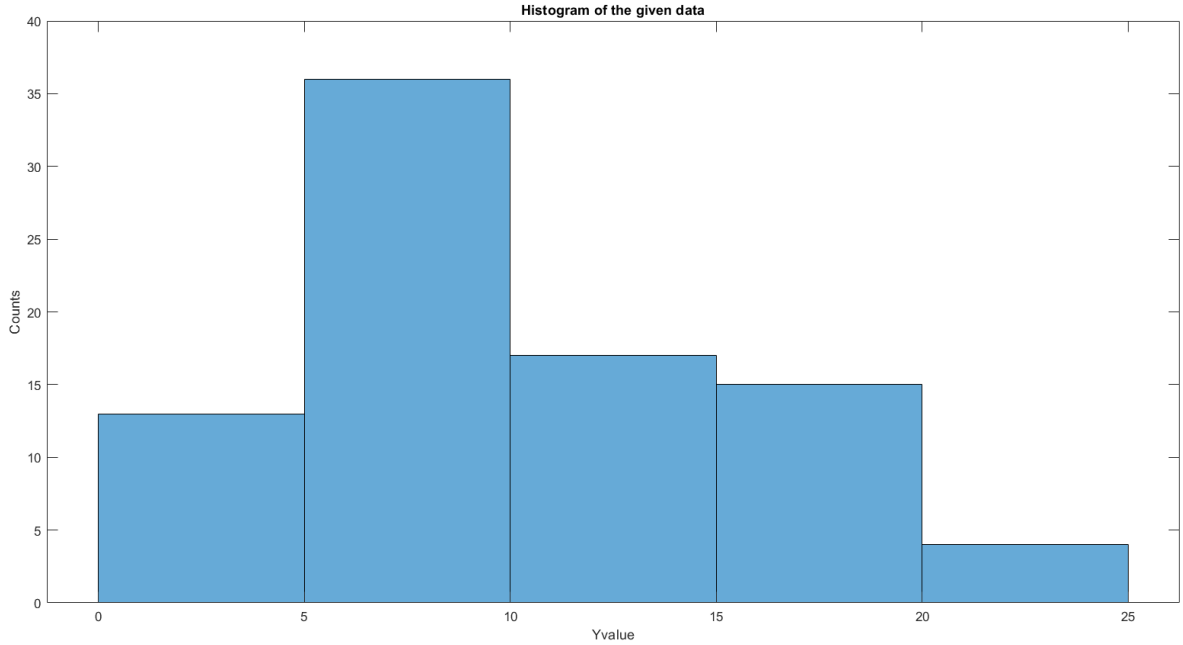
BY: S.VISHAL
CH18B020

CH5115 Assignment-4

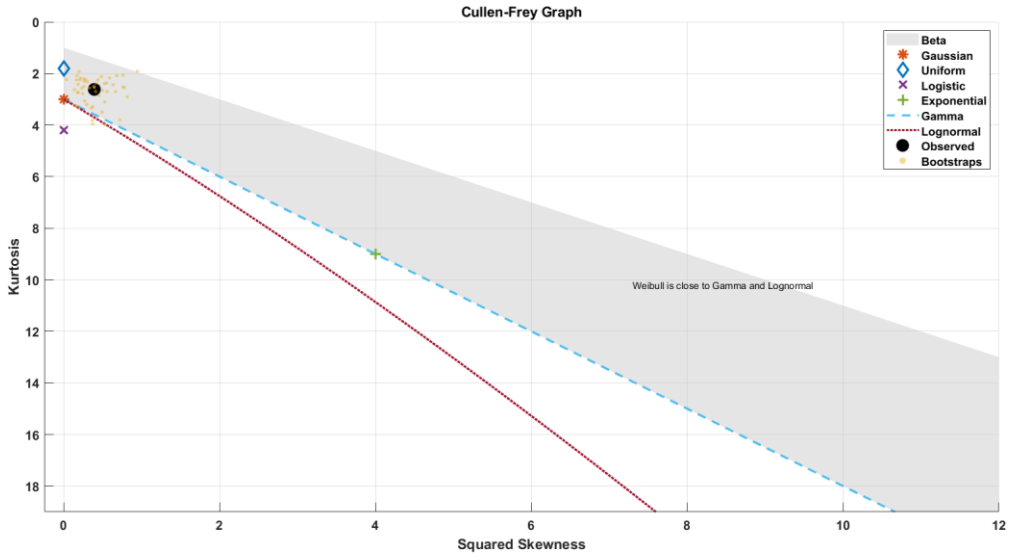
Question 1) a)

Observing the data we can conclude that it is most likely to be generated by a continuous distribution.



Since all Y values are **non-zero** we could say it should be from a distribution like chi-square, Weibull, Rayleigh and others where the realisations are non-negative.

We can plot the **Cullen-Frey graph** (Skewness vs Kurtosis) to make a better guess for the distribution. We should also note that the CF graph estimates are susceptible to errors.



BY: S.VISHAL
CH18B020

Possibilities(as observed from the graph):

1. Beta
2. Gamma
3. Weibull

However note that the random variable in a *beta distribution can have values only between zero and one.* So we eliminate that option.

We fit distributions using **fitdist**() for both **Weibull** and **Gamma**. Negative Log Likelihood is obtained and is used in the **aicbic** function to obtain the AIC (and BIC) score.

	Negative Log Likelihood	**No. of parameters**	**AIC**
Gamma	254.7378	2	513.4755
Weibull	255.2957	2	514.5915

Parameters(the numbers in brackets indicate the **95% confidence interval**)**:**

Gamma distribution

a = 3.57672 [2.68242, 4.76917]

b = 2.83105 [2.0787, 3.8557]

Weibull distribution

A = 11.4721 [10.2897, 12.7904]

B = 2.06503 [1.75126, 2.43502]

From the Negative Log Likelihood and AIC score, we see Gamma distribution wins since it has lower AIC as well as Negative Log Likelihood. But the shape parameter B is close to 2 for Weibull. So, it could be that the data is obtained from a **Rayleigh distribution**.

	Negative Log Likelihood	**No. of parameters**	**AIC**
Rayleigh	255.3670	1	512.7340

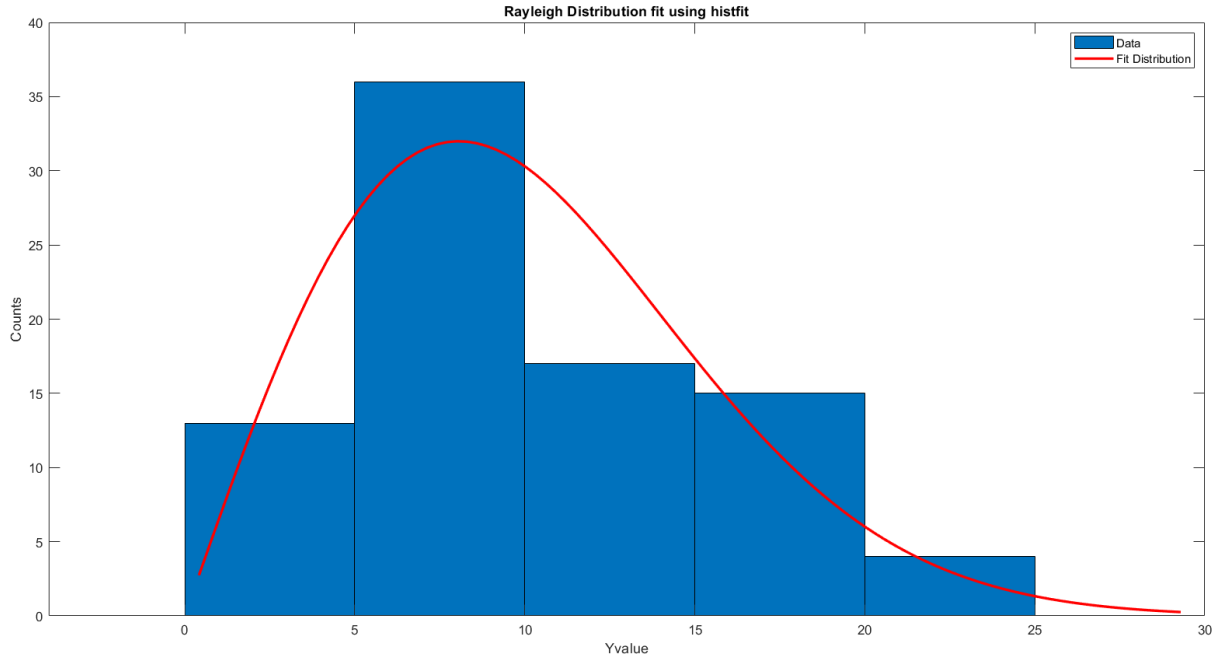
Parameter value**:** B = 8.05729 [7.28428, 9.01529]

Since Rayleigh distribution has the lowest AIC score, we can declare Rayleigh as the best fit.

AD test (for Weibull; since Rayleigh is a type of Weibull distribution) is further performed and the **null hypothesis** that the data is from a Weibull distribution is **not rejected**.(Obtained p value = 0.2092).

histfit is used to fit and visualize the distribution.

BY: S.VISHAL
CH18B020



Thus the fit distribution:

$$f(y) = \frac{y}{b^2} e^{-\frac{x^2}{2b^2}}$$

Where estimate of b = 8.05729.

Time to publication can be imagined as a sort of a lifetime value (where lifetime is the time it takes to work on the article, complete it and get it published). So Rayleigh distribution making the best fit makes sense since it is used to model lifetime of objects.

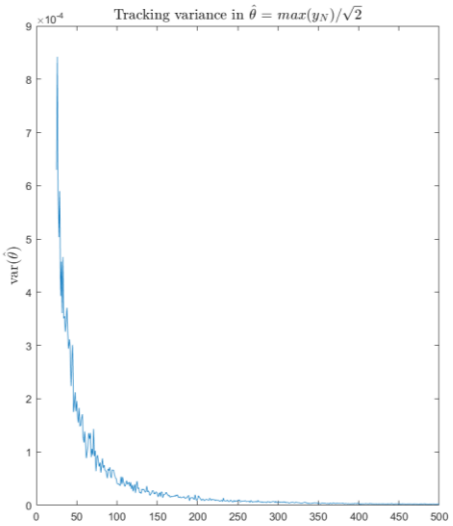
BY: S.VISHAL
CH18B020

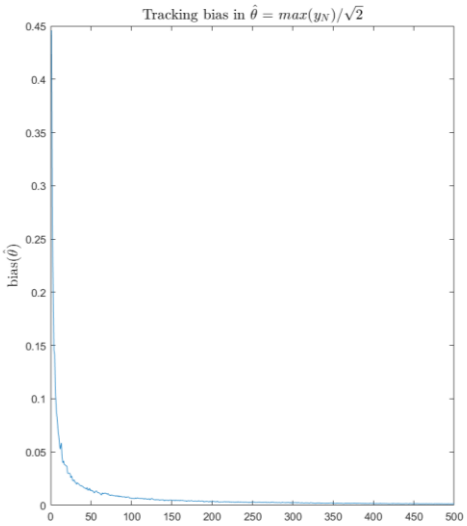
Question 2) d)

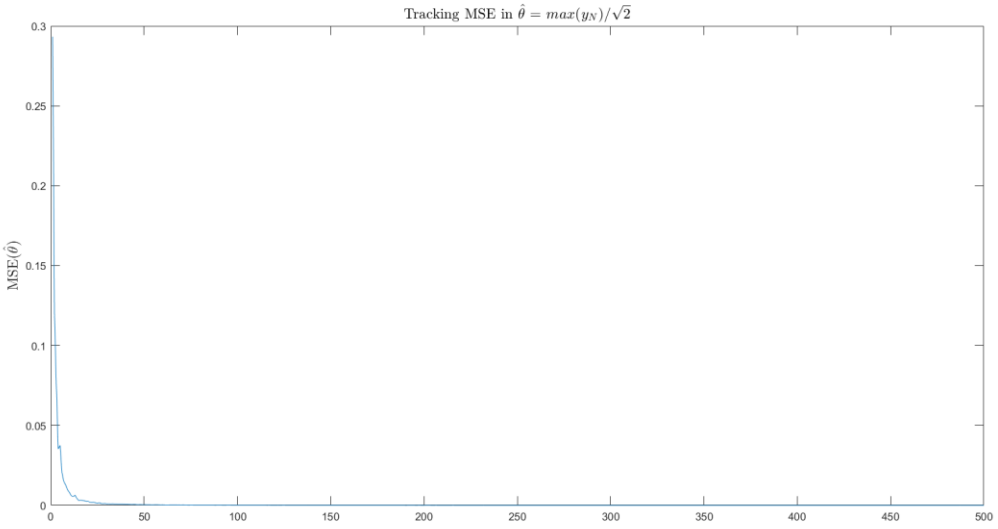
Data sampled from the given distribution using Inverse CDF method. Bias, variance and mse was observed as the number samples increased for both the MLE estimate as well as the corrected estimate.

As expected from the part c), both of them converge in the mean squared sense and hence they are consistent.

MLE estimate:

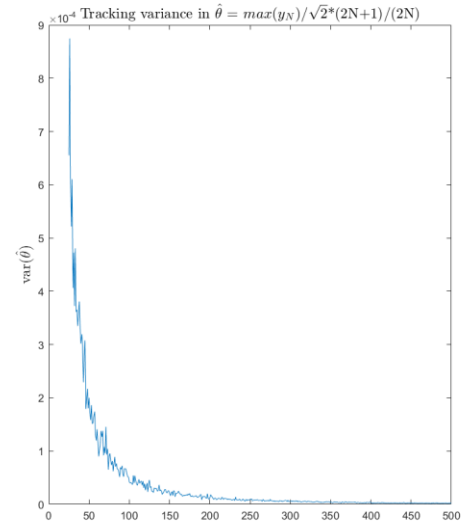


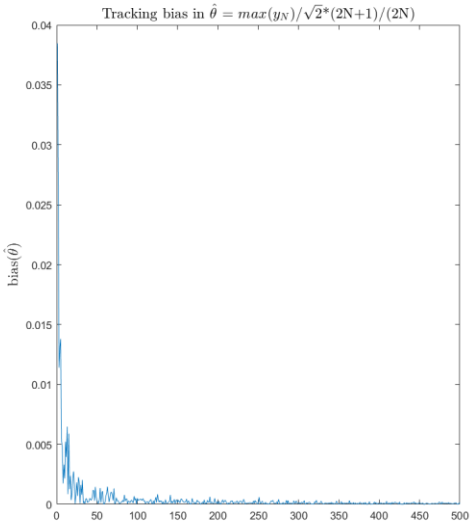


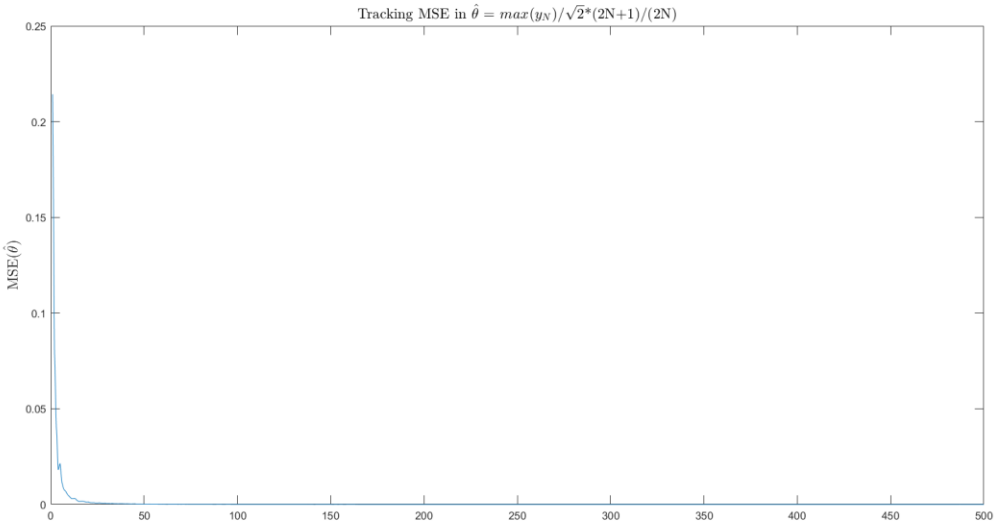


BY: S.VISHAL
CH18B020

Modified Estimate:

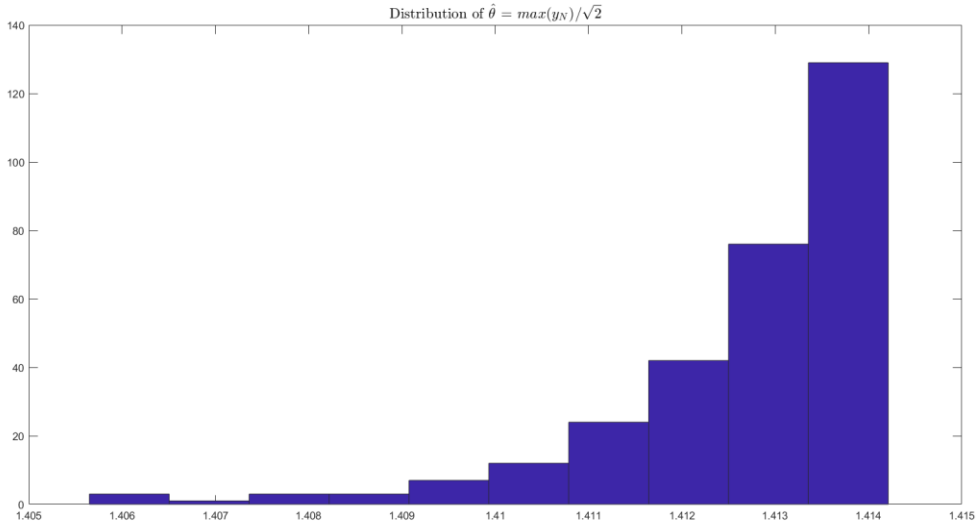


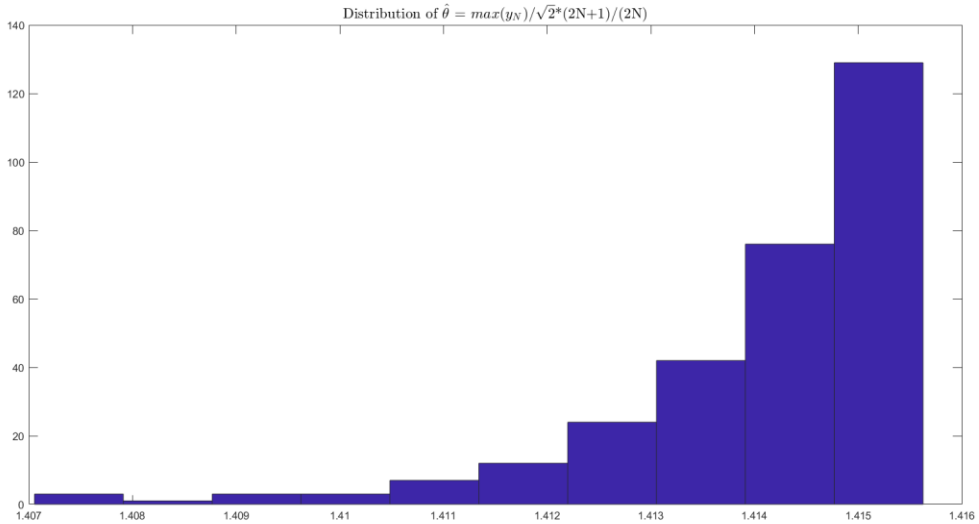




We see that the mean squared errors go to zero as we increase N for both estimators. **Therefore, we can conclude that the estimates exhibit mean squared convergence.**

BY: S.VISHAL
CH18B020





From the above histograms, we can see that the data doesn't resemble the normal distribution at all. To mathematically confirm this AD test is done.

Null hypothesis that the distribution is Gaussian is rejected in both cases.(level of significance 95%)

As mentioned in the handwritten part, the estimates are not Gaussian because the PDF doesn't satisfy the regularity conditions- the support of the PDF depends on the estimate.

BY: S.VISHAL
CH18B020

Question 3)

a) Determining correlation of regressor with output

Correlations were determined using **corr**() function

- correlation between regressor 1 and y is 0.9197
- correlation between regressor 2 and y is 0.8755
- correlation between regressor 3 and y is 0.3998

Regressors 1 & 2 are reasonably correlated with y(close to 0.9). The third regressor is not so well correlated with y (0.4). A linear model must be sufficient to explain the output y. (And in fact, regressor 3 might not be useful)

b) Fit model and compute goodness of fit measures

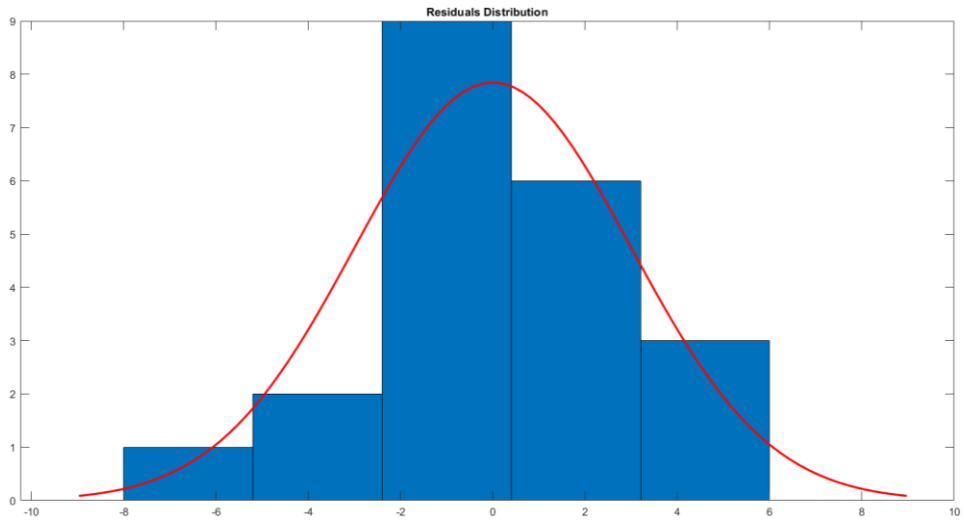
Linear Regression was performed using **fitlm**().

i) $R^2 = 0.9136$; Adjusted $R^2 = 0.8983$

ii) P-value for significance of regression = $3.01633*10^{-09}$. Since p-value< 0.05, **the model is significant**.

iii) **Residual distribution**:
AD-test was performed and the null hypothesis that the residuals are normal was **not rejected**.

Distribution of Residuals



All of the above parameters indicate that the model is good.

c) Test of significance for each coefficient

Coefficient_Name	pValue	Significance
'(Intercept)'	0.0037503	Yes
'x1'	5.799e-05	Yes
'x2'	0.0026301	Yes

'x3'	0.34405	No

The **intercept**, **x1** (air flow), **x2** (temperature) coefficients are **significant**. The **coefficient of x3** (acid concentration) is **insignificant**. This agrees with the correlations which we examined earlier. Air flow and temperature were highly correlated with y (amount of ammonia escaped during its oxidation), x3 (acid concentration) was not.

d) 95% Confidence interval on mean stack loss at α = 0.05

predict() method was used

Prediction of y = 36.0227

Confidence interval for y : $32.2188 \leq y \leq 39.8265$

e) 95% Prediction interval on mean stack loss at α = 0.05

predict() method was used

Prediction of y = 36.0227

Confidence interval for y : $28.1936 \leq y \leq 43.8518$

As expected, the prediction interval is larger.

Note:

Correlation Matrix of Regressors

	X1	**X2**	**X3**
X1	1	0.7819	0.5001
X2	0.7819	1	0.3909
X3	0.5001	0.3909	1

We can see that x3 is correlated with x2 and x1. Also the coefficient of x3 in the model is insignificant. So maybe, x3 can be reconstructed using x1 and x2 (linearly).

Nonetheless, model for y **built using only x1 & x2** gives a slightly lower R^2 but **improved adjusted-R^2** and a **lower AIC**. So we can say that such a model would be better, rather than modelling using all the three regressors.

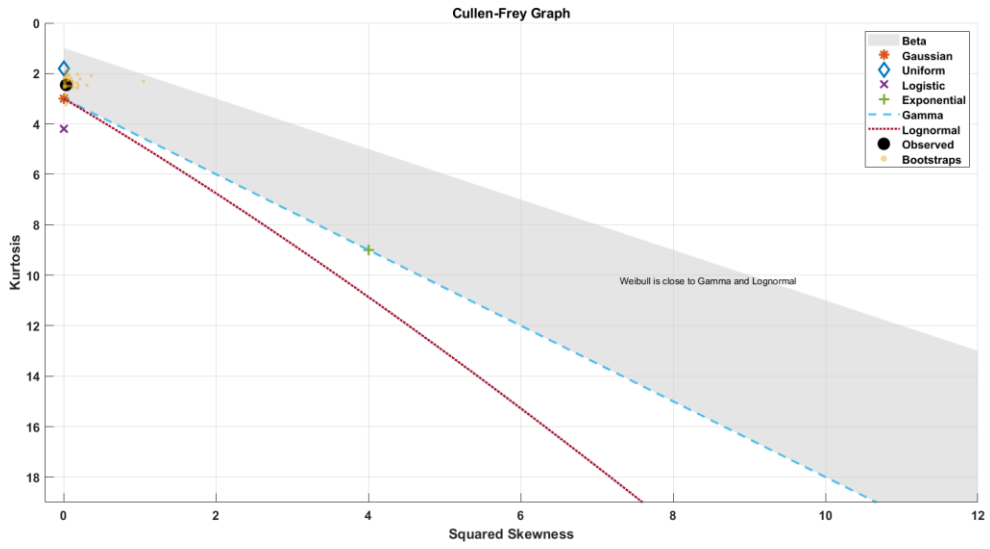
BY: S.VISHAL
CH18B020

Question 4)

a) Performing Linear Regression and model diagnostics

Assumptions in OLS estimators (which are made to ensure good properties):

1. *There is a linear relationship between the response variable and the regressors.* This can be seen from io_corr matrix that has been generated. The first 5 regressors have high correlation with the response variable. (Correlations: 0.9950, 0.9760, 0.9288, 0.9951, 0.8717, -0.1474)
2. *Regressor matrix should be full rank*. The rank of phi matrix is 6, which is same as number of regressors. So this condition is satisfied.
3. *Regressors should be non-constant and have finite excitation.* This is satisfied as seen from the matrix phi.
4. *Regressors are free of errors.* This condition is assumed to hold as there is no way of testing this without the knowledge of measurement mechanisms.
5. *The observations constitute a random sample.* This is again based on how the data is measured/obtained. We shall assume this to be true.
6. *The sequence of equation errors are i.i.d.* We are able to verify that the errors obtained by fitting the model are WN using **lbqtest**. CF graph is also observed and the bootstrap estimates are close to Gaussian. And **ADtest** confirms that they are indeed Gaussian (with 95% significance). This implies that the errors are indeed iid (GWN =>iid).



7. *Residuals are zero-mean conditioned on x.* The estimate of the mean of residuals is very close to zero (= $2.3*10^{-12}$). So, this condition is also satisfied.

Therefore, all the assumptions made on OLS estimators for Linear Regression model are satisfied. And it is further supported by the test of whiteness of residuals.

Model fit using fitlm() method:

Root Mean Squared Error: 26.5

R-squared: 0.998, **Adjusted R-Squared**: 0.997

F-statistic vs. constant model: 2.35e+03, **p-value = 6.07e-42**

The R^2 values are high and p-value is sufficiently low indicating that the **model is significant.**

b) Eliminating the insignificant terms and redoing the regression

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-4738	2444.7	-1.938	0.061213
x1	1.1185	0.28647	3.9045	0.00044089
x2	-0.030184	0.038234	-0.78946	0.43548
x3	0.23062	0.11803	1.9539	0.059231
x4	3.8495	2.6862	1.4331	0.16125
x5	0.82186	0.35075	2.3432	0.025298
x6	-16.946	2.6201	-6.4679	2.4504e-07

We can see that intercept term, x2, x3, x4 are insignificant (if α=0.05). So we build a new model using only **x1, x5** and **x6**.

New model:

Estimated Coefficients:

	Estimate	SE	tStat	pValue
x1	1.6101	0.043065	37.388	5.3825e-31
x5	1.5025	0.13287	11.308	1.4513e-13
x6	-15.318	1.5483	-9.8936	6.1356e-12

Root Mean Squared Error: 28.8

R-squared = 0.997, Adjusted R-squared = 0.997

Once again lbq test is performed, and the residuals are **Gaussian White Noise**. The estimate of the mean is also close to zero (~10^{-13}). All coefficients are significant as seen in the above table. For the sake of parsimony we have lost in terms of the RMSE. As we can see, the RMSE has slightly increased.

c) Model 1 vs Model 2 & Stepwise Regression

Model 1 and Model 2 comparison

We see that the RMSE has slightly increased because we have dropped several regressors. However to have a fair comparison we can compare the AIC/BIC scores which takes into account both the number of regressors to be estimated as well as the error accrued.

	Model-1	Model-2(after dropping regressors)
AIC	382.018	385.2788
AICc	385.518	385.9455
BIC	393.8405	390.3454

CAIC	400.8405	393.3454

Model1 has the better AIC score (lower). But the corrected AICs are very close. If we look at BIC and Consistent AIC (CAIC) which impose more penalty for complex models, we see that Model2 seems to be better.

We conclude that model-2 is the better model. However, this model could be improved by including x3 and the intercept term which despite having p-values greater than alpha, the values are very close to alpha=0.05.

Stepwise Regression

Obtained using **stepwiselm ()** method

Root Mean Squared Error: 26.7

R-squared: 0.997, **Adjusted R-Squared**: 0.997

F-statistic vs. constant model: 3.47e+03, **p-value** = 6.04e-45

Thus the model is significant. The RMSE is close to Model-1 (26.5) but greater than it.

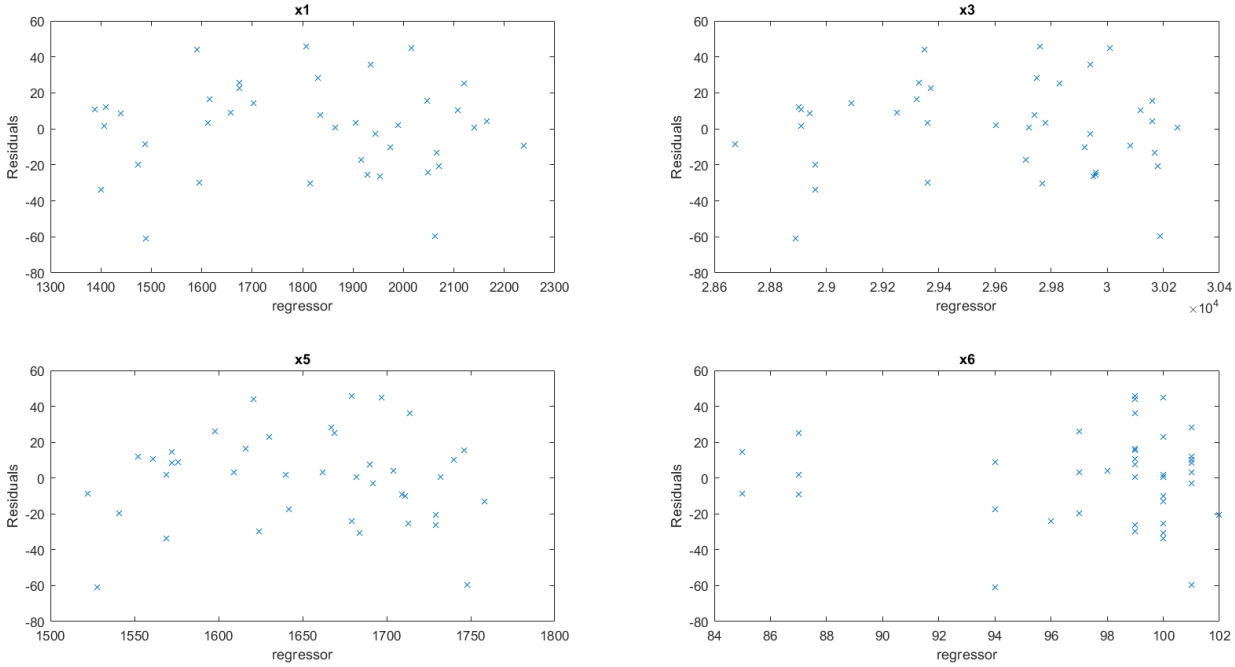
Both **AIC** and **BIC** scores are less than that of model-1 and model-2. This offers an improvement over model-2 because the level of significance for coefficients is relaxed a bit, i.e., α is > 0.05. It was 0.05 earlier. So this allows us to include some more useful regressors, making it a better model.

	Estimate	SE	tStat	pValue
(Intercept)	-4280.2	2257.5	-1.896	0.066245
x1	1.442	0.14258	10.114	6.3049e-12
x3	0.20982	0.10157	2.0657	0.046322
x5	0.64674	0.32624	1.9824	0.055331
x6	-17.51	2.336	-7.4958	8.8577e-09

Conclusion: Stepwise regression model is the best model overall (as accounted by AIC & BIC scores).

BY: S.VISHAL
CH18B020

d) Residuals vs Regressors plot- Checking for non-linearities



By plotting each regressor vs residual, we can check whether the residual (and hence y) has any relationship of the form f(xi) where xi is the regressor on x-axis.

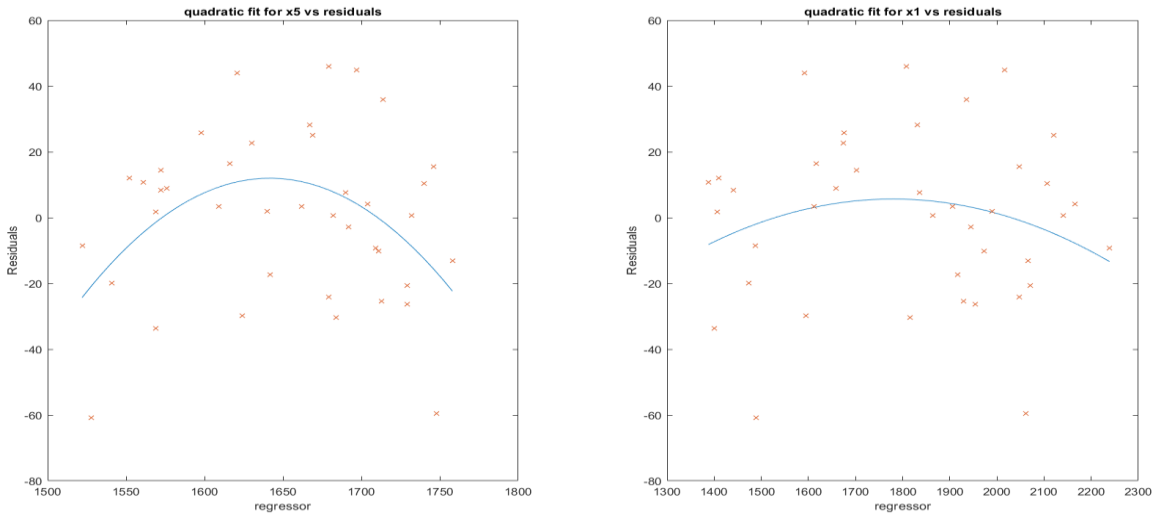
$$\epsilon = f(x1) + g(x3) + h(x5) + j(x6) + \delta$$

$$=> \; y = \; Ax + f(x1) + g(x3) + h(x5) + j(x6) + \delta$$

*Note: The residual can also be function of two or more variables simultaneously, such as x3*x5 or even x1*sin(x2*b/x3) etc. I am not sure how to check for such relationships. (One can go for 3D plots but they don't seem as informative as 2D, and beyond 3D it is not possible to visualize)*

Regressors x3 and x6 don't seem to have any particular relationship with residuals but x1 and x5 seem to have very rough quadratic relationship (inverted parabola).

A quadratic fit is performed using polyfit and the graph is attached below.



BY: S.VISHAL
CH18B020

e) Fit non-linear model & compare with the linear model

A new model of the form `'y ~ x1 + x5 + x6 + x3 + x1^2 + x5^2 - x2 - x4 - 1'` is built using fitlm(). (The model is still linear in coefficients, so it is still linear regression). (variable name: mdl5)

```
Linear regression model:
    y ~ x1 + x3 + x5 + x6 + x1^2 + x5^2
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
x1	2.2915	0.34731	6.5978	1.4569e-07
x3	-0.027427	0.095977	-0.28576	0.77679
x5	1.8015	2.8232	0.63811	0.52768
x6	-13.159	2.1656	-6.0765	6.8533e-07
x1^2	-0.00014683	8.6287e-05	-1.7017	0.097942
x5^2	-0.00023239	0.00078768	-0.29503	0.76977

```
Number of observations: 40, Error degrees of freedom: 34
Root Mean Squared Error: 27.3
```

A stepwise lm (Variable name: mdl4) is also performed, where interaction terms (xi*xj) and quadratic terms (xi*xi) are allowed. Significance levels for entry and exit are same as those in part c). This yields a model of the form:

```
Linear regression model:
    y ~ 1 + x1 + x5 + x6 + x5^2
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-8386	2641.1	-3.1752	0.0031184
x1	1.6739	0.059395	28.182	1.2021e-25
x5	11.812	3.2338	3.6526	0.00084167
x6	-14.675	1.486	-9.8756	1.1751e-11
x5^2	-0.0032273	0.00097167	-3.3214	0.0021041

```
Number of observations: 40, Error degrees of freedom: 35
Root Mean Squared Error: 24.7
R-squared: 0.998,  Adjusted R-Squared 0.998
F-statistic vs. constant model: 4.07e+03, p-value = 3.74e-46
```

As we can see the final model is better in terms of number of parameters estimated as well as the RMSE. So, obviously it is the winner of AIC/BIC tests too. Mdl4 is the better non-linear model.

BY: S.VISHAL
CH18B020

	RMSE	**No. of parameters**	**AIC**
Linear model	26.7	4	381.021
Non-Linear model (technically it is linear)	24.7	4	374.663

The incorporation of x5^2 term has offered some improvement over the vanilla linear model.

To conclude, the model mdl4 is declared the best model!