STAT 3503/8109 Lecture 9 Notes

Edoardo Airoldi Scribe: Srikar Katta*

Fall 2020: October 26, 2020

Introduction

Today, we will give a brief recap of modeling concepts. Today, we will start to discuss strategies for estimating latent variables and unknown constants. Specifically, we will discuss two estimation strategies for unknown constants in models with proper likelihoods available: maximum likelihood estimation and method of moments estimation. Under maximum likelihood, we will consider the invariance property and the idea that the maximum likelihood estimator always lives in the parameter space. We will also discuss the pros and cons of these two strategies.

Review

The likelihood, a mathematical expression in its most root for, is simply a way of representing how probable the empirical distribution (i.e., the observed data) is given a set of constants. For example, suppose we go to a shady gambling ring to try and win some money to pay off our ever-increasing pile of college debt. We see a very simple coin-flipping game: flip the coin 20 times and win \$5 by flipping heads and lose \$1 by flipping tails. Naively, one might believe this is a great deal because if this were a fair coin, after an infinite number of flips, the player would win a positive number. In other words, the theoretical distribution of flipping a heads in a single game might be a Bernoulli random variable with probability of success p. So, the final winnings would be a linear transformation of a random variable with a Binomial distribution with parameters $N = 100$ and probability of success p.

Technically, the probability of success is unknown (i.e., the probability of flipping a heads is unknown). If this were a fair coin, then it would be known, but the fairness of the coin is not guaranteed. After watching many people play and recording the game results, one notices that players *always* lose \$20 exactly. The likelihood is a way of representing the probability of perceiving this outcome given a set of constants. In this case, the known constants would be the number of games someone plays and the game outcomes. The unknown constants would be the probability of success p. The proper likelihood is the probability of the empirical distribution given the known and unknown constants. Mathematically, *Likelihood* $= \mathbb{P}($ game outcomes $|100, p)$. This section will demonstrate strategies for estimating p.

[*] Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

Maximum Likelihood Estimation

Let us motivate this section with an example.

Example 0.1. *Suppose* $x_i \overset{\text{iid}}{\sim} Normal(\mu, 1)$ *and represents the expression of gene called TPS1 in sample* i*. Assume we observe* $x_1, ..., x_n$ *(i.e., data).*

Then, the 2x2 table of this example would look as follows:

	Observed	**Unobserved**
Variable	$x_1, ..., x_n$	NA
Constant	NA	μ

And the likelihood would look as follows:

$$\begin{aligned} likelihood &= \prod_{i=1}^{n} Normal(x_i|\mu, 1) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \prod_{i=1}^{n} e^{\frac{-(x_i-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2}}, \text{ because the product of exponents is the exponent of sums.} \end{aligned}$$

The likelihood is an expression of the unknown constant μ given all the constants. Simply, the likelihood is a function of $x_1, ..., x_n, \mu$, namely the probability of $x_1, ..., x_n$ given μ. Since μ is the only unknown, this is the same as saying the likelihood is a function of μ. Suppose the histogram of our data looks as follows:

We want to *estimate* μ, an unknown constant that describes the normal distribution for the data. In statistics, the estimator is a function of the random variables denoted with a hat, so the *estimator* of μ would be $\hat{\mu}(X_1, ..., X_n)$. The *estimate* of μ on the other hand is a function of the observed data and is denoted as $\hat{\mu} = \hat{\mu}(x_1, ..., x_n)$. In other words, the estimator is a random variable and the estimate is the realization of this random variable. Suppose we randomly select a μ_1, μ_2, μ_3. Each μ is a parameter for a distribution with their own forms, drawn below:

We have three random estimates for $\hat{\mu}$. Maximum likelihood estimation will allow us to find the best estimate for this distribution. Recall, these are ways of representing the likelihood:

$$likelihood = \mathbb{P}(x_1, ..., x_n|\mu) = function(\mu)$$

And with these three guesses for μ, we want to find the *likelihood* of seeing this empirical distribution given each of μ_1, μ_2, μ_3. In other words, the likelihood can be thought of as a score of the fit of a distribution for each of the unknown constants. And we want to maximize this score. Without μ, $\frac{1}{\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2}}$ is just a mathematical expression. However, if we consider specific $\hat{\mu}$, the *likelihood*$(\hat{\mu})$ is a specific number.

Now, the maximum likelihood estimate is the value of $\hat{\mu}$ that makes $x_1, ..., x_n$ most probable. In other words, the maximum likelihood estimator is a function of the observed random variables that maximizes the likelihood. Mathematically,

$$\hat{\mu}(x_1, ..., x_n) = \hat{\mu}_{MLE} = \arg\max_{\mu} likelihood(\mu).$$

Because the likelihood is essentially a function of μ, we can plot this function and find the maximum, as seen below:

So, to recap, the maximum likelihood estimator is the argument that maximizes the likelihood of the unknown constants given the data. So, in a more general case, the maximum likelihood estimator of the unknown constants θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta} likelihood(\theta).$$

Log Likelihood

One thing to recognize when maximizing likelihoods is that the argument that maximizes the likelihood is the same argument that maximizes the log of the likelihood because the logarithmic function is an always increasing function (further details in Appendix). Additionally, because the log of products is the sum of logs, the log likelihood is often times computationally easier to compute than the likelihood, which tends to have many products. The log likelihood is simpler and easier to maximize. We denote likelihood of θ as $L(\theta)$ and the log likelihood of θ as $l(\theta)$.

Example 0.2. *Find the log likelihood of*

$$L(\mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\sum(x_i - \mu)^2}.$$

Solution: So,

$$\begin{aligned}
l(\mu) &= logL(\mu) \\
&= log\left[\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}\sum(x_i-\mu)^2}\right] \\
&= log\frac{1}{\sqrt{2\pi}} + log\left[e^{\frac{-1}{2}\sum(x_i-\mu)^2}\right] \\
&= log\frac{1}{\sqrt{2\pi}} + \frac{-1}{2}\sum(x_i-\mu)^2 \\
&= log(1) - log(\sqrt{2\pi}) + \frac{-1}{2}\sum(x_i-\mu)^2 \\
&= 0 - \frac{1}{2}log(2\pi) + \frac{-1}{2}\sum(x_i-\mu)^2 \\
&= -\frac{1}{2}[log(2)+log(\pi)] + \frac{-1}{2}\sum(x_i-\mu)^2 \\
&= -\frac{1}{2}[log(2)+log(\pi)] + \frac{-1}{2}\sum x_i^2 + \mu^2 - 2x_i\mu \\
&= -\frac{1}{2}[log(2)+log(\pi)] + \frac{-1}{2}\sum x_i^2 + \sum \mu^2 - \sum 2x_i\mu \\
&= -\frac{1}{2}[log(2)+log(\pi)] + \frac{-1}{2}\sum x_i^2 + n\mu^2 - 2\mu\sum x_i,
\end{aligned}$$

which is obviously easier to deal with analytically. The log likelihood is more desirable than the likelihood. □

Invariance of Maximum Likelihood Estimator

One key strength of maximum likelihood estimation is invariance. In short, it says that if $\hat{\theta}_{MLE}$ is the maximum likelihood estimator for $L(\theta)$, then $g(\hat{\theta}_{MLE})$ is the maximum likelihood estimator for $L(g(\theta))$.

Example 0.3. *Consider the random variables $X_1, ..., X_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$. In this case, σ^2 is unknown. What is $\hat{\sigma^2}_{MLE}$.*

Consider the following two distributions for our data; what is a better distribution to describe the data? Based on these two proposed distributions, the red distribution is a significantly better fit for the data because it follows it more closely.

Now, let us consider the likelihood function. Because we are looking for the σ^2 that maximizes the likelihood, which also maximizes the log likelihood, which ever one we consider really does not matter. While we can compute this by hand, consider the following plot of the likelihood/log likelihood over different values of σ^2:

So, $\hat{\sigma}^2_{MLE}$ is the value that maximizes the likelihood of the model in the first place. Notice, if the estimate if below 0, there is no likelihood associated with that value. That is because the likelihood estimator will follow the same support as the constant we are hoping to estimate. Since variance is always greater than or equal to 0, the maximum likelihood estimator for the variance will always be greater than or equal to 0 as well. Because the likelihood is defined over the exact parametric space that defines the model (here, the model is $x_i, ..., x_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$), the possible for the maximum likelihood estimate will always like in the same parametric space. This property of maximum likelihood is one of the reasons it is so popular.

Example 0.4. *Suppose $x_1, ..., x_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$. Find $\sigma^2 - MLE$.*

Solution: So here, we want to estimate a function of σ^2. For example, we want to estimate $\hat{\theta}_{MLE} = \frac{\sqrt{\sigma^2}}{\mu} = \frac{\sigma}{3}$. This computation is complicated. If we find the argument that maximizes $Likelihood(\sigma^2)$ known as $\hat{\sigma}^2_{MLE}$, then $\hat{\theta}_{MLE} = \frac{\sigma}{3} = \frac{\sqrt{\hat{\sigma}^2_{MLE}}}{3}$. In other words, if $\theta = \frac{\sqrt{\sigma^2}}{3}$, then $(3\theta)^2 = \sigma^2$. To find $\hat{\theta}_{MLE}$ without the invariance principle, we would have to maximize the likelihood with respect to θ: $\arg\max_\theta L(\theta)$. But we no longer have to do that computation because we can utilize the relationship between θ and σ^2, which will be the same relationship as $\hat{\theta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$. □

Example 0.5. *Suppose $x_1, ..., x_n \overset{\text{iid}}{\sim} Poisson(\lambda)$, where λ is unknown. Because x_i is from a Poisson distribution, the support of x_i is just the natural numbers (i.e., 0, 1, 2, 3,,...,) and the support of $\lambda > 0$. Find $\hat{lambda}_{MLE}$.*

Solution: First, we need to find the likelihood of this model:

$$\begin{aligned}
Likelihood(\lambda) &= \mathbb{P}(x_1, ..., x_n|\lambda) \\
&= \mathbb{P}(x_1|\lambda)*, ..., *\mathbb{P}(x_n|\lambda) \\
&= \prod_{i=1}^{n} \mathbb{P}(x_i|\lambda) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!} \\
&= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}.
\end{aligned}$$

To make the computation simpler, it makes sense to find the log liikelihood:

$$\begin{aligned}
log - Likelihood(\lambda) &= log\left[\frac{e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\right] \\
&= log\left[e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n}x_i}\right] - log\prod_{i=1}^{n}x_i! \\
&= log\left[e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n}x_i}\right] - log\prod_{i=1}^{n}x_i! \\
&= log\left[e^{-\sum_{i=1}^{n}\lambda}\right] + log\left[\lambda^{\sum_{i=1}^{n}x_i}\right] - log\prod_{i=1}^{n}x_i! \\
&= -\sum_{i=1}^{n}\lambda + \sum_{i=1}^{n}x_i log(\lambda) - \sum_{i=1}^{n}log(x_i!\,) \\
&= -n\lambda + log(\lambda)\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}log(x_i!\,).
\end{aligned}$$

Now, we can find the critical points of the log likelihood by taking the first derivative:

$$\begin{aligned}
\frac{d}{d\lambda}l(\lambda) &= \frac{d}{d\lambda} - n\lambda + log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} log(x_i!\,) \\
&= \frac{d}{d\lambda} - n\lambda + \frac{d}{d\lambda}log(\lambda)\sum_{i=1}^{n} x_i - \frac{d}{d\lambda}\sum_{i=1}^{n} log(x_i!\,) \\
&= -n\frac{d}{d\lambda}\lambda + \sum_{i=1}^{n} x_i \frac{d}{d\lambda}log(\lambda) - \frac{d}{d\lambda}\sum_{i=1}^{n} log(x_i!\,) \\
&= -n(1) + \sum_{i=1}^{n} x_i \frac{1}{\lambda} - 0,
\end{aligned}$$

because $\frac{d}{d\lambda}\lambda = 1$, $\frac{d}{d\lambda}log(\lambda) = \frac{1}{\lambda}$, and the derivative with respect to λ of a term without λ is 0. Now, by setting the derivative equal to 0, we can find the critical points:

$$\begin{aligned}
0 &= l(\lambda) \\
&= -n + \sum_{i=1}^{n} x_i \frac{1}{\lambda} \\
\iff n &= \sum_{i=1}^{n} x_i \frac{1}{\lambda} \\
\iff n\lambda &= \sum_{i=1}^{n} x_i \\
\iff \lambda^* &= \frac{\sum_{i=1}^{n} x_i}{n}.
\end{aligned}$$

Now, we have to check if λ^* is a minimum or a maximum by finding the second derivative, which is just the first derivative of the first derivative. So,

$$\begin{aligned}
\frac{d^2}{d\lambda^2}Likelihood(\lambda) &= \frac{d}{d\lambda}\frac{d}{d\lambda}Likelihood(\lambda) \\
&= \frac{d}{d\lambda} - n + \sum_{i=1}^{n} x_i \frac{1}{\lambda} \\
&= \frac{d}{d\lambda} - n + \frac{d}{d\lambda}\sum_{i=1}^{n} x_i \frac{1}{\lambda} \\
&= 0 + \sum_{i=1}^{n} x_i \frac{d}{d\lambda}\frac{1}{\lambda} \\
&= -\sum_{i=1}^{n} x_i \frac{1}{\lambda^2}.
\end{aligned}$$

Since $\lambda \geq 0$ and $x_i \geq 0$ always, the second derivative must be less than 0. So, that means λ^* is a maximum. So, $\hat{\lambda}_{MLE} = \lambda^* = \frac{\sum_{i=1}^{n} x_i}{n}$. If we were interested in the maximum of the likelihood, then we simply find $Likelihood(\hat{\lambda}_{MLE})$. □

In summary, to compute the maximum likelihood estimator for θ, an unknown constant, we follow these steps:

- Get likelihood function: $L(\theta)$
- Get log of the likelihood function: $l(\theta) = logL(\theta)$
- Find the first derivative of the log likelihood: $\frac{dl}{d\theta} = 0 \implies \hat{\theta}$
- Find second derivative of the log likelihood and check that $\hat{\theta}$ is in fact a maximum: $\frac{d^2l}{d\theta^2} < 0$
- If $\frac{d^2l}{d\theta^2} < 0$, then $\hat{\theta}_{MLE} = \hat{\theta}$.

Notes About Maximizing $l(\theta)$

First, it is important to know that we can find a global maximum, local maxima, and local/global minimum. We want to ensure that we are careful in checking our work and finding the *global* maximum. Second, it is possible for the maximum to live on the boundary, for which we will not observe as maximum. For example, suppose the likelihood looks as follows:

In this situation, because the global maximum is at a boundary point (1), the maximum likelihood estimator will return the *local* maximum instead. Additionally, if a practitioner did not check the second derivative of the likelihood, then they might use the *minimum* instead of the maximum because the first derivative of the likelihood is 0 at both points.

Another convenient feature of maximum likelihood estimation is that the the maximum likelihood estimator will always exist in the parameter estimate of the value it is estimating. For example, if the unknown constant was a variance (which is always greater than or equal to 0), the maximum likelihood estimator will always be greater than or equal to 0. Even though MLE might seem somewhat formulaic, it still requires care or else someone could fall into these problems.

Method of Moments Estimator (MOME)

When computing the likelihood, it may be challenging for more complex problems, especially if done by hand. However, computing the “moment” is not. The moments are the same as those in probability theory: the mean is the first central moment and the variance is the second central moment. These are typically denoted as follows:

Name	Uncentral Moments	Central Moments
Mean	$\mu_1 = \frac{1}{n}\sum x_i$	$\mu_1 = \frac{1}{n}\sum x_i$
Variance	$\mu_2 = \frac{1}{n}\sum x_i^2$	$\sigma^2 = \frac{1}{n}\sum(x_i - \mu_1)^2$

In some situations, we may be unable to compute the whole likelihood, but we will at least be able to compute the empirical moments, which are going to be the method of moments estimators. To find the method of moment estimators, we equate the first theoretical moment to the first empirical moment and the second central theoretical moment to the second empirical moment. So, with many constants, we set up as many systems as we unknown constants. And we continuously equate the empirical moment to the corresponding theoretical moment. In general, this is the following approach for find MOME:

1. Equate empirical and theoretical moments together

2. Represent the data as a function of unknowns. For example $f(x_1, ..., x - n) \sim g(\mu, \sigma^2)$
3. Set up a system of equations where each function of our data corresponds to a different function for the unknowns
4. Solve the system for unknown constants

Let us demonstrate these ideas with examples.

Example 0.6. *Suppose* $x_1, ..., x_n \overset{\text{iid}}{\sim} Normal(\mu, 1)$*. Find* $\hat{\mu}_{MOME}$

Solution: We can find the empirical first moment of our data:

$$\mathbb{E}[X] = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\implies \hat{\mu}_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

□

Let's consider a slightly more complicated example:

Example 0.7. *Suppose* $x_1, ..., x_n \overset{\text{iid}}{\sim} Normal(\mu, \sigma^2)$*. Suppose* μ, σ^2 *are unknown. Find* $\hat{\mu}_{MOME}$ *and* $\hat{\sigma}^2_{MOME}$*.*

Solution: Recall that the first moment (mean) is $\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the second central moment (variance) is $\mathbb{V}[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)$. We want to estimate μ and σ^2, so the method of moments estimators for each of these will be $\hat{\mu}_{MOME} = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$ and $\hat{\sigma}^2_{MOME} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i) = \sigma^2$.

Because we have multiple moments, we can also set this up a system of equations. For example, the first empirical moment $\frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[X] = \mu$ and the second empirical moment $\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \mathbb{E}[X^2] = \sigma^2 + \mu^2$. We do not know what σ^2 is but we can find $\hat{\mu}_{MOME}$ very easily and use that as our estimate to find σ^2. So $\hat{\mu}_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, $\hat{\sigma}^2_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}_{MOME}$. □

Drawbacks of MOME

The method of moments, while simple, does not involve the likelihood, which is a function defined on the parametric space. Our MOME has no such relationship, so it does not have to exist in the support of the unknown constants we hope to estimate.

Example 0.8. *Suppose* $x_1, ..., x_n \overset{\text{iid}}{\sim} Binomial(\theta, N)$ *where* θ *is the probability of success and* N *is the number of independent Bernoulli trials. Suppose* θ, N *are both unknown but we observe* $x_1, ..., x_n$*. Estimate* θ, N*.*

Solution: We know that the first moment for a Bernoulli random variable is $N\theta$. So, our theoretical first moment is $\mathbb{E}[X] = N\theta$. And our empirical first moment is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Additionally, the second central theoretical moment is $\mathbb{V}[X] = N\theta(1-\theta)$, and the second central empirical moment is $S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2$. Then, we can set up a system of equations by equating the theoretical and empirical moments together. So,

$$\begin{aligned} N\theta &= \bar{X}_n \\ N\theta(1-\theta) &= S_n. \end{aligned}$$

Then, we can solve for $\hat{\theta}$: $\hat{\theta} = \frac{\bar{X}_n}{N}$. And using our second moment, we can find N:

$$\begin{aligned} S_n &= N\theta(1-\theta) \\ &= N\left(\frac{\bar{X}_n}{N}\right)\left(1-\frac{\bar{X}_n}{N}\right) \\ &= \left(\bar{X}_n\right)\left(\frac{N-\bar{X}_n}{N}\right) \\ \iff NS_n &= N\bar{X}_n - \bar{X}_n^2 \\ \iff \bar{X}_n^2 &= N\bar{X}_n - NS_n \\ \iff \bar{X}_n^2 &= N(BarX_n - S_n) \\ \iff \frac{\bar{X}_n^2}{(BarX_n - S_n)} &= N. \end{aligned}$$

So, $\hat{\theta}_{MOME} = \frac{\bar{X}_n}{N}$ and $\hat{N}_{MOME} = \frac{\bar{X}_n^2}{(BarX_n - S_n)}$. So, we can plug N into

$$\begin{aligned} \hat{\theta}_{MOME} &= \frac{\bar{X}_n}{N} \\ &= \bar{X}_n \frac{(BarX_n - S_n)}{\bar{X}_n^2} \\ &= \frac{(BarX_n - S_n)}{\bar{X}_n}. \end{aligned}$$

So, the final solution reads as follows:

$$\begin{aligned} \hat{\theta}_{MOME} &= \frac{(BarX_n - S_n)}{\bar{X}_n} \\ \hat{N}_{MOME} &= \frac{\bar{X}_n^2}{(BarX_n - S_n)}. \end{aligned}$$

□

MOME Issues

The MOME has one key issue: its estimators do not follow the same parametric space as the unknown constants, which could lead to wrong analyses. For example, under the Binomial distribution, $0 \leq \theta \leq 1$, so $0 \leq 1-\theta \leq 1$. So, $N\theta > N\theta(1-\theta)$. However, let us sample $y_1, ..., y_n \overset{\text{iid}}{\sim} Binomial(N = 10, \theta = 0.1)$:

Sample 1: $x_1^{(1)}, ..., x_n^{(1)} \to \bar{X}_n^{(1)}, S_n^{(1)} \to \hat{N}_{MOME}^{(1)}$
Sample 2: $x_1^{(2)}, ..., x_n^{(2)} \to \bar{X}_n^{(2)}, S_n^{(2)} \to \hat{N}_{MOME}^{(2)}$
$, ...,, ...,$ Sample S: $x_1^{(S)}, ..., x_n^{(S} \to \bar{X}_n^{(S)}, S_n^{(S)} \to \hat{N}_{MOME}^{(S)}$
$, ...,, ...,$

Each of the empirical means and variances from these samples will lead to different values of $\hat{N}_{MOME}^{(S)}$, so $\bar{X}_n^{(S)} > S_n^{(S)}$ does not necessarily need to hold. Then, if $\bar{X}_n < S_n$ (the empirical mean is less than the theoretical variance which is possible), then $\hat{N}_{MOME} < 0$, which is not possible. So, the estimator may fail. However, if N is very large and the data is really sampled from the theoretical distribution, then MOME gives a good estimate. Now, imagine if the Binomial distribution is mis-specified out of convenience: then MOME may be incorrect.

This leads to a discussion of different samples. Often times, in estimation, people say there are empirical moments computed directly from the data and the theoretical moments. However, in practice, there are empirical moments that summarize the data, true theoretical moments from the true DGP that are accessible, and the model specified theoretical moments from the model/DGP approximation. In reality, the specified model may not be the true DGP but hopefully approximates the true DGP; if the specified model is not properly specified, there are many issues that may arise. So, careful precision is required in MOME.

Conclusion

Today, we started discussing estimation strategies, namely maximum likelihood estimation and method of moments estimation. Both of these models require care and precision, especially in specifying the model. While defining the model is not discussed here, understanding whether the proposed theoretical distributions and models are apt is incredibly important and should be taken into consideration in any data science discussion.

Appendix

Fact 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function (takes real-valued inputs and spits out real valued inputs). If x^* maximizes $f(x)$, then x^* also maximizes $log(f(x))$.*

Proof. Let f be a real valued function and let x^* be the argument that maximizes f. Consider the logarithmic function: it is continuously increasing. So, in other words if $a > b$, $log(a) > log(b)$. By definition, $f(x^*) \geq f(x)$ for all real numbers x (definition of $\arg\max$). So, $f(x^*) > f(x)$ implies $log(f(x^*)) \geq log(f(x))$ for all real numbers x. So,

$$\arg\max_{x} f(x) \equiv \arg\max_{x} log(f(x)).$$

□

Note About Invariance

Suppose σ^2 is some unknown constant and $\theta = \frac{\sigma^2}{3}$. Then, if we found $\hat{\sigma}^2_{MLE}$, by the invariance property, we have also found $\hat{\theta}_{MLE} = \frac{\hat{\sigma}^2_{MLE}}{3}$. Note however, that this is *not* the same as the transformation theorem. In this situation, a constant is transformed into another constant whereas in transformation theorem, one *random variable* is transformed into another random variable.

In this situation, because the global maximum is at a boundary point (1), the maximum likelihood estimator will return the *local* maximum instead. Additionally, if a practitioner did not check the second derivative of the likelihood, then they might use the *minimum* instead of the maximum because the first derivative of the likelihood is 0 at both points.

Another convenient feature of maximum likelihood estimation is that the the maximum likelihood estimator will always exist in the parameter estimate of the value it is estimating. For example, if the unknown constant was a variance (which is always greater than or equal to 0), the maximum likelihood estimator will always be greater than or equal to 0. Even though MLE might seem somewhat formulaic, it still requires care or else someone could fall into these problems.

Method of Moments Estimator (MOME)

When computing the likelihood, it may be challenging for more complex problems, especially if done by hand. However, computing the "moment" is not. The moments are the same as those in probability theory: the mean is the first central moment and the variance is the second central moment. These are typically denoted as follows:

Name	Uncentral Moments	Central Moments
Mean	$\mu_1 = \frac{1}{n}\sum x_i$	$\mu_1 = \frac{1}{n}\sum x_i$
Variance	$\mu_2 = \frac{1}{n}\sum x_i^2$	$\sigma^2 = \frac{1}{n}\sum(x_i - \mu_1)^2$

In some situations, we may be unable to compute the whole likelihood, but we will at least be able to compute the empirical moments, which are going to be the method of moments estimators. To find the method of moment estimators, we equate the first theoretical moment to the first empirical moment and the second central theoretical moment to the second empirical moment. So, with many constants, we set up as many systems as we unknown constants. And we continuously equate the empirical moment to the corresponding theoretical moment. In general, this is the following approach for find MOME:

1. Equate empirical and theoretical moments together
2. Represent the data as a function of unknowns. For example $f(x_1, ..., x - n) \sim g(\mu, \sigma^2)$
3. Set up a system of equations where each function of our data corresponds to a different function for the unknowns
4. Solve the system for unknown constants

Let us demonstrate these ideas with examples.

Example 0.9. *Suppose* $x_1, ..., x_n \overset{\text{iid}}{\sim} Normal(\mu, 1)$. *Find* $\hat{\mu}_{MOME}$

Solution: We can find the empirical first moment of our data:

$$\mathbb{E}[X] = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\implies \hat{\mu}_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

□

Let's consider a slightly more complicated example:

Example 0.10. *Suppose $x_1, ..., x_n \overset{\text{iid}}{\sim} Normal(\mu, \sigma^2)$. Suppose μ, σ^2 are unknown. Find $\hat{\mu}_{MOME}$ and $\hat{\sigma}^2_{MOME}$.*

Solution: Recall that the first moment (mean) is $\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the second central moment (variance) is $\mathbb{V}[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)$. We want to estimate μ and σ^2, so the method of moments estimators for each of these will be $\hat{\mu}_{MOME} = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$ and $\hat{\sigma}^2_{MOME} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i) = \sigma^2$.

Because we have multiple moments, we can also set this up a system of equations. For example, the first empirical moment $\frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[X] = \mu$ and the second empirical moment $\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \mathbb{E}[X^2] = \sigma^2 + \mu^2$. We do not know what σ^2 is but we can find $\hat{\mu}_{MOME}$ very easily and use that as our estimate to find σ^2. So $\hat{\mu}_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, $\hat{\sigma}^2_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}_{MOME}$. □

Drawbacks of MOME

The method of moments, while simple, does not involve the likelihood, which is a function defined on the parametric space. Our MOME has no such relationship, so it does not have to exist in the support of the unknown constants we hope to estimate.

Example 0.11. *Suppose $x_1, ..., x_n \overset{\text{iid}}{\sim} Binomial(\theta, N)$ where θ is the probability of success and N is the number of independent Bernoulli trials. Suppose θ, N are both unknown but we observe $x_1, ..., x_n$. Estimate θ, N.*

Solution: We know that the first moment for a Bernoulli random variable is $N\theta$. So, our theoretical first moment is $\mathbb{E}[X] = N\theta$. And our empirical first moment is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Additionally, the second central theoretical moment is $\mathbb{V}[X] = N\theta(1-\theta)$, and the second central empirical moment is $S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2$. Then, we can set up a system of equations by equating the theoretical and empirical moments together. So,

$$N\theta = \bar{X}_n$$
$$N\theta(1-\theta) = S_n.$$

Then, we can solve for $\hat{\theta}$: $\hat{\theta} = \frac{\bar{X}_n}{N}$. And using our second moment, we can find N:

$$\begin{aligned}
S_n &= N\theta(1-\theta) \\
&= N\left(\frac{\bar{X}_n}{N}\right)\left(1-\frac{\bar{X}_n}{N}\right) \\
&= (\bar{X}_n)\left(\frac{N-\bar{X}_n}{N}\right) \\
\iff NS_n &= N\bar{X}_n - \bar{X}_n^2 \\
\iff \bar{X}_n^2 &= N\bar{X}_n - NS_n \\
\iff \bar{X}_n^2 &= N(BarX_n - S_n) \\
\iff \frac{\bar{X}_n^2}{(BarX_n - S_n)} &= N.
\end{aligned}$$

So, $\hat{\theta}_{MOME} = \frac{\bar{X}_n}{N}$ and $\hat{N}_{MOME} = \frac{\bar{X}_n^2}{(BarX_n - S_n)}$. So, we can plug N into

$$\begin{aligned}
\hat{\theta}_{MOME} &= \frac{\bar{X}_n}{N} \\
&= \bar{X}_n\frac{(BarX_n - S_n)}{\bar{X}_n^2} \\
&= \frac{(BarX_n - S_n)}{\bar{X}_n}.
\end{aligned}$$

So, the final solution reads as follows:

$$\begin{aligned}
\hat{\theta}_{MOME} &= \frac{(BarX_n - S_n)}{\bar{X}_n} \\
\hat{N}_{MOME} &= \frac{\bar{X}_n^2}{(BarX_n - S_n)}.
\end{aligned}$$

□

MOME Issues

The MOME has one key issue: its estimators do not follow the same parametric space as the unknown constants, which could lead to wrong analyses. For example, under the Binomial distribution, $0 \leq \theta \leq 1$, so $0 \leq 1-\theta \leq 1$. So, $N\theta > N\theta(1-\theta)$. However, let us sample $y_1, ..., y_n \overset{\text{iid}}{\sim} Binomial(N = 10, \theta = 0.1)$:

$$\begin{aligned}
&\text{Sample 1: } x_1^{(1)}, ..., x_n^{(1)} \to \bar{X}_n^{(1)}, S_n^{(1)} \to \hat{N}_{MOME}^{(1)} \\
&\text{Sample 2: } x_1^{(2)}, ..., x_n^{(2)} \to \bar{X}_n^{(2)}, S_n^{(2)} \to \hat{N}_{MOME}^{(2)} \\
&\qquad ...... \\
&\text{Sample S: } x_1^{(S)}, ..., x_n^{(S} \to \bar{X}_n^{(S)}, S_n^{(S)} \to \hat{N}_{MOME}^{(S)} \\
&\qquad ......
\end{aligned}$$

Each of the empirical means and variances from these samples will lead to different values of $\hat{N}_{MOME}^{(S)}$, so $\bar{X}_n^{(S)} > S_n^{(S)}$ does not necessarily need to hold. Then, if $\bar{X}_n < S_n$ (the empirical mean is less than the theoretical variance which is possible), then $\hat{N}_{MOME} < 0$, which is not possible. So, the estimator may fail. However, if N is very large and the data is really sampled from the theoretical distribution, then MOME gives a good estimate. Now, imagine if the Binomial distribution is mis-specified out of convenience: then MOME may be incorrect.

This leads to a discussion of different samples. Often times, in estimation, people say there are empirical moments computed directly from the data and the theoretical moments. However, in practice, there are empirical moments that summarize the data, true theoretical moments from the true DGP that are accessible, and the model specified theoretical moments from the model/DGP approximation. In reality, the specified model may not be the true DGP but hopefully approximates the true DGP; if the specified model is not properly specified, there are many issues that may arise. So, careful precision is required in MOME.

Conclusion

Today, we started discussing estimation strategies, namely maximum likelihood estimation and method of moments estimation. Both of these models require care and precision, especially in specifying the model. While defining the model is not discussed here, understanding whether the proposed theoretical distributions and models are apt is incredibly important and should be taken into consideration in any data science discussion.

Appendix

Fact 2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function (takes real-valued inputs and spits out real valued inputs). If x^* maximizes $f(x)$, then x^* also maximizes $log(f(x))$.*

Proof. Let f be a real valued function and let x^* be the argument that maximizes f. Consider the logarithmic function: it is continuously increasing. So, in other words if $a > b$, $log(a) > log(b)$. By definition, $f(x^*) \geq f(x)$ for all real numbers x (definition of $\arg\max$). So, $f(x^*) > f(x)$ implies $log(f(x^*)) \geq log(f(x))$ for all real numbers x. So,

$$\arg\max_x f(x) \equiv \arg\max_x log(f(x)).$$

□

Note About Invariance

Suppose σ^2 is some unknown constant and $\theta = \frac{\sigma^2}{3}$. Then, if we found $\hat{\sigma}^2_{MLE}$, by the invariance property, we have also found $\hat{\theta}_{MLE} = \frac{\hat{\sigma}^2_{MLE}}{3}$. Note however, that this is *not* the same as the transformation theorem. In this situation, a constant is transformed into another constant whereas in transformation theorem, one *random variable* is transformed into another random variable.