Probability and Statistics with R

Assignment 2

Arvind Raghavendran MDS202214 Swastika Mohapatra MDS202245

16 November, 2022

Note: Below I explain how you collaborate on `GitHub`.

1. It will be group assignment.
2. A group would be of size at most 3. If you want to create a group size more than 3, you must take permission.
3. Decide among yourself and one of you create a GitHub repository for Probability Statistics Assignments.
4. In that repository add your group members as collaborator
5. Once you add your collaborator (or group members), create a folder and name it as `Assignment_2`
6. In that folder you should have 2 folders `code` and `report`. And one `README.md` file. Write a brief report in `README.md` file.
7. For each problem, you should create a separate GitHub `issue`. All your discussion should be documented in the `issue`.
8. In the issue mention clearly, which group member is taking ownership of what problem?
9. The other member should `fork` the repository in their GitHub account.
10. Once you have your forked the main repository in your GitHub account - you should clone the repository in you local laptop or just download it as zip.
11. Once you develop the code - you should `commit` the code first in your repository and then `push` it.
12. Finally you make the `pull-request` in the final repository.
13. Once a member make a pull request, the other members have to review the code.
14. While reviewing the code the reviewer may have to download the code and run the code in his or her system and reproduce the result.
15. If the result is reproduced then she or he would accept and merge the code in final repository.
16. At the end you submit the link of the repository in the moodle.
17. The entire process will be evaluated.

Problem 1

Suppose X denote the number of goals scored by home team in premier league. We can assume X is a random variable. Then we have to build the probability distribution to model the probability of number of goals. Since X takes value in $\mathbb{N} = \{0, 1, 2, \cdots\}$, we can consider the geometric progression sequence as possible candidate model, i.e.,

$$S = \{a, ar, ar^2, ar^3, \cdots\}.$$

But we have to be careful and put proper conditions in place and modify S in such a way so that it becomes proper probability distributions.

Q1 Figure out the necessary conditions and define the probability distribution model using S.

A1 We need the following conditions for S to be a pdf:

1. $0 <= ar^n <= 1 \;\; \forall n$.
2. $\sum_{n=0} \mathbb{P}(X = n)$ converges to 1, hence we must have that $r < 1$.
3. Since $\sum_{n=0} ar^n = \frac{a}{1-r}$ and using the above, we must have $\mathbb{P}(X = n) = (1 - r)r^n$. (It is clear that $= X \sim Geom(1 - r)$)

Q2 Check if mean and variance exists for the probability model.

A2 Yes, they exist, since we know the mean and variance of Geometric random variable as done in class. we will derive them now.

Q3 Can you find the analytically expression of mean and variance?

1. Mean

$$\begin{aligned}
E[X] &= \sum_{n=0} n\mathbb{P}(X = n) \\
&= \sum_{n=0} n(1 - r)r^n \\
&= r\sum_{n=0} n(1 - r)r^{n-1} \\
&= r\sum_{n=1}(n - 1)(1 - r)r^{n-1} + r\sum_{n=1}(1 - r)r^{n-1} \\
&= rE[X] + r(1 - r)(\frac{1}{1 - r}) \\
&= rE[X] + r \\
\Rightarrow E[X] &= \frac{r}{1 - r}
\end{aligned}$$

2. Variance

$$
\begin{aligned}
E[X^2] &= \sum_{n=0} n^2 \mathbb{P}(X=n) \\
&= \sum_{n=0} n^2(1-r)r^n \\
&= r\sum_{n=0} n^2(1-r)r^{n-1} \\
&= r\sum_{n=1}(n-1)^2(1-r)r^{n-1} + \sum_{n=1} 2n(1-r)r^n - r\sum_{n=1}(1-r)r^{n-1} \\
&= rE[X^2] + 2E[X] - r(1-r)\frac{1}{1-r} \\
&= rE[X^2] + \frac{2r}{1-r} - r \\
\Rightarrow E[X^2] &= \frac{r^2+r}{(1-r)^2}. \\
\text{Therefore, } Var(X) &= E[X^2] - (E[X])^2 \\
&= \frac{r^2+r}{(1-r)^2} - \frac{r^2}{(1-r)^2} \\
&= \frac{r}{(1-r)^2}
\end{aligned}
$$

Q4 From historical data we found the following summary statistics Using the summary statistics and your

mean	median	variance	total number of matches
1.5	1	2.25	380

newly defined probability distribution model find the following:

- The above summary statistics do not seem to follow from a Geometric distribution since if we assume mean is true and solve for r, the variance is incorrect, and vice versa. But for this question, we will model our parameters using the mean. So if $\frac{r}{1-r} = 1.5$, then $r = 0.6$, meaning that our model has a variance of 3.5.

1. What is the probability that home team will score at least one goal?

Ans

$$
\begin{aligned}
\mathbb{P}(\text{atleast one goal}) &= 1 - \mathbb{P}(\text{no goal}) \\
&= 1-(1-r) \\
&= r \\
&= 0.6
\end{aligned}
$$

2. What is the probability that home team will score at least one goal but less than four goal?

Ans

$$\begin{aligned}
\mathbb{P}(1 <= X < 4) &= \sum_{i=1}^{3} \mathbb{P}(X = i) \\
&= \sum_{i=1}^{3} (1-r)r^i \\
&= r(1-r^3) \\
&= 0.47
\end{aligned}$$

Q5 Suppose on another thought you want to model it with off-the shelf Poisson probability models. Under the assumption that underlying distribution is Poisson probability find the above probabilities, i.e.,

– Once again, we will model it using the mean as parameter. The given summary statistics cannot be that of a Poisson distribution, since it has unequal mean and variance, but like in the previous question, we will assume mean is correct and hence, $\lambda = 1.5$. So, $\mathbb{P}(X = n) = \frac{e^{-\lambda}\lambda^n}{n!}$

1. What is the probability that home team will score at least one goal?

Ans

$$\begin{aligned}
\mathbb{P}(\text{atleast one goal}) &= 1 - \mathbb{P}(\text{no goal}) \\
&= 1 - e^{-\lambda} \\
&= 0.77
\end{aligned}$$

2. What is the probability that home team will score at least one goal but less than four goal?

Ans

$$\begin{aligned}
\mathbb{P}(1 <= X < 4) &= \sum_{i=1}^{3} \mathbb{P}(X = i) \\
&= \sum_{i=1}^{3} \frac{e^{-\lambda}\lambda^i}{i!} \\
&= 0.758
\end{aligned}$$

Q6 Which probability model you would prefer over another?

Model	Mean	Median	Variance
Expected	1.5	1	2.25
Geometric	1.5	2	3.5
Poisson	1.5	1	1.5

A6 Poisson has the least variance and also seems to have central tendencies closer to the expected values, hence it could be a better fit.

Q7 Write down the likelihood functions of your newly defined probability models and Poisson models.

A7 (a) $\mathcal{L}(r|(x_1, x_2, .., x_n)) = \Pi_i (1-r) r^{x_i} = (1-r) r^{\sum x_i}$

(b) $\mathcal{L}(\lambda|(x_1, x_2, .., x_n)) = \Pi_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\Pi_{i=1}^{n} x_i!}$

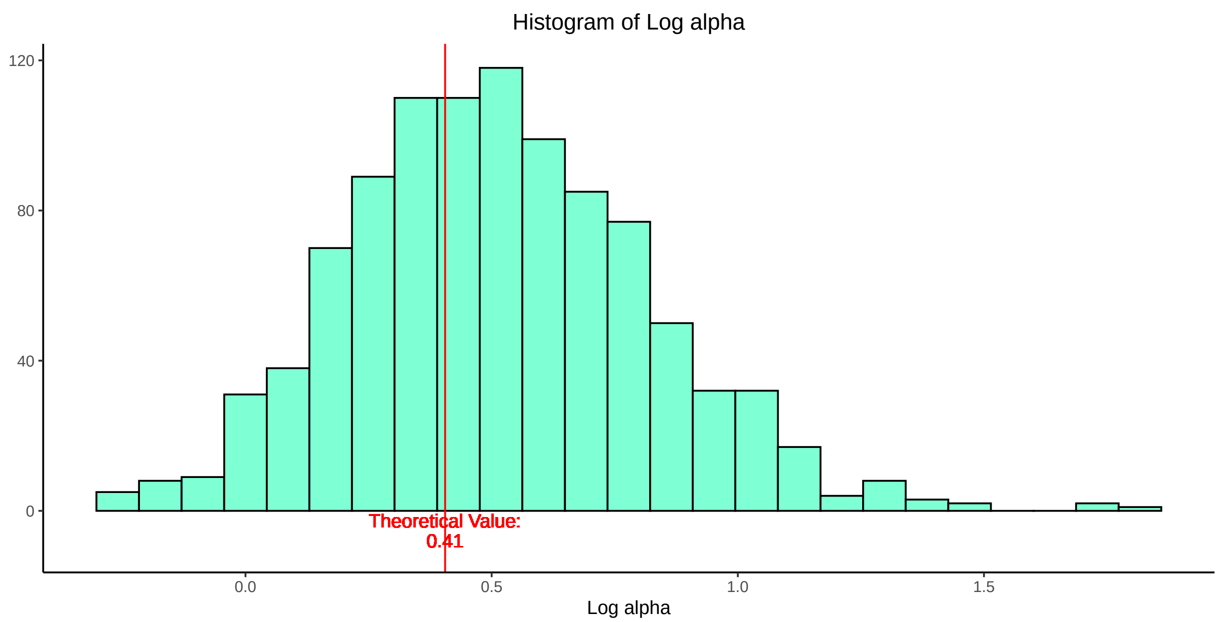
Problem 2 : Simulation Study to Understand Sampling Distribution

Part A Suppose $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Gamma(\alpha, \sigma)$, with pdf as

$$f(x|\alpha, \sigma) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} e^{-x/\sigma} x^{\alpha-1}, \quad 0 < x < \infty,$$

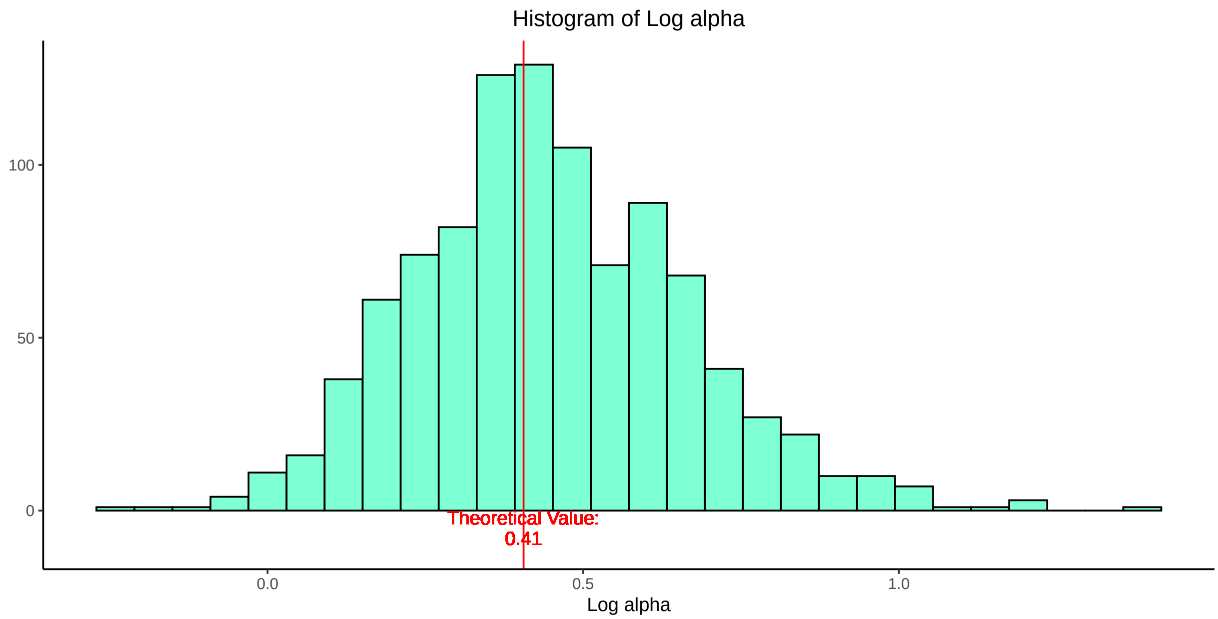
The mean and variance are $E(X) = \alpha\sigma$ and $Var(X) = \alpha\sigma^2$. Note that `shape` = α and `scale` = σ.

1. Write a `function` in `R` which will compute the MLE of $\theta = \log(\alpha)$ using `optim` function in `R`. You can name it `MyMLE`
2. Choose `n=20`, and `alpha=1.5` and `sigma=2.2`
   (i) Simulate $\{X_1, X_2, \cdots, X_n\}$ from `rgamma(n=20,shape=1.5,scale=2.2)`
   (ii) Apply the `MyMLE` to estimate θ and append the value in a vector
   (iii) Repeat the step (i) and (ii) 1000 times
   (iv) Draw histogram of the estimated MLEs of θ.
   (v) Draw a vertical line using `abline` function at the true value of θ.
   (vi) Use `quantile` function on estimated θ's to find the 2.5 and 97.5-percentile points.



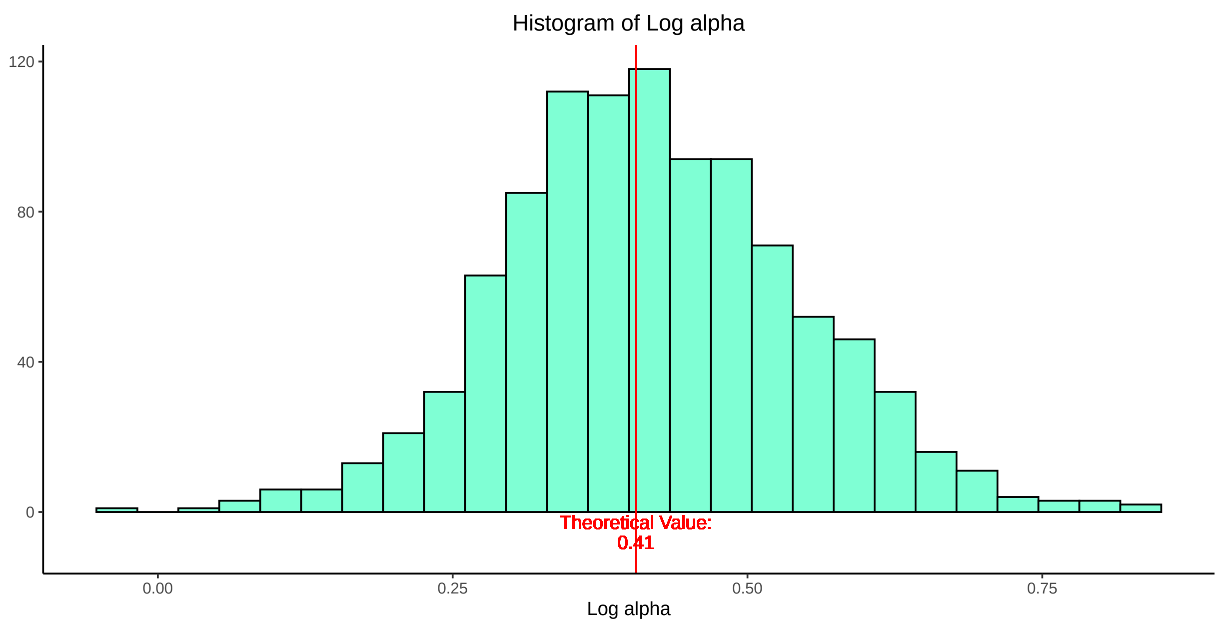

```
## The 2.5th quantile is -0.03511931
## The 97.5th quantile is 1.11573
## The gap between the 2.5th and 97.5th quantile is 1.150849
```

3. Choose `n=40`, and `alpha=1.5` and repeat the (2).




```
## The 2.5th quantile is 0.062348
## The 97.5th quantile is 0.8939899
## The gap between the 2.5th and 97.5th quantile is 0.8316419
```

4. Choose `n=100`, and `alpha=1.5` and repeat the (2).



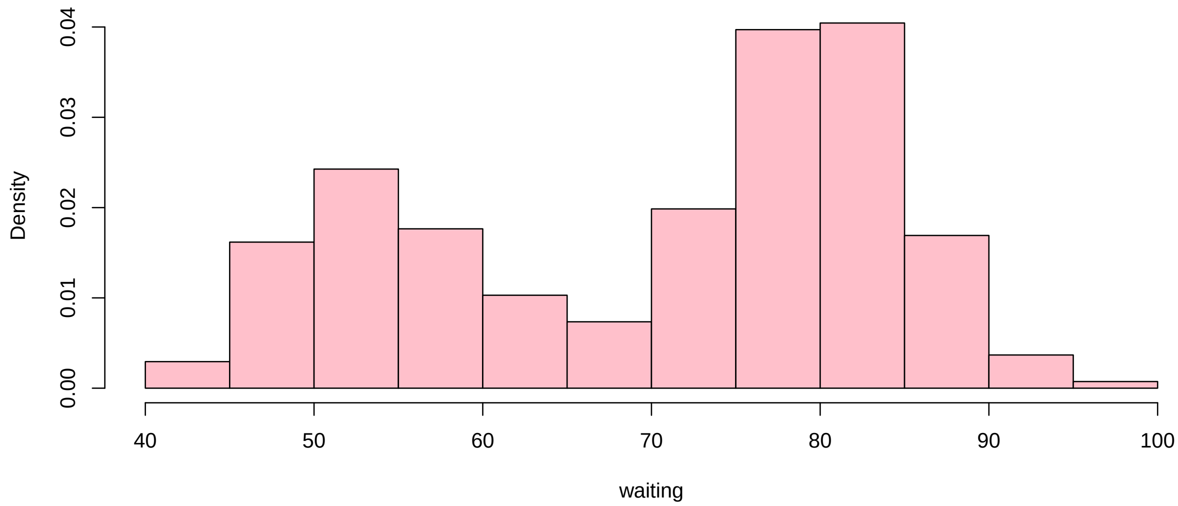

```
## The 2.5th quantile is 0.1834856
## The 97.5th quantile is 0.6670532
## The gap between the 2.5th and 97.5th quantile is 0.4835677
```

5. Check if the gap between 2.5 and 97.5-percentile points are shrinking as sample size `n` is increasing.

Yes, the gap between the 2.5 and 97.5 quantiles is decreasing as we increase the sample size.

Problem 3: Analysis of `faithful` datasets.

Consider the `faithful` datasets:



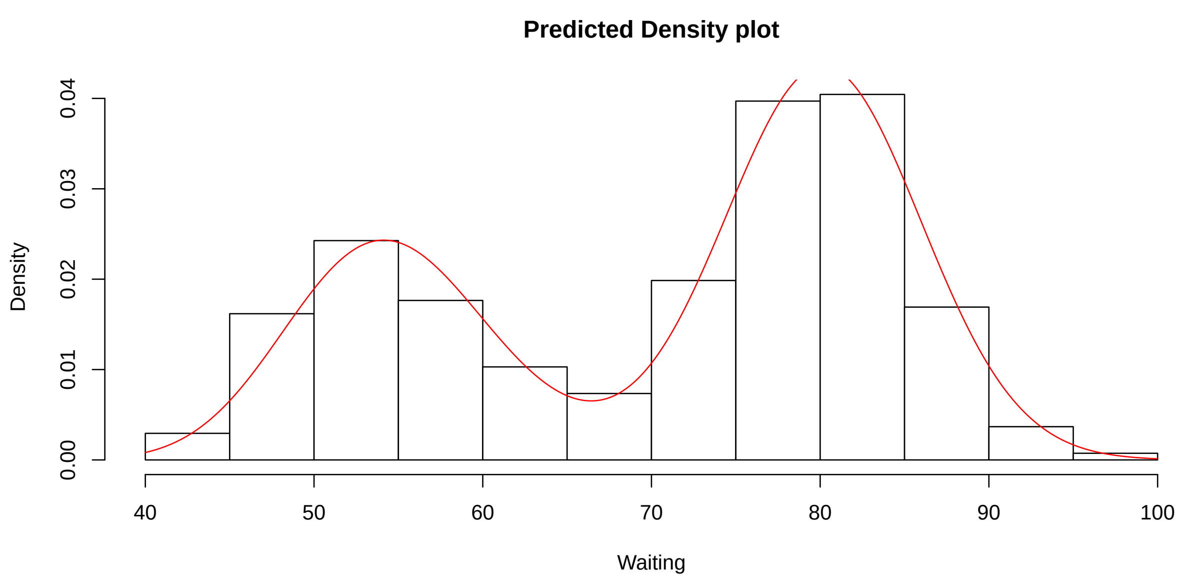
Fit following three models using MLE method and calculate **Akaike information criterion** (aka., AIC) for each fitted model. Based on AIC decides which model is the best model? Based on the best model calculate the following probability

$$\mathbb{P}(60 < \texttt{waiting} < 70)$$

(i) **Model 1**:

$$f(x) = p * Gamma(x|\alpha, \sigma_1) + (1-p) * N(x|\mu, \sigma_2^2), \quad 0 < p < 1$$

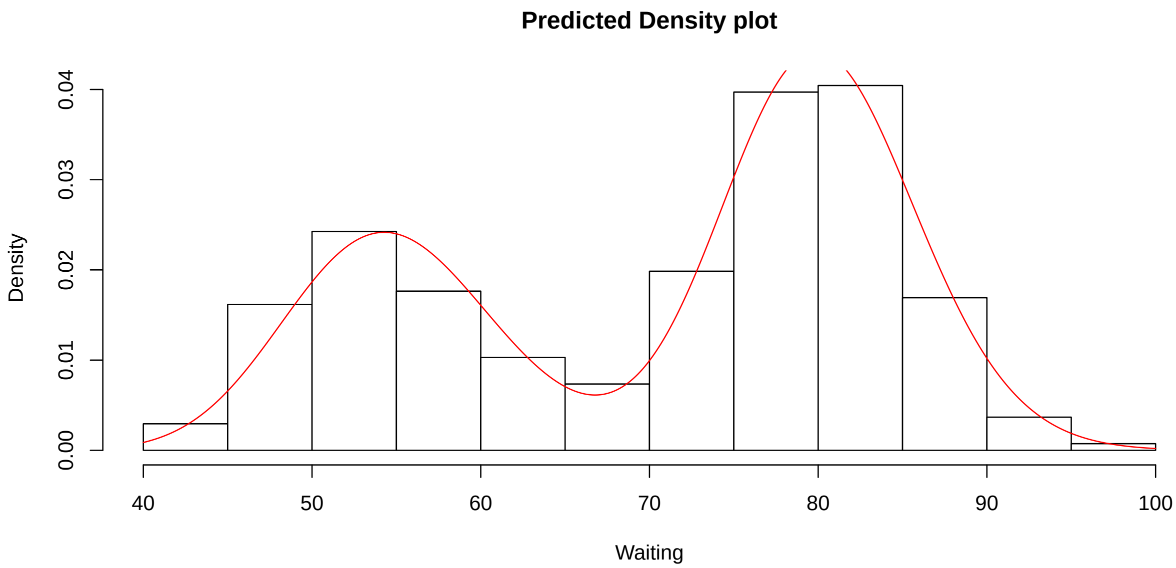
```
## Optimal (p,alpha,sigma_1,mu,sigma_2) is 0.3652286 82.77782 0.661879 80.16515 5.809759
## AIC is 2076.18
```



(ii) **Model 2**:

$$f(x) = p * Gamma(x|\alpha_1, \sigma_1) + (1-p) * Gamma(x|\alpha_2, \sigma_2), \quad 0 < p < 1$$

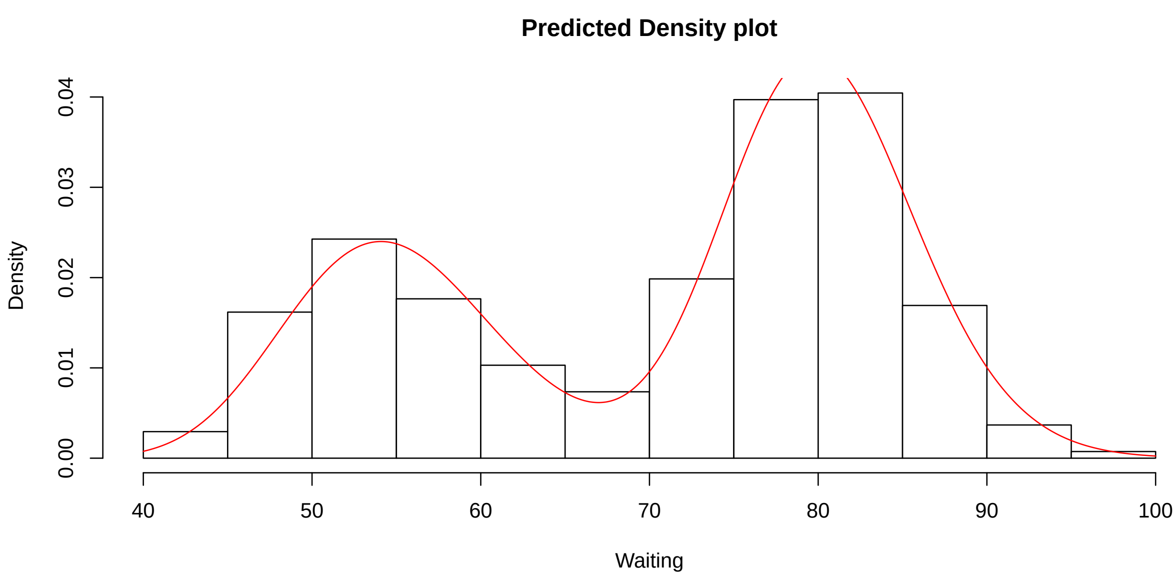
```
## Optimal (p,alpha_1,sigma_1,alpha_2,sigma_2) is 0.3708364 79.77377 0.6890328 199.2932 0.402859
## AIC is 2076.117
```



(iii) **Model 3**:

$$f(x) = p * logNormal(x|\mu_1, \sigma_1^2) + (1-p) * logNormal(x|\mu_1, \sigma_1^2), \quad 0 < p < 1$$

```
## Optimal (p,alpha_1,sigma_1,alpha_2,sigma_2) is 0.3761462 4.003843 0.1148572 4.384302 0.06973756
## AIC is 2075.42
```



The AIC of the three models are 2076.18, 2076.117, 2075.42. Since AIC is a measure of information lost by a model, lower the AIC, better the model. So, given the data, Model 3 is the best fit for the data. We will use this model to predict $\mathbb{P}(60 < X < 70)$.

```
## [1] 0.09082032
```

Problem 4: Modelling Insurance Claims

Consider the `Insurance` data sets in the `MASS` package. The data given in data frame `Insurance` consist of the numbers of policyholders of an insurance company who were exposed to risk, and the numbers of car insurance claims made by those policyholders in the third quarter of 1973.

This data frame contains the following columns:

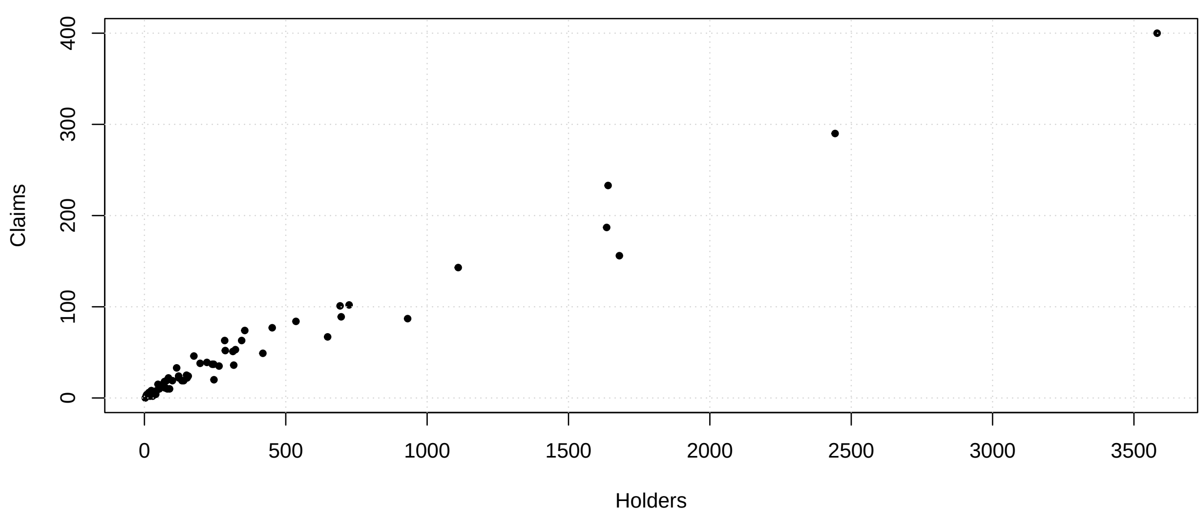
`District` (factor): district of residence of policyholder (1 to 4): 4 is major cities.

`Group` (an ordered factor): group of car with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.

`Age` (an ordered factor): the age of the insured in 4 groups labelled <25, 25–29, 30–35, >35.

`Holders` : numbers of policyholders.

`Claims` : numbers of claims



Note: If you use built-in function like `lm` or any packages then no points will be awarded.

Part A: We want to predict the `Claims` as function of `Holders`. So we want to fit the following models:

$$\texttt{Claims}_i = \beta_0 + \beta_1 \ \texttt{Holders}_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

Assume : $\varepsilon_i \sim N(0, \sigma^2)$. Note that $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

The above model can alse be re-expressed as,

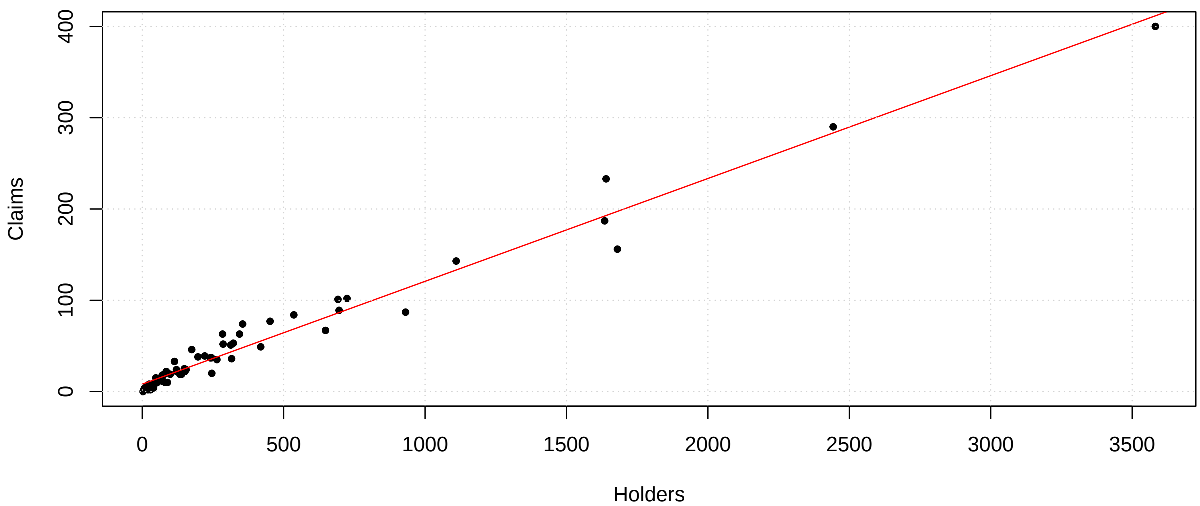
$$\texttt{Claims}_i \sim N(\mu_i, \sigma^2), \quad \textit{where}$$

$$\mu_i = \beta_0 + \beta_1 \ \texttt{Holders}_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

(i) Clearly write down the negative-log-likelihood function in `R`. Then use `optim` function to estimate MLE of $\theta = (\beta_0, \beta_1, \sigma)$

(ii) Calculate **Bayesian Information Criterion** (BIC) for the model.

```
## Optimal parameters are 8.120581 0.1126441 11.86623
## The BIC for this model 510.7587
```



Part B: Now we want to fit the same model with change in distribution:

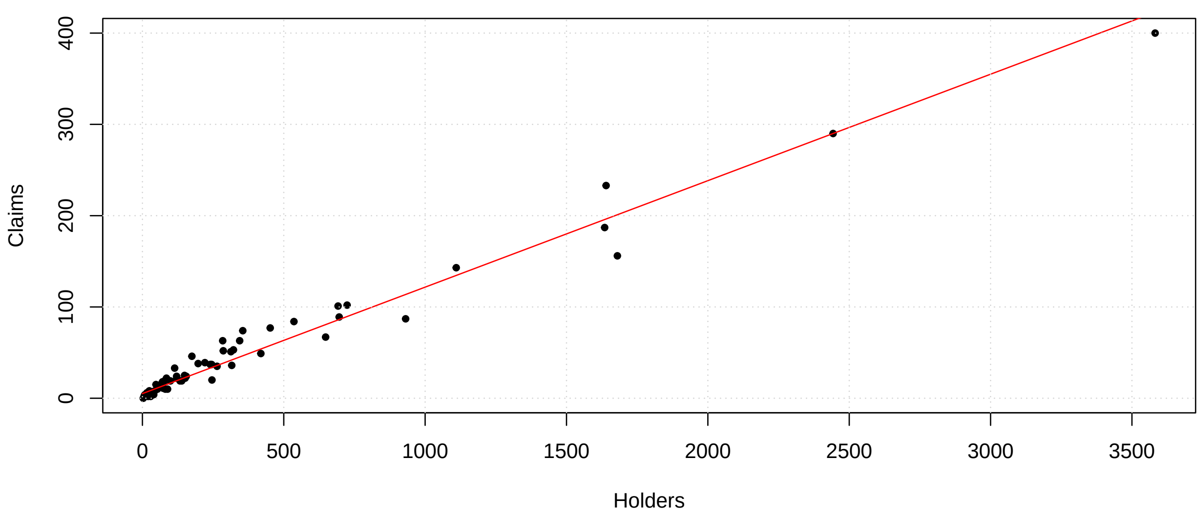
$$\texttt{Claims}_i = \beta_0 + \beta_1 \ \texttt{Holders}_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

Assume : $\varepsilon_i \sim Laplace(0, \sigma^2)$. Note that $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

(i) Clearly write down the negative-log-likelihood function in `R`. Then use `optim` function to estimate MLE of $\theta = (\beta_0, \beta_1, \sigma)$

(ii) Calculate **Bayesian Information Criterion** (BIC) for the model.

```
## Optimal parameters are 5.084375 0.1166253 2.863707
## The BIC for this model 498.687
```



Part C: We want to fit the following models:

$$\texttt{Claims}_i \sim LogNormal(\mu_i, \sigma^2), where$$

$$\mu_i = \beta_0 + \beta_1 \log(\texttt{Holders}_i), \quad i = 1, 2, ..., n$$

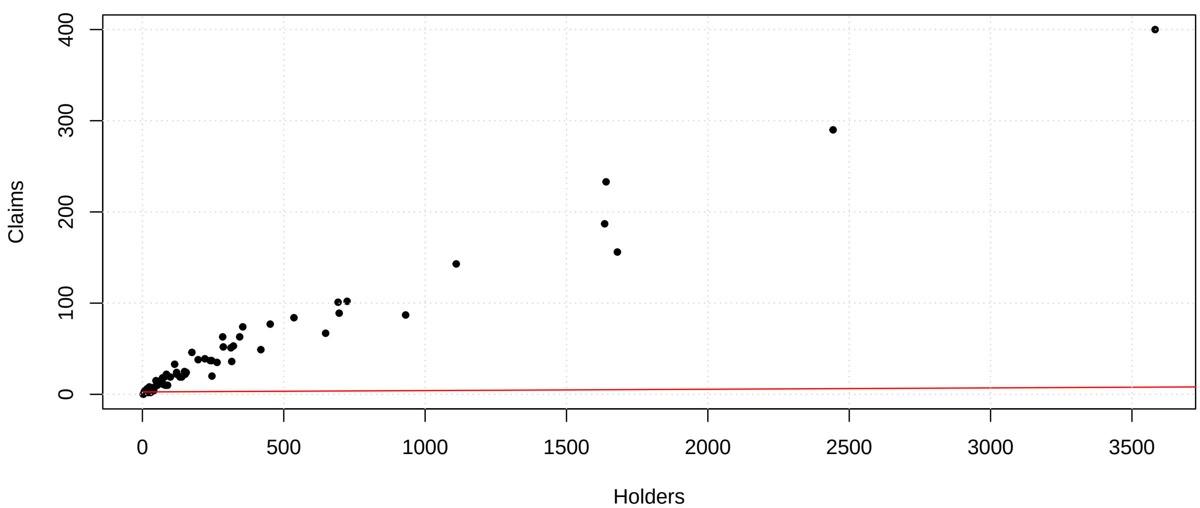
Note that $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

(i) Clearly write down the negative-log-likelihood function in `R`. Then use `optim` function to estimate MLE of $\theta = (\alpha, \beta, \sigma)$

(ii) Calculate **Bayesian Information Criterion** (BIC) for the model.

Note that we have one point in our data set where `Claims`=0. Since the support of Lognormal distribution is (0,∞), we remove this point.

```
## Optimal parameters are 2.639779 0.001472126 0.8229641
## The BIC for this model 568.0199
```

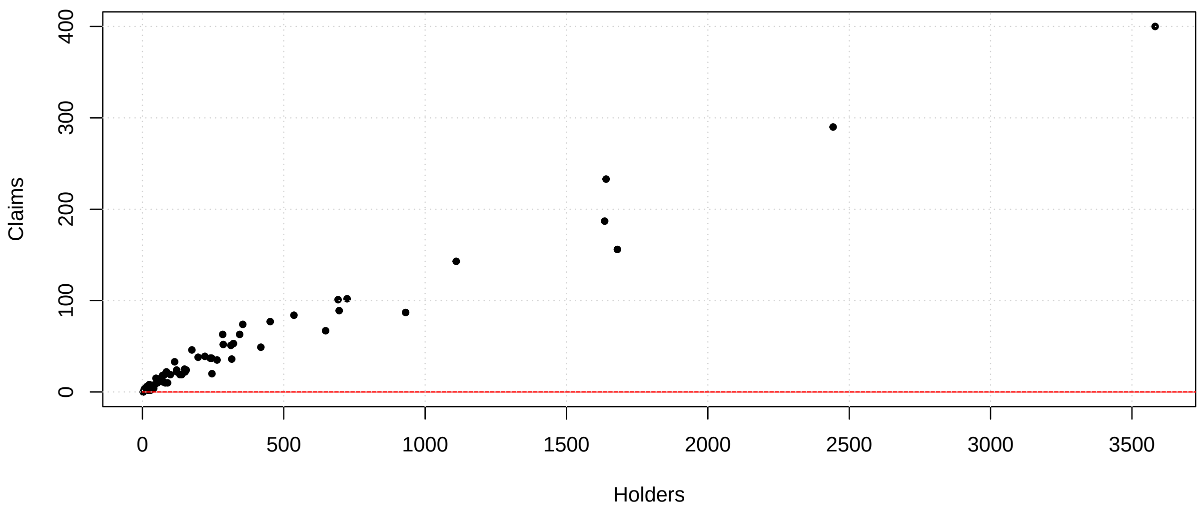


Part D: We want to fit the following models:

$$\texttt{Claims}_i \sim Gamma(\alpha_i, \sigma), where$$

$$log(\alpha_i) = \beta_0 + \beta_1 \log(\texttt{Holders}_i), \quad i = 1, 2, ..., n$$

```
## Optimal parameters are 0 0 49.23438
## The BIC for this model 639.2404
```



Q Compare the BIC of all three models.

A The BIC of the models are 510.76, 498.69, 568.02 and 639.24 respectively. Since the number of parameters being estimated are the same in all cases (3), lower the BIC, higher the likelihood and hence better the model. Based on this remark, the second model, i.e., $\epsilon_i \sim \text{Laplace}(0, \sigma^2)$ is the best fit for the data. The Gauss-Markov assumptions ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$) follow closely as the second best model. The 4th model ($\epsilon_i \sim \Gamma(0, \sigma^2)$) fits the worst.

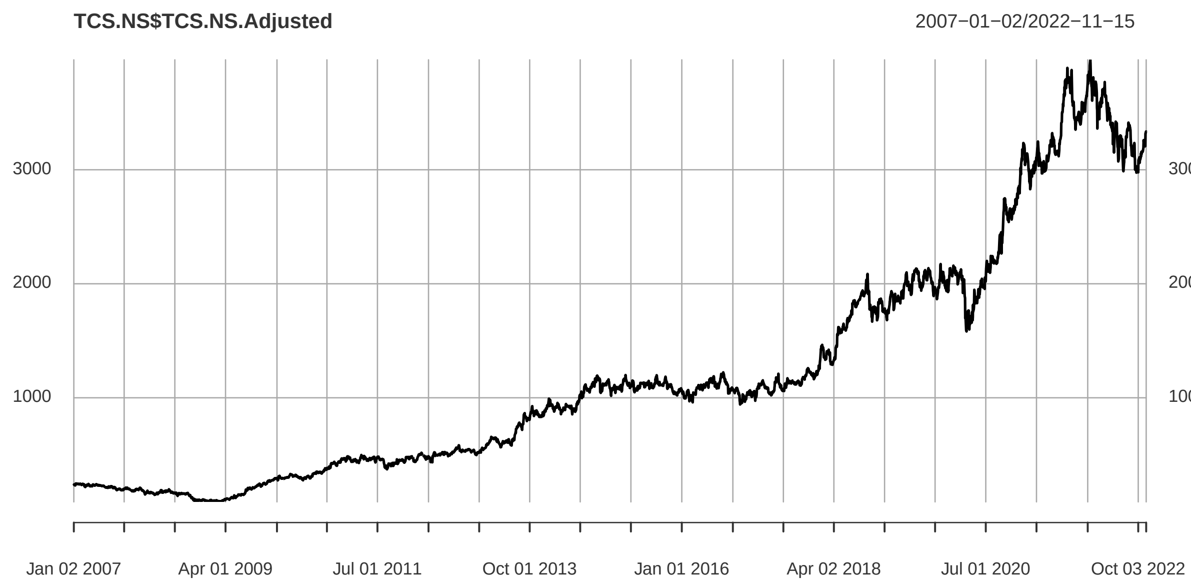
Problem 5: Computational Finance - Modelling Stock prices

Following piece of code download the prices of TCS since 2007

```
## [1] "TCS.NS"
```

```
##            TCS.NS.Open TCS.NS.High TCS.NS.Low TCS.NS.Close TCS.NS.Volume
## 2022-11-07      3229.0     3242.80    3195.10      3233.70       1474498
## 2022-11-09      3249.8     3249.80    3201.65      3216.05       1162267
## 2022-11-10      3170.0     3225.00    3170.00      3205.65       1573092
## 2022-11-11      3269.6     3341.60    3255.05      3315.95       3265394
## 2022-11-14      3324.0     3349.00    3309.00      3335.50       1342074
## 2022-11-15      3321.0     3339.95    3292.00      3332.60       1400708
##            TCS.NS.Adjusted
## 2022-11-07         3233.70
## 2022-11-09         3216.05
## 2022-11-10         3205.65
## 2022-11-11         3315.95
## 2022-11-14         3335.50
## 2022-11-15         3332.60
```

Plot the adjusted close prices of TCS

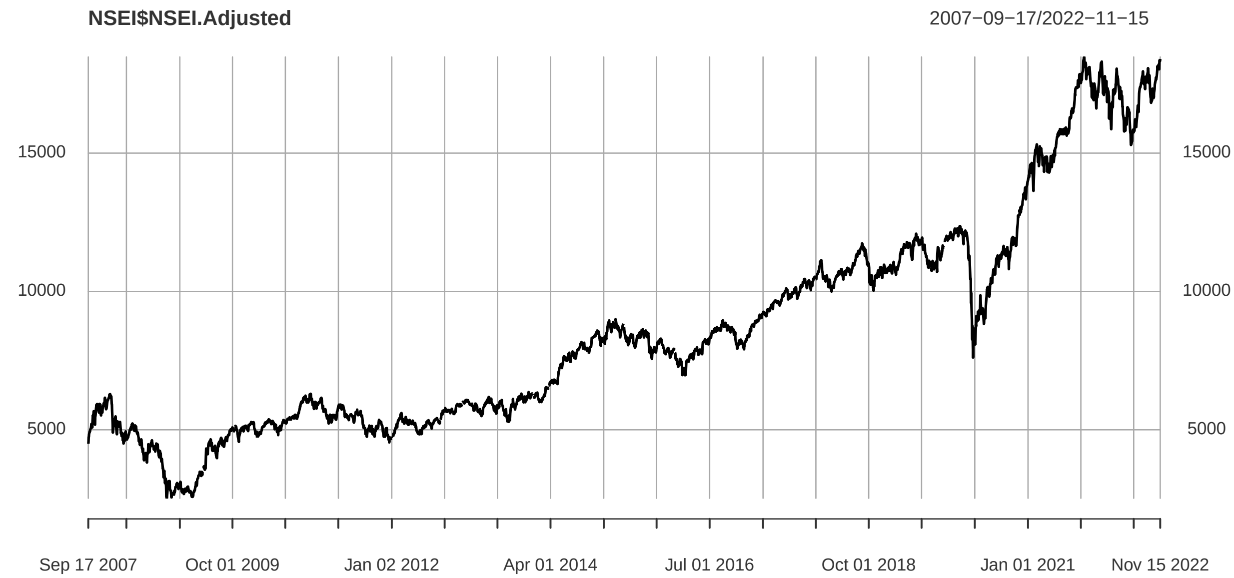


Download the data of market index Nifty50. The Nifty 50 index indicates how the over all market has done over the similar period.

```
## [1] "^NSEI"
```

```
##            NSEI.Open NSEI.High NSEI.Low NSEI.Close NSEI.Volume NSEI.Adjusted
## 2022-11-07  18211.75  18255.50 18064.75   18202.80      314800      18202.80
## 2022-11-09  18288.25  18296.40 18117.50   18157.00      307200      18157.00
## 2022-11-10  18044.35  18103.10 17969.40   18028.20      256500      18028.20
## 2022-11-11  18272.35  18362.30 18259.35   18349.70      378500      18349.70
## 2022-11-14  18376.40  18399.45 18311.40   18329.15      301400      18329.15
## 2022-11-15  18362.75  18427.95 18282.00   18403.40      250900      18403.40
```

Plot the adjusted close value of Nifty50



Log-Return

We calculate the daily log-return, where log-return is defined as

$$r_t = \log(P_t) - \log(P_{t-1}) = \Delta \log(P_t),$$

where P_t is the closing price of the stock on t^{th} day.

- Consider the following model:

$$r_t^{TCS} = \alpha + \beta r_t^{Nifty} + \varepsilon,$$

where $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}ar(\varepsilon) = \sigma^2$.

1. Estimate the parameters of the models $\theta = (\alpha, \beta, \sigma)$ using the method of moments type plug-in estimator discussed in the class.

```
## Optimal parameters are 0.0004616526 0.7436615 0.01618262
```

2. Estimate the parameters using the `lm` built-in function of `R`. Note that `lm` using the OLS method.

```
## Optimal parameters are 0.0004616526 0.7436615 0.01618482
```

3. Fill-up the following table

Parameters	Method of Moments	OLS
α	0.0004611214	0.0004611214
β	0.7436967	0.7436967
σ	0.01618653	0.01618873

4. If the current value of Nifty is 18000 and it goes up to 18200. The current value of TCS is Rs. 3200/-. How much you can expect TCS price to go up?

```
## New TCS price would be expected to be Rs. 3227.894
```