous Gamma Distribution-

Variable Interval	Gamma Frequency	Normal Frequency
[0 , 2)	63	87
[2 , 4)	305	239
[4 , 6)	319	318
[6 , 8)	199	231
[8 , 10)	81	91
[10 , 12)	24	15
[12 , 14)	7	0
[14 , 16)	1	0
[16 , 18)	1	0
Total	1000	1000

(Table generated from R (gamma_norm_table.csv) available in the folder)

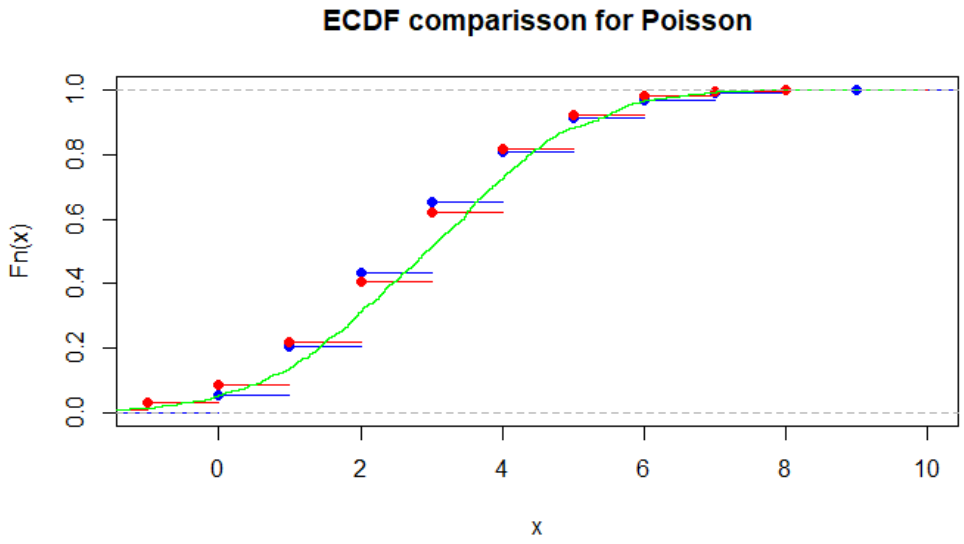


Histogram of Poisson with fitted Normal curve (red)



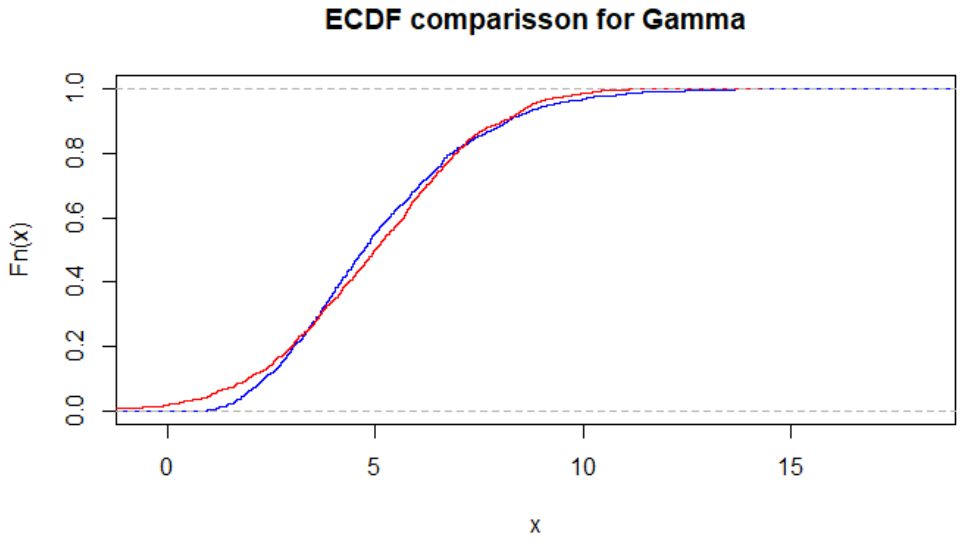
(b) We will use the Kolmogorov-Smirnov test as goodness-of-fit test:

(i) Poisson Distribution (*poi_fit*) – **D = 0.034** **p-value = 0.6099**



(Blue – Sample Distribution CDF Red – Fitted Discrete Normal Distribution CDF Green – Continuous Normal Distribution CDF)

(ii) Gamma Distribution (*gamma_fit*) – **D = 0.055** **p-value = 0.0971**

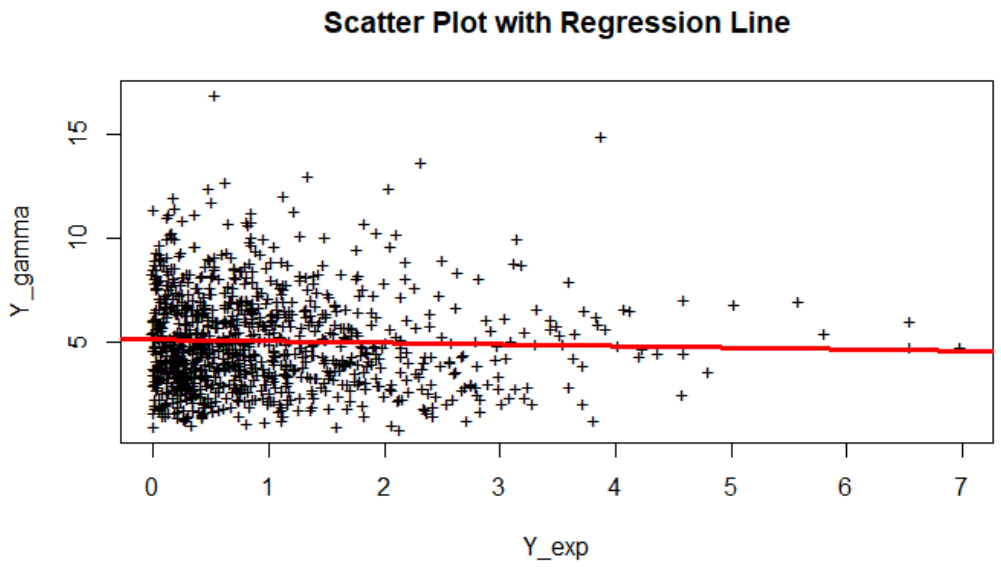


(Blue – Sample Distribution CDF Red – Fitted Normal Distribution CDF)

The null hypothesis here is that both distributions are similar. From the above results, since the p-value for both the Poisson and Gamma Distribution fitted Normal curve is more than **0.05 (5%)** which is considered a norm, we cannot reject the null hypothesis. That is, we conclude that the fitted **Normal Curve is a good fit** (good approximation) for the generated Poisson and Gamma Distributions. This is evident from ECDF plots also since normal CDF and distribution CDF are very close.

Solution to question 3 :-

(i) Regression line (*regress*) scatter plot for Gamma Distribution and Exponential Distribution from the generated data:



From visual inspection of the scatter plot, it is quite evident that the two distributions **do not follow a linear relationship**. That is, fitting a regression line between the two data generates a lot of error values.

(ii) F-statistic (*fstat*) for the linear regression is as below:

Value = **1.1375** Degree of Freedom = {1, 998}

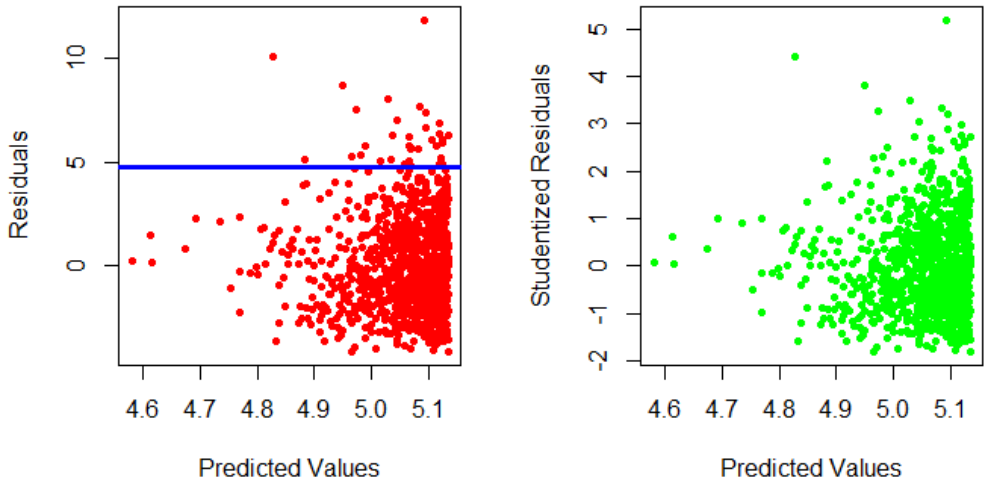
ANOVA Table for this test -

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Y_exp	1	6.10324	6.10324	1.137491	0.286441
Residuals	998	5354.796	5.365527	NA	NA

(Table generated from R (anova.csv) available in the folder) The value of **F-statistic is very low** (high p-value). This means the null hypothesis that the two distributions are similar **can be rejected** with very high evidence. This is in par with observation from scatter plot.

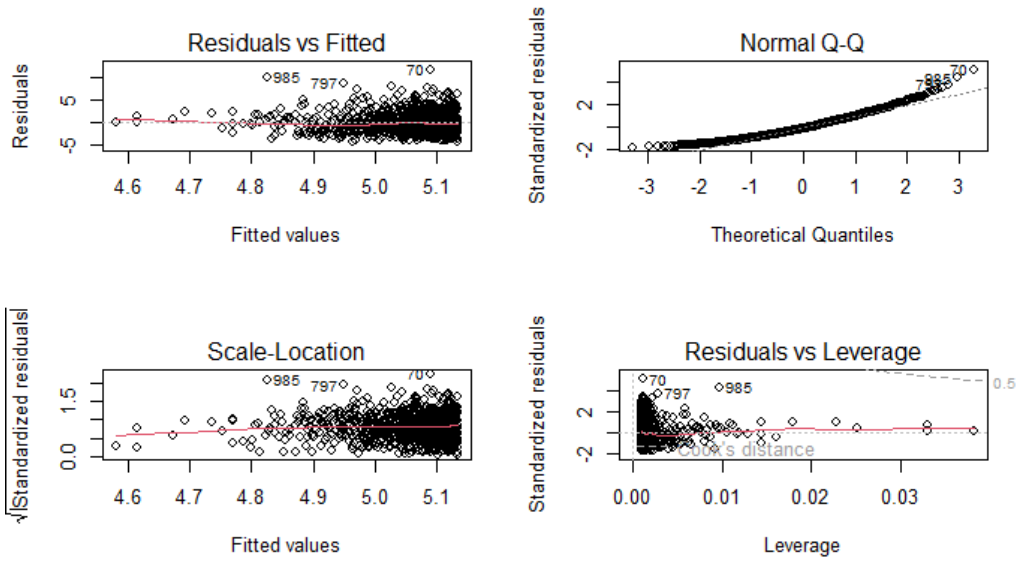
(iii) The coefficient of determination (*rsquared*) R^2 for the fitted model is: **0.001138**. The value of R^2 is very low (close to 0) which tells us that the linear regression model **does not fit** the two generated data. This observation is on par with that of F-statistic test and scatter plot.

(iv) The Residual and Studentized Residual plots for the fitted line is as below:



Both the residuals and studentized residuals plot shows **high amount of residuals** for the fitted linear regression model. The residuals are **biased as well as heteroscedastic**. This shows that the fit is not good. This observation is on par with that from R-squared, F-statistic and scatter plot.

(v) The regression summary plots is as follows:



From the plot of studentized residuals, we can see presence of many influential points. We will define influential points as those which have Cooke's distance (*cooks_dist*) of more than 3 times the mean Cooke's distance for all the observations taken together. We get **66 such influential points** (*influential_points*) (list of values and there distance available in the folder – influential_points.csv). The presence of so many influential points suggests that the linear regression **fit is not good**. This observation is same as from residuals plot, R-squared, F-statistic and scatter plot.

Therefore, from all the above tests, both **visual and heuristic statistical methods**, we conclude that the linear regression line is **not a good fit** from generated data on Exponential Distribution and Gamma distribution.