Homework 1 – Maximum Likelihood Estimator

STA 211 – The Mathematics of Regression – Spring 2022

Erik Novak

2022 January 25

Instructions

This assignment covers methods for maximum likelihood estimation for general statistical models. Upload a scanned or word-processed version of the assignment to the Assignments folder on Sakai. For problems where you are asked to derive or show a result, include all intermediate steps in your answer for full credit.

Problems

1. Watch the Panopto video on "Properties of Estimator" posted in the Panopto folder on the Sakai site. This video covers Section 7 of the STA 211 Supplement, which you are welcome to read in addition. For those learning maximum likelihood estimation for the first time, or needing a review, read Section 8 of the STA 211 Supplement. For this problem, write a sentence indicating whether or not you viewed the video and what sections of the STA 211 Supplement you read, if any. Watching the video earns 2 points on this assignment.

I watched the Panopto video, and have read all of the supplement sections, the most relevant of which for this homework were certainly sections 7 and 8.

2. A commonly used probability distribution for monetary random variables is the Pareto distribution. Assume all values of the random variable are greater than or equal to some baseline value k, which is fixed and known. The Pareto probability density function is

$$f(y) = \begin{cases} \theta k^{\theta} \left(\frac{1}{y}\right)^{\theta+1} & y \geq k,\, \theta > 0 \\ 0 & \text{otherwise.} \end{cases}$$

   Suppose you have a random sample of n independent measurements. **Find an expression for the MLE for** θ.

Given the assumption that the n measurements are i.i.d, then let $f(y_i)$, for $1 \leq i \leq n$, be the probability density function for one of the n measurements.

From that, the Likelihood function $L(\theta)$ is defined as the product of the probability density functions for each measurement. That is,

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} (f y_i) \\
&= \prod_{i=1}^{n} \theta k^{\theta} \left(\frac{1}{y_i}\right)^{\theta+1} \\
&= \left(\theta k^{\theta}\right)^{n} \prod_{i=1}^{n} \frac{1}{y_i^{\theta+1}} \\
&= \theta^{n} k^{n\theta} \prod_{i=1}^{n} y_i^{-(\theta+1)}
\end{aligned}$$

Now, having the expression for the likelihood function we need to find the values for its variables that maximize it. That is, the maximum likelihood estimator for θ will be the value of θ that maximizes $L(\theta)$

To solve this optimization problem, we need only differentiate $L(\theta)$ with respect to θ, set it equal to 0, and solve for θ.

However, to differentiate this expression, we notice it is the product of $n+2$ different simpler functions; thus, to find the derivative, we are best served by first taking the natural logarithm of the function, and thus finding the log-likelihood function:

$$\begin{aligned}
l(\theta) = \ln(L(\theta)) &= \ln\left(\theta^{n} k^{n\theta} \prod_{i=1}^{n} y_i^{-(\theta+1)}\right) \\
&= \ln(\theta^{n}) + \ln(k^{n\theta}) + \ln\left(\prod_{i=1}^{n} y_i^{-(\theta+1)}\right) \\
&= n\ln(\theta) + n\ln(k)\theta + \sum_{i=1}^{n} \ln\left(y_i^{-(\theta+1)}\right) \\
&= n\ln(\theta) + n\ln(k)\theta - (\theta+1)\sum_{i=1}^{n} \ln\left(y_i\right) \\
&= n\ln(\theta) + n\ln(k)\theta - \theta\sum_{i=1}^{n} \ln\left(y_i\right) - \sum_{i=1}^{n} \ln\left(y_i\right)
\end{aligned}$$

Note that that our values for y_i remain the same, as the logarithm was taken over the function as a whole; so, the window for which the distributions is defined as nonzero will stay the same.

Now, the value that maximizes the natural logarithm of a function is the same as the value that maximizes the original function, so we need only consider the log-likelihood function $l(\theta)$ and differentiate it with respect to θ, and finally find the value of θ for which the derivative gives 0. That value will be the same for $L(\theta)$ too.

So, we differentiate:

$$\begin{aligned}\frac{d}{d\theta}l(\theta) &= \frac{d}{d\theta}\left(n\ln(\theta) + n\ln(k)\theta - \theta\sum_{i=1}^{n}\ln(y_i) - \sum_{i=1}^{n}\ln(y_i)\right)\\ &= \frac{d}{d\theta}(n\ln(\theta)) + \frac{d}{d\theta}(n\ln(k)\theta) - \frac{d}{d\theta}\left(\theta\sum_{i=1}^{n}\ln(y_i)\right) - \frac{d}{d\theta}\left(\sum_{i=1}^{n}\ln(y_i)\right)\\ &= \frac{n}{\theta} + n\ln(k) - \sum_{i=1}^{n}\ln(y_i)\end{aligned}$$

As regards this result, we note that by definition $\theta > 0$ so there will be no division by 0 issue. Additionally, $k > 0$ necessarily too, since our random variables are monetary. Finally, $n > 0$ as well, since it is the number of observations. Hence, this function will be properly defined for any possible fixed and random variables used.

Now, we set the derivative of $l(\theta)$ equal to 0, and find the values of θ that ensure that result; those will be our MLE values.

For that, we have

$$\begin{aligned}\frac{n}{\theta} + n\ln(k) &= \sum_{i=1}^{n}\ln(y_i)\\ \iff n\left(\frac{1}{\theta} + \ln(k)\right) &= \ln\left(\prod_{i=1}^{n}y_i\right)\\ \iff \frac{1}{\theta} + \ln(k) &= \frac{1}{n}\ln\left(\prod_{i=1}^{n}y_i\right)\\ \iff \frac{1}{\theta} + \ln(k) &= \ln\left(\left(\prod_{i=1}^{n}y_i\right)^{1/n}\right)\end{aligned}$$

Now, recall that all y_i were assumed to be i.i.d. This means their geometric mean is simply multiplying all of them together and then taking the n-th root. This is precisely what the inside of the left-hand side of the equation above indicates.

So, define y_{GM} as the geometric mean of all y_i. From that, we have

$$\begin{aligned}
\iff \frac{1}{\theta} + \ln(k) &= \ln\left(\left(\prod_{i=1}^{n} y_i\right)^{1/n}\right) \\
\iff \frac{1}{\theta} + \ln(k) &= \ln(y_{GM}) \\
\iff \frac{1}{\theta} &= \ln(y_{GM}) - \ln(k) \\
\iff \frac{1}{\theta} &= \ln\left(\frac{y_{GM}}{k}\right) \\
\iff \theta &= \frac{1}{\ln\left(\frac{y_{GM}}{k}\right)}
\end{aligned}$$

We have isolated a value for θ that ensures the derivative of $l(\theta)$ is 0, and thus that the derivative of $L(\theta)$ is also 0.

Note this value has to be a maximum, and not a minimum, because the first derivative yielded a single inflection point possible. Now, because our likelihood function is not constant, and is positive, this inflection point must be a maximum.

Because we have already drawn the n independent variables, and because k is assumed to be fixed, this is a final constant value, our MLE.

So, we have

$$\theta = \frac{1}{\ln\left(\frac{y_{GM}}{k}\right)}$$

3. Suppose that in a company all individuals earn at least $k = \$20,000$ per year. Suppose you have $n = 5$ individuals' salaries, $y_1 = \$110,501$, $y_2 = \$45,662$, $y_3 = \$89,680$, $y_4 = \$1,658,909$, and $y_5 = \$20,218$. (Professor Reiter drew these randomly from a Pareto distribution using R.) Find the MLE of θ.

We plug in the values from this question to find our answer. Firstly, note that $k = 20000$. Now, we find the geometric mean of all the five y values obtained:

$$y_{GM} = \sqrt[5]{y_1 y_2 y_3 y_4 y_5} = 108701.49037577$$

This means the geometric mean salary is just under $\$110,000$.

Now, we divide this number by k: $y_G M/k = 108701.49037577/20000 = 5.43507451879$.

Now, we take the natural logarithm of this number: $\ln(y_G M/k) = 1.69287323139$.

Finally, we take the reciprocal of this number: $1.69287323139^{-1} = 0.59071168558$.

So, for this dataset, we have $\theta = 0.59071168558$.

4. Using the data from Problem 3 and the model from Problem 2, use R to plot the likelihood function over the range $0 < \theta < 1$ using step sizes of 0.01. Turn in the plot and code. Label the maximum likelihood estimate on the plot, and confirm that it approximately matches your answer in Problem 3.

From question 2, we got that

$$L(\theta) = \theta^n k^{n\theta} \prod_{i=1}^{n} y_i^{-(\theta+1)}$$

Knowing there are 5 values for y, we get $n = 5$. We also know $k = 20000$.

So, we can rewrite this as follows:

$$L(\theta) = \frac{\theta^5 (20000^5)^\theta}{(y_1 y_2 y_3 y_4 y_5)^{\theta+1}}$$

I did not specify the value of the product of all the y as it would be a number that is long and whose exact value is only tangentially relevant here.

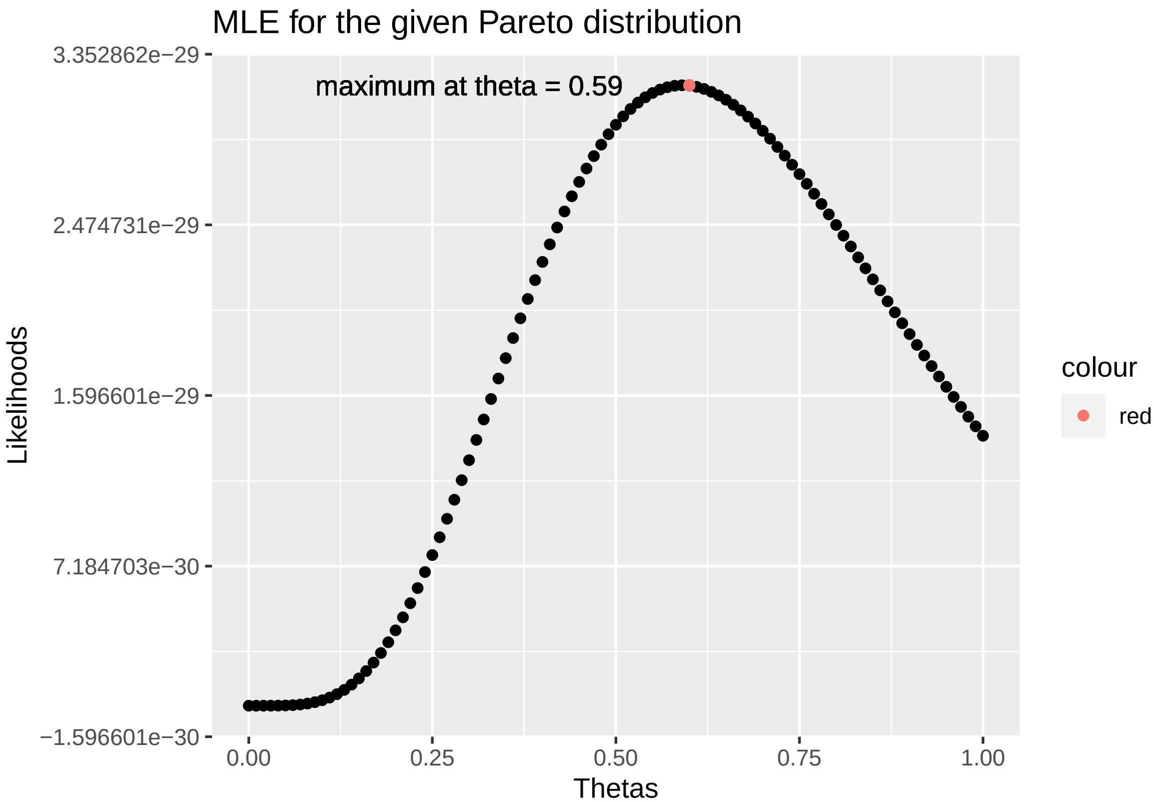
Now, we plot this function, using step sizes of 0.01 over the interval between 0 and 1 for values of θ.

Below follows my code chunk:

```r
library(tidyverse);
```

```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1

-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
y_1 <- 110501;
y_2 <- 45662;
y_3 <- 89680;
y_4 <- 1658909;
y_5 <- 20218;

k <- 20000;

y_vec <- c(y_1, y_2, y_3, y_4, y_5);

y_prod <- exp(sum(log(y_vec)));

k_fifth <- k**5

likelihood_func <- function(t)
  {(  (t**5) * (k_fifth**t)  )  /  (y_prod)**(t + 1)}

likelihoods <- numeric();
thetas_all <- numeric();


for (t in 0:100){
  theta <- t/100
  likelihood <- likelihood_func(theta);
  thetas_all <- c(thetas_all, theta);
  likelihoods <- c(likelihoods, likelihood);
}

plot_values <- data.frame(likelihoods, thetas_all);

ggplot(data = plot_values, aes(x = thetas_all, y = likelihoods)) +
  geom_point()+
  labs(title = "MLE for the given Pareto distribution",
       x = "Thetas", y = "Likelihoods") +
  geom_point(aes(which.max(likelihood_func(thetas_all))/100,
                 max(likelihood_func(thetas_all)), color = "red")) +
  geom_text(aes(which.max(likelihood_func(thetas_all))/100 - 0.3,
                max(likelihood_func(thetas_all)), label = "maximum at theta = 0.59"))
```



Clearly, the maximum height is attained by this function at $\theta = 0.59$ for the function. This approximately matches my answer from question 3, which was around $\theta = 0.5907$.