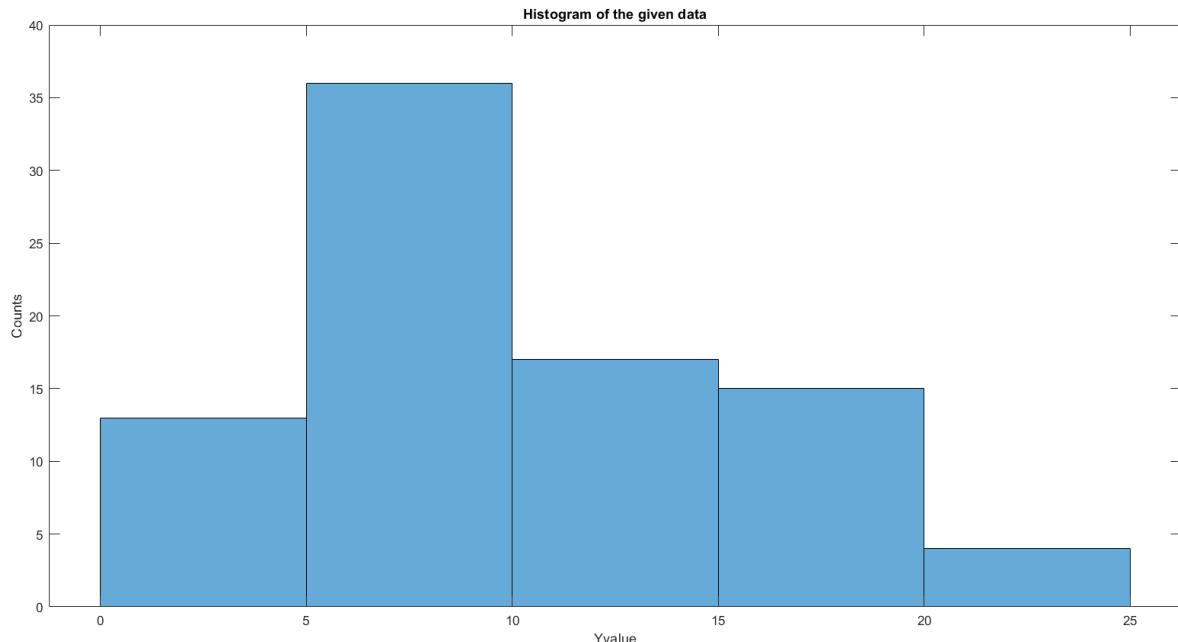


CH5115 Assignment-4

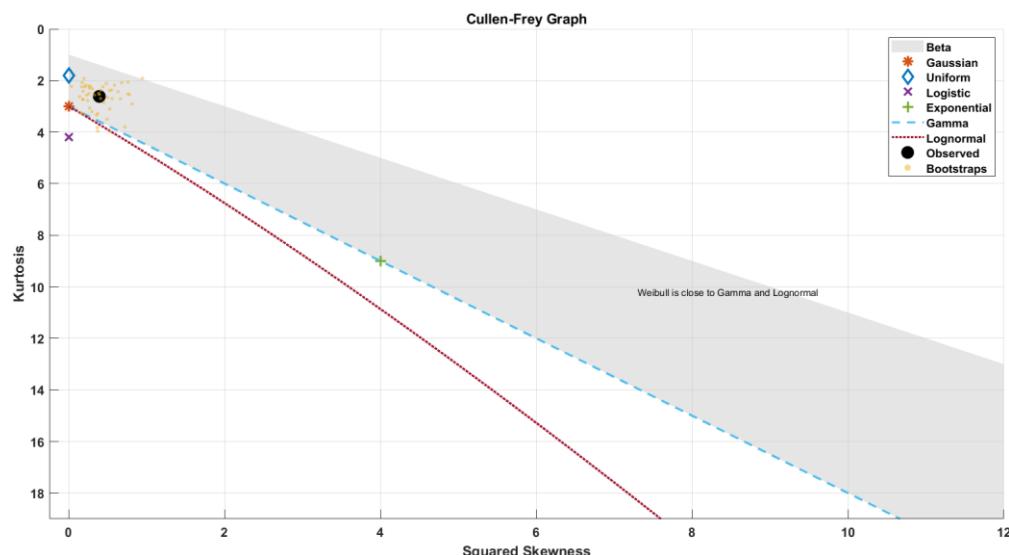
Question 1) a)

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.



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**.

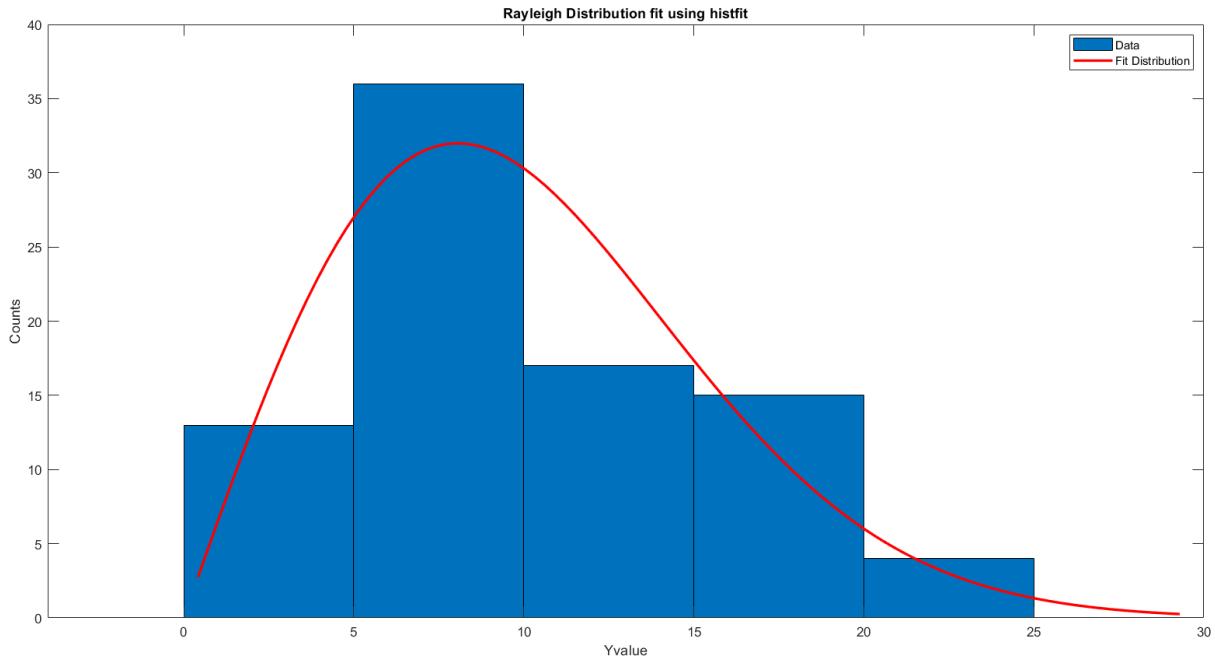
	Negative Log Likelihood	No. of parameters	AIC
Rayleigh	255.3670	1	512.7340

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.



Thus the fit distribution:

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

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.

① b) i) Systematic error:

- These errors shifts or elongates the distribution by a constant factor.
- That is, X_{measured} could be of form $aX + b$ where X is the ~~true~~ value of the realization of X after and X_m is the measured value of X .
- This error affects the mean and variance of the distribution. However the distribution of X_m is likely to be same as distribution of X .
- a: multiplication / scaling factor.
- b: offset / zero-setting error.

ii) Random error:

- These errors occur independent of the realization X .
- $X_{\text{measured}} = X + \epsilon$.
- Some they are random & they tend to have zero mean and they can be averaged out by increasing the sample size.
- The distribution of X_{measured} depends on the Signal to Noise Ratio. If SNR is very high, X_{measured} essentially follows distribution of X . If it is very low (near 0), X_{measured} is distributed as determined by ϵ .
- In most cases, SNR can be increased by increasing sample size.

However, systematic errors are difficult to detect.
unless we have some ^{prior} information on the true value.

CH5115 ASSIGNMENT-4

②

$$a) f(y; \theta) = \begin{cases} \frac{2y}{\theta^2} & 0 \leq y \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} l &= \prod_{k=1}^N \frac{2y[k]}{\theta^2} && \text{if } 0 \leq y[k] \leq \theta \\ &= \frac{2^N}{\theta^{2N}} \prod_{k=1}^N y[k] \\ \Rightarrow L &= \log \left(\frac{2^N}{\theta^{2N}} \prod_{k=1}^N y[k] \right) \\ &= \sum \log y[k] + N \log 2 - 2N \log \theta \end{aligned}$$

$$\Rightarrow \frac{\partial L}{\partial \theta} = \frac{-2N}{\theta}$$

We notice that there is no maxima/minima in the domain of theta.

So we just have to check for the extreme values

Also $\frac{\partial L}{\partial \theta} < 0 \forall \theta (\because \theta \in \mathbb{R}[0, \infty))$

\Rightarrow Likelihood is a decreasing function of θ

\Rightarrow We need to choose the minimum possible value of θ . — ① ~~(not if θ can be ∞)~~

Also note $l = 0$ if $y[k]$ doesn't belong to $[0, \theta]$

For all $y[k]$ to be included in this range,

$$\theta \geq \max(y[k]) — ②$$

From ① & ②, $\hat{\theta}_{MLE} = \max(\underline{y}_N)$ — ③

Checking for bias.

We need the PDF of $\max_y(\underline{y}_N)$ to estimate the mean.

$$F(Y \leq y) = \int_0^y \frac{2y}{\theta^2} = \frac{y^2}{\theta^2}$$

If $\max(\underline{y}_N) \leq y$, then $y[k] \leq y + k$

$$\Rightarrow F(\max(\underline{y}_N) \leq y) = (F(Y \leq y))^N$$

$$\Rightarrow F(\max(\underline{y}_N) \leq y) = \left(\frac{y^2}{\theta^2}\right)^N = \frac{y^{2N}}{\theta^{2N}}$$

$$\Rightarrow f(\max(\underline{y}_N) = y) = \frac{2N y^{2N-1}}{\theta^{2N}} \quad — ④$$

$$E(\max(\underline{y}_N)) = \int_0^\theta 2N y \times \frac{y^{2N-1}}{\theta^{2N}} dy$$

$$= \left(\frac{2N}{2N+1}\right) \frac{\theta^{2N+1}}{\theta^{2N}} = \left(\frac{2N}{2N+1}\right) \theta$$

$$\Rightarrow E(\hat{\theta}_{MLE}) = \left(\frac{2N}{2N+1}\right) \theta \quad — ⑤$$

So the MLE estimator is biased. Correcting

for the bias, $\hat{\theta}_{unbiased} = \left(\frac{2N+1}{2N}\right) (\max(\underline{y}_N))$ — ⑥

b)

Let x be the median of Y

$$\int_0^x \frac{2y}{\theta^2} dy = \frac{1}{2} \quad (\because F(Y \leq y) = F(Y \geq y)) \\ = \frac{1}{2}$$

for median)

$$\Rightarrow \frac{x^2}{\theta^2} = \frac{1}{2}$$

$$\Rightarrow x = \frac{\theta}{\sqrt{2}} \quad \text{--- (7)}$$

($\because x \geq 0$) [Technically, the $\pm \theta$ -sign depends on whether θ itself is +ve or negative. But without loss of generality we can assume $\theta > 0$ because it is present as θ^2 in the PDF]

Now, we know MLE has the following property,

$$\text{if } \phi = g(\theta) \text{ then } \hat{\phi}_{MLE} = g(\hat{\theta}_{MLE}) \quad \text{--- (8)}$$

$$\Rightarrow \hat{\phi}_{MLE} = \frac{\hat{\theta}_{MLE}}{\sqrt{2}} \Rightarrow$$

$$\Rightarrow \hat{x}_{MLE} = \frac{\max(y_N)}{\sqrt{2}} \quad \text{--- (9)}$$

However since the estimator is biased,
 $E(\hat{x}_{MLE}) = \frac{E(\max(y_N))}{\sqrt{2}} = \frac{2N}{2N+1} \frac{\theta}{\sqrt{2}} = \frac{2N}{2N+1} x$

we propose a modification,

$$\boxed{\hat{x} = \left(\frac{2N+1}{2N} \right) \left(\frac{\max(y_N)}{\sqrt{2}} \right)} \quad \text{--- (10)}$$

c) We have already shown that the estimator is unbiased

$$\text{var}(\hat{\theta}_{\text{MLE}}) = \left(\frac{2N+1}{2N}\right)^2 \times \frac{1}{2} \times \text{var}(\max(y_N))$$

$$\text{var}(\max(y_N)) = E((\max(y_N))^2) - (E(\max(y_N)))^2$$

From the PDF in eqn ④,

$$E((\max(y_N))^2) = \int_0^\Theta y^2 \left(\frac{2N y^{2N-1}}{\Theta^{2N}} \right) dy \\ = \frac{2N}{\Theta^{2N}} \int_0^\Theta y^{2N+1} dy = \frac{2N \times \Theta^{2N+2}}{(2N+2) \Theta^{2N}}$$

$$\Rightarrow E((\max(y_N))^2) = \left(\frac{2N}{2N+2} \right) \Theta^2$$

$$\therefore \text{var}(\max(y_N)) = \Theta^2 \left(\left(\frac{2N}{2N+2} \right) - \left(\frac{2N}{2N+1} \right)^2 \right)$$

$$= \Theta^2 \left(\frac{1}{1 + \frac{1}{N}} - \left(\frac{1}{1 + \frac{1}{2N}} \right)^2 \right)$$

$$= \Theta^2 \frac{1 + \frac{1}{4N^2} + \frac{1}{N} - \frac{1}{N} - \frac{1}{N}}{\left(1 + \frac{1}{N}\right)\left(1 + \frac{1}{2N}\right)^2}$$

$$= \Theta^2 \left(\frac{\frac{1}{4N^2}}{\left(1 + \frac{1}{N}\right)\left(1 + \frac{1}{2N}\right)^2} \right) \quad \text{--- ⑦}$$

$$\lim_{N \rightarrow \infty} \text{var}(\text{max}(y_N)) = \lim_{N \rightarrow \infty} \frac{\theta^2 \left(\frac{1}{4N^2} \right)}{\left(1 + \frac{1}{N} \right) \left(1 - \frac{1}{N} \right)^2}$$

$$\Rightarrow \lim_{N \rightarrow \infty} \text{var}(\text{max}(y_N)) = 0 \quad \text{--- (8)}$$

$$\begin{aligned} \lim_{N \rightarrow \infty} (E((\hat{x} - x)^2)) &= (\Delta \hat{x})^2 + \text{var}(\hat{x}) \\ &= \lim_{N \rightarrow \infty} 0 + \left(\frac{2N+1}{2N} \right)^2 + \frac{1}{2} \times \text{var}(\text{max}(y_N)) \\ &= \lim_{N \rightarrow \infty} \left(\frac{\text{var}(\text{max}(y_N))}{2} \right) \\ \Rightarrow \lim_{N \rightarrow \infty} E((\hat{x} - x)^2) &= 0 \end{aligned}$$

the estimator exhibits mean squared convergence

$\Rightarrow \hat{x} = \left(\frac{2N+1}{2N} \right) \left(\frac{\text{max}(y_N)}{\sqrt{2}} \right)$
 is a consistent estimator

d) Consistency is verified numerically.
 However, we note that the estimator doesn't
don't doesn't show asymptotic normality (asymptotically Gaussian)
 because the PDF doesn't satisfy all the regularity conditions

It violates the condition that the support of
 the PDF is independent of θ .

Here, the support: $[0, \theta]$ depends on
 θ itself.

So from theory we conclude that our observations from numerical simulations that the estimates are not asymptotically Gaussian is correct.

$$\text{contd. } \lim_{N \rightarrow \infty} E((x - \hat{x}_{MLE})^2) = \lim_{N \rightarrow \infty} (x - E(\hat{x}_{MLE}))^2 + \text{var}(\hat{x}_{MLE})$$

Also, even if we consider the unbiased estimator,

$$\hat{x}_{MLE} = \frac{\max(y_N)}{\sqrt{2}}$$

$$\text{we still notice } \lim_{N \rightarrow \infty} \text{var}(\hat{x}_{MLE}) = \lim_{N \rightarrow \infty} \frac{\text{var}(\max(y_N))}{2} = 0 \quad \text{--- (1)}$$

$$\text{ie } \lim_{N \rightarrow \infty} x - E(\hat{x}_{MLE}) = \lim_{N \rightarrow \infty} x - \left(\frac{2N}{2N+1} \right) x.$$

$$= \lim_{N \rightarrow \infty} \frac{x}{2N+1} \rightarrow 0 \quad \text{--- (2)}$$

$\text{① } \text{ie } \text{②} \Rightarrow \hat{x}_{MLE}$ is also consistent in the mean squared sense.

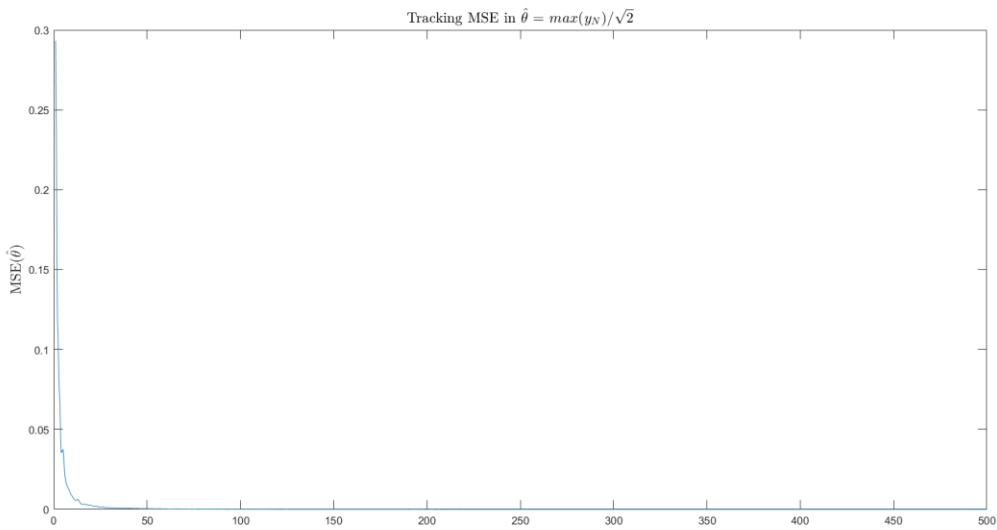
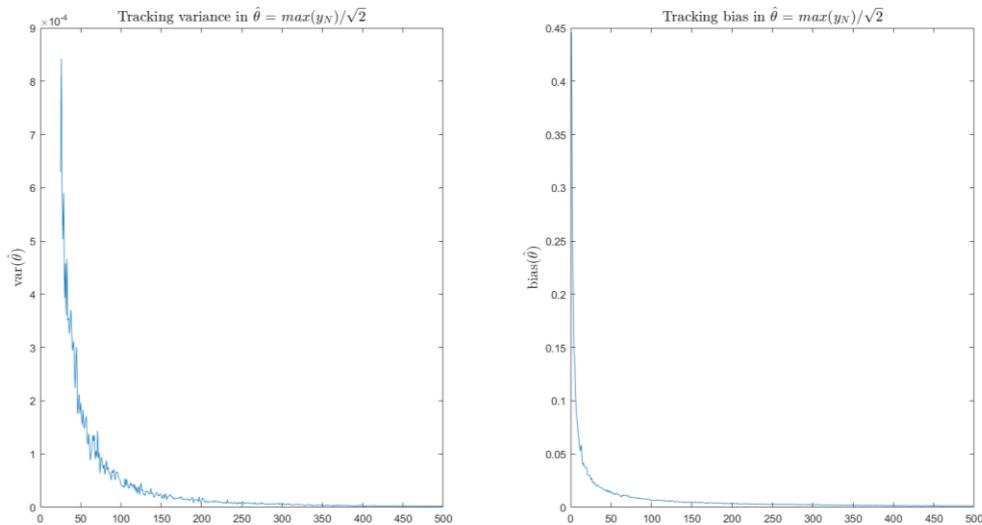
(because this even though it is biased, it is asymptotically unbiased)

Question 2) d)

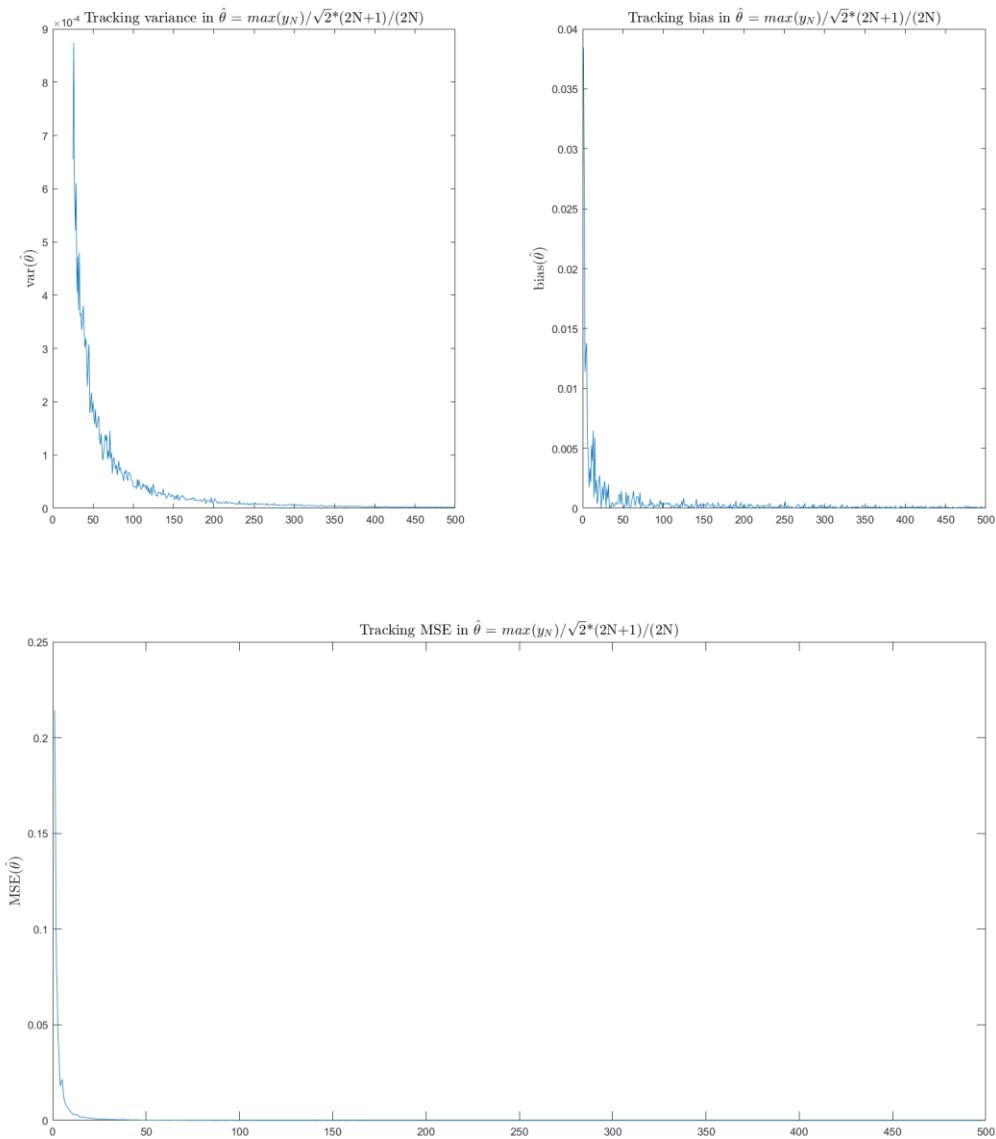
Data sampled from the given distribution using Inverse CDF method. Bias, variance and mse was observed as the number samples increased for both the MLE estimate as well as the corrected estimate.

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

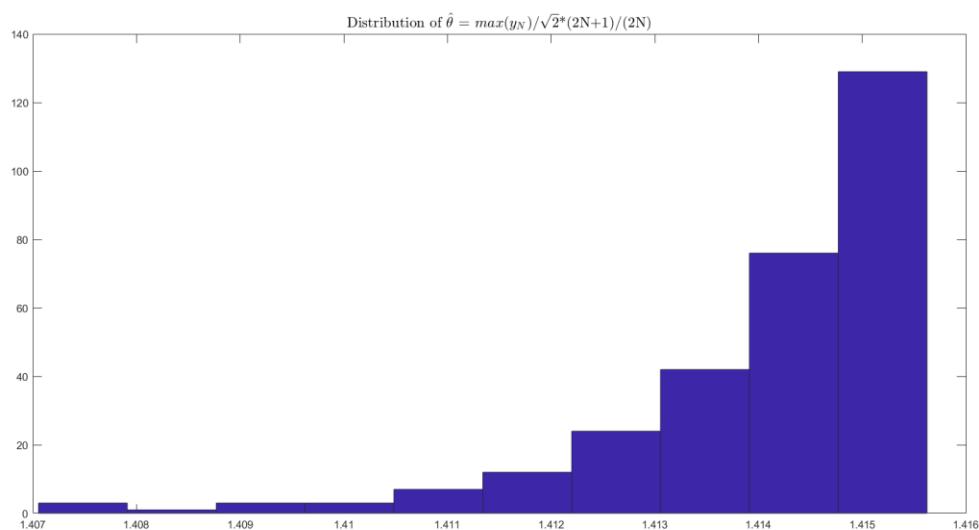
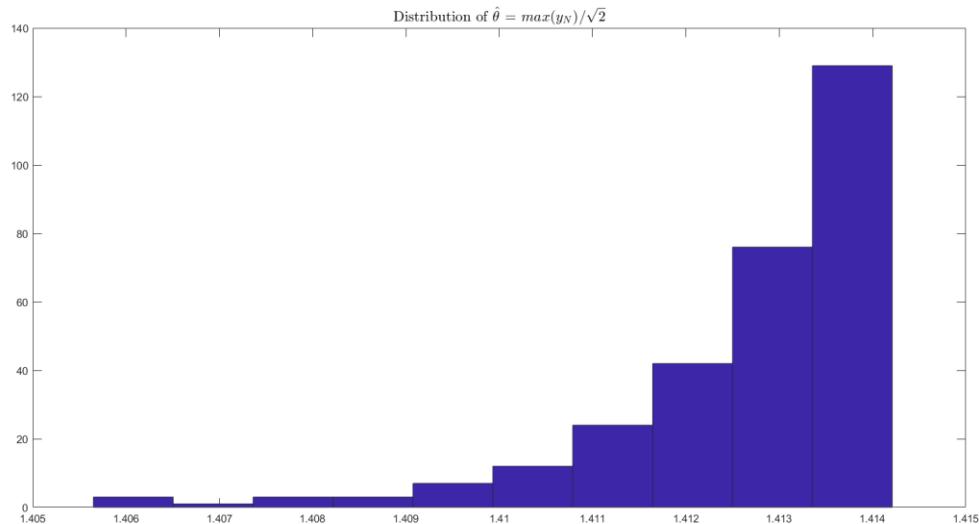
MLE estimate:



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.**



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.

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

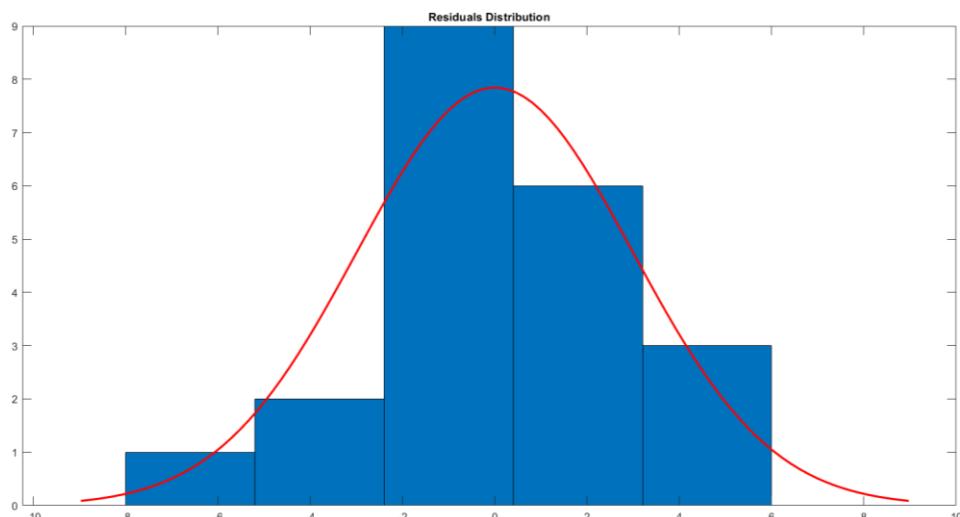
Regressors 1 & 2 are reasonably correlated with y (close to 0.9). The third regressor is not so well correlated with y (0.4). A linear model must be sufficient to explain the output y. (And in fact, regressor 3 might not be useful)

b) Fit model and compute goodness of fit measures

Linear Regression was performed using `fitlm()`.

- $R^2 = 0.9136$; Adjusted $R^2 = 0.8983$
- P-value for significance of regression = 3.01633×10^{-9} . Since p-value < 0.05, **the model is significant.**
- 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 $\alpha = 0.05$

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 $\alpha = 0.05$

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).

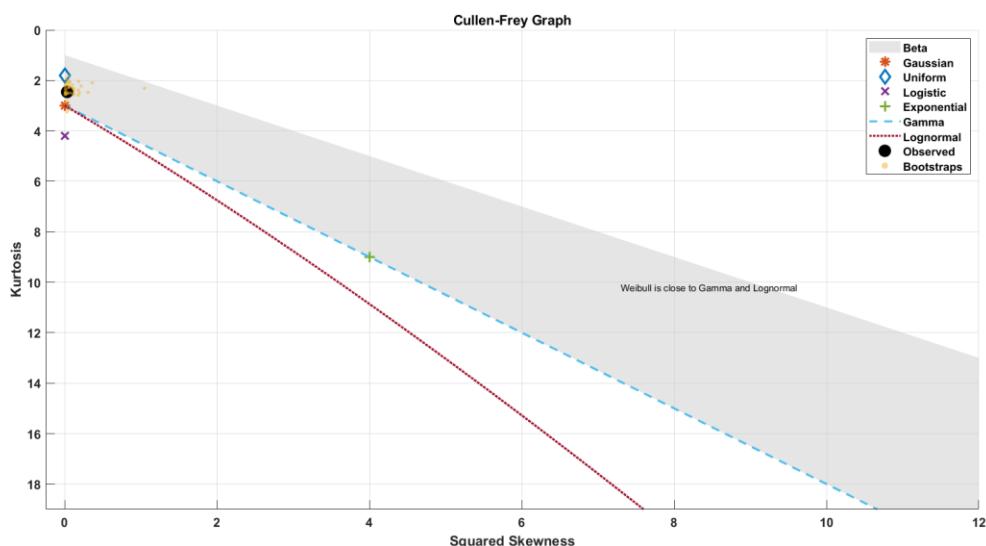
Nonetheless, model for y **built using only x1 & x2** gives a slightly lower R^2 but **improved adjusted-R²** and a **lower AIC**. So we can say that such a model would be better, rather than modelling using all the three regressors.

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 ($= 2.3 \times 10^{-12}$). 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 R² 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 $\alpha=0.05$). 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 ($\sim 10^{-13}$). 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

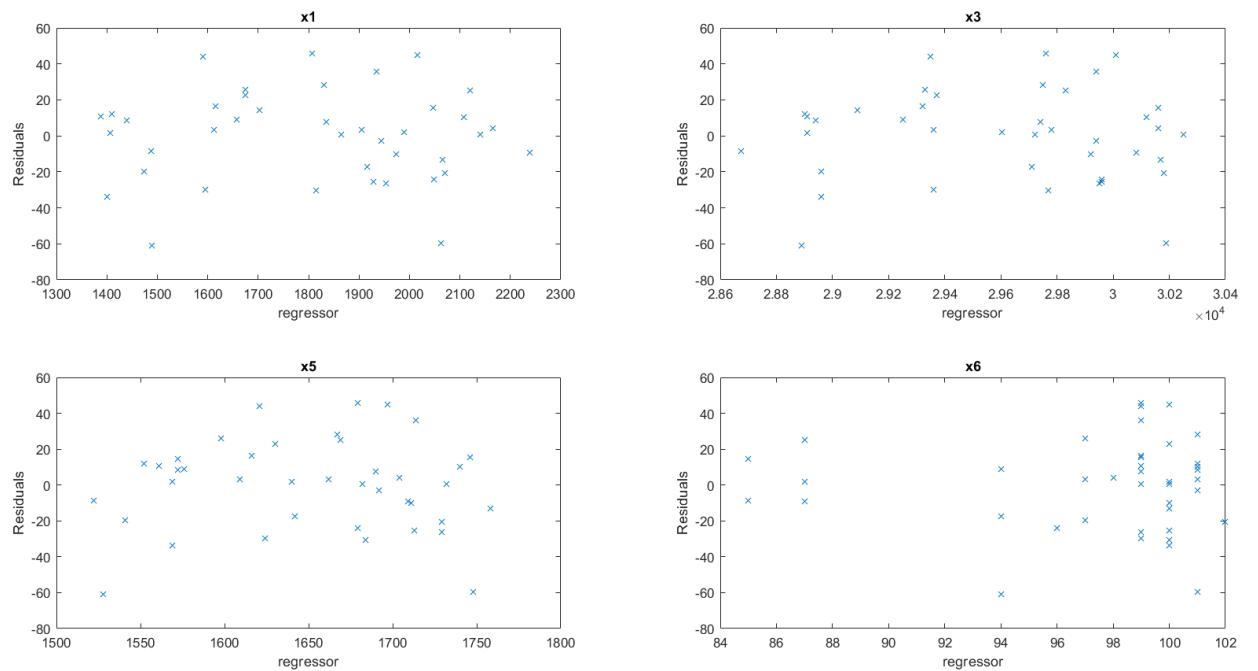
Thus the model is significant. The RMSE is close to Model-1 (26.5) but greater than it.

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).

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



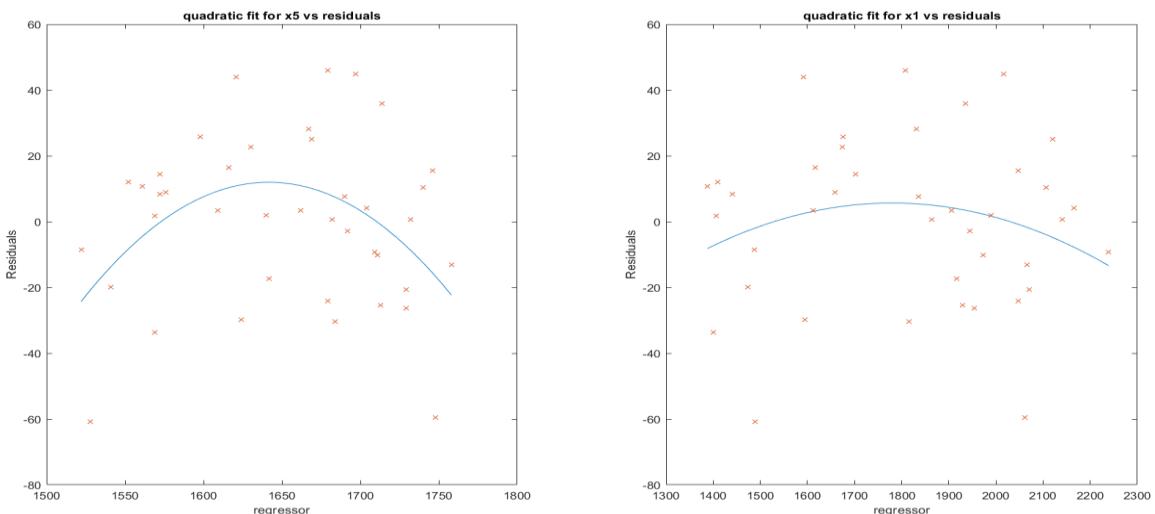
By plotting each regressor vs residual, we can check whether the residual (and hence y) has any relationship of the form $f(x_i)$ where x_i is the regressor on x-axis.

$$\begin{aligned} \epsilon &= f(x_1) + g(x_3) + h(x_5) + j(x_6) + \delta \\ \Rightarrow y &= Ax + f(x_1) + g(x_3) + h(x_5) + j(x_6) + \delta \end{aligned}$$

*Note: The residual can also be function of two or more variables simultaneously, such as x_3*x_5 or even $x_1*\sin(x_2*b/x_3)$ 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 x_3 and x_6 don't seem to have any particular relationship with residuals but x_1 and x_5 seem to have very rough quadratic relationship (inverted parabola).

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



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:

$$y \sim x1 + x3 + x5 + x6 + x1^2 + x5^2$$

Estimated Coefficients:

	Estimate	SE	tStat	pValue
<code>x1</code>	2.2915	0.34731	6.5978	1.4569e-07
<code>x3</code>	-0.027427	0.095977	-0.28576	0.77679
<code>x5</code>	1.8015	2.8232	0.63811	0.52768
<code>x6</code>	-13.159	2.1656	-6.0765	6.8533e-07
<code>x1^2</code>	-0.00014683	8.6287e-05	-1.7017	0.097942
<code>x5^2</code>	-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 ($x_i \cdot x_j$) and quadratic terms ($x_i \cdot x_i$) 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:

$$y \sim 1 + x1 + x5 + x6 + x5^2$$

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-8386	2641.1	-3.1752	0.0031184
<code>x1</code>	1.6739	0.059395	28.182	1.2021e-25
<code>x5</code>	11.812	3.2338	3.6526	0.00084167
<code>x6</code>	-14.675	1.486	-9.8756	1.1751e-11
<code>x5^2</code>	-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.

	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!

(5) a) Ridge regression:

$$\hat{\underline{\theta}}_{RR} = \arg \min_{\underline{\theta}} \|\underline{y} - \underline{\Phi} \underline{\theta}\|_2^2 + \lambda \|\underline{\theta}\|_2^2 \quad \text{--- (1)}$$

Bayesian estimation:

i) Gaussian prior: $f(\underline{\theta}) = \frac{1}{(\sqrt{2\pi} \sigma_B)^p} \exp\left(-\frac{1}{2} \frac{\sum_{i=1}^p \underline{\theta}_i^2}{\sigma_B^2}\right)$

where p is the length of $\underline{\theta}$ vector.

--- (2)

ii) $f(\underline{y}_N | \underline{\theta}) = \prod_{k=1}^N \frac{1}{\sqrt{2\pi} \sigma_e} \exp\left(-\frac{1}{2} \frac{(\underline{y}[k] - \underline{\Phi}^T[k] \underline{\theta})^2}{\sigma_e^2}\right)$

$$(\because e(k) = \underline{y}(k) - \underline{\Phi}^T(k) \underline{\theta} \quad \& \quad e(k) \sim \mathcal{N}(0, \sigma_e^2))$$

$$\Rightarrow f(\underline{y}_N | \underline{\theta}) = a \exp\left(-\frac{1}{2\sigma_e^2} \sum_{k=1}^N (\underline{y}[k] - \underline{\Phi}^T(k) \underline{\theta})^2\right) \quad \text{--- (3)}$$

where 'a' is a constant

iii) $f(\underline{\theta} | \underline{y}_N) = \frac{f(\underline{\theta}) f(\underline{y}_N | \underline{\theta})}{f(\underline{y}_N)}$

Substituting from eqns (2) & (3) and collecting the constant terms in one place.

$$f(\underline{\theta} | \underline{y}_N) = C \exp\left(-\frac{1}{2} \left[\frac{\sum_{i=1}^p \underline{\theta}_i^2}{\sigma_B^2} + \frac{\sum_{k=1}^N (\underline{y}[k] - \underline{\Phi}^T(k) \underline{\theta})^2}{\sigma_e^2} \right]\right)$$

$$= C \cdot \exp\left(-\frac{1}{2} \left(\frac{\|\underline{\theta}\|_2^2}{\sigma_B^2} + \frac{\|\underline{y}_N - \underline{\Phi} \underline{\theta}\|_2^2}{\sigma_e^2} \right)\right)$$

where C is just a constant.

$$\Rightarrow f(\underline{\theta} | \underline{y}_N) = C \exp \left(-\frac{\sigma_e^2}{2} \left\{ \left(\frac{\sigma_e^2}{\sigma_B^2} \right) \|\underline{\theta}\|_2^2 + \|\underline{y}_N - \Phi \underline{\theta}\|_2^2 \right\} \right) \quad (4)$$

Consider the MAP estimate :

$$\hat{\underline{\theta}}_{BE} = \arg \max_{\underline{\theta}} f(\underline{\theta} | \underline{y}_N)$$

From (4),

$$\begin{aligned} \hat{\underline{\theta}}_{BE} &= \arg \max_{\underline{\theta}} \exp \left(-\frac{\sigma_e^2}{2} \left\{ \left(\frac{\sigma_e^2}{\sigma_B^2} \right) \|\underline{\theta}\|_2^2 + \|\underline{y} - \Phi \underline{\theta}\|_2^2 \right\} \right) \\ &= \arg \min_{\underline{\theta}} \left(- \left\{ \frac{\sigma_e^2}{\sigma_B^2} \|\underline{\theta}\|_2^2 + \|\underline{y} - \Phi \underline{\theta}\|_2^2 \right\} \right) \end{aligned}$$

$$\Rightarrow \hat{\underline{\theta}}_{BE} = \arg \min_{\underline{\theta}} \|\underline{y} - \Phi \underline{\theta}\|_2^2 + \frac{\sigma_e^2}{\sigma_B^2} \|\underline{\theta}\|_2^2 \quad (5)$$

We find that the Bayesian estimate (eqn 5)
is identical to the Ridge Regression estimate (eqn 11)

$$\boxed{\lambda = \frac{\sigma_e^2}{\sigma_B^2}}$$

$$\begin{aligned}
 \textcircled{5} \quad b) \quad & \min_{\theta} \|y - \Phi \theta\|_2^2 + \lambda [\alpha \|\theta\|_2^2 + (1-\alpha) \|\theta\|_1] \\
 & = \min_{\theta} \|y - \Phi \theta\|_2^2 + \lambda [\alpha \|\theta\|_2^2 + (1-\alpha) \|\theta\|_1] \quad (\text{Elastic net}) \\
 & = \min_{\theta} \|y - \Phi \theta\|_2^2 + (\lambda \alpha) \|\theta\|_2^2 + \lambda (1-\alpha) \|\theta\|_1 \\
 & = \min_{\theta} \sum_{k=1}^N (y[k] - \Phi^T(k) \theta)^2 + (\lambda \alpha) \sum_{i=1}^p \theta_i^2 \\
 & \quad + \lambda (1-\alpha) \|\theta\|_1 \\
 & = \min_{\theta} \sum_{k=1}^N (y[k] - \Phi^T(k) \theta)^2 + \sum_{i=1}^p (\lambda \alpha \theta_i^2) + \lambda (1-\alpha) \|\theta\|_1 \\
 & = \min_{\theta} \sum_{k=1}^N (y[k] - \Phi^T(k) \theta)^2 + \sum_{i=1}^p (\theta_i - \lambda \alpha \theta_i)^2 \\
 & \quad + \lambda (1-\alpha) \|\theta\|_1 \\
 & = \min_{\theta} \sum_{k=1}^N (y[k] - \Phi^T(k) \theta)^2 + \sum_{i=1}^p (\theta_i - \lambda \alpha \beta_i^T \theta)^2 \\
 & \quad + \lambda (1-\alpha) \|\theta\|_1
 \end{aligned}$$

where $\beta_j^T(i) = \begin{cases} \beta_{j(i)} & \text{if } k_j = i \\ 0 & \text{otherwise} \end{cases}$

(β is i th element of $\beta^T(i)$ in 1, rest of the elements are zeroes. $\beta^T(i) \in \mathbb{R}^{p \times 1}$)

$$\begin{aligned}
 & = \min_{\theta} \sum_{k=1}^N (y[k] - \Phi^T(k) \theta)^2 + \sum_{i=1}^p (\theta_i - \lambda \alpha \beta_i^T \theta)^2 \\
 & \quad + \lambda (1-\alpha) \|\theta\|_1
 \end{aligned}$$

The second summation is of the form,

$$\sum (y[k] - \psi^T(k) \underline{\theta})^2 = 0$$

$$\text{where } y[k] = 0 \text{ & } \psi(k) = \lambda \alpha \beta(k)$$

So accordingly we augment the matrix $\underline{y} \in \underline{\Phi}$

such that they become $(N+P) \times 1$ and

$$(N+P) \times P$$

respectively

$$\underline{y}^* = \begin{bmatrix} \underline{y} \\ 0_{P \times 1} \end{bmatrix}$$

where $0_{P \times 1}$ is a column vector of zeroes of size $P \times 1$

$$\underline{\Phi}^* = \begin{bmatrix} \underline{\Phi} \\ \lambda \alpha I_{P \times P} \end{bmatrix}$$

where $I_{P \times P}$ is the identity matrix with dimensions $P \times P$.

$$(\beta^{(1)}, \dots, \beta^{(P)})]$$

is the elastic net optimization problem as

Now we can write the

$$\min_{\underline{\theta}} \sum_{k=1}^{N+P} (y^*(k) - \psi^*(k) \underline{\theta})^2 + \lambda(1-\alpha) \|\underline{\theta}\|_1$$

$$= \min_{\underline{\theta}} \|\underline{y}^* - \underline{\Phi}^* \underline{\theta}\|_2^2 + \lambda(1-\alpha) \|\underline{\theta}\|_1$$

which is a LASSO problem.

⑥ a) Jeffrey's prior, $\pi(\theta) \sim \{I(\theta)\}^{1/2}$

$$f(y[k]|\theta) = \begin{cases} \frac{\theta^{y[k]} e^{-\theta}}{(y[k])!} & \text{if } y[k] \text{ is a non-negative integer} \\ 0 & \text{otherwise} \end{cases}$$

$$L = \log \left(\prod_{k=1}^N \frac{\theta^{y[k]} e^{-\theta}}{(y[k])!} \right)$$

$$= \log (\bar{e}^{-\theta})^N + \log \theta^{\sum y[k]} + C$$

where C is a constant

$$\Rightarrow L = \sum_{k=1}^N y[k] \log \theta - N\theta + C \quad (\text{term independent of } \theta)$$

$$S = \frac{\partial L}{\partial \theta} = \frac{\sum y[k]}{\theta} - N$$

$$\Rightarrow \frac{\partial^2 L}{\partial \theta^2} = -\frac{\sum y[k]}{\theta^2}$$

$$E\left(-\frac{\partial^2 L}{\partial \theta^2}\right) = \frac{\sum E(y[k])}{\theta^2} = \frac{N\theta}{\theta^2} \quad (\because E(y[k]) = \theta)$$

$$\Rightarrow I(\theta) = \frac{1}{\theta}$$

$$\therefore \text{Jeffrey's prior, } \pi(\theta) \sim \left(\frac{1}{\theta}\right)^{1/2}$$

$$\sim \boxed{\theta^{-1/2}}$$

So a Dabaja's prior is in the class of
Jeffrey's priors.

$$b) f(\theta | y_N) = \frac{(f(y_N | \theta)) f(\theta)}{f(y_N)}$$

$$= C \times \theta^{\sum_{k=1}^N y[k]} \frac{e^{-\theta}}{\prod_{k=1}^N (y[k])!} \times \theta^{-1/2}$$

$$= C \times \theta^{\sum_{k=1}^N y[k]} e^{-N\theta} \times \theta^{-1/2}$$

$$\frac{1}{\prod_{k=1}^N (y[k])!}$$

for given y_N , $\prod_{k=1}^N y[k]!$ is a constant

$$\Rightarrow f(\theta | y_N) = C' \times \theta^{\sum_{k=1}^N y[k]} \times e^{-N\theta} \times \theta^{-1/2}$$

Jeffrey's prior doesn't change upon transforming

$$\text{the RV } \Rightarrow f(2N\theta) \sim f(\theta) \sim \theta^{-1/2}$$

$$f(2N\theta | y_N) = \frac{f(y_N | 2N\theta) f(2N\theta)}{f(y_N)}$$

$$= C' \times \left(\frac{2N\theta}{2N}\right)^{\sum_{k=1}^N y[k]} e^{-\frac{2N\theta}{2}} \times \left(\frac{2N\theta}{2N}\right)^{-1/2}$$

$$\text{Let } x = 2N\theta, \quad f(x|y_N) = c'' x^{\frac{(2\sum y_k)}{2}} e^{-x/2} \sim \chi_{1/2}^{1/2}$$

$$\bar{y} = \frac{\sum y_k}{N} \Rightarrow \sum y_k = N\bar{y}$$

$$\Rightarrow f(x|y_N) = c'' x^{\left(\frac{2N\bar{y}+1-1}{2}\right)} e^{-x/2} \sim \chi_{1/2}^{1/2}$$

$$= c'' x^{\left(\frac{2N\bar{y}+1}{2}-1\right)} e^{-x/2}$$

This is of the χ^2 distribution form.

$$\text{dof} = (2N\bar{y} + 1). \text{ So } c'' = \frac{2^{-\frac{(2N\bar{y}+1)}{2}}}{\Gamma\left(\frac{2N\bar{y}+1}{2}\right)}$$

$$\Rightarrow f(x|y_N) = \frac{1}{\frac{(2N\bar{y}+1)}{2} \Gamma\left(\frac{2N\bar{y}+1}{2}\right)} x^{\left(\frac{2N\bar{y}+1}{2}-1\right)} e^{-x/2}$$

$$= \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad \text{where } k = \boxed{2N\bar{y}+1}$$

Mean of this distribution, $E(x) = E(2N\theta)$ is the def k.

$$\Rightarrow E(2N\theta) = (2N\bar{y} + 1)$$

$$\therefore \boxed{f(2N\theta|y_N) \sim \chi^2(2N\bar{y} + 1)}$$

c)

$$\hat{\theta}_{\text{MMSE}} = \min_{\theta} E((\theta - \hat{\theta})^2)$$

$$= \min_{\theta} E((2N\theta - 2N\hat{\theta})^2) + \frac{1}{(2N)^2}$$

$$\Rightarrow \hat{\theta}_{\text{MMSE}} = \min_{\theta} (E(2N\theta - 2N\hat{\theta})^2) + \frac{1}{(2N)^2}$$

$$\hat{\theta}_{\text{MMSE}} = \frac{1}{2N} \times \min_{\theta} \left(\frac{E((x - 2N\hat{\theta})^2)}{(2N)^2} \right)$$

$\because 2N$ is a constant for given data

$$\hat{x} = \frac{\hat{\theta}_{\text{MMSE}}}{2N} \Rightarrow \hat{\theta}_{\text{MMSE}} = \frac{\hat{x}_{\text{MMSE}}}{2N}$$

\hat{x}_{MMSE} is the mean of the posterior $\chi^2(k)$

$$\hat{x}_{\text{MMSE}} = 2N\bar{y} + 1$$

$$\therefore \boxed{\hat{\theta}_{\text{MMSE}} = \bar{y} + \frac{1}{2N}}$$

- d) Let a be the lower end and b be the higher end of the credible interval. Let F be the CDF of X^2
- If a symmetric credible interval over the MMSE estimate is required:
- $$F(b) - F(a) = 1 - \alpha \quad \text{--- ①}$$

$$\frac{a+b}{2} = \bar{x} (2N\bar{y} + 1) \quad \textcircled{2}$$

Solving eqns $\textcircled{1}$ & $\textcircled{2}$ for a and b gives the symmetric credible interval around mean estimate of $2N\theta$

ii) Unsymmetric credible interval (when $\hat{\theta}$ is not median est.)

- This interval will be unsymmetric when the MMSE estimate is used
- If Median estimate is used the interval is symmetric.

→ Solve the eqns $\textcircled{3}$ & $\textcircled{4}$

$$F(a) = 0.025 \quad \textcircled{3} \quad F(b) = 0.975 \quad \textcircled{4}$$

To get a & b which will represent the end points of the interval for $2N\theta$ estimate

$$\text{CDF}, F(u=2N\theta) = \frac{\gamma\left(\frac{k}{2}, \frac{u}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}, k = 2N\bar{y} + 1$$

$\gamma(\cdot)$ → lower incomplete gamma fn.; Γ → upper incomplete gamma function

∴ Interval for θ is $\left[\frac{a}{2N}, \frac{b}{2N} \right]$

where a & b are obtained by solving $\textcircled{1}$ & $\textcircled{2}$ or $\textcircled{3}$ & $\textcircled{4}$ as needed.