Methods of Generating Parametric Estimators: The MLE and Others

Ryan Ren, Barry Yao

Abstract

Our project is like a mixture of discussing a ”textbook” topic and results from several research papers. The reason why we incorporate such wide scope of materials is because behind every aspect of MLE, there will be bottomless theorems. If we choose only dig into one aspect of MLE, we might end up with pages of mathematical proofs without any practical use. In the meantime, we need to take tons of theorems as proven. That's definitely not the thing we want to explore.

Since we've already know that MLE shares numerous good properties in most of cases, digging into the cases which MLE faces challenges might actually turn out to be a more instructive way of studying. We derived most of the properties, such as the expectations and variances by ourselves.

Also, one thing needs to make it clear is that we are not trying to provide a instruction manual of when and how to implement MLE but providing some unusual cases that we need to go in with eyes open, especially when we hold a high expectation of MLE.

Contents

Introduction 1

1 Analysis 2

1.1 Parameter Estimation of Simple Linear Regression Using MLE . . . 2

1.2 MLE of Exponential Distribution . . . 2

1.3 Unbiased Estimator for Exponential . . . 3

1.4 Simulation of Different estimators for exponential 4

Accuracy • Different Pursuit

1.5 Inconsistency of MLE . . . 5

Background • Sum of Normal Distribution

2 Application of the MLE 6

2.1 The Likelihood Ratio Test . . . 6

Example 1 of LRT(simple)

2.2 Logistic regression . . . 7

Conclusion 8

Appendix 9

References 9

Introduction

A common challenge of doing a statistical analysis is figuring out the parameter, which contains the crucial and special characteristic of a given data set. If we already have a named distribution with unknown parameters, the task left for us is just figuring out the parameters of these distributions. In a real world case, it's nearly impossible to calculate the true parameter of a data set. But, the thing we could at least try to do is estimating these parameters by incorporating some well founded methods. In this project, we will dive into different kinds of methods of generating parametric estimators and make comparisons between them to find out the degree of adaptations of these methods with different kinds of data set and with different statistical persuasions.

As the name of the project suggests, our discussion of different parametric estimator will be expanded with the focus of the Maximum likelihood estimator(MLE). As it turns out, though the capacity in data collection has grown exponentially, the demand for data has always managed to beat the technological advancement in data collection. Using data efficiently is only becoming more necessary.In practice, the MLE is widely used in machine learning, econometrics, neural science, psychophysics, and so on because it has been shown to require less data and trials, converge quickly, involve a manageable computational cost.

In this project, we want to closely examine the MLE's possible short-comings or special cases in lieu of only discussing its various advantages. Specifically, we delve into the biased-ness and consistency/inconsistency of the MLE. We demonstrate cases where the MLE is biased as well as cases where it is inconsistent. In the following sections of discussion, the MLE is also compared to another parameter estimator to understand the extent which the MLE might underperform. Lastly, two instances of application of the MLE is included.

1. Analysis

In this section, we further our discussion of the several qualities Maximum Likelihood Estimator with the specificity of examples. Under some conditions (or simply in most of the cases), MLE is a consistent, asymptotically normal, and asymptotically efficient estimator. It is worth noting that the MLE works very well when the sample size is very large. That is, the MLE could be biased but its bias converges to zero when sample size goes to infinity (consistency).

1.1 Parameter Estimation of Simple Linear Regression Using MLE

Suppose the linear relation:

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i \tag{1}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$
To conduct maximum likelihood estimation, we first generate the likelihood function of ε_i.

$$L(\mu,\sigma^2|\varepsilon_1,\varepsilon_2,\varepsilon_3,...,\varepsilon_n) \tag{2}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} \qquad ,\varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2) \tag{3}$$

Per the goal of the MLE, we want to derive estimators for all three parameters in the simple linear regression β_o, β_1, and σ^2. Thus, observe from equation (1):

$$\varepsilon_i = Y_i - \beta_o - \beta_1 X_i \tag{4}$$

Plugging (4) into the likelihood function:

$$L(\beta_0,\beta_1,\sigma^2|(X_i,Y_i)) \tag{5}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(Y_i-\beta_o-\beta_1X_i)^2}{2\sigma^2}) \tag{6}$$

For easy of computation, take the log of (6):

$$\ell(\beta_0,\beta_1,\sigma^2) \tag{7}$$

$$= \frac{n}{2} ln(2\pi) + \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}((Y_i-\beta_o-\beta_1X_i)^2 \tag{8}$$

To find our estimators, we maximize (8) by taking the partial derivative with respect to each parameter and setting them to zero:

$$\frac{\partial\ell}{\partial\beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_o-\beta_1X_i) = 0 \tag{9}$$

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1\sum_{i=1}^{n} X_i = 0 \tag{10}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1\bar{X} \tag{11}$$

We realize that our maximum likelihood estimator for β_0 is the same as the Ordinary Least Square estimator.

Similarly, we can show that the MLE also estimates β_1 the same as the OLS:

$$\frac{\partial\ell}{\partial\beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_o-\beta_1X_i)X_i = 0 \tag{12}$$

$$\sum_{i=1}^{n} Y_iX_i - \sum_{i=1}^{n}(\bar{Y}-\beta_1\bar{X})X_i - \beta_1\sum_{i=1}^{n} X_i^2 = 0 \tag{13}$$

$$\sum_{i=1}^{n} Y_iX_i - \bar{Y}\sum_{i=1}^{n} X_i + \beta_1\bar{X}\sum_{i=1}^{n} X_i - \beta_1\sum_{i=1}^{n} X_i^2 = 0 \tag{14}$$

$$\frac{\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \hat{\beta}_1 \tag{15}$$

Lastly, for σ^2 we have:

$$\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i-\beta_o-\beta_1X_i)^2 = 0 \tag{16}$$

$$\frac{1}{2\sigma^4}\sum_{i=1}^{n}\varepsilon^2 = \frac{n}{2\sigma^2} \tag{17}$$

$$\sigma^2\sum_{i=1}^{n}\varepsilon^2 = n\sigma^4 \tag{18}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon^2 \tag{19}$$

Importantly, we realized we have shown a situation where the MLE is biased since the unbiased estimator for σ^2 from OLS is

$$\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}\varepsilon^2 \tag{20}$$

where n-p represent the degrees of freedom we have. Thus, observe that our estimator from the maximum likelihood produces underestimates. Nonetheless, as we take the sample size large enough,

$$\lim_{n\to\infty}\frac{1}{n-p}\sum_{i=1}^{n}\varepsilon^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon^2 = 0 \tag{21}$$

consistency still hold. [1, 2]

1.2 MLE of Exponential Distribution

Suppose $X_1, X_2, X_3.....X_n \sim Exp(\lambda)$, we first want to find $\hat{\lambda}_{MLE}$.

$$L_n(\lambda) = \prod_{i=1}^{n}\lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda\sum_{i=1}^{n} X_i} \tag{22}$$

$$\ell_n(\lambda) = n \cdot ln(\lambda) - \lambda\sum_{i=1}^{n} X_i \tag{23}$$

In order to find the maximizer λ for equation (23), we take the derivative of it.

$$\frac{\partial \ell_n}{\partial \lambda} = n \cdot \frac{1}{\lambda} - \sum_{i=1}^{n} X_i \tag{24}$$

Since function(24) is a function of $\frac{1}{\lambda}$, it's not hard to show that it will reach its maximum when the first derivative equals to 0.

$$n \cdot \frac{1}{\lambda} - \sum_{i=1}^{n} X_i = 0 \tag{25}$$
$$n \cdot \frac{1}{\lambda} = \sum_{i=1}^{n} X_i \tag{26}$$
$$\frac{1}{\lambda} = \overline{X} \tag{27}$$
$$\hat{\lambda}_{MLE} = \frac{1}{\overline{X}} \tag{28}$$

Next, we want to test whether the MLE estimator we got for exponential distribution is biased or not.

$$\mathbb{E}\left[\hat{\lambda}\right] = E\left[\frac{1}{\overline{X_i}}\right] = \mathbb{E}\left[\frac{n}{\sum_{i=1}^{n} X_i}\right] \tag{29}$$

We notice that the denominator is actually a sum of n independent identically distributed $Exp(\lambda)$ random variables, and we use the fact that it equals to $Gamma(n, \lambda)$. Then we use the pdf of the Gamma distribution to help us solve this equation.

Let $a = \sum_{i=1}^{n} X_i$, (23) can be rewrite into:

$$\mathbb{E}\left[\frac{n}{a}\right] = n \cdot \mathbb{E}\left[\frac{1}{a}\right] \tag{30}$$
$$= n \int_{o}^{\infty} \frac{1}{a} \cdot \frac{\lambda^n}{\Gamma(n)} a^{n-1} e^{-\lambda a} da \tag{31}$$
$$= n \cdot \frac{\lambda^n}{\Gamma(n)} \int_{o}^{\infty} a^{n-2} e^{-\lambda a} da \tag{32}$$
$$= n \cdot \frac{\lambda^n}{\Gamma(n)} \cdot \frac{1}{\lambda} \int_{o}^{\infty} \left(\frac{u}{\lambda}\right)^{n-2} e^{-u} du \tag{33}$$
$$= n \cdot \frac{\lambda^n}{\Gamma(n)} \cdot \frac{1}{\lambda} \frac{1}{\lambda^{n-2}} \int_{o}^{\infty} u^{n-2} e^{-u} du \tag{34}$$

Since $\int_{o}^{\infty} u^{n-2} e^{-u} du = \Gamma(n-1)$,

$$(34) = n \frac{\lambda^n}{(n-1)\Gamma(n-1)} \frac{1}{\lambda^{n-1}} \Gamma(n-1) \tag{35}$$
$$= \frac{n}{n-1} \lambda. \tag{36}$$

Clearly, we have shown that $\hat{\lambda}_{MLE}$ is a biased estimator. But, will this estimator still be consistent? First we need to calculate the standard error for $\hat{\lambda}_{MLE}$.

$$se(\hat{\lambda}_{MLE}) = \sqrt{\mathbb{V}(\hat{\lambda}_{MLE})} \tag{37}$$
$$= \sqrt{\mathbb{V}\left(\frac{n}{\sum_{i=1}^{n} X_i}\right)} \tag{38}$$

From previous example of calculating the expectation, we could use the similar approach to calculate $\mathbb{E}\left[\frac{n^2}{a^2}\right]$.

$$\mathbb{V}(\hat{\lambda_{MLE}}) = \mathbb{E}\left[\frac{n^2}{a^2}\right] - \mathbb{E}^2\left[\frac{n}{a}\right] \tag{39}$$
$$= \frac{n^2\lambda^2}{(n-1)(n-2)} \cdot \frac{n^2\lambda^2}{(n-1)^2} \tag{40}$$
$$= \frac{n^2\lambda^2}{(n-1)^2(n-2)} \tag{41}$$
$$se(\hat{\lambda_{MLE}}) = \frac{1}{\sqrt{n-2}} \cdot \frac{n\lambda}{n-1} \tag{42}$$

To be noticed that,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n-2}} \cdot \frac{n\lambda}{n-1} = 0 \tag{43}$$
$$\lim_{n\to\infty} \frac{n\lambda}{n-1} = 0 \tag{44}$$

Therefore, we finally proved that $\hat{\lambda_{MLE}}$ is a consistent estimator after all. So far, we know that when the number of data set is large enough, MLE should be no problem for exponential distribution. However, what if we don't have a large data set? Is there any other method can generate unbiased estimator for exponential distribution?

1.3 Unbiased Estimator for Exponential

Right now, we already had an estimator for exponential distribution which is a biased one. Can we try to make some small twist to it in order to make an unbiased one?

We know from previous section that $\mathbb{E}[\hat{\lambda}_{MLE}] = \frac{n}{n-1}\lambda$. By definition of an unbiased estimator, we need $\mathbb{E}[\hat{\lambda} - \lambda] = 0$. Therefore, we can simply times $\frac{n-1}{n}$ to the $\hat{\lambda}_{MLE}$, then we have an unbiased estimator for exponential distribution:

$$\hat{\lambda}_{UB} = \frac{n-1}{n} \frac{1}{\overline{X}} \tag{45}$$

To be noticed, there is a method, "Uniformly Variance-Minimize Unbiased Estimator (UVMUE)", can actually generate the same unbiased estimator for exponential distribution without the need of tracing backward. However, the proof of how to find this estimator would involve too much extra information which would digress from the main focus of this report.

When we first get a new estimator, the first thing to do is to check whether this estimator is consistent or not.

$$\mathbb{V}\left[\frac{n-1}{n}\frac{1}{\overline{X}}\right] = \left(\frac{n-1}{n}\right)^2 \cdot \mathbb{V}\left[\frac{1}{\overline{X}}\right] \tag{46}$$

$$= \left(\frac{n-1}{n}\right)^2 \cdot \frac{n^2\lambda^2}{(n-1)^2(n-2)} \tag{47}$$

$$= \frac{\lambda^2}{n-2} \tag{48}$$

$$se(\hat{\lambda}_{UB}) = \frac{\lambda}{\sqrt{n-2}} \tag{49}$$

We also have,

$$\lim_{n\to\infty} \frac{\lambda}{\sqrt{n-2}} = 0 \tag{50}$$

Therefore, our new estimator $\hat{\lambda}_{UB}$ is also an consistent estimator.

Interestingly, we can easily see that $se(\hat{\lambda}_{UB}) < se(\hat{\lambda}_{MLE})$. So, we don't even need to actually calculate MSE to proof that the unbiased estimator will be a more accurate estimator than MLE for exponential distribution in theory.

1.4 Simulation of Different estimators for exponential

First of all, we use exponential distribution when we have a data set which smaller values are more easily to appear than the larger values. Another key property is memorylessness, the event will be happened next has no relationship with the historical events.

One practical example is that suppose I'm the owner of a convenient store, and I wish to record how much does each of my customer spend in my convenient. It's reasonable to say that data set which I have recorded follows an exponential distribution since first, the amount for each customer spend will not depend on how much does others spend in my store; since this is a convenient store, most of people will only spend little money (It's quite unlikely that some people spend few hundreds of dollars in a convenient store).

1.4.1 Accuracy

In this section, we want to visualize the performances of these two estimators for λ we found in previous section when we don't have a large data set.

Suppose the amount of money people spend in this convenient store follows an exponential distribution with real $\lambda = \frac{1}{10}$.

We use simulations to check what will happen to these two estimators from the case which we have only 2 data to the case which we have 60 data.

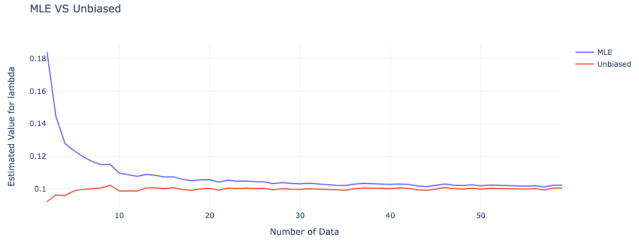

Figure 1. (2,60)

The main idea of this simulation is to first use exponential random numbers generator to generate n data, then using these methods to find their corresponding $\hat{\lambda}$. In order to avoid the randomness, we will do this experiment for 1000 times and find the average $\hat{\lambda}$ for each n. In the end, we want to see how these two estimators would be changed as n grows from 2 to 60. (Codes can be found in Appendix)

From figure 1, we can easily see that the unbiased estimator outperformed MLE in accuracy. On average, when $n = 5$, the unbiased estimator have already reached to a steady state close to the real $\lambda = \frac{1}{10}$. However, MLE remained to deviate from the real λ even when $n = 60$.

From this visualized result, we showed again that the unbiased estimator should be a more accurate estimator when n is small.

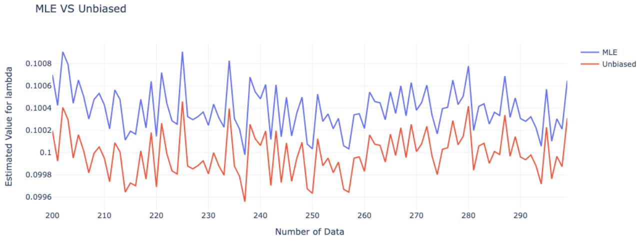

Figure 2. (100,200)

Also, another thing to be noticed about from figure 2 is that when we do the simulation from $n = 100$ to $n = 200$, we could see that MLE have already oscillated in a small interval of $(0.1, 0.1008)$. This reflects the consistency of MLE for exponential distribution.

1.4.2 Different Pursuit

Normally, we definitely prefer a more accurate estimator. But in Statistics, we might have different pursuits which the accurate one might not be the most ideal choice.

To be continued, with our previous example of the convenient store, suppose right now we have generated both $\hat{\lambda}_{MLE}$ and $\hat{\lambda}_{UB}$. Also we know that for an exponential distribution, mean equals to $\frac{1}{\lambda}$. Therefore, we could calculate:

$$\hat{\mu}_{MLE} = \frac{1}{\hat{\lambda}_{MLE}} \tag{51}$$

$$\hat{\mu}_{UB} = \frac{1}{\hat{\lambda}_{UB}} \tag{52}$$

Bringing these two parameters into the example of convenient store, we could find the estimated average spending for each customer per each visit. So, for this particular estimator, except for accuracy, is there any other quality which we need to care about? One of the thing is that we may not want to have an overestimated mean value. Thinking about this realistically that the owner of the store overestimates the average sale for each customer, it might mislead the owner to be too satisfied and make some wrong decisions. Therefore, when we talk about earning money, an underestimated but relatively accurate estimator may sometimes perform better than the overestimate one.

If we take a closer look of the Figure 1, we will find out that MLE tends to overestimate λ, which implies it underestimate μ. On the other hand, the unbiased estimator tends to oscillate around the real λ which make it unavoidable to overestimate the λ more easily. By showing this, we take another data simulation.

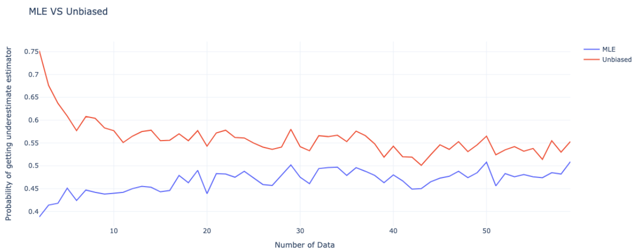

Figure 3. Underestimating Probability of $\hat{\lambda}$

Basically, We followed the same approach of simulating the estimator values, but rather than simply calculating $\hat{\lambda}$, we take a further step to check whether $\hat{\lambda}$ is larger than the real λ or not for each one of the 1000 trials for each n, the number of available data in the data set. Then, among 1000 trials, we calculate the weight of trials which produce overestimated $\hat{\lambda}$ as the probability of getting an overestimated estimator.

From Figure 3, we could easily see the difference between the probabilities of getting an overestimated $\hat{\lambda}$ from these two methods. So, if I'm the owner of the convenient store, I would give away some accuracy to exchange for a smaller probability to get an underestimated $\hat{\lambda}$ and an overestimated $\hat{\mu}$.

Important Notes: One with watchful eagle eyes would ask what am I doing here? Since exponential distribution also has a form of taking $\frac{1}{\lambda}$ as the scale parameter, if we just want $\hat{\mu}$, why did we go around such a big circle to calculate $\hat{\lambda}$?
The answer is that it will be a completely different story. The MLE for $\hat{\mu}$ is actually an unbiased estimator! It's also very easy to show why $\hat{\lambda}_{MLE}$ is biased since when we try to find $\mathbb{E}\left[\frac{1}{\overline{X}}\right]$, the unbiasedness didn't pass through the non-linear transformation. But, if we are dealing with the scale parameter $\frac{1}{\lambda}$, MLE will simply be $\overline{X}$, which will be an unbiased one for sure.

The reason why we choose to talk about $\hat{\lambda}$ is again to deliberately look for the weakness of Maximum Likelihood Estimator. If we want to estimate the parameter of exponential practically, it's suggested to just use MLE for μ. Since when we literally use exponential distribution in our life, the parameter we care the most should be μ not the λ we spend so much paper to discuss about.

1.5 Inconsistency of MLE

In this section, we re-evaluate the inconsistent example proposed by ReadFord Neal.

1.5.1 Background

Among all of the properties of an estimator, consistency should be the most important one. Since if the estimator doesn't hold this property, there is always a chance that our estimated value will be different with the real value, no matter how big the data set we have. That's should be case which we always need to avoid.
On the other hand, it shows that why MLE has the popularity since in most of general cases, MLE will always be consistent; also, most of inconsistent cases for MLE are quite pathological.

1.5.2 Sum of Normal Distribution

For some $\theta \in \mathbb{R}$ and $\theta > 0$,

$$F|\theta \sim \frac{1}{2}N(0,1) + \frac{1}{2}N(\theta, e^{(-1/\theta^2)^2}) \tag{53}$$

This equation was first posted by ReadFord Neal[3] in his blog, and we use a slightly different approach to find the MLE for this problem.

First, we noticed that this distribution consists a standard normal distribution and one normal distribution depending on variable θ. Normally, for the sum of two normal random variables we need to use convolution integral to calculate the pdf. In this case, it will end up with an integral without closed form which is quite messy. However, there is not any outstanding evidence to stop us from treating them as independent, so it's intuitively to take the shortcut to combine this two normal distributions into one since we know that the sum of two normal will still be normal if they are independent. Therefore, we change the original distribution to:

$$F|\theta \sim \frac{1}{2}N(\theta, 1 + e^{(-1/\theta^2)^2}) \tag{54}$$

One proper question to ask is why MLE loses its consistency for this particular normal distribution. If it's a regular normal distribution which has separate variables of μ and σ^2, MLE will hold consistency. But, in this case, both the variance and mean are actually depending on θ.

The most important part is calculating $\hat{\theta}$, and in this problem, we will use a different approach to solve for $\hat{\theta}_{MLE}$ which is solving by brute force.

The basic idea is that we first construct the likelihood function, and plug in random X_i generated by following the distribution. The most important part is trying different θ and plug into the likelihood function. In the end, we will get a list or a distribution of data generated by likelihood function by plugging in different θs. To be noticed, the θ we plug in will be discrete, which means that the estimated value we have generated might not be accurate, especially when we are taking 0.01, a relatively large increment for each step. In addition, in order to minimize the complexity of the calculation, we take the log likelihood function.

```
def inconsistentexp(N,arr):
    t = 0.01 #theta starts from 0.01
    finallist=[]
    j=0
    for j in range(0,200):
        t+=j*0.01 #increment equals 0.01
        list=[]
        t_pdf=0
        for i in range(0,N): #caculating the log likelihood
            X_i = arr[i]
            list.append(np.log(0.5*(sc.norm(t,np.sqrt(1+(np.exp(-1/t**2))**2)).pdf(X_i))))
        t_pdf = sum(list)
        finallist.append(t_pdf)
    index=np.argmax(finallist) #find the index of theta which produce the maximum like hood
    return index*0.01

result=[]
num_data=[]

for i in range(10,50): #do this simulation from 10 data point to 50 data point
    t=1.5
    a=0.5*np.random.normal(0,1,i)
    b=0.5*np.random.normal(t,np.exp(-1/t**2),i)
    arr=[]
    for x in range(0,i):
        arr.append(a[x]+b[x])
    result.append(inconsistentexp(i,arr))
    num_data.append(i)
```

Listing 1. Codes for inconsistent simulation

By using this algorithm, from Figure 4, we could easily find out that no matter how large our data set is, $\hat{\theta}$ still can not asymptotically converge to the real θ.

Extending from this point, there are some interesting questions could be raised which deserve further discussions beyond this project: What exactly causes inconsistency for this distribution? Two parameters depending on one single variable or the exponential term on the variance? If it's the first case, is it suggesting that for any parametric distribution which has two parameters depended on one single variable will also make MLE become inconsistent?

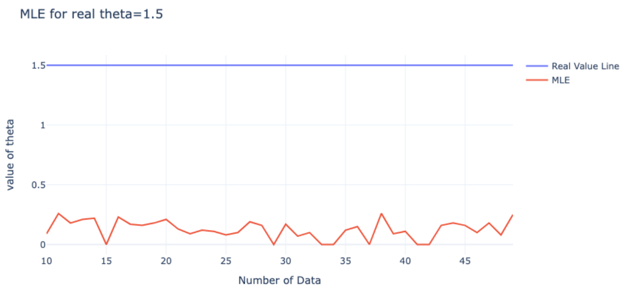

Figure 4. Simulation Result for Inconsistent MLE

2. Application of the MLE

In this section, two applications of the MLE are introduced.

2.1 The Likelihood Ratio Test

The Likelihood Ratio Test is a method of hypothesis test which uses a function of maximum likelihood as the test statistic to determine whether a parameter θ is within a subset Θ_0(the null hypothesis) of the parameter space Θ against the alternative hypothesis that the parameter at interest is in Θ_0^c. The test statistic is given by a ratio of maximum likelihood, it is called λ_{LR} (likelihood ratio):

$$\lambda_{LR} = -2ln\left(\frac{\sup_{\theta\in\Theta_0}\mathscr{L}(\theta)}{\sup_{\theta\in\Theta}\mathscr{L}(\theta)}\right) \tag{55}$$

Notice the test statistics is bounded between 0 and 1 as the fraction of the two supremum is bounded between 0 and 1. To be specific, the supremum of the likelihood function over the entire parameter space Θ is always larger than the likelihood function of the same parameter over the subset Θ_0 of the parameter space.

The upshot of this testing method is to evaluate the weight of the maximum likelihood of our null hypothesis $\theta \in \Theta_0$ over that of all of the parameter space. That is, the larger the value of $\sup_{\theta\in\Theta_0}\mathscr{L}(\theta)$ is compared to $\sup_{\theta\in\Theta}\mathscr{L}(\theta)$, the more likely that our null hypothesis is correct. Hence, we choose an α value to determine a critical c to compare to our test statistics.

The frame of the likelihood ratio test is the following:

$$\begin{aligned} &H_0 : \theta \in \Theta_0 \\ &H_1 : \theta \in \Theta_0^c \end{aligned} \tag{56}$$

We obtain $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$,..., $X_n = x_n \overset{\text{iid}}{\sim} \mathscr{F}(\theta)$. Calculating (22), the test statistic λ_{LR} will reveal.

- if $\lambda_{LR} \geq c$, we reject H_0, concluding $\theta \in \Theta_0^c$
- if $\lambda_{LR} < c$, we fail to reject H_0

Where c is determined through the preset α.
It is worth noting that an identical test is to use the log likelihood functions to calculate the test statistic. Here, we call it λ_{LLR}:

$$\lambda_{LLR} = -2(\sup_{\theta\in\Theta_0} \mathscr{L}(\theta) - \sup_{\theta\in\Theta} \mathscr{L}(\theta)) \tag{57}$$

However, regardless of the the version of the test statistics, according to the Wilk's Theorem, the test statistic λ will converge asymptotically to the chi-square distribution $\mathscr{X}^2$.

$$\lambda_{LR} = -2ln\left(\frac{\sup_{\theta\in\Theta_0} \mathscr{L}(\theta)}{\sup_{\theta\in\Theta} \mathscr{L}(\theta)}\right) \sim \mathscr{X}_q^2 \tag{58}$$

where q denote the degrees of freedom equal to the different in dimensionalities between Θ and Θ_0
As implied by the asymptotic convergence, the Wilk's Theorem is not significant in the most strict sense. Rather, it offers us guidance to whether the ratio of the maximum likelihood is too large, allowing statisticians to compare the test statistic λ_{LR} with the $\mathscr{X}_q^2$ value to arrive at an approximate test result.

2.1.1 Example 1 of LRT(simple)

In this example, we make our setup as simple as possible, assuming we have only 1 data, X=x, from the random variable:

$$X = \theta + \varepsilon, \qquad \varepsilon \sim \mathscr{N}(0,1) \tag{59}$$

We want to design a likelihood ratio test using X=x of if $\theta = 0$ or $\theta = 1$ with $\alpha = 0.05$.

$$H_0 : \theta_0 = 0 \qquad H_1 : \theta_1 = 1 \tag{60}$$

By the null setup,

$$H_0 = 0 \implies X \sim \mathscr{N}(0,1) \tag{61}$$

Under the null hypothesis, the likelihood functions are following:

$$\mathscr{L}(X = x|\theta_0) = \frac{1}{\sqrt{2\pi}} exp(\frac{-x^2}{2}) \tag{62}$$

$$\mathscr{L}(X = x|\theta_1) = \frac{1}{\sqrt{2\pi}} exp(\frac{-(x-1)^2}{2}) \tag{63}$$

Now, we are able to compute λ_{LR}:

$$\frac{\mathscr{L}(X = x|\theta_0)}{\mathscr{L}(X = x|\theta_1)} = exp(\frac{-x^2}{2} + \frac{(x-1)^2}{2}) \tag{64}$$

$$= exp(\frac{1-2x}{2}) \tag{65}$$

We have our test statistic λ_{LR}. To finish our test, we look for the critical value c which corresponds to the our level $\alpha = 0.05$.
We first find an expression of c by separating x from

$$exp(\frac{1-2x}{2}) \geq z \tag{66}$$

Note (33) represents the domain where we fail to reject the null hypothesis.

$$x \leq \frac{1}{2}(1 - 2ln(z)) = c \tag{67}$$

Equivalently, (34) is the condition which we fail to reject H_0 with our eventual expression of c.
To find our critical value, recall that an α value represents the probability of making a type I error. Hence:

$$\mathbb{P}(\text{Reject } H_0|H_0 \text{ is true}) = \mathbb{P}(X > c|\theta = 0) \tag{68}$$

Notice here the Wilk's theorem doesn't apply as we clearly don't have anything close to an asymptotic convergence. However, we set up the null hypothesis such that we can use (28) to continue our test.

$$\mathbb{P}(X > c|\theta = 0) = 1 - \phi(c) = \alpha \tag{69}$$

$$\implies c = \phi^{-1}(1-\alpha) = 1.6449 \tag{70}$$

$$\implies z = 0.318 \tag{71}$$

[1, 4]

2.2 Logistic regression

In this section, we decode how parameters for a logistic regression are determined by MLE.
First, we recall the formula for a standard logistic regression:

$$f(z) = \frac{e^z}{1+e^z} \tag{72}$$

In (72), a quick note is if we take z to ∞ and $-\infty$, we see it is bounded by 0 and 1. That is, as we input data, z, into a logistic function, it returns a probability measure of the data, allowing us to solve a supervised classification problem.

Much like simple linear regression, we characterize our data z as:

$$z = \beta_0 + \beta_1 X_i \tag{73}$$

$$\implies f(\beta_0 + \beta_1 X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} \tag{74}$$

Note we take the simple case where our data is characterized by only 1 feature. However, there could be any number of features and as many βs. The calculation doesn't change.

To recall, logistic regression uses a Bernoulli distribution when it predicts the probability of an event Y, where p is the parameter for the Bernoulli distribution:

$$p = f(z) = P(Y = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} \tag{75}$$

The PMF is therefore

$$
\begin{aligned}
& p^{y_i}(1-p)^{(1-y_i)} && (76)\\
&= \left(\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}}\right)^{(1-y_i)} && (77)\\
&= \left(\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1X_i}}\right)^{(1-y_i)} && (78)\\
&= \frac{(e^{\beta_0+\beta_1X_i})^{y_i}}{(1+e^{\beta_0+\beta_1X_i})^{y_i}(1+e^{\beta_0+\beta_1X_i})^{(1-y_i)}} && (79)\\
&= \frac{e^{y_i(\beta_0+\beta_1X_i)}}{1+e^{\beta_0+\beta_1X_i}} && (80)
\end{aligned}
$$

y take value 0 or 1 of course.

Thus, using this fact logistic regressions are generated by MLE to produce the parameters β_0 and β_1. The likelihood function for β_0 and β_1 is:

$$
\prod_{i=1}^{n}\frac{e^{y_i(\beta_0+\beta_1X_i)}}{1+e^{\beta_0+\beta_1X_i}} \qquad (81)
$$

Log likelihood:

$$
\begin{aligned}
\ell(\beta_0,\beta_1,X_i,Y_i) &= \prod_{i=1}^{n} log\left(\frac{e^{y_i(\beta_0+\beta_1X_i)}}{1+e^{\beta_0+\beta_1X_i}}\right) && (82)\\
&= \sum_{i=1}^{n} y_i(\beta_0+\beta_1X_i)-\sum_{i=1}^{n} log(1+e^{\beta_0+\beta_1X_i}) && (83)
\end{aligned}
$$

To actually find the MLE, we take the partial derivatives of (83) and find the maximum.

$$
\begin{aligned}
\frac{\partial\ell}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left(\sum_{i=1}^{n} y_i(\beta_0+\beta_1X_i)-\sum_{i=1}^{n} log(1+e^{\beta_0+\beta_1X_i})\right) && (84)\\
&= \frac{\partial}{\partial\beta_1}\left(\sum_{i=1}^{n} y_i\beta_0+\beta_1X_iy_i\right)-\sum_{i=1}^{n}\frac{X_ie^{(\beta_0+\beta_1X_i)}}{1+e^{(\beta_0+\beta_1X_i)}} && (85)\\
&= \sum_{i=1}^{n} X_iy_i-\sum_{i=1}^{n}\frac{X_ie^{(\beta_0+\beta_1X_i)}}{1+e^{(\beta_0+\beta_1X_i)}} && (86)\\
\frac{\partial\ell}{\partial\beta_0} &= \frac{\partial}{\partial\beta_0}\left(\sum_{i=1}^{n} y_i(\beta_0+\beta_1X_i)-\sum_{i=1}^{n} log(1+e^{\beta_0+\beta_1X_i})\right) && (87)\\
&= \sum_{i=1}^{n} y_i-\sum_{i=1}^{n}\frac{e^{(\beta_0+\beta_1X_i)}}{1+e^{(\beta_0+\beta_1X_i)}} && (88)
\end{aligned}
$$

By setting (86) and (88) to 0, we complete our derivation of the MLE. (Note there is closed form solution for the estimators)[5, 6]

Conclusion

The project is an extension to the introductory presentation of the maximum likelihood estimation in textbook *All of Statistics: A Concise Course in Statistical Inference* by Larry A. Wasserman. The discussion of the MLE is centered around its consistency, asymptotic normality, and efficiency under certain conditions.

So much as our discussion mention several specific instances where consistency would break or the MLE could be biased through mathematical analysis and simulations using algorithms, this paper is the opposite of an account of the disadvantages of the MLE. As we are find marginal cases where the MLE might underperform, we stumble upon ingenious uses of the MLE in a wide range of industries and very important use of the MLE such as the logistic regression. It is in turn demonstrative how powerful the MLE is. Our project therefore serves as a further exploration to the less known aspect of the MLE as an addition to our undergraduate study of mathematical statistics.

Appendix

```
import numpy as np
from scipy.stats import norm

import matplotlib.pyplot as plt
from scipy.stats import expon
from statistics import mean

N=30
def do_experiments(N):
    lamda=1/10
    mle = 0
    count_mle=0
    count_ub=0
    ub = 0
    num = 1000
    for i in range(0,num):
        X=expon.rvs(scale = 1/lamda,size = N)
        x_mle = (1/mean(X))
        if(x_mle<lamda): #Statement used for check prob. of underestimate of lambda
            count_mle+=1
        x_ub = (((N-1)/N)*(1/mean(X)))
        if(x_ub<lamda):
            count_ub+=1

        mle += x_mle
        ub+= x_ub
    list = [(mle/num),(ub/num),count_mle/num,count_ub/num]
    return list
print("MLE:",do_experiments(N)[0])
print("UVMUE:", do_experiments(N)[1])

num_data=[]
y_mle=[]
y_ub=[]
y_count_mle=[]
y_count_ub=[]
for i in range(200,300):
    waitedforadd=do_experiments(i)
    num_data.append(i)
    y_mle.append(waitedforadd[0])
    y_ub.append(waitedforadd[1])
    y_count_mle.append(waitedforadd[2])
    y_count_ub.append(waitedforadd[3])

import plotly.graph_objects as go

fig = go.Figure()

fig.add_trace(go.Scatter(
    x=num_data,
    y=y_mle,
    name="MLE"
))
fig.add_trace(go.Scatter(
    x=num_data,
    y=y_ub,
    name="Unbiased"
))
fig.update_layout(template=’plotly_white’,
                  xaxis=dict(title=’Number of Data’),
                  yaxis=dict(title=’Estimated Value for lambda’),
                  title=’MLE VS Unbiased’)
fig.show()

import plotly.graph_objects as go

fig = go.Figure()

fig.add_trace(go.Scatter(
    x=num_data,
    y=y_count_mle,
    name="MLE"
))
fig.add_trace(go.Scatter(
    x=num_data,
    y=y_count_ub,
    name="Unbiased"
))
fig.update_layout(template=’plotly_white’,
                  xaxis=dict(title=’Number of Data’),
                  yaxis=dict(title=’Probability of getting overestimate estimator’),
                  title=’MLE VS Unbiased’)
fig.show()
```

Listing 2. Codes for Simulation of ”MLE vs. Unbiased” for exponential

A nicer and with more explanation code file could be accessed here![Code]

References

[1] Jian-Xin Pan and Kai-Tai Fang. Maximum likelihood estimation. *Growth Curve Models and Statistical Diagnostics*, March 2002.

[2] Ryan P. Adams. Linear regression via maximization of the likelihood. *Elements of Machine Learning Princeton University*, September 2018.

[3] Radford Neal. Inconsistent maximum likelihood estimation: An “ordinary” example. *Radford Neal’s blog*, August 2008.

[4] Hossein Pishro-Nik. Likelihood ratio tests. *Introduction to Probability, Statistics, and Random Processes*, March 2014.

[5] Naman Agrawal. Decoding logistic regression using mle. *Analytics Vidhya*, March 2022.

[6] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology,Volume 47, Issue 1*, 2003.