1) Illustrate the Topic

1.1) What is Maximum Likelihood Estimation (MLE)

In econometrics, maximum likelihood estimation is a parameter estimation technique through maximizing a likelihood function given an observed data.

1.2) What is a likelihood function?

A likelihood function is the joint probability of a dataset which shows the probability of a particular situation to be realized in a sample space. Different from the probability distribution function (PDF), it highlights the joint density rather than the probability of only one event in a dataset.

We may need to dive deeper into the concept of joint probability in order to understand likelihood function better. Joint probability expresses the probability of multiple independent events. Joint probability tracks each data point observation and treats the observation in the collected dataset as a series of events. This is how joint probability is represented;

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times f(x_n|\theta) = \prod_{i}^{n} f(x_i|\theta)$$

Mathematical notation for likelihood varies; $f_n(y|\theta) = \mathscr{L}_n(y|\theta) = \mathscr{L}_n(\theta)$. In the notation, $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ stands for vector for joint distribution of parameters, $x = (x_1, x_2, \ldots, x_n)$ stands for sample data and $i = (1,2,3,\ldots,n)$ stands for each observation in the dataset.

Parameters we are dealing with here are important elementary units of probability density function (PDF), and parametrization is the act of inputting and/or playing with some of the features that define and shape a probability distribution. For instance, parametrization for normal distribution is mean and standard deviation whereas beta distribution is parametrized with two shape parameters that define the shape of the distribution.

1.3) What is the difference between likelihood function and probability distribution function?

Both likelihood function and probability distribution function (PDF) serve for similar purposes but there are differences between two functions. PDF is used for demonstrating the probability given the parameters. Hence, parameters are known in a PDF. On the other hand, a likelihood function shows the likelihood of parameters taking place given the data. Thus, in a likelihood function, parameter values are unknown.

1.4) Why is the natural log of the likelihood function taken?

Naturally, the joint distribution function is burdensome to deal with since it is an exponential function which implies its value is volatile/sensitive to changes in the sample size. For that reason, we take the natural logarithm of the function. By this way, our function becomes a monotonically increasing function; a property of which provides that the maximum point of the function realizes at the same point (i.e., value of θ^*) with the original model.

2) Details

2.1) What is MLE in detail?

Maximum likelihood estimation (MLE) is a technique to estimate the parameters of a distribution. It estimates the values of the parameters which maximize the likelihood function within the possible parameter space. Expressing it with notations, θ is the value that maximizes the likelihood; $\mathscr{L}_n$.

Further, i.i.d assumptions are needed for MLE. In other words, the data should be independently and identically distributed. Specifically, i.i.d assumptions obligate that none of the data points are dependent on another data point (i.e., independence) and the data points should be taken from the same distribution source (identicality, indistinguishability).

There are some other preliminaries for utilizing MLE. Firstly, a data generating process is required and also one needs to be able to derive the corresponding likelihood function. Data generating process is a process that leads the dataset to come into existence and in the end, some kind of data distribution is found suitable to the dataset, such as Normal distribution, Poisson Distribution, etc.

2.2) MLE derivation

To better understand the derivation, we will make use of an intuitive example. Say we have some continuous data in our hands yet we don’t know what kind of distribution it follows. Hence, we would like to fit a distribution to it. As we don’t know what kind of distribution our data follows (Poisson, Exponential, Normal, etc.), we need to find and use a distribution function which fits our data well.

For that, one may use fitting methods and measure how good the fit is. Besides, one may also go traditional and benefit from the fact that, most of the time, there exists a distribution which is most likely to fit our type of data best. Therefore, that distribution is commonly used for such types of data. To instantiate, Gaussian Distribution is commonly used for features like height or temperature whereas

Exponential Distribution for features in regards to duration of phone calls, bacterial population growth, and Poisson Distribution for quantity of houses purchased within a specific period, etc.

As this is not within the scope of our topic, we will assume that we already have found the best distribution for our dataset. Note that we may also have collected our data drawn from a known distribution but then we may still not know the parameters that determine its nature.

Therefore, once a distribution is found appropriate and suitable for our dataset, we shall estimate the specific parameters of the chosen distribution that would best fit our data. This is where MLE comes into play.

With the aim of making the explanation intuitive and uncomplicated, we will assume that our data follows a Normal Distribution (i.e., Gaussian Bell Curve). This assumption is often called the "normality assumption". Thereby, the parameters to be estimated will be mean (μ) and standard deviation (σ). And what we estimate would determine how loose the curve is and at what point exactly the peak of the curve realizes.

The parameters μ (mean) and σ (standard deviation) are usually indicated as a set of parameters. We name that set θ;

$$\theta = \{\mu, \sigma\}$$

Since we are dealing with a continuous probability distribution, the probability of observing a specific data point observation is equal to 0. We can verify this very easily. We can put an infinite number of points in the domain, which would imply a sample size of infinity, that will be used in the denominator of any probability expression for any continuous distribution. Thus, the probability of any specific data point would be equal to 0.

For that matter, we need to deal with probability density rather than probability itself because in a continuous domain, probability density is analogous (similar) to probability in a discrete sample space. Using probability density in MLE derivation;

$$f(x_1, x_2,..., x_n|\theta)$$

In the MLE of the parameters of a normally distributed dataset, we try to maximize the probability density of the data as a function of θ. In the maximization process, we need to consider θ as our independent variable instead of our observed data. This is different from what we are used to seeing in common optimization problems. In this case, each observation x_i can be treated as constant given that observed data is unchanging. Hence, we can find the optimal θ by maximizing the joint probability density. Herein, argmax function of this joint probability density function with respect to θ will be sought:

$$\widehat{\theta}_{\text{MLE}} = \text{argmax}_{\theta} \prod_{i}^{n} f(\text{x}_{\text{i}}|\theta)$$

As a side note, argmax function is a mathematical operation which finds the argument for the maximum values of the target function. We use argmax instead of max function because bear in mind that there may be multiple optimal θ values. In other words, there may be multiple θ values that give the same maximum value for a joint probability density function.

As we are trying to find the maximum value, we should take the derivative with respect to θ and look for what θ is equal to when the first derivative is equal to 0. Observe that argmax of joint probability density basically implies the first derivative of the joint probability density with respect to θ being equal to 0;

$$\text{argmax}_{\theta} \prod_{i}^{n} f(\text{x}_{\text{i}}|\theta) \rightarrow \frac{\partial}{\partial\theta}\prod_{i}^{n} f(\text{x}_{\text{i}}|\theta) = 0$$

This is a complex derivative to find a solution for. Herein, we take advantage of the natural logarithm and its properties in order to make the calculations simpler for this equation. Taking advantage of one of the logarithmic properties, we can do the following modification:

$$ln(\prod f) \;=\; \sum ln(f)$$

As the logarithmic transformation is utilized, the function is now treated as a monotonic function (i.e., the value of y always goes up whenever the value of x increases). In this way, the maximum value of the logarithmic function and the maximum value of the original function will be different yet occur at the same value of the independent variable.

Let us write down the problem again. The proof of the third and fourth equations are omitted;

$$\frac{\partial}{\partial\theta}\prod_{i}^{n} f(x_{i}|\theta) \sim \frac{\partial}{\partial\theta} ln\langle\prod_{i}^{n} f(x_{i}|\theta)\rangle$$

$$= \frac{\partial}{\partial\theta}\sum_{i}^{n} ln\langle f(x_{i}|\theta)\rangle \;=\; \sum_{i}^{n}\frac{\partial}{\partial\theta} ln\langle f(x_{i}|\theta)\rangle \;=\; 0$$

We want to find the optimal θ (i.e., optimal parameters) that would fit our observed data. To solve with respect to the mentioned two parameters, we will switch to gradient notation.

Gradient of a scalar-valued (associates a single number for every point in a space) function f is the vector-valued (i.e., a function of one or more variables whose range is a set of multidimensional vectors) function where its value at a particular point indicates the direction and the rate of the fastest increase. At a point p where the gradient of f is not equal to zero, the direction of the gradient is the one that function goes up most quickly from the point p. Further, the magnitude of a gradient is the rate of increase in that direction.

The gradient of a function $f(x)$ is represented below:

$$df/dx \ = \ \nabla f$$

Switching to gradient notation:

$$\sum_i^n \frac{\partial}{\partial\theta} ln\langle f(x_i|\theta)\rangle \ = \sum_i^n \nabla_{\mu,\sigma} ln\langle f(x_i|\mu,\sigma)\rangle \ = \ 0$$

Beginning with taking the gradient of the function w.r.t. μ, we can replace $f(x_i|\mu,\sigma)$ with the PDF of the normal distribution;

$$\sum_i^n \nabla_\mu ln\langle \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \rangle \ = \ 0$$

Then, we make use of the natural logarithm in order to simplify the calculations;

$$\sum_i^n \nabla_\mu ln\langle \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \rangle \ = \ \sum_i^n \nabla_\mu -\frac{1}{2} ln(2\pi\sigma^2) \ - \ \frac{(x_i-\mu)^2}{2\sigma^2}$$

Continues as;

$$\sum_i^n \nabla_\mu - \frac{1}{2} ln(2\pi\sigma^2) \ - \ \frac{(x_i-\mu)^2}{2\sigma^2} \ = \ -\frac{1}{2\sigma^2}\sum_i^n \nabla_\mu (x_i - \mu)^2 \ = \ -\frac{1}{2\sigma^2}\sum_i^n - 2(x_i - \mu) = \frac{1}{\sigma^2}\sum_i^n (x_i - \mu)$$

Equating the ultimate expression to zero, we obtain the estimator for the optimal μ.

$$\frac{1}{\sigma^2}\sum_i^n (x_i - \mu) \ = \ 0 \rightarrow \widehat{\mu}_{MLE} = \frac{1}{n}\sum_i^n x_i$$

Observe that the estimator of optimal μ ($\widehat{\mu}_{MLE}$) is independent from the other parameter, σ.

Now, we take the gradient with respect to σ in order to solve for the optimal σ.

$$\sum_i^n \nabla_\sigma ln\langle \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \rangle \ = \sum_i^n \nabla_\sigma - \frac{1}{2} ln(2\pi\sigma^2) \ - \ \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$=-\frac{n}{2}\nabla_{\sigma} ln\sigma^{2} - \frac{1}{2}\nabla_{\sigma}\langle\frac{1}{2\sigma^{2}}\sum_{i}^{n}(x_{i} - \mu)^{2}\rangle =-\frac{n}{\sigma} + \frac{1}{\sigma^{3}}\sum_{i}^{n}(x_{i} - \mu)^{2}$$

Equating the last term to zero will give the estimator for the optimal σ.

$$-\frac{n}{\sigma} + \frac{1}{\sigma^{3}}\sum_{i}^{n}(x_{i} - \mu)^{2} = 0$$

$$\sum_{i}^{n}(x_{i} - \mu)^{2} = n\sigma^{2} \rightarrow \hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i}^{n}(x_{i} - \mu)^{2}}$$

What we found for these two parameters are exactly the same formulas of the general mean and standard deviation formulas. This outcome is not a result of coincidence but a result of i.i.d. assumptions we assumed previously.

Note that these values should not be perceived only as the mean and the standard deviation of the dataset, but also as the parameters of the Normal Distribution that have the largest likelihood to fit a dataset.

2.3) Why are we dealing with maximum likelihood rather than maximum probability?

As we discussed in section 1.2, probability and likelihood are not the same notion although what they express are similar. We can show the difference between them using the following equation;

$$L(\mu, \sigma|data) = P(data|\mu, \sigma)$$

This identity holds for any number of parameters and for any distribution. The proof of this identity is omitted. The right hand side expresses that probability density of data given its mean (μ) and its standard deviation (σ) and the left hand side represents the likelihood of the parameters (mean and standard deviation) taking certain values given the observed data. We can see the similarities and differences of these two expressions by looking at this equation. Simply put, the probability function requires the data as its input whereas the likelihood function requires the values of the parameters as its input.

2.4) Properties of MLE

Maximum likelihood estimator is normally distributed, consistent and efficient when large samples are used. In the case of small samples, it is a function of a sufficient number of statistics. In other words, it doesn't require any additional statistics to bring any further information about the sample. Further, with small samples, MLE is invariant and, in specific cases, is unbiased and unique.

When applied on finite samples, MLE doesn't have optimum properties and thus, other estimating methods may prevail on estimating the accurate parameter values. However, as sample size goes to infinity, MLE possesses some properties;

The first one is consistency; as the number of observations goes to infinity, the value of estimator $\widehat{\theta}$ converges to the true value, that is, θ.

Consistency also has two sufficient conditions; identifiability of the model and compactness of the parameter space. Identifiability expresses that it’s possible to detect true parameter values of a model given that we can possess infinitely many observations of this model. Compactness is a property of a subset that implies the subset is closed and bounded.

The second property is functional equivariance. As mentioned above, MLE tries to find the largest possible joint probability by picking parameter values. However, in the case that the parameter contains several components, we need to characterize different MLEs for each component and these MLEs together will determine the whole parameter.

The last one is efficiency. As the sample size approaches infinity, MLE would give a lower mean squared error than any other consistent estimator.

2.5) When is least squares minimization the same as maximum likelihood estimation?

One other method that is used commonly is the Least Squares Minimization (LSM). It is true that if one assumes that their dataset follows a normal distribution, the MLE estimations and the LSM estimations will be equivalent. However, even if both of the techniques give equivalent results, they have some differences in approaching the problem. In Least Squares Minimization, we try to minimize the total sum of distances between data points/observations and the fitted curve.

Given the normality assumption just as we did in our MLE derivation, one can identify the parameters that would give the maximum likelihood by locating the distribution’s mean as close as possible to as many observations as possible. Bearing in mind that the Normal Distribution is symmetrical, one can notice that the minimization of the distances between the mean value and the data points would yield the same estimators and thereby the same estimates. Note that this is a specific example where the dataset is assumed to follow as Normal Distribution. Needless to say, these two methods do not necessarily yield the same estimators.

3) Application of MLE on Panel Data

We use the -helpml- command to get information about the built-in MLE functions in Stata 17. We possess a time-series cross-sectional dataset of 18 countries for the period 1960-1978.
The following generalized command will be used to conduct MLE:

```
xtmixed dependent_variable independent_variables || grouping_variable: , mle
```

In our example, “independent_variables” are the variables that we use for estimation whereas “dependent_variable” is the variable that we’d like to estimate given the independent variables. Apart from these, note that “grouping_variable” classifies the groups in the dataset. The “||” sign indicates that the random effects model is estimated for each “grouping_variable” and “mle” expression points out that the model is estimated with Maximum Likelihood Estimation.
In our case, the command becomes the following:

```
xtmixed LGASPCAR LINCOMEP LRPMG LCARPCAP || numerical_country: , mle
```

This command performs a gradient-based optimization. At the bottom of the output, one can observe the Likelihood Ratio (LR) test. The test statistic turned out significant, which implies MLE yields better results than OLS.

We store this as MLE_RE_Amemiya_M1 and then output a table of the estimation along with other information about the estimation.

REFERENCES

Brooks-Bartlett, Jonny. “Probability Concepts Explained: Introduction.” *Medium*, Towards Data Science, 6 Jan. 2018, https://towardsdatascience.com/probability-concepts-explained-introduction-a7c0316de465.

Brooks-Bartlett, Jonny. “Probability Concepts Explained: Maximum Likelihood Estimation.” *Medium*, Towards Data Science, 31 Jan. 2018, https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1.

Eppes, Marissa. “Maximum Likelihood Estimation Explained - Normal Distribution.” *Medium*, Towards Data Science, 21 Sept. 2019, https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f.

“Maximum Likelihood Estimation.” *Wikipedia*, Wikimedia Foundation, 30 Apr. 2023, https://en.wikipedia.org/wiki/Maximum_likelihood_estimation.

Naqvi, Asjad. “Maximum Likelihood Estimation (MLE).” *Medium*, The Stata Guide, 5 July 2021, https://medium.com/the-stata-guide/maximum-likelihood-estimation-mle-88b869158a7d.

Taskesen, Erdogan. “How to Find the Best Theoretical Distribution for Your Data.” *Medium*, Towards Data Science, 30 Mar. 2023, https://towardsdatascience.com/how-to-find-the-best-theoretical-distribution-for-your-data-a26e5673b4bd.

Wilkinson, Philip. “Maximum Likelihood Estimation and OLS Regression.” *Medium*, Towards Data Science, 11 Feb. 2021, https://towardsdatascience.com/maximum-likelihood-estimation-and-ols-regression-36c049c94a48.