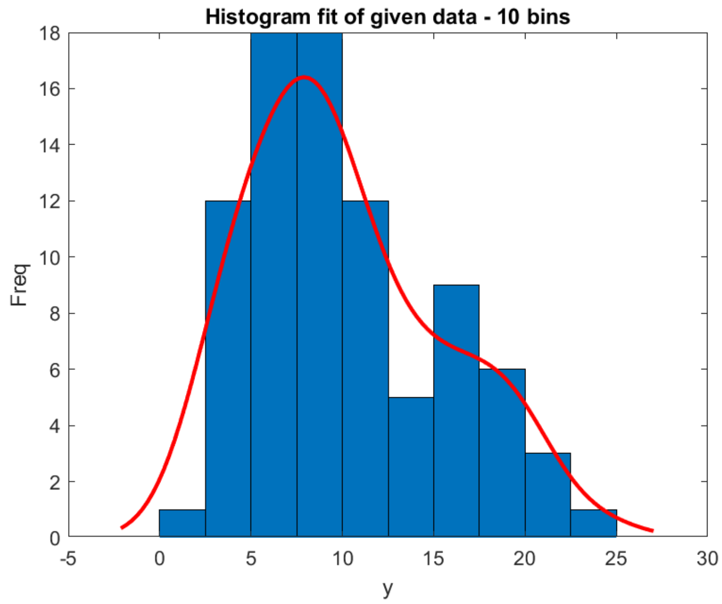
CH5115 – Assignment 4   G PRASHANT (BS17B011)

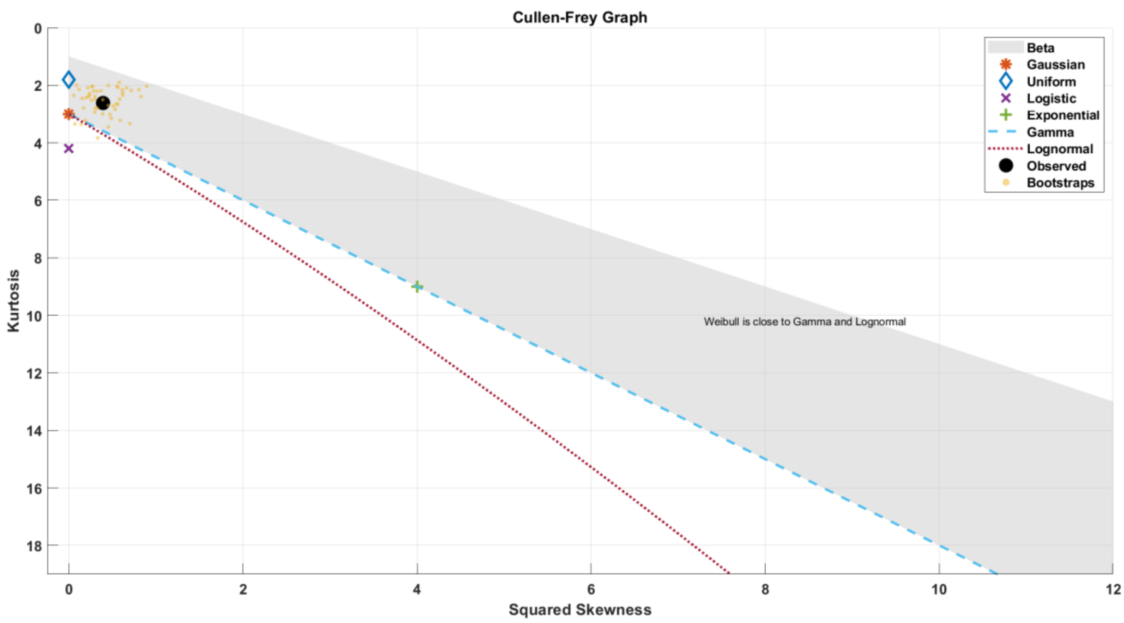
Question 1

Part (a)

The histogram fit of the data is shown below. It can be observed that the histogram does not exactly appear to take the form of a Gaussian distribution.



The Cullen-Frey Graph showing the regions/curves/points of multiple distributions (Gaussian, Uniform, Logistic, Exponential, Gamma and Lognormal) along with the empirical estimates (observed and bootstrapped) is given below:



From the plot above, it can be observed that the Kurtosis vs Squared Skewness point of the observed data falls in the region of Beta distribution. However, the values in the data do not fall in the support region of beta distribution (0 to 1).

To further narrow down upon the distribution, we perform AD test using the distributions Normal, Lognormal, exponential and Weibull. The significance values are shown below:

Distributions	**p-value**	**H_0 (alpha = 0.05)**
Normal	0.0011	1
Exponential	5e-4	1
Lognormal	0.1846	0
Weibull	0.2092	0

This gives a clue that the distribution can be either of Lognormal or Weibull. We hence fit both distributions by finding parameters using MLE (using fitdist).

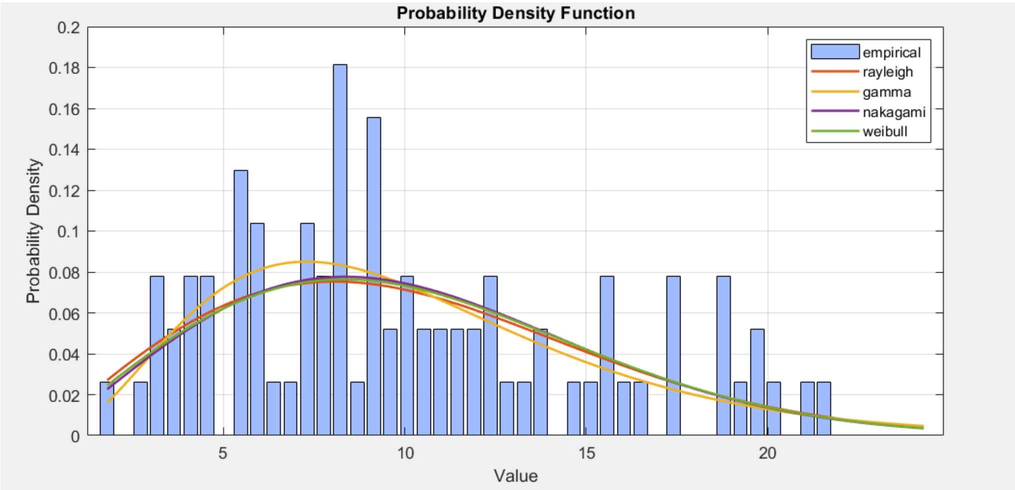
- The negative log likelihood of Weibull fit is 255.2957
- The negative log likelihood of Lognormal fit is 256.5944

Since the Weibull distribution fit has a stronger significance and the MLE negative loglikelihood is lower, fitting a Weibull distribution is more appropriate. The MLE estimates of the parameters of the Weibull distribution are:

$$A = 11.4721$$

$$B = 2.06503$$

To further validate the claim, we make use of the FitDistributionGUI to obtain the following results



	Distribution 1	Distribution 2	Distribution 3	Distribution 4
Dist Name	rayleigh	gamma	nakagami	weibull
NLogL	255.367	254.7378	255.2024	255.2957
BIC	515.1767	518.3608	519.29	519.4768
AIC	512.734	513.4755	514.4047	514.5915
AICc	512.7822	513.6219	514.5511	514.7378
Params Nemas	B,	a, b,	mu, omega,	A, B,
Params Values	8.0573	3.5767 2.8311	1.08208 129.8399	11.4721 2.06503

It can be observed that Rayleigh distribution fit has the lowest AIC and BIC scores, as well as reasonably low negative log likelihood value. Furthermore, this distribution has only 1 parameter to estimate, thereby making it a simpler model compared to all others.

Henceforth, after all the analysis, it is justified to choose Rayleigh as the appropriate distribution for the data.

Part (b)

The two types of errors, namely, systematic and random errors are explained and their impacts are mentioned below:

- Systematic Errors: Systematic errors occur due to inaccuracies that are prevalent in the system, thereby causing a statistical bias. Sources of systematic error includes imperfect calibration of measurement device and incorrect usage of instrument. This type of error affects the results of the analysis, however, in a predictable direction. In other words, in most cases it is easy to determine the source of the error and make appropriate corrections because it is characterized by a consistent departure from the truth. Example: Batch effects in gene expression data. Systematic error, as mentioned before, introduces a bias (eg. a mean shift) in the distribution of the observed data when compared to that of error-free data.

- Random Errors: Random errors, unlike systematic errors occur due to unpredictable fluctuations while performing an experiment. They are unavoidable and it is never possible to accurately characterize the source of random errors. This leads to uncertainty in measurements. To resolve this issue, multiple observations are usually considered. Since the mean of the random error is mostly 0, this does not induce a mean shift to the distribution of the observed data. However, the variance of the error will be nonzero, and this induces a change in variance of the distribution of observed data compared to that of error-free data.

Question 2

Given pdf

$$f(y;\theta) = \frac{2y}{\theta^2} \quad , \quad 0 \le y \le \theta$$

Part (a)

For N observations

$$f(y_N;\theta) = \frac{2^N}{\theta^{2N}} \prod_{K=0}^{N-1} y[K] = l(\theta; y_N)$$

↳ likelihood

We know,

$$y[K] \le \theta$$

$$\prod_{K=0}^{N-1} y[K] \le \theta^N$$

$$\frac{2^N}{\theta^{2N}} \prod_{K=0}^{N-1} y[K] \le \frac{2^N}{\theta^{N}}$$

$$\Rightarrow \quad l(\theta; y_N) \le \left(\frac{2}{\theta}\right)^N$$

∴ The likelihood takes a maximum value of $\left(\frac{2}{\theta}\right)^N$ when $\theta = \left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}$

$$\Rightarrow \quad \hat{\theta}_{ML} = \left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}$$

$$E(\hat{\theta}_{ML}) = E\left(\left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}\right)$$

$$= E\left(\prod_{K=0}^{N-1} (y[K])^{1/N}\right)$$

As they are i.i.d

$$E(\hat{\theta}_{ML}) = \prod_{K=0}^{N-1} E(y[K]^{1/N})$$

$$E(y[K]^{1/N}) = \int_0^{\theta} y[K]^{1/N} f(y[K];\theta)\, dy[K]$$

$$E(y[K]^{1/N}) = \int_0^{\theta} \frac{2}{\theta^2} y[K]^{1/N} . y[K]\, dy[K]$$

$$= \frac{2}{\theta^2}\int_0^{\theta} (y[K])^{1/N+1}\, dy[K]$$

$$= \frac{2}{\theta^2} \left. \frac{y[K]^{1/N+2}}{\frac{1}{N}+2} \right|_0^{\theta}$$

$$= \frac{2}{\theta^2} \frac{\theta^{1/N} . \theta^2}{\frac{1}{N}+2} = \frac{2N}{1+2N}\theta^{1/N}$$

$$\Rightarrow E(\hat{\theta}_{ML}) = \left(E(y[K]^{1/N})\right)^N$$

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

$$\Delta\theta : E(\hat{\theta}_{ML}) - \theta = \cancel{\frac{\theta}{1+1/2N}} \neq 0$$

Hence, the estimator is biased. To correct the bias term, we multiply $\left(\frac{2N+1}{2N}\right)^N$

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

$$\hat{\theta}^*_{ML} = \left(\frac{2N+1}{2N}\right)^N \left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}$$

Part (b)

cdf

$$F(y;\theta) = \int_0^y f(y;\theta)\,dy = \frac{2}{\theta^2}\int_0^y y\,dy$$

$$= \frac{y^2}{\theta^2}, \quad y \leq \theta$$

For median, $F(y;\theta) = 0.5$

$$\frac{y^2}{\theta^2} = \frac{1}{2}$$

$$\tilde{y}_m = \frac{\theta}{\sqrt{2}}$$

$\therefore$ ~~Medi~~ Theoretical median $\tilde{y} = \frac{\theta}{\sqrt{2}}$

ML estimate of median

$$\hat{\tilde{y}}_{ML} = \frac{\hat{\theta}^*_{ML}}{\sqrt{2}} = \frac{\left(\frac{2N+1}{2N}\right)^N \left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}}{\sqrt{2}}$$

Scanned with CamScanner

W

Part (c)

We examine the mean square consistency

$$E\left((\tilde{y}_{ML} - \tilde{y})^2\right) \cancel{= E\left(\left(\frac{2N+1}{2\sqrt{2}N} \prod_{K=0}^{N-1} y[K]\right.\right.}$$

$$= E(\tilde{y}_{ML}^2 + \tilde{y}^2 - 2\tilde{y}_{ML}\tilde{y})$$

$$= E(\tilde{y}_{ML}^2) + \tilde{y}^2 - 2\tilde{y}\,E(\tilde{y}_{ML})$$

$$E(\tilde{y}_{ML}) = \tilde{y} \qquad \text{(unbiased)}$$

$$\Rightarrow MSE = E((\tilde{y}_{ML} - \tilde{y})^2) = E(\tilde{y}_{ML}^2) - \tilde{y}^2$$

Scanned with CamScanner

$$= E\left(\left(\frac{2N+1}{2N}\right)^{2N} \left(\frac{1}{\sqrt{2}}\right)^{2} \prod_{K=0}^{N-1} y[K]^{2/N} \right) - \tilde{y}^{2}$$

$$= \left(\frac{1}{2}\right) \left(1 + \frac{1}{2N}\right)^{2N} E\left(\prod_{K=0}^{N-1} y[K]^{2/N} \right) - \tilde{y}^{2}$$

$$E\left(y[K]^{2/N} \right) = \frac{2}{\theta^{2}} \cdot \frac{\theta^{2/N} \cdot \theta^{2}}{\frac{2}{N} + 2}$$

$$= \frac{N \, \theta^{2/N}}{N+1}$$

$$\Rightarrow E\left(\prod_{K=0}^{N-1} y[K]^{2/N} \right) = \theta^{2} \left(\frac{N}{N+1}\right)^{N}$$

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

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

$$\Rightarrow \lim_{N\to\infty} MSE = \frac{1}{2}\lim_{N\to\infty}\left(1+\frac{1}{2N}\right)^{2N} \cdot \lim_{N\to\infty}\left(1-\frac{1}{N+1}\right)^{N}\theta^2 - \tilde{y}^2$$

$$= \frac{1}{2}(e)\left(\frac{1}{e}\right)\theta^2 - \tilde{y}^2$$

$$= \frac{\theta^2}{2} - \tilde{y}^2$$

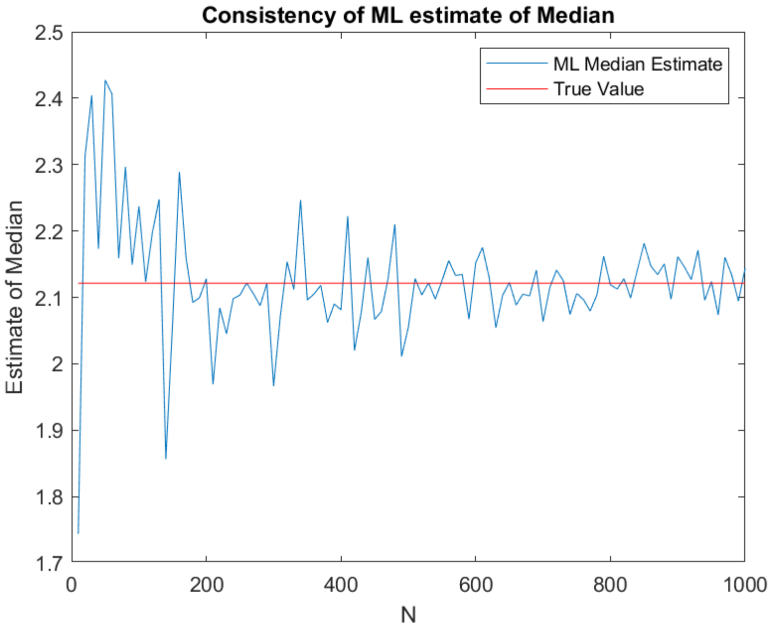
$$= \tilde{y}^2 - \tilde{y}^2$$

$$= 0$$

$\therefore$ As MSE tends to 0 ~~as~~ when $N\to\infty$, $\tilde{y}_{ML} = \frac{\hat{\theta}^*_{ML}}{\sqrt{2}}$ is a consistent estimator of median.

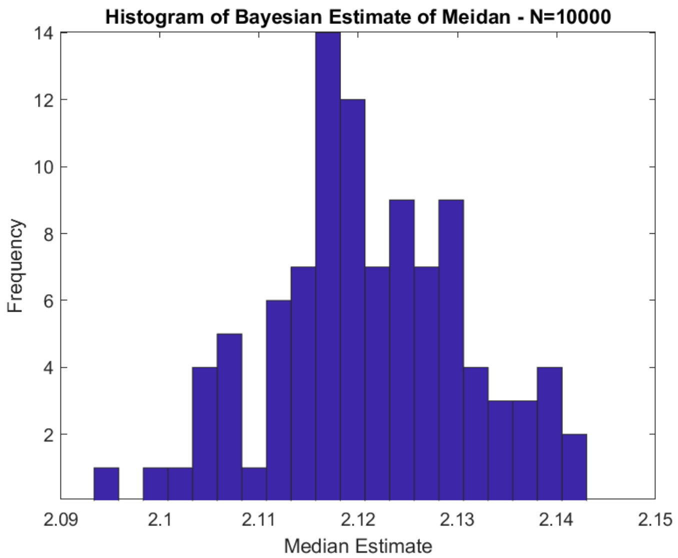
Question 2

Part (d)



From the above plot, it can be observed that the maximum likelihood estimate of the median of the distribution is consistent, since it converges to the true value as the number of observations, N increases. This agrees with what was theoretically proved in the previous part.

Further, to examine the consistency, 10000 observations of Bayesian Median Estimate was generated and a histogram was plotted



Although the above plot slightly resembles Gaussian, it cannot be said that the Bayesian median estimates are asymptotically Gaussian. This is because neither the random variable y[k] nor the median estimate takes negative values. Moreover, the median estimator is a function of the Geometric Mean of the sample, which unlike arithmetic mean, does not converge to Gaussian distribution for large N.

Part 1(d) (contd)

$$\hat{y}_{ML} = \frac{\left(1+\frac{1}{2N}\right)^N}{\sqrt{2}} \left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}$$

The asymptotic distribution of $\hat{y}_{ML}$ depends on the asymptotic distribution of the geometric mean (sample), i.e. $\left(\prod_{K=0}^{N-1} y[K]\right)^{1/N}$, which is not necessarily Gaussian, unlike sample mean.

Question 3

Part (a)

The correlation between the response and each of the regressors is given below:

$$corr(\psi_1, y) = 0.9197$$

$$corr(\psi_2, y) = 0.8755$$

$$corr(\psi_3, y) = 0.3998$$

The above values indicate that the regressors ψ_1 and ψ_2 are highly linearly correlated with the response variable, and the regressor ψ_3 is weakly linearly correlated. Overall, a linear model well qualified between y and the regressors, especially ψ_1 and ψ_2.\

Part (b)

The following linear model was fit

$$y = c_0 + c_1\psi_1 + c_2\psi_2 + c_3\psi_3 + \epsilon$$

"fitlm" implementation in MATLAB was used for linear regression, and the estimates of the regression parameters are given below:

$$c_0 = -39.92$$

$$c_1 = 0.71564$$

$$c_2 = 1.2953$$

$$c_3 = -0.15212$$

Metrics for assessing goodness of fit

$$R^2 = 0.914$$

Adjusted $R^2 = 0.898$

p-value = $3.02 * 10^{-9}$

As the p-value<0.05, the linear model is statistically significant

Part (c)

We perform statistical significance tests for each coefficient with alpha = 0.05, and the results are tabulated below:

Coefficient	t_{stat}	P-value	Significant (alpha=0.05)
c_0	-3.3557	0.00375	Yes
c_1	5.3066	5.799e-05	Yes
c_2	3.5196	0.00263	Yes
c_3	-0.97331	0.344	No

Part (d)

The 95% confidence interval of the conditioned mean stack loss was calculated using the expression below:

$$\mu_{Y|\psi} = \hat{y} \pm t_{\frac{\alpha}{2},(n-4)} s_e \sqrt{\psi_0 (\psi^T \psi)^{-1} \psi_0^T}$$

Where,

$$\psi_0 = [1\ \psi_1\ \psi_2\ \psi_3\ ]^T = [1\ 80\ 25\ 90\ ]^T$$

$$\hat{y} = \psi_0 C = 36.0227$$

$$s_e = \sqrt{sum\left((y - \hat{y}_\psi)^2\right)/(N-4)} = 3.2434$$

$$\alpha = 0.05$$

Hence, we get the bounds of the mean stack loss as

$$lb = 32.2188$$

$$ub = 39.8265$$

Therefore, the 95% Confidence Interval for the mean stack loss is

$$32.2188 < \mu_{Y|\psi} < 39.8265$$

Part (e)

The 95% prediction interval of the conditioned stack loss was calculated using the expression below:

$$y(\psi_0) = \hat{y}(\psi_0) \pm t_{\frac{\alpha}{2},(n-4)} s_e \sqrt{1 + \psi_0 (\psi^T \psi)^{-1} \psi_0^T}$$

Where,

$$\psi_0 = [1\ \psi_1\ \psi_2\ \psi_3\ ]^T = [1\ 80\ 25\ 90\ ]^T$$

$$\hat{y} = \psi_0 C = 36.0227$$

$$s_e = \sqrt{sum\left(\left(y - \hat{y}_{\psi}\right)^2\right)/(N-4)} = 3.2434$$

$$\alpha = 0.05$$

Hence, we get the bounds of the mean stack loss as

$$lb = 28.1936$$

$$ub = 43.8518$$

Therefore, the 95% prediction Interval for the stack loss is

$$28.1936 < y(\psi_0) < 43.8518$$

Question 4

Part (a)

We are given regressors of 6 dimensions. We fit the linear regression model using all the six regressor variables.

$$y = c_0 + c_1\psi_1 + c_2\psi_2 + c_3\psi_3 + c_4\psi_4 + c_5\psi_5 + c_6\psi_6 + \epsilon$$

The following estimates are obtained

Coefficient	Estimate	t_{stat}	P-value	Significant (alpha=0.05)
c_0	-4738	-1.938	0.061213	No
c_1	1.1185	3.9045	0.00044089	Yes
c_2	-0.030184	-0.78946	0.43548	No
c_3	0.23062	1.9539	0.059231	No
c_4	3.8495	1.4331	0.16125	No
c_5	0.82186	2.3432	0.025298	Yes
c_6	-16.946	-6.4679	2.4504e-07	Yes

<u>Model Diagnostics:</u>

$$R^2 = 0.998$$

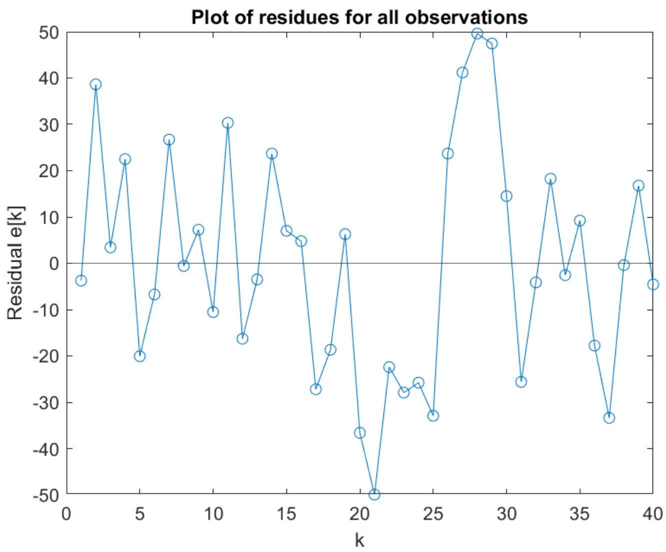
Adjusted $R^2 = 0.997$

F-test p-value = $6.07 * 10^{-42}$

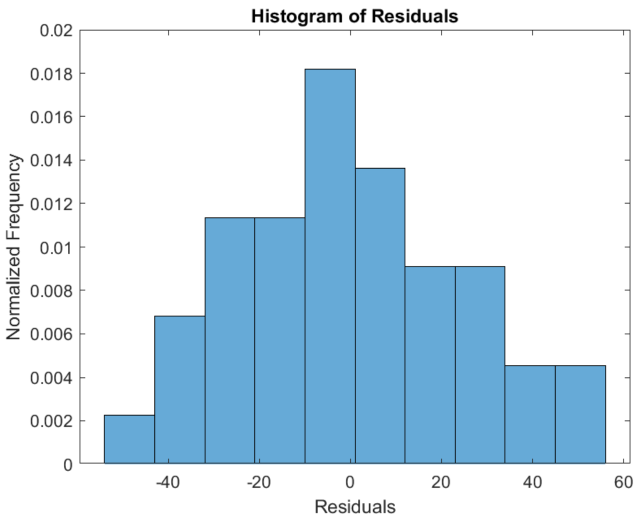
<u>Residual Analysis:</u>

The residual order plot and the histogram of residuals are shown below:

The errors (residuals) $\epsilon[k], k = 1, \ldots, N$ are i.i.d



The above plot suggests that the residuals do not follow a trend, and are completely random.

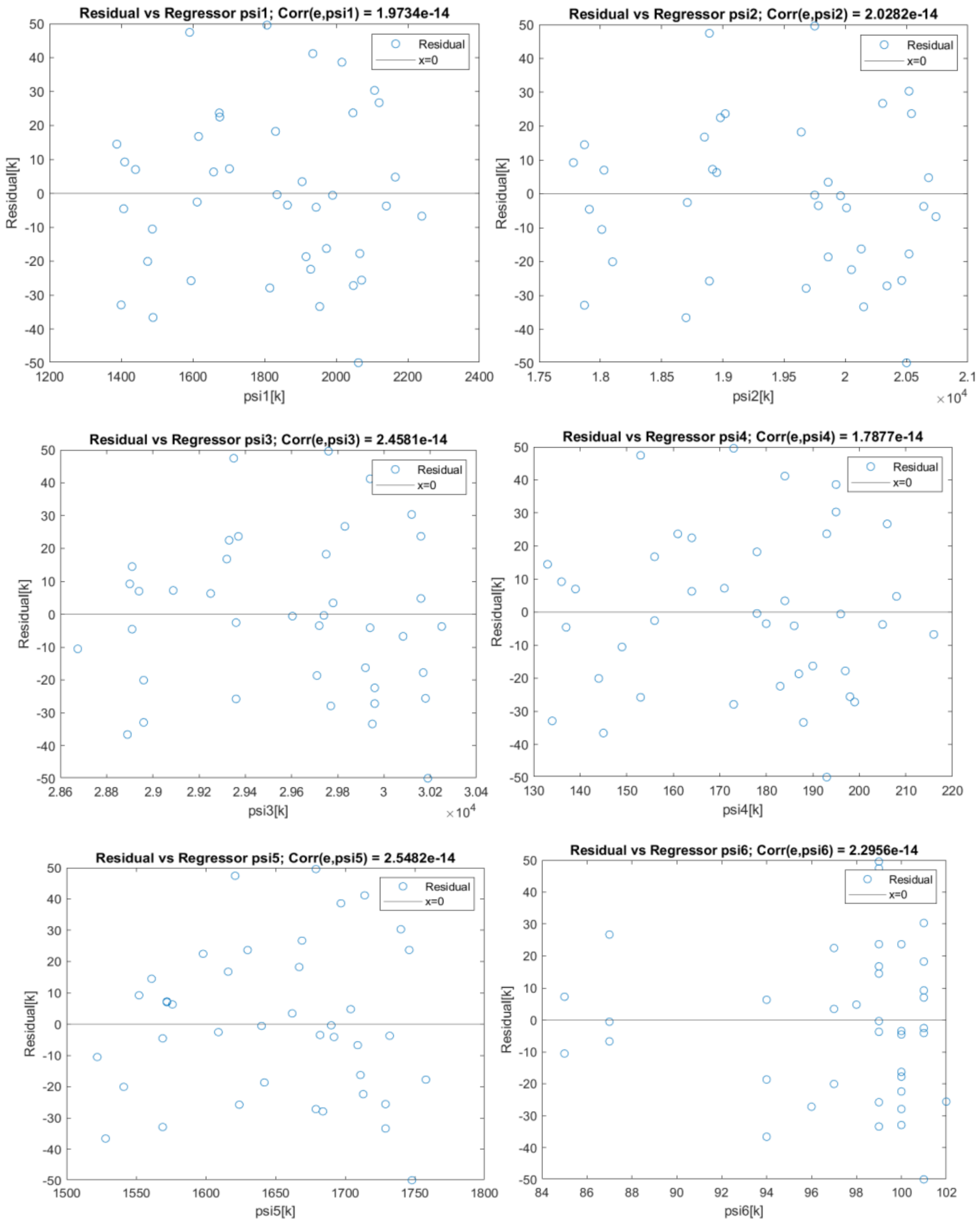


We perform KS test and fail to reject the Null hypotheses that the errors come from Gaussian Distribution. Further, we perform Ljung Box Test to again fail to reject the null hypothesis that the residuals are white. This indicates that the errors are i.i.d.

By fitting a Gaussian distribution to the conditioned residuals, we obtain the following residual parameter estimates, where the mean is almost zero, and the variance is finite. Hence, the condition of normality and non serial correlation of residuals is also satisfied.

After fitting a normal distribution to the residuals, the following mean and sigma estimates were obtained, and the mean is very close to 0.

```
Normal distribution
     mu = 2.55795e-12   [-7.79856, 7.79856]
  sigma =     24.3845   [19.9749, 31.3106]
```



Finally, we have also shown that the residual plots of all regressors indicate random deviation from the x axis, with very low correlation between regressors and residuals. Therefore, all the assumptions are satisfied.

Part (b)

From the table in part (a), we notice that the insignificant coefficients are c_0, c_2, c_3 and c_4. We therefore reperform the linear regression after removing the regressors 2, 3, 4 and the intercept.

Coefficient	Estimate	t_{stat}	P-value	Significant (alpha=0.05)
c_1	1.6101	37.388	5.3825e-31	Yes
c_5	1.5025	11.308	1.4513e-13	Yes
c_6	-15.318	-9.8936	6.1356e-12	Yes

$$R^2 = 0.9969$$

Adjusted $R^2 = 0.9967$

In this case, we can notice that all the coefficients are significant, and hence is a better model than the previous.

Part (c)

The model obtained in 4(b) is better than that obtained in Part 4(a), as the latter had coefficients that were not statistically significant, while all coefficients of the former were significant (having p-value less than 0.05).

Stepwise linear regression was performed using the stepwiselm function in matlab. This resulted in estimating the coefficients of the following model.

$$y = c_0 + c_1\psi_1 + c_5\psi_5 + c_6\psi_6 + c_{15}\psi_1\psi_5 + \epsilon$$

The following estimates were obtained.

Coefficient	Estimate	t_{stat}	P-value	Significant (alpha=0.05)
c_0	-2420.2	-2.5765	0.014357	Yes
c_1	3.3026	6.2105	4.0813e-07	Yes
c_5	2.7787	4.6174	5.0663e-05	Yes
c_6	-13.693	-9.0955	9.5355e-11	Yes
c_{15}	-0.00096816	-3.022	0.0046725	Yes

$$R^2 = 0.998$$

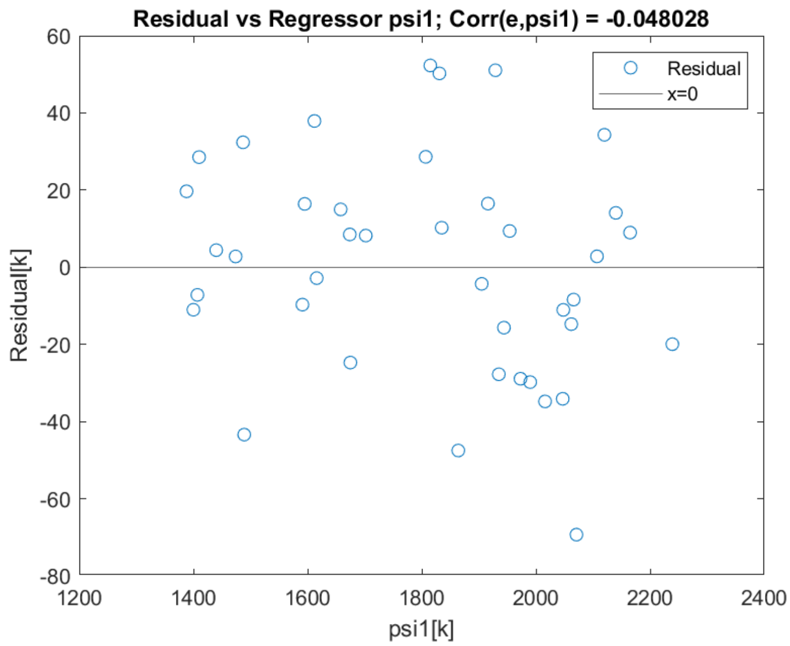
Adjusted $R^2 = 0.998$

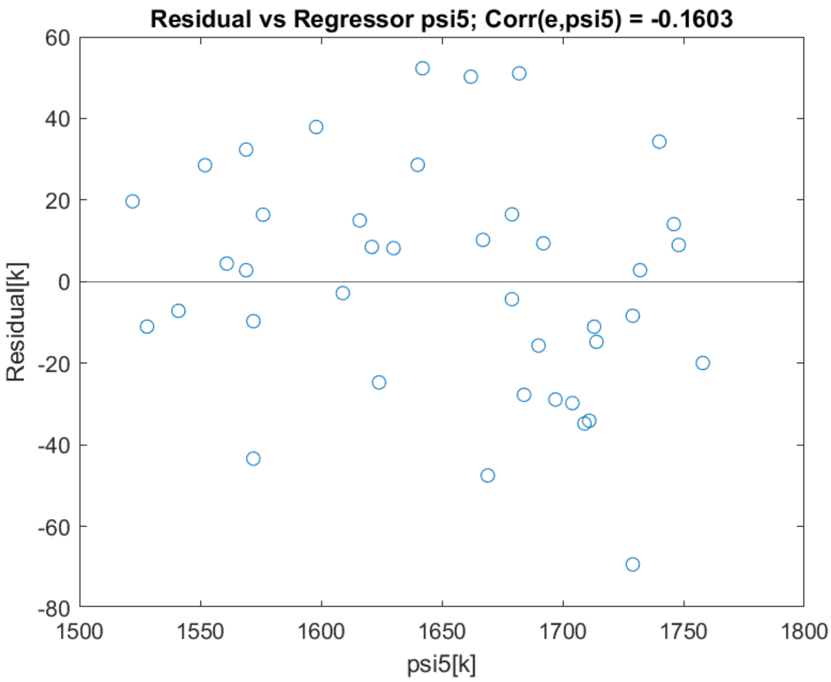
Ftest p-value = $7.82 * 10^{-46}$

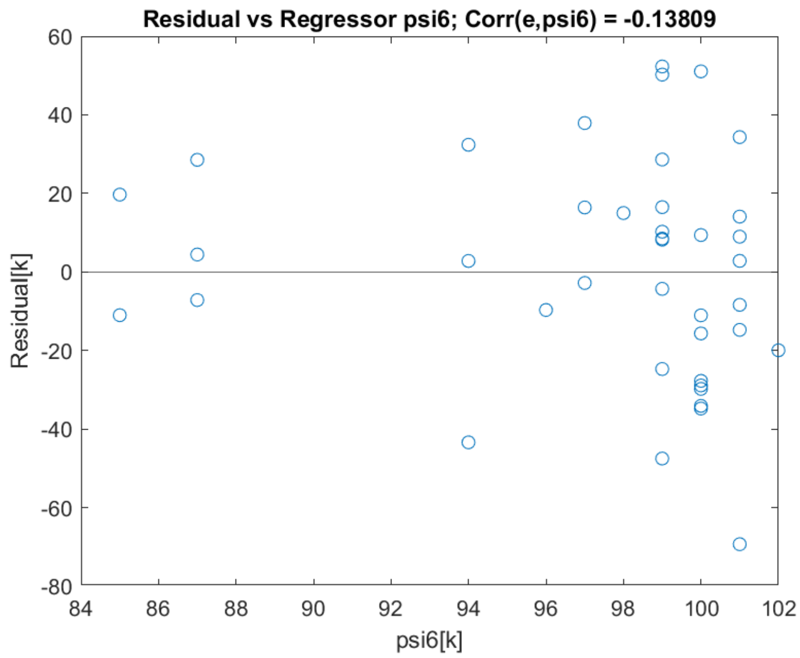
It can be observed that the model returned by the stepwise linear regression consists of only the regressors 1, 5 and 6, in addition to the intercept. This very well agrees with the best-chosen model in 4(b), which also consists of only these regressors.

Part (d)

The residual plots with the regressors ψ_1, ψ_5 and ψ_6 are shown below:


Residual vs Regressor psi1; Corr(e,psi1) = -0.048028
Residual
x=0
Residual[k]
psi1[k]
60
40
20
0
-20
-40
-60
-80
1200
1400
1600
1800
2000
2200
2400


Residual vs Regressor psi5; Corr(e,psi5) = -0.1603
Residual[k]
psi5[k]
60
40
20
0
-20
-40
-60
-80
1500
1550
1600
1650
1700
1750
1800



As observed, in all the plots, the points are equally scattered/deviated above and below the x axis, indicating no observable relationship (or trend) between the residuals and each of the three regressors.

Part (e)

In part 4(d), we have seen that there was no observable trend between the residuals and each of significant regressors ψ_1, ψ_5 and ψ_6. This suggests that the linear model is already well suited for this data.

However, the model obtained in part 4(c), after stepwise regression, which contains a nonlinear term $\psi_5\,\psi_6$ seems to give better (hence lower) Ftest p-value for the significance of regression.

Question 5

Part (a)

$$y[k] = \psi^T[k]\theta + e[k]$$

$$e[k] \sim N(0, \sigma_e^2)$$

$$\theta \sim N(0, \sigma_\beta^2)$$

$$\Rightarrow \quad f(y[k]/\theta) = \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left(-\frac{(y[k]-\psi^T[k]\theta)^2}{2\sigma_e^2}\right)$$

For N observations

$$f(y_N|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma_e}\right)^N \exp\left(-\sum_{k=0}^{N-1} \frac{(y[k]-\psi^T[k]\theta)^2}{2\sigma_e^2}\right)$$

$$f_\theta(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\beta^2} \exp\left(\frac{-\theta^T\theta}{2\sigma_\beta^2}\right)$$

Posterior

$$f(\theta/y_N) = \frac{f(y_N|\theta)\, f(\theta)}{f(y_N)} = C\, f(y_N|\theta)\cdot f(\theta)$$

where $C \to$ constant independent of θ

$$\hat{\theta}^*_{MAP} = \arg\max_\theta f(\theta/y_N) = \arg\max_\theta f(y/\theta)\, f(\theta).$$

$$= \arg\max_\theta \underbrace{\left(\ln f(y_N|\theta) + \ln(f(\theta))\right)}_{J}$$

$$J = \ln(f_Y(y_N|\theta)) + \ln(f_\theta(\theta))$$

$$\Rightarrow J = N\ln\left(\frac{1}{\sqrt{2\pi}\sigma_e}\right) - \sum_{K=0}^{N-1}\frac{(y[K] - \psi^T[K]\theta)^2}{2\sigma_e^2}$$

$$+ \ln\left(\frac{1}{\sqrt{2\pi}\sigma_\beta}\right) - \frac{\theta^T\theta}{2\sigma_\beta^2}$$

We remove terms /constants independent of θ

$$\Rightarrow J_m = -\sum_{K=0}^{N-1}\frac{(y[K] - \psi^T[K]\theta)^2}{2\sigma_e^2} - \frac{\theta^T\theta}{2\sigma_\beta^2}$$

$$\hat{\theta}^*_{MAP} = \underset{\theta}{\operatorname{argmax}}\,(J_m) = \underset{\theta}{\operatorname{argmin}}\,(-J_m)$$

$$= \underset{\theta}{\operatorname{argmin}}\left(\sum_{K=0}^{N-1}\frac{(y[K] - \psi^T[K]\theta)^2}{2\sigma_e^2} + \frac{\theta^T\theta}{2\sigma_\beta^2}\right)$$

or

$$\hat{\theta}^*_{MAP} = \underset{\theta}{\operatorname{argmin}}\Big(\underbrace{\sum_{K=0}^{N-1}(y[K] - \psi^T[K]\theta)^2 + \lambda\,\theta^T\theta}_{J^*}\Big)$$

where $\lambda = \dfrac{\sigma_e^2}{\sigma_\beta^2}$

The final cost function

$$J^* = \sum_{K=0}^{N-1}(y[K] - \psi^T[K]\theta)^2 + \lambda\,\|\theta\|_2^2$$

which is the same as ridge regression (Tikhonov regularization).

Here $\lambda = \dfrac{\sigma_e^2}{\sigma_\beta^2}$

part (b)

The cost function for naive elastic net is given by

$$J = \|y - \Phi\theta\|_2^2 + \lambda(\alpha\|\theta\|_2^2 + (1-\alpha)\|\theta\|_1)$$

We augment the ~~a~~ number of training samples by adding p (number of features) additional points.

To cast the elastic net regression into LASSO, we have enforce that the ridge term in the cost function is zero, or in other words, should be absorbed in the least squares term.

We perform the following transformation

$$\Phi_{n\times p},\ y_n \longrightarrow \Phi^*_{(n+p)\times p},\ y^*_{(n+p)}$$

such that

$$\Phi^* = \frac{1}{\sqrt{1+\alpha\lambda}}\begin{bmatrix}\Phi \\ \sqrt{\alpha\lambda}\, I_p\end{bmatrix},\quad y^* = \begin{bmatrix} y_{n\times 1} \\ 0_{p\times 1}\end{bmatrix}$$

And $\theta^* = \sqrt{1+\alpha\lambda}\ \theta$

$$J = \| y - \Phi\theta \|_2^2 + \lambda(\alpha \|\theta\|_2^2 + (1-\alpha)\|\theta\|_1)$$

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

let $\gamma = \frac{\lambda(1-\alpha)}{\sqrt{1+\alpha\lambda}}$.

$\Rightarrow$ The cost function can hence be cast into LASSO such that

$$J^* = \|y^* - \Phi^* \theta^*\|_2^2 + \gamma \|\theta^*\|_1$$

$$\hat{\theta}^* = \underset{\theta^*}{\operatorname{argmin}} \; J^*$$

We can retreive the original weights that were supposed to be predicted

$$\hat{\theta} = \frac{\hat{\theta}^*}{\sqrt{1+\alpha\lambda}}$$

$\therefore$ Augmenting the data ~~and~~ to cast elastic net into LASSO is a more efficent way to solve the problem.

NOTE: $\sqrt{1+\alpha\lambda}$ is multiplied for computational (numerical) stability.

The Solution was adopted from the paper "Regularization and Variable selection via Elastic Net"

Question 6

Given

$$f(y[k]|\theta) = \frac{\theta^{y[k]} e^{-\theta}}{y[k]!} = l(\theta / y[k])$$

$$\pi(\theta) \sim \frac{1}{\sqrt{\theta}}$$

Part (a)

For N observations,

$$l(\theta / y_N) = \frac{\theta^{\sum_{k=0}^{N-1} y[k]} e^{-N\theta}}{\prod_{k=0}^{N-1} y[k]!}$$

Log-likelihood

$$L(\theta / y_N) = \sum_{k=0}^{N-1} y[k] \ln\theta - N\theta + \sum_{k=0}^{N-1} \ln(y[k]!)$$

$$S = \frac{\partial L}{\partial \theta} = \frac{\sum_{k=0}^{N-1} y[k]}{\theta} - N$$

$$\frac{\partial S}{\partial \theta} = \frac{-1}{\theta^2} \sum_{k=0}^{N-1} y[k]$$

$$I(\theta) = -E\left(\frac{\partial S}{\partial \theta}\right) = \frac{1}{\theta^2} \sum_{k=0}^{N-1} E(y[k]) = \frac{N\theta}{\theta^2} = \frac{N}{\theta}$$

$$\Rightarrow \pi(\theta) \propto \sqrt{I(\theta)}, \text{ as } \pi(\theta) \propto \frac{1}{\sqrt{\theta}}$$

This proves that $\pi(\theta)$ is in the class of Jeffery's priors.

Part 1(b)

Posterior θ / y_N is

$$f(\theta / y_N) \propto f(y_N / \theta)\, \pi(\theta)$$

$$\Rightarrow f(\theta / y_N) = c\, f(y_N / \theta)\, \pi(\theta). \quad , \quad \pi(\theta) = \frac{1}{\sqrt{\theta}}$$

From the property

$$\int_0^\infty f(\theta / y_N)\, d\theta = 1$$

$$\Rightarrow c \int_0^\infty f(y_N / \theta)\, \pi(\theta)\, d\theta = 1$$

$$\Rightarrow c \int_0^\infty \frac{\theta^{N\bar{y} - \frac{1}{2}} \cdot e^{-N\theta}}{\prod_{K=0}^{N-1} y[K]}\, d\theta = 1$$

let $Z = N\bar{y} + \frac{1}{2}$, $N\theta = u$

$$\Rightarrow d\theta = \frac{du}{N}.$$

$$\Rightarrow \frac{c}{N^Z} \int_0^\infty \frac{u^{Z-1} \cdot e^{-u}}{\prod_{K=0}^{N-1} y[K]}\, du = 1$$

$$\Rightarrow \frac{C}{N^{Z}\prod_{K=0}^{N-1} y[K]} \Gamma(Z) = 1$$

$$\Rightarrow C = \frac{N^{Z}\prod_{K=0}^{N-1} y[K]}{\Gamma(Z)}$$

$$\cancel{\therefore f(\theta|y_N) = \frac{N^{N\bar{y}+\frac{1}{2}}}{\Gamma(Z)}}$$

$$\therefore f(\theta|y_N) = \frac{N^{N\bar{y}+\frac{1}{2}}}{\Gamma(N\bar{y}+\frac{1}{2})} \theta^{N\bar{y}-\frac{1}{2}} . e^{-N\theta}$$

$$= \frac{N}{\Gamma(N\bar{y}+\frac{1}{2})} (N\theta)^{N\bar{y}-\frac{1}{2}} e^{-N\theta}$$

$$= \frac{N}{\Gamma(N\bar{y}+\frac{1}{2})} \left(\frac{2N\theta}{2}\right)^{N\bar{y}-\frac{1}{2}} . e^{-\frac{2N\theta}{2}} .$$

$$\left[\text{let } K = 2N\bar{y}+1\right]$$

$$= \frac{N}{2^{K/2-1}\,\Gamma(K/2)} (2N\theta)^{K/2-1} e^{-(2N\theta)/2}$$

$$= \frac{2N}{2^{K/2}\,\Gamma(K/2)} (2N\theta)^{K/2-1} e^{-(2N\theta)/2}$$

$$= g(2N\theta / y_N)$$

Therefore,

$$2N\theta \sim \chi^2(K)$$

or $$2N\theta \sim \chi^2(2N\bar{y}+1)$$

Part (c)

Bayesian estimate $\hat{\theta}^*_{BA}$

$$\hat{\theta}^*_{BA} = E(\theta / y_N)$$

$$= \frac{1}{2N} \underbrace{E(2N\theta / y_N)}$$

↳ mean of χ^2 distribution derived above

$$\hat{\theta}^*_{BA} = \frac{K}{2N} = \frac{2N\bar{y}+1}{2N}$$

$$\Rightarrow \hat{\theta}^*_{BA} = \bar{y} + \frac{1}{2N}$$

Part (d)

We need to find a, b such that

$$\Pr(a < \theta < b) = 1-\alpha$$

or $$\Pr(2Na < 2N\theta < 2Nb) = 1-\alpha$$

$\Rightarrow$ ~~$\Pr(2N\theta > 2Na)$~~

$\Rightarrow$ $$\Pr(2N\theta < 2Na) = \alpha/2 = F_{2N\theta}(2Na).$$

and

$$\Pr(2N\theta < 2Nb) = 1-\alpha/2 = F_{2N\theta}(2Nb)$$

Question 6

Part (d)

Since it is not really possible to invert the cdf function of chi squared distribution, a simulation was performed in MATLAB to determine the credible interval of θ.

N = 10000 points were sampled from a Poisson distribution, have a true mean of $\theta_0 = 10$. The Bayesian Estimate is therefore,

$$\widehat{\theta^*}_{BA} = \bar{y} + \frac{1}{N} = 9.9825$$

By setting $\alpha = 0.05$, i.e., 95% credibility, the lower bound and upper bound (a, b) of the credible interval were determined as follows:

$$a = 9.9206$$

$$b = 10.0445$$

Therefore, the $(1 - \alpha)100\%$ credible interval for θ is

$$9.9206 < \theta < 10.0445$$