The Method of Maximum Likelihood

References

"Econometrics", Hayashi, F., *Princeton University Press*, 2000.

"Econometric theory and methods", Davidson, R., and MacKinnon, J., *Oxford University Press*, 2004.

1 Introduction

The method of moments is not the only fundamental principle of estimation, even though the estimation methods for regression models discussed up to this point (ordinary, generalised least squares, and instrumental variables) can all be derived from it. In this note we introduce another fundamental method of estimation, namely, the method of maximum likelihood. For regression models, if we make the assumption that the error terms are normally distributed, then the maximum likelihood (ML) estimators coincide with the various least squares estimators with which we are already familiar with. So why bother with maximum likelihood? Because ML estimators can also be applied to an extremely wide variety of models other than regression models, and they generally have excellent asymptotic properties. The major disadvantage of ML estimation is that it requires stronger distributional assumptions than does the method of moments.

1.1 A word on extremum estimators

Extremum estimators include least squares, generalised method of moments (GMM), and ML as special cases. Considering them in this unified approach is useful for bringing out the common structure that underlies these apparently diverse estimation principles.

An estimator $\hat{\boldsymbol{\theta}}$ is called an extremum estimator if there is a scalar objective function $Q_n(\boldsymbol{\theta})$ such that

$$\hat{\boldsymbol{\theta}} \text{ maximises } Q_n(\boldsymbol{\theta}) \text{ subject to } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p, \tag{1}$$

where Θ, called the parameter space, is the set of possible parameter values. Here, we restrict our attention to the case where Θ is a subset of the finite dimensional Euclidean space $\mathbb{R}^p$. The objective function $Q_n(\boldsymbol{\theta})$ depends not only on $\boldsymbol{\theta}$ but also on the sample or the data, $\{w_t\}_{t=1}^n$ where $\mathbf{w}_i$ is the t-th observation and the sample size is n. The ML and GMM estimators are particular extremum estimators.

1.1.1 Measurability of $\hat{\boldsymbol{\theta}}$

The maximisation problem (1) may not necessarily have a solution. But recall from calculus the following fact:

Let $h : \mathbb{R}^p \to \mathbb{R}$ be continuous and $A \subset \mathbb{R}^p$ be a compact (closed and bounded) set. Then h has a maximum on the set A. That is, there $\exists$ an $\mathbf{x}^* \in A$ such that $h(\mathbf{x}^*) \geq h(\mathbf{x})$ $\forall$ $\mathbf{x} \in A$.

Therefore, if $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for any data $\{\mathbf{w}_t\}_{t=1}^n$ and Θ is compact, then there exists a $\boldsymbol{\theta}$ that solves the maximisation problem (1) for any given data $\{\mathbf{w}_t\}_{t=1}^n$. In the event of multiple solutions, we would choose one from them. So $\hat{\boldsymbol{\theta}}$, uniquely determined for any data $\{\mathbf{w}_t\}_{t=1}^n$, is a function of the data.

Strictly speaking, however, being a function of the vector of random variables $\{\mathbf{w}_t\}_{t=1}^n$ is not enough to make $\hat{\boldsymbol{\theta}}$ a well-defined random variable; $\hat{\boldsymbol{\theta}}$ needs to be a 'measurable' function of $\{\mathbf{w}_t\}_{t=1}^n$.[1] The following lemma shows that $\hat{\boldsymbol{\theta}}$ is measurable if $Q_n(\boldsymbol{\theta})$ is

Lemma (existence of extremum estimators) (7.1 in Hayashi): Suppose that i) the parameter space Θ is a compact subset of $\mathbb{R}^p$, ii) $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for any data $\{\mathbf{w}_t\}_{t=1}^n$, and iii) $Q_n(\boldsymbol{\theta})$ is a measurable function of the data for all $\boldsymbol{\theta}$ in Θ. Then, there exists a measurable function $\hat{\boldsymbol{\theta}}$ of the data that solves (1).

In most applications, we do not know the upper or lower bound for the true parameter vector. Even if we do, those bounds are not included in the parameter space, so the parameter space is not closed. So the compactness assumption for Θ is something we wish to avoid. We can replace the compactness assumption by some other conditions that are satisfied in many applications when we resort to asymptotic theory.

1.1.2 Two classes of extremum estimators

1. **M-Estimators**: An extremum estimator is an M-estimator if the objective function is a sample average:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} m(\mathbf{w}_t;\boldsymbol{\theta}), \tag{2}$$

   where m is a real-valued function of $(\mathbf{w}_t, \boldsymbol{\theta})$. Two examples of an M-estimator we study are the ML and nonlinear least squares (NLS) estimators.

2. **GMM**: An extremum estimator is a GMM estimator if the objective function can be written as:

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})^\top \hat{\mathbf{W}} \mathbf{g}_n(\boldsymbol{\theta}), \tag{3}$$

   with

$$\underset{K\times 1}{\mathbf{g}_n(\boldsymbol{\theta})} \equiv \frac{1}{n}\sum_{t=1}^{n} \mathbf{g}(\mathbf{w}_t;\boldsymbol{\theta}),$$

   and where $\hat{\mathbf{W}}$ is a $K \times K$ symmetric and positive definite matrix that defines the distance of $\mathbf{g}_n(\boldsymbol{\theta})$ from zero. It can depend on the data. Maximising this objective function is equivalent to minimising the distance $\mathbf{g}_n(\boldsymbol{\theta})^\top \hat{\mathbf{W}} \mathbf{g}_n(\boldsymbol{\theta})$, which is the definition of GMM.

Note, however, that there are extremum estimators that do not fall into either class. A prominent example is classical minimum distance estimators. But that's beyond the scope of the course.

[1] If a function is continuous it is measurable.

2 Basic concepts of ML estimation

Models that are estimated by maximum likelihood must be fully specified parametric models. So far, everything we have considered os far are sometimes termed 'semi-parametric', because they do not require the joint distribution of the data to be specified completely. For example, OLS estimation in the linear regression model:

$$y_t = \mathbf{X}_t \boldsymbol{\beta} + u_i,$$

we assumed $\mathbb{E}\left[\mathbf{X}_t^\top u_t\right] = 0$ (or possibly $\mathbb{E}\left[u_t|\mathbf{X}_t\right] = 0$), but said nothing about the distribution of the u_t's and the $\mathbf{X}_t$'s. Semi-parametric estimation methods are often attractive, precisely because they allow us to remain relatively agnostic about those aspects of the model (e.g. the distribution of the errors in a linear regression model) that are not of interest to us.

Parametric (or 'fully' parametric) estimation methods, by contrast, require the entire joint distribution of the data – or, at a minimum, certain conditional distributions — to be completely specified. This means that there is a correspondingly higher risk, when using these methods, of misspecifying [some components of] the model, which may lead to the parameters of interest being inconsistently estimated. On the other hand, if we are prepared to make these assumptions, then we can lever off these to obtain estimates of the parameters of interest that are often much more efficient than provided by semi-parametric methods.

As usual, we denote the dependent variable by the n-vector $\mathbf{y}$. For a given k-vector $\boldsymbol{\theta}$ of parameters, let the joint PDF of $\mathbf{y}$ be written as $f(\mathbf{y}, \boldsymbol{\theta})$. This joint PDF constitutes the specification of the model. Since a PDF provides an unambiguous recipe for simulation, it suffices to specify the vector $\boldsymbol{\theta}$ in order to give a full characterisation of a DGP in the model. Thus, there is a one-to-one correspondence between the DGPs of the model and the admissible parameter vectors.

Maximum likelihood estimation is based on the specification of the model through the joint PDF $f(\mathbf{y}, \boldsymbol{\theta})$. When $\boldsymbol{\theta}$ is fixed, the function $f(\cdot, \boldsymbol{\theta})$ of $\mathbf{y}$ is interpreted as the PDF of $\mathbf{y}$. But if instead $f(\mathbf{y}, \boldsymbol{\theta})$ is evaluated at the n-vector $\mathbf{y}$ found in a given data set, then the function $f(\mathbf{y}, \cdot)$ of the model parameters can no longer be interpreted as a PDF. Instead, it is referred to as the likelihood function of the model for the given data set. ML estimation then amounts to maximising the likelihood function with respect to the parameters. A parameter vector $\hat{\boldsymbol{\theta}}$ at which the likelihood takes on its maximum value is called a maximum likelihood estimate (MLE) of the parameters.

In many cases, the successive observations in a sample are assumed to be statistically independent. In that case, the joint density of the entire sample is just the product of the densities of the individual observations. Let $f(y_t, \boldsymbol{\theta})$ denote the PDF of a typical observation, y_t. Then the joint density of the entire sample $\mathbf{y}$ is

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t, \boldsymbol{\theta}). \tag{4}$$

Working with products such as (4) is difficult, however, since it may be very large or small. For this reason, we work with the log-likelihood function:

$$l(\mathbf{y}, \boldsymbol{\theta}) \equiv \log f(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(y_t, \boldsymbol{\theta}), \tag{5}$$

where $l_t(y_t, \boldsymbol{\theta})$, the contribution to the log-likelihood function made by observation t, is equal to $\log f_t(y_t, \boldsymbol{\theta})$. The t subscripts on f_t and l_t have been added to allow for the possibility that the density of y_t may vary from observation to observation, perhaps because there are exogenous variables in the model. Since logarithms are just a monotonic transformation, whatever value of $\boldsymbol{\theta}$ which maximises $f(\mathbf{y}, \boldsymbol{\theta})$ must also maximise $l(\mathbf{y}, \boldsymbol{\theta})$. The MLE of $\boldsymbol{\theta}$, therefore, is an M-estimator with

$$\begin{aligned} m(y_t; \boldsymbol{\theta}) &= l(y_t, \boldsymbol{\theta}), \\ \implies Q_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{t=1}^{n} l(y_t; \boldsymbol{\theta}). \end{aligned}$$

3 Examples of ML estimation

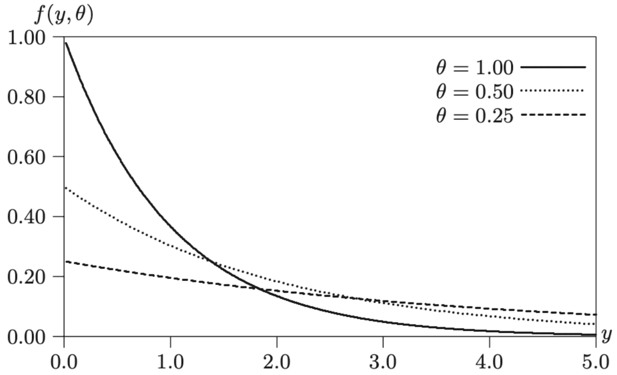
3.1 The exponential distribution

As a simple example of ML estimation, suppose that each observation y_t is generated by the density

$$f(y_t, \theta) = \theta \exp(-\theta y_t), \; y_t > 0, \; \theta > 0. \tag{6}$$

This is the PDF of what is called the exponential distribution. There are assumed to be n independent observations from which to calculate the log-likelihood function.

Figure 1: **The Exponential Distribution** (Davidson & MacKinnon)



Taking the logarithm of the density (6), we find that the contribution to the log-likelihood from observation t is $l_t(y_t, \theta) = \log\theta - \theta y_t$. Therefore,

$$l(\mathbf{y}, \theta) = \sum_{t=1}^{n} (\log\theta - \theta y_t) = n\log\theta - \theta\sum_{t=1}^{n} y_t.$$

To maximise this log-likelihood function with respect tot he single unknown parameter θ, we differentiate it with respect to θ and set the derivative equal to 0. The result is

$$\frac{\partial l(\mathbf{y}, \theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{t=1}^{n} y_t = 0,$$

which can be solved to yield:

$$\hat{\theta} = \frac{n}{\sum_{t=1}^{n} y_t}. \tag{7}$$

This solution is clearly unique, because the second derivative of $l(\mathbf{y}, \theta)$ is always negative:

$$\frac{\partial^2 l(\mathbf{y}, \theta)}{\partial \theta^2} = -\frac{n}{\theta^2},$$

and because it is unique, the estimator defined in (7) can be called *the* MLE that corresponds to the log-likelihood function, $l(\mathbf{y}, \theta)$.

In this case, interestingly, the MLE $\hat{\theta}$ is exactly the same as the MM estimator. As we now show, the expected value of y_t is $1/\theta$. By definition, this expectation is:

$$\mathbb{E}[y_t] = \int_0^\infty y_t \theta \exp(-\theta y_t) dy_t.$$

Since $-\theta \exp(-\theta y_t)$ is the derivative of $\exp(-\theta y_t)$ with respect to y_t, we may integrate by parts[2] to obtain

$$\begin{aligned}
\int_0^\infty y_t \theta \exp(-\theta y_t) dy_t &= \left[y_t \left(-\exp(-\theta y_t)\right)\right]_0^\infty - \int_0^\infty \left(-\exp(-\theta y_t)\right) dy_t \\
&= -\left[y_t \exp(-\theta y_t)\right]_0^\infty + \int_0^\infty \exp(-\theta y_t) dy_t \\
&= \int_0^\infty \exp(-\theta y_t) dy_t \\
&= \left[-\frac{1}{\theta} \exp(-\theta y_t) dy_t\right]_0^\infty - \int_0^\infty 0 \times \left(-\frac{1}{\theta} \exp(-\theta y_t)\right) dy_t \\
&= [0] - \left[-\frac{1}{\theta}\right] \\
&= \frac{1}{\theta}.
\end{aligned}$$

The most natural MM estimator of θ is the one that matches θ^{-1} to the empirical analogue of $\mathbb{E}[y_t]$, which is $\bar{y}$, the sample mean. This estimator of θ is therefore $1/\bar{y}$, which is identical to the ML estimator (7).

[2] Recall that integration by parts is:

$$\int_a^b u dv = [uv]_a^b - \int_a^b v du,$$

where we choose u by the followed order: LIATE: Logs, inverse, algebraic, trigonometric, and exponential.

It is not uncommon for an MLE to coincide with an MM estimator, as happens in this case. This may suggest that maximum likelihood is not a very useful addition to the econometrician's toolkit, but such an inference would be unwarranted! Even in this simple case, the MLE was considerably easier to obtain than the MM estimator, because we did not need to calculate an expectation. In more complicated cases, this advantage of the MLE is often much more substantial.

3.2 Regression models with normal errors

It is interesting to see what happens when we apply the method of maximum likelihood to the classical normal linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \ \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{8}$$

For this model, the explanatory variables in the matrix $\mathbf{X}$ are assumed to be exogenous. Consequently, in constructing the likelihood function, we may use the density of $\mathbf{y}$ conditional on $\mathbf{X}$. The elements of u_t of the vector $\mathbf{u}$ are independently distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$, and so y_t is distributed, conditionally on $\mathbf{X}$, as $N(\mathbf{X}_t\boldsymbol{\beta}, \sigma^2)$. Thus, the PDF of y_t is:

$$f_t(y_t, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_t - \mathbf{X}_t\boldsymbol{\beta})^2}{2\sigma^2}\right). \tag{9}$$

The contribution to the log-likelihood function made by the t-th observation is the logarithm of (9). Since $\log\sigma = \frac{1}{2}\log\sigma^2$, this can be written as

$$l_t(y_t, \boldsymbol{\beta}, \sigma) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi - \frac{(y_t - \mathbf{X}_t\boldsymbol{\beta})^2}{2\sigma^2}.$$

Since the observations are assumed to be independent, the log-likelihood function is just the sum of these contributions over all t, or:

$$\begin{aligned} l(\mathbf{y}, \boldsymbol{\beta}, \sigma) &= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \mathbf{X}_t\boldsymbol{\beta})^2, \\ &= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{10}$$

In the second line, we rewrite the sum of squared residuals (SSR) as the inner product of the residual vector with itself. To find the MLE, we need to maximise (10) with respect to the unknown parameters $\boldsymbol{\beta}$ and σ.

The first step in maximising $l(\mathbf{y}, \boldsymbol{\beta}, \sigma)$ is to concentrate it with respect to the parameter σ, solving the resulting FOC for σ as a function of the data and the remaining parameters, and then substituting the result back into (10). This will yield the concentrated log-likelihood function. The second step is to maximise this function with respect to $\boldsymbol{\beta}$. For models that involve variance parameters, it is very often convenient to concentrate the log-likelihood function in this way.

Differentiating (10) with respect to σ and equating the derivative to zero yields the FOC:

$$\frac{\partial l(\mathbf{y}, \boldsymbol{\beta}, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

and solving this for σ, yields:

$$\begin{aligned} n &= \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \implies \hat{\sigma}^2(\boldsymbol{\beta}) &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Here, the notation $\hat{\sigma}^2(\boldsymbol{\beta})$ indicates that the value of σ^2 that maximises (10) depends on $\boldsymbol{\beta}$. Substituting $\hat{\sigma}^2(\boldsymbol{\beta})$ back into (10) yields:

$$\begin{aligned} l^c(\mathbf{y}, \boldsymbol{\beta}) &= -\frac{n}{2}\log\left(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) - \frac{n}{2}\log 2\pi - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\left(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)} \qquad (11) \\ &= -\frac{n}{2}\log\left(\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) - \frac{n}{2}\log 2\pi - \frac{n}{2}. \qquad (12) \end{aligned}$$

The first term here is minus $n/2$ times the logarithm of the SSR, and the other two terms do not depend on $\boldsymbol{\beta}$. Thus we see that maximising the concentrated log-likelihood (12) is equivalent to minimising the SSR as a function $\boldsymbol{\beta}$. Therefore, the MLE $\hat{\boldsymbol{\beta}}$ must be identical to the OLS estimator.

Once $\hat{\boldsymbol{\beta}}$ has been found, the MLE $\hat{\sigma}^2$ of σ^2 is $\hat{\sigma}^2(\boldsymbol{\beta})$, and the MLE of σ is the positive square root of $\hat{\sigma}^2$. Thus, the MLE $\hat{\sigma}^2$ is biased downward. The actual maximised value of the log-likelihood function can be written in terms of the SSR function SSR evaluated at $\hat{\boldsymbol{\beta}}$. From (12) we have:

$$l(\mathbf{y}, \hat{\boldsymbol{\beta}}, \hat{\sigma}) = -\frac{n}{2}\left(1 + \log 2\pi - \log n\right) - \frac{n}{2}\log \text{SSR}(\hat{\boldsymbol{\beta}}),$$

where $\text{SSR}(\hat{\boldsymbol{\beta}})$ denotes the minimised SSR.[3]

The fact that the ML and OLS estimators of $\boldsymbol{\beta}$ are identical depends critically on the assumption that the error terms in (8) are normally distributed. If we had started with a different assumption about their distribution, we would have obtained a different MLE. The asymptotic efficiency result to be discussed later would the imply that the least squares estimator is asymptotically less efficient then the MLE whenever the two do not coincide.

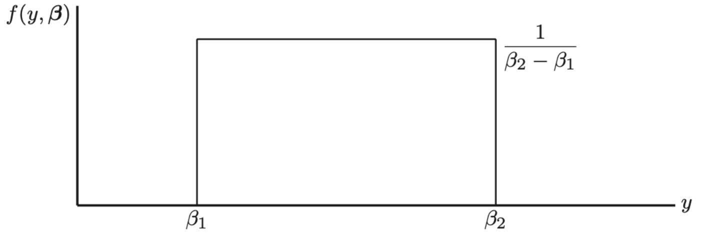
3.3 The uniform distribution

Another example of ML estimation, which we consider a somewhat pathological, but rather interesting, example, is the uniform distribution. Suppose that the y_t are generated as independent realisations from the uniform distribution with parameters β_1 and β_2, which can be written as a vector $\boldsymbol{\beta}$. The density for y_t, which is graphed in Figure 2, is:

$$f(y_t, \boldsymbol{\beta}) = \begin{cases} 0 & \text{if } y_t < \beta_1, \\ \frac{1}{\beta_2 - \beta_1} & \text{if } \beta_1 \leq y_t \leq \beta_2, \\ 0 & \text{if } y_t > \beta_2. \end{cases}$$

[3] See the third problem set on how to solve this problem.

Figure 2: **The Uniform Distribution** (Davidson & MacKinnon)



Provided that $\beta_1 < y_t < \beta_2$ for all observations, the likelihood function is equal to:

$$f(y_t, \boldsymbol{\beta}) = \frac{1}{(\beta_2 - \beta_1)^n},$$

and the log-likelihood function is therefore

$$l(\mathbf{y}, \boldsymbol{\beta}) = -n \log(\beta_2 - \beta_1).$$

It is easy to verify that this function cannot be maximised by differentiating it with respect to the parameters and setting the partial derivatives to zero. Instead, the way to maximise $l(\mathbf{y}, \boldsymbol{\beta})$ is to make $\beta_2 - \beta_1$ as small as possible. But we cannot make β_1 larger than the smallest observed y_t, and we cannot make β_2 smaller than the largest observed y_t. Otherwise, the likelihood function would be equal to 0. It follows that the MLEs are:

$$\hat{\beta}_1 = \min(y_t), \ \hat{\beta}_2 = \max(y_t). \tag{13}$$

These estimators are rather unusual. For one thing, they always lie on one side of the true value. Because all the y_t must lie between β_1 and β_2, it must be the case that $\hat{\beta}_1 \geq \beta_1^0$ and $\hat{\beta}_2 \leq \beta_2^0$. However, despite this, these estimators turn out to be consistent. Intuitively, this is because, as the sample size gets large, the observed values of y_t fill up the entire space between β_1^0 and β_2^0.

The ML estimators in (13) are super-consistent, which means that they approach the true values of the parameters they are estimating at a rate faster than the usual rate of $1/\sqrt{n}$. Formally, $\sqrt{n}(\hat{\beta}_1 - \beta_1^0)$ tends to 0 as $n \to \infty$, while $n(\hat{\beta}_1 - \beta_1^0)$ tends to a limiting random variable. Now consider the parameter $\gamma \equiv \frac{1}{2}(\beta_1 + \beta_2)$. One way to estimate it is to use the MLE:

$$\hat{\gamma} = \frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2).$$

Another approach would simply be to use the sample mean, say $\bar{\gamma}$, which is a least squares estimator. But the MLE $\hat{\gamma}$ is super-consistent, while $\bar{\gamma}$ is only root-n consistent. This implies that, except perhaps for very small sample sizes, the MLE is very much more efficient than the least squares estimator.

4 Two types of MLEs

There are two different ways of defining the MLE, although most MLEs actually satisfy both definitions:

Type 1 MLE: Maximises the log-likelihood function over the set Θ, in which the parameter vector $\boldsymbol{\theta}$ lies, which is generally assumed to be a subset of $\mathbb{R}^p$. This is the natural meaning of an MLE, and all three of the MLEs we just discussed are Type 1 estimators.

If the log-likelihood function is differentiable and attains an interior maximum in the parameter space, then the MLE must satisfy the FOCs for a maximum.

Type 2 MLE: A solution to the likelihood equations, which are just the following FOCs:

$$\mathbf{g}(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{14}$$

where here $\mathbf{g}(\mathbf{y}, \boldsymbol{\theta})$ is the gradient vector, or score vector, which has typical elements:

$$g_i(\mathbf{y}, \boldsymbol{\theta}) \equiv \frac{\partial l(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i} = \sum_{t=1}^{n} \frac{\partial l_t(y_t, \boldsymbol{\theta})}{\partial \theta_i}.$$

Because there may be more than one value of $\boldsymbol{\theta}$ that satisfies the likelihood equations (14), the definition further requires that the Type 2 estimator $\hat{\boldsymbol{\theta}}$ be associated with a local maximum of $l(\mathbf{y}, \boldsymbol{\theta})$ and that, as $n \to \infty$, the value of the log-likelihood function associated with $\hat{\boldsymbol{\theta}}$ be higher than the value associated with any other root of the likelihood equations.

Most MLEs are both Type 1 and Type 2 (e.g. the MLE for the exponential and OLS estimators of $\boldsymbol{\beta}$ and σ^2). But some MLEs, such as the parameters of the uniform distribution are Type 1 but not Type 2, because they are not solutions to any set of likelihood equations.

5 Asymptotic properties of MLEs

One of the attractive features of ML estimation is that the MLEs are consistent under quite weak regularity conditions and asymptotically normally distributed under somewhat stronger conditions. Therefore, if an estimator is an MLE and the regularity conditions are satisfied, it is not necessary to show that it is consistent or derive its asymptotic distribution.

5.1 Consistency of the MLE

Since almost all MLEs are of Type 1, we will discuss consistency only for this type of MLE. We first show that the expectation of the log-likelihood function is greater when it is evaluated at the true values of the parameters than when it is evaluated at any other values. For consistency, we also need both a finite sample identification condition and an asymptotic identification condition. The former requires that the log-likelihood be different for different sets of parameter values. If, contrary to this assumption, there were two distinct parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, such that $l(\mathbf{y}, \boldsymbol{\theta}_1) = l(\mathbf{y}, \boldsymbol{\theta}_2)$ for all $\mathbf{y}$, then it would obviously be impossible to distinguish between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus, a finite sample identification condition is necessary for the model to make sense. The role of the asymptotic identification condition will be discussed below.

Let $L(\boldsymbol{\theta}) = \exp(l(\boldsymbol{\theta}))$ denote the likelihood function, where the dependence on $\mathbf{y}$ of both L and l has been suppressed for notational simplicity. We wish to apply a result known as Jensen's Inequality[4] to the ratio

$$\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)},$$

where $\boldsymbol{\theta}_0$ is the true parameter vector and $\boldsymbol{\theta}^*$ is any other vector in the parameter space of the model.

Since the logarithm is a strictly concave function over the nonnegative real line, and since likelihood functions are nonnegative, we can conclude from Jensen's Inequality that:

$$\mathbb{E}_0 \log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) < \log \mathbb{E}_0\left[\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right], \tag{15}$$

with strict inequality for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, on account of the finite-sample identification condition. Here the notation $\mathbb{E}_0$ means the expectation taken under the DGP characterised by the true parameter vector $\boldsymbol{\theta}_0$. Since the joint density of the sample is simply the likelihood function evaluated at $\boldsymbol{\theta}_0$, the expectation on the RHS of (15) can be expressed as an integral over the support of the vector random variable $\mathbf{y}$. We have:

$$\mathbb{E}_0\left[\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right] = \int \frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)} L(\boldsymbol{\theta}_0) d\mathbf{y} = \int L(\boldsymbol{\theta}^*) d\mathbf{y} = 1,$$

where the last equality here holds because every density must integrate to 1. Therefore, because $\log 1 = 0$, the equality (15) implies:

$$\mathbb{E}_0 \log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) = \mathbb{E}_0 l(\boldsymbol{\theta}^*) - \mathbb{E}_0 l(\boldsymbol{\theta}_0) < 0. \tag{16}$$

In words, this says that the expectation of the log-likelihood function when evaluated at the true parameter vector, $\boldsymbol{\theta}_0$, is strictly greater than its expectation when evaluated at any other parameter vector, $\boldsymbol{\theta}^*$.

If we can apply an LLN to the contributions to the log-likelihood function, then we can assert that

$$\operatorname{plim} n^{-1} l(\boldsymbol{\theta}) = \lim n^{-1} \mathbb{E}_0 l(\boldsymbol{\theta}).$$

Then, (16) implies that:

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\boldsymbol{\theta}^*) \leq \operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\boldsymbol{\theta}_0),\ \forall \boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0, \tag{17}$$

where the inequality is not necessarily strict, because we have taken a limit. Since the MLE $\hat{\boldsymbol{\theta}}$ maximises $l(\boldsymbol{\theta})$, it must be the case that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\hat{\boldsymbol{\theta}}) \geq \operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\boldsymbol{\theta}_0). \tag{18}$$

[4] Jensen's Inequality tells us that, if X is a real valued random variable, then

$$\mathbb{E}\left[h(X)\right] \leq h(\mathbb{E}[X]),$$

whenever h is strictly concave over at least part of the support of the random variable X. That is, the set of real numbers for which the density of X is nonzero, and the support contains more than one point.

The only way that that (17) and (18) can both be true is if

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\hat{\boldsymbol{\theta}}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} l(\boldsymbol{\theta}_0).$$

In words, this says that the plim of $1/n$ times the log-likelihood function must be the same when it is evaluated at the MLE $\hat{\boldsymbol{\theta}}$ as when it is evaluated at the true parameter vector $\boldsymbol{\theta}_0$.

By itself, this result does not prove that $\hat{\boldsymbol{\theta}}$ is consistent, because the weak inequality does not rule out the possibility that there may be many values of $\boldsymbol{\theta}^*$ for which $\operatorname{plim} n^{-1} l(\boldsymbol{\theta}^*) = \operatorname{plim} n^{-1} l(\boldsymbol{\theta}_0)$. We must therefore explicitly assume that:

$$\operatorname*{plim}_{n\to\infty} l(\boldsymbol{\theta}^*) \neq \operatorname*{plim}_{n\to\infty} l(\boldsymbol{\theta}_0),\ \forall \boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0. \tag{19}$$

This is a form of asymptotic identification condition.

5.2 Dependent observations

Before discussing asymptotics of the MLE, we need to introduce some notation and terminology, and we need to establish a few preliminary results. First, consider the case when successive observations are not independent (e.g. when the regression function involves lags of the dependent variable).

Recall that the conditional density of a random variable is defined as:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)},$$

which we can rewrite as:

$$f(x_1, x_2) = f(x_1) f(x_2|x_1),$$

or, using y_1 and y_2 notation as:

$$f(y_1, y_2) = f(y_1) f(y_2|y_1).$$

We can also apply this to situations in which y_1 and y_2 are really vectors of random variables. Accordingly, consider the joint density of three random variables, and group the first two together. We then have:

$$f(y_1, y_2, y_3) = f(y_1, y_2) f(y_3|y_1, y_2).$$

For a sample of size n, it is easy to see that this last result generalises to:

$$f(y_1, ..., y_n) = f(y_1) f(y_2|y_1) \cdots f(y_n|y_1, ..., y_{n-1}).$$

This result can be written as:

$$f(\mathbf{y}^n) = \prod_{t=1}^{n} f(y_t|\mathbf{y}^{t-1}),$$

where the vector $\mathbf{y}^t$ is a t-vector with components $y_1, y_2, ..., y_t$. One can think of $\mathbf{y}^t$ as the subsample consisting of the first t observations of the full sample .For a model that is to be estimated by maximum likelihood, the density $f(\mathbf{y}^n)$ depends on a k-vector of parameters $\boldsymbol{\theta}$, and we can then write:

$$f(\mathbf{y}^n, \boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t | \mathbf{y}^{t-1}; \boldsymbol{\theta}).$$

The log-likelihood function corresponding to the above expression is therefore:

$$l(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\mathbf{y}^t, \boldsymbol{\theta}), \tag{20}$$

where we omit the superscript n from $\mathbf{y}$ for the full sample. In addition, in the contributions $l_t(\cdot)$ to the log-likelihood, we do not distinguish between the current variable y_t and the lagged variables in the vector $\mathbf{y}^{t-1}$. In this way, (20) has exactly the same structure as (5).

5.3 The gradient

The gradient, or score, vector $\mathbf{g}(\mathbf{y}, \boldsymbol{\theta})$ is a k-vector that was defined in (14). As that equation makes clear, each component of the gradient vector is itself a sum of n contributions, and this remains true when the observations are dependent; the partial derivative of l_t with respect to θ_i now depends on $\mathbf{y}^t$ rather than just y_t. It is convenient to group these partial derivatives into a matrix. We define the $n \times k$ matrix $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$ so as to have typical element:

$$G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) \equiv \frac{\partial l_t(\mathbf{y}^t, \boldsymbol{\theta})}{\partial \theta_i}. \tag{21}$$

This matrix is called the matrix of contributions to the gradient, because:

$$g_i(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}).$$

Thus, each element of the gradient vector is the sum of the elements of one of the columns of the matrix $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$.

A crucial property of the matrix $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$ is that, if $\mathbf{y}$ is generated by the DGP characterised by $\boldsymbol{\theta}$, then the expectations of all the elements of the matrix, evaluated at $\boldsymbol{\theta}$, are zero. This result is a consequence of the fact that all densities integrated to 1. Since l_t is the log of the density f_t of y_t conditional on $\mathbf{y}^{t-1}$, we see that, for all t and for all $\boldsymbol{\theta}$,

$$\int \exp l_t(\mathbf{y}^t, \boldsymbol{\theta}) dy_t = \int f_t(\mathbf{y}^t, \boldsymbol{\theta}) dy_t = 1,$$

where the integral is over the support of y_t. Since this relation holds identically in $\boldsymbol{\theta}$, we can differentiate it with respect to the components of $\boldsymbol{\theta}$ and obtain a further set of identities. Under weak regularity conditions, it can be shown that the derivatives of the integral on the LHS are the integrals of the derivatives of the integrand. Thus, since the derivative of the constant 1 is 0, identically in $\boldsymbol{\theta}$ and for $i = 1, ..., k$,

$$\int \exp\left(l_t(\mathbf{y}^t, \boldsymbol{\theta})\right) \frac{\partial l_t(\mathbf{y}^t, \boldsymbol{\theta})}{\partial \theta_i} dy_t = 0.$$

Since $\exp(l_t(\mathbf{y}^t, \boldsymbol{\theta}))$ is, for the DGP characterised by $\boldsymbol{\theta}$, the density of y_t conditional on $\mathbf{y}^{t-1}$, this last equation, along with the definition (21) gives

$$\mathbb{E}_{\boldsymbol{\theta}}\left[G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) | \mathbf{y}^{t-1}\right] = 0, \ \forall t = 1, ..., n, \ i = 1, ..., k. \tag{22}$$

The notation $\mathbb{E}_{\boldsymbol{\theta}}$ here means that the expectation is being taken under the DGP characterised by $\boldsymbol{\theta}$. Take unconditional expectations, and sum over $t = 1, ..., n$ and we have:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}}\left[g_i(\mathbf{y}, \boldsymbol{\theta})\right] = 0, \ i = 1, ..., k, \\ \Leftrightarrow\ &\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{g}(\mathbf{y}, \boldsymbol{\theta})\right] = \mathbf{0}. \end{aligned}$$

In addition to the conditional expectations of the elements of the matrix $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$, we can compute the covariances of these elements. Let $t \neq s$, and suppose, without loss of generality, that $t < s$. Then the covariance under the DGP characterised by $\boldsymbol{\theta}$ of the ti-th and sj-th elements of $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$ is:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) G_{sj}(\mathbf{y}^s, \boldsymbol{\theta})\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) G_{sj}(\mathbf{y}^s, \boldsymbol{\theta}) | \mathbf{y}^t\right]\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}\left[G_{sj}(\mathbf{y}^s, \boldsymbol{\theta}) | \mathbf{y}^t\right]\right] = 0, \end{aligned}$$

since $G_{ti}(\cdot)$ is a deterministic function of $\mathbf{y}^t$ and because from (22) we have $\mathbb{E}_{\boldsymbol{\theta}}\left[G_{sj}(\mathbf{y}^s, \boldsymbol{\theta}) | \mathbf{y}^t\right] = 0$ since $t < s$.

5.4 The information matrix and the Hessian

The covariance matrix of the elements of the t-th row $\mathbf{G}_t(\mathbf{y}^t, \boldsymbol{\theta})$ of $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$ is the $k \times k$ matrix $\mathbf{I}_t(\boldsymbol{\theta})$, of which the ij-th element is $\mathbb{E}_{\boldsymbol{\theta}}\left[G_{ti}(\mathbf{y}^t, \boldsymbol{\theta}) G_{tj}(\mathbf{y}^t, \boldsymbol{\theta})\right]$. As a covariance matrix, $\mathbf{I}_t(\boldsymbol{\theta})$ is normally positive definite. The sum of the matrices $\mathbf{I}_t(\boldsymbol{\theta})$ over all t is the $k \times k$ matrix:

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \sum_{t=1}^{n} \mathbf{I}_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} \mathbb{E}_{\theta}\left[\mathbf{G}_t(\mathbf{y}, \boldsymbol{\theta})^{\top} \mathbf{G}_t(\mathbf{y}, \boldsymbol{\theta})\right], \tag{23}$$

which is called the formation matrix. The matrices $\mathbf{I}_t(\boldsymbol{\theta})$ are the contributions to the information matrix made by the successive observations. An equivalent definition of the information is

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) \mathbf{g}(\mathbf{y}, \boldsymbol{\theta})^{\top}\right],$$

where the information matrix is the expectation of the outer product of the gradient with itself. Less exotically, it is just the covariance matrix of the score vector. As we will soon see, the information matrix is a measure of the total amount of information about the parameters in the sample. The requirement that it should be positive definite is a condition for strong asymptotic identification of those parameters.

Closely related to the information matrix is the asymptotic information matrix:

$$\mathscr{J}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{\boldsymbol{\theta}, n \to \infty} \frac{1}{n} \mathbf{I}(\boldsymbol{\theta}), \tag{24}$$

which measures the average amount of information about the parameters that is contained in the observations of the sample. As with the notation $\mathbb{E}_{\boldsymbol{\theta}}$, we use $\operatorname{plim}_{\boldsymbol{\theta}}$ to denote the plim under the DGP characterised by $\boldsymbol{\theta}$.

For asymptotic analysis, we are also interested in the asymptotic Hessian:

$$\mathcal{H}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{\boldsymbol{\theta}, n\to\infty} \frac{1}{n}\mathbf{H}(\mathbf{y}, \boldsymbol{\theta}). \tag{25}$$

The asymptotic Hessian is related to the ordinary Hessian in exactly the same way as the asymptotic information matrix is related to the information matrix.

There is a very important relationship between the asymptotic information matrix and the asymptotic Hessian. One version of this relationship is the information matrix equality:

$$\mathcal{J}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta}). \tag{26}$$

Both the Hessian and the information matrix measure the amount of curvature in the log-likelihood function. Although they are both measuring the same thing, the Hessian is negative definite, at least in the neighbourhood of $\hat{\boldsymbol{\theta}}$, while the information matrix is always positive definite.

5.5 Asymptotic normality of the MLE

An MLE must be a Type 2 MLE in order to be asymptotically normally distributed, as well as satisfying regularity conditions.[5] The Type 2 requirement arises because the proof asymptotic normality is based on the likelihood equations (14), which apply only to Type 2 estimators.

The first step of the proof is to perform a Taylor expansion of the likelihood equations around $\boldsymbol{\theta}_0$:

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{\theta}_0) + \mathbf{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}, \tag{27}$$

where we suppress the dependence on $\mathbf{y}$ for notational simplicity. The notation $\bar{\boldsymbol{\theta}}$ is a shorthand notation for Taylor expansions of vector expressions. We may therefore write:

$$||\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| \leq ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||.$$

The fact that the MLE $\hat{\boldsymbol{\theta}}$ is consistent then implies that $\bar{\boldsymbol{\theta}}$ is also consistent.

If we solve (27) and insert the factors of powers of n that are needed for asymptotic analysis, we obtain:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -(n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}}))^{-1}) \left(\frac{1}{\sqrt{n}}\mathbf{g}(\boldsymbol{\theta}_0) \right).$$

Because $\bar{\boldsymbol{\theta}}$ is consistent, the matrix $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$ must tend to the same non-stochastic limiting matrix as $n^{-1}\mathbf{H}(\boldsymbol{\theta}_0)$, namely, $\mathcal{H}(\boldsymbol{\theta}_0)$. Therefore, the above equation implies that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{=} -\mathcal{H}(\boldsymbol{\theta}_0)^{-1}\frac{1}{\sqrt{n}}\mathbf{g}(\boldsymbol{\theta}_0). \tag{28}$$

If the information matrix equality holds, then this result can equivalently be written as:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{=} \mathcal{J}(\boldsymbol{\theta}_0)^{-1}\frac{1}{\sqrt{n}}\mathbf{g}(\boldsymbol{\theta}_0). \tag{29}$$

[5] See Davidson and MacKinnon (1993, Section 8.5)

Since the information matrix equality holds only if the model is correctly specified, the above result is not in general valid for misspecified models!

The asymptotic normality of the Type 2 MLE follows immediately from the asymptotic equalities above if we can show that the vector $n^{-\frac{1}{2}}\mathbf{g}(\boldsymbol{\theta}_0)$ is asymptotically distributed as multivariate normal. As can be see from the matrix of contributions to the gradient (21), each element of $\mathbf{g}_i(\boldsymbol{\theta}_0)$ is $n^{-\frac{1}{2}}$ times a sum of n random variables, each of which has mean 0 by (22).

Under standard regularity conditions[6], a multivariate CLT can be applied to this vector. For finite n, the covariance matrix of the score vector is, by definition, the information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$. Thus the covariance matrix of the vector $n^{-\frac{1}{2}}\mathbf{g}(\theta_0)$ is $n^{-1}\mathbf{I}(\boldsymbol{\theta}_0)$, of which, as $n \to \infty$ in the limit is the the asymptotic information matrix, $\mathcal{J}(\boldsymbol{\theta}_0)$. It follows that

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{\sqrt{n}}\mathbf{g}(\boldsymbol{\theta}_0)\right) \overset{a}{\sim} N(\mathbf{0}, \mathcal{J}(\boldsymbol{\theta}_0)). \tag{30}$$

This result, when combined with (28) or (29), implies that the Type 2 MLE is asymptotically normally distributed.

5.6 Variance-Covariance matrix of the MLE

For Type 2 ML estimators, we can obtain the asymptotic distribution of the estimator by combining (30) with (28). The asymptotic distribution of the estimator is the distribution of the random variable plim $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. This distribution is normal, with mean vector zero and variance-covariance matrix

$$\text{Var}\left(\operatorname*{plim}_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) = \mathcal{H}(\boldsymbol{\theta}_0)^{-1}\mathcal{J}(\boldsymbol{\theta}_0)\mathcal{H}(\boldsymbol{\theta}_0)^{-1}, \tag{31}$$

which is of the sandwich covariance matrix form. When the information matrix equality (26) holds, the sandwich simplifies to:

$$\text{Var}\left(\operatorname*{plim}_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) = \mathcal{J}(\boldsymbol{\theta}_0)^{-1}.$$

Thus, the asymptotic information matrix is seen to be the asymptotic precision matrix of a Type 2 ML estimator. This shows why the matrices $\mathbf{I}$ and $\mathcal{J}$ are called information matrices of various sorts.

Clearly, any method that allows us to estimate $\mathcal{J}(\boldsymbol{\theta}_0)$ consistently can be used to estimate the covariance matrix of the ML estimates. In fact, several different methods are widely used, because each has advantages in certain situations.

5.6.1 Empirical Hessian estimator

The first method is just to use minus the inverse of the Hessian, evaluated at the vector of ML estimates. Because these estimates are consistent, it is valid to evaluate the Hessian at $\hat{\boldsymbol{\theta}}$ rather than at $\boldsymbol{\theta}_0$. This yields the estimator:

$$\widehat{\text{Var}}_{\mathbf{H}}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}, \tag{32}$$

[6] Which is beyond the scope of this course.

which is referred to as the empirical Hessian estimator. Notice that, since it is the covariance matrix of $\hat{\boldsymbol{\theta}}$ in which we are interested, the factor of $n^{1/2}$ is no longer present. This estimator is easy to obtained whenever Newton's Method, or some sort of quasi-Newton method that uses second derivatives, is used to maximise the log-likelihood function.

5.6.2 Information matrix estimator

Although the empirical Hessian estimator often works well, it does not use all the information we have about the model. Especially for simpler models, we may actually be able to find an analytic expression for $\mathbf{I}(\boldsymbol{\theta})$. If so, we can use the inverse of $\mathbf{I}(\boldsymbol{\theta})$, evaluated at the ML estimates. This yields the information matrix, or IM, estimator:

$$\widehat{\text{Var}}_{\text{IM}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}). \tag{33}$$

The advantage of this estimator is that it normally involves few random terms than does the empirical Hessian, and it may therefore be somewhat more efficient. In the case of the classical normal linear model, it is not at all difficult to obtain $\mathbf{I}(\boldsymbol{\theta})$, and the IM estimator is therefore the one that is normally used.

5.6.3 Outer-Product-of-the-Gradient estimator

The third method is based on (23), from which we see that:

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{G}(\boldsymbol{\theta}_0)^\top \mathbf{G}(\boldsymbol{\theta}_0)\right].$$

We can therefore estimate $n^{-1}\mathbf{I}(\boldsymbol{\theta}_0)$ consistently by $n^{-1}\mathbf{G}(\hat{\boldsymbol{\theta}})^\top\mathbf{G}(\hat{\boldsymbol{\theta}})$. The corresponding estimator of the covariance matrix, which is usually called the outer-product-of-the-gradient (OPG) estimator is:

$$\widehat{\text{Var}}_{\text{OPG}}(\hat{\boldsymbol{\theta}}) = (\mathbf{G}(\hat{\boldsymbol{\theta}})^\top\mathbf{G}(\hat{\boldsymbol{\theta}}))^{-1}. \tag{34}$$

The OPG estimator has the advantage of being very easy to calculate. Unlike the empirical Hessian, it depends solely on first derivatives. unlike the IM estimator, it requires no theoretical calculations. However, it tends to be less reliable in finite samples than either of the other two. The OPG estimator is sometimes called the BHHH estimator.[7]

5.6.4 Sandwich estimator

A fourth covariance matrix estimator, which follows directly from (31), is the sandwich estimator:

$$\widehat{\text{Var}}_{\text{S}}(\hat{\boldsymbol{\theta}}) = \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{G}(\hat{\boldsymbol{\theta}})^\top\mathbf{G}(\hat{\boldsymbol{\theta}})\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}. \tag{35}$$

In normal circumstances, this estimator is not recommended. It is harder to compute than the OPG estimator and can be just as unreliable in finite samples. However, unlike the other three, it will be valid even when the information matrix equality does not hold. Since this equality will generally fail to hold when the model is misspecified, it may be desirable to compute the sandwich estimator and compare it with the others.

[7] After Berndt, Hall, Hall, and Hausman (1974).

When an MLE is applied to a model which is misspecified in ways that do not affect the consistency of the estimator, it is said to be a quasi-ML estimator, or QMLE. In general, the sandwich covariance matrix estimator is valid for QMLEs, but the other covariance matrix estimators, which depend on the information matrix equality, are not valid.

We have see that the MLE for a regression model with normal errors is just the OLS estimator. But we know that the latter is consistent under conditions which do not require normality. If the error terms are not normal, therefore, the MLE is a QMLE.

5.7 The classical normal linear model

It's better to do an example in order to understand our theoretical results we just derived.

For the classical normal linear model, the contribution to the log-likelihood function made by the t-th observation is given by expression

$$l_t(y_t, \boldsymbol{\beta}, \sigma) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_t - \mathbf{X}_t\boldsymbol{\beta})^2. \tag{36}$$

There are $k+1$ parameters. The first k of them are the elements of the vector $\boldsymbol{\beta}$, and the last one is σ. A typical element of any of the first k columns of the matrix $\mathbf{G}$, indexed by i, is:

$$G_{ti}(\boldsymbol{\beta}, \sigma) = \frac{\partial l_t}{\partial \beta_i} = \frac{1}{\sigma^2}(y_t - \mathbf{X}_t\boldsymbol{\beta})X_{ti},\ i = 1, ..., k, \tag{37}$$

and a typical element of the last column is:

$$G_{t,k+1}(\boldsymbol{\beta}, \sigma) = \frac{\partial l_t}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(y_t - \mathbf{X}_t\boldsymbol{\beta})^2. \tag{38}$$

These two equations give us everything we need to calculate the information matrix. For $i, j = 1, ..., k$, the ij-th element of $\mathbf{G}^\top\mathbf{G}$ is:

$$\sum_{t=1}^{n}\frac{1}{\sigma^4}(y_t - \mathbf{X}_t\boldsymbol{\beta})^2 X_{ti}X_{tj}. \tag{39}$$

This is just the sum over all t of $G_{ti}(\boldsymbol{\beta}, \sigma)$ times $G_{tj}(\boldsymbol{\beta}, \sigma)$ as defined in (37). When we evaluate at the true values of $\boldsymbol{\beta}$ and σ, we have that $y_t - \mathbf{X}_t\boldsymbol{\beta} = u_t$ and $\mathbb{E}[u_t^2] = \sigma^2$, and so the expectation of this matrix element is easily seen to be:

$$\sum_{t=1}^{n}\frac{1}{\sigma^2}X_{ti}X_{tj}. \tag{40}$$

In matrix notation, the whole $\boldsymbol{\beta}$-$\boldsymbol{\beta}$ block of $\mathbf{G}^\top\mathbf{G}$ has expectation $\mathbf{X}^\top\mathbf{X}/\sigma^2$. The $(i, k+1)$-th element of $\mathbf{G}^\top\mathbf{G}$ is

$$\begin{aligned}&\sum_{t=1}^{n}\left(-\frac{1}{\sigma} + \frac{1}{\sigma^3}(y_t - \mathbf{X}_t\boldsymbol{\beta})^2\right)\left(\frac{1}{\sigma^2}(y_t - \mathbf{X}_t\boldsymbol{\beta})X_{ti}\right)\\ &= -\sum_{t=1}^{n}\frac{1}{\sigma^3}(y_t - \mathbf{X}_t\boldsymbol{\beta})X_{ti} + \sum_{t=1}^{n}\frac{1}{\sigma^5}(y_t - \mathbf{X}_t\boldsymbol{\beta})^3 X_{ti}.\end{aligned} \tag{41}$$

This is the sum over all t of the product of expressions (37) and (38). We know that $\mathbb{E}[u_t] = 0$, and, if the error terms u_t are normal, we also know that $\mathbb{E}[u_t^3] = 0$. Consequently, the expectation of this sum is 0. This result depends critically on the assumption, following from normality, that the distribution of the error terms is symmetric around zero. For a skewed distribution, the third moment would be nonzero, and therefore the above expression would not have mean 0.

Finally, the $(k+1),(k+1)$-th element of $\mathbf{G}^\top\mathbf{G}$ is

$$\begin{aligned}&\sum_{t=1}^{n}\left(-\frac{1}{\sigma}+\frac{1}{\sigma^3}(y_t-\mathbf{X}_t\boldsymbol{\beta})^2\right)^2\\&=\frac{n}{\sigma^2}-\sum_{t=1}^{n}\frac{2}{\sigma^4}(y_t-\mathbf{X}_t\boldsymbol{\beta})^2+\sum_{t=1}^{n}\frac{1}{\sigma^6}(y_t-\mathbf{X}_t\boldsymbol{\beta})^4. \qquad (42)\end{aligned}$$

This is the sum over all t of the square of expression (38). To compute its expectation, we replace $y_t - \mathbf{X}_t\boldsymbol{\beta}$ by u_t and use the result that

$$\mathbb{E}[u_t^4] = 3\sigma^4.$$

Then, we can see that (42) has expectation $2n/\sigma^2$. Once more, this result depends crucially on the normality assumption. If the kurtosis of the error terms were greater (or less) than that of the normal distribution, the expectation of (42) would be larger (or smaller) than $2n/\sigma^2$.

Combining results (40), (41), and (42), the asymptotic information matrix for $\boldsymbol{\beta}$ and σ jointly is seen to be:

$$\mathscr{J}(\boldsymbol{\beta},\sigma) = \operatorname*{plim}_{n\to\infty}\begin{bmatrix}\frac{\mathbf{X}^\top\mathbf{X}}{n\sigma^2} & \mathbf{0}\\ \mathbf{0} & \frac{2}{\sigma^2}\end{bmatrix}. \qquad (43)$$

Inverting this matrix, multiplying the inverse by n^{-1}, and replacing σ by $\hat{\sigma}$, we find that the IM estimator of the covariance matrix of all the parameter estimates is:

$$\widehat{\mathrm{Var}}_{\mathrm{IM}}(\hat{\boldsymbol{\beta}},\hat{\sigma}) = \begin{bmatrix}\hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{\hat{\sigma}^2}{2n}\end{bmatrix}. \qquad (44)$$

The upper left-hand block of this matrix would be the familiar OLS covariance matrix if we had used s instead of $\hat{\sigma}$ to estimate σ. The lower right-hand element is the approximate variance of $\hat{\sigma}$, under the assumption of normally distributed error terms.

Finally, note that the asymptotic information matrix and its inverse are block-diagonal. This implies that there is no covariance between $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. This is a property of all regression models – nonlinear as well as linear, and it is responsible for much of the simplicity of these models. The block-diagonality of the information matrix means that we can make inferences about $\boldsymbol{\beta}$ without taking account of the fact that σ has also been estimated, and vice versa.

5.8 Asymptotic efficiency of the MLE

A Type 2 MLE must be at least as asymptotically efficient as any other root-n consistent estimator that is asymptotically unbiased.

Consider any other root-n consistent and asymptotically unbiased estimator, say $\tilde{\boldsymbol{\theta}}$. It can be show that:

$$\operatorname*{plim}_{n\to\infty} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \operatorname*{plim}_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{v}, \tag{45}$$

where $\mathbf{v}$ is a random k-vector that has mean zero and is uncorrelated with the vector $\text{plim}\,\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. This means we have:

$$\text{Var}\left(\operatorname*{plim}_{n\to\infty} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) = \text{Var}\left(\operatorname*{plim}_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) + \text{Var}(\mathbf{v}). \tag{46}$$

Since $\text{Var}(\mathbf{v})$ must be a positive semidefinite matrix, we conclude that the asymptotic covariance matrix of the estimator $\tilde{\boldsymbol{\theta}}$ must be larger than that of $\hat{\boldsymbol{\theta}}$, in the usual sense. This is akin to the Gauss-Markov Theorem: any unbiased linear estimator can be written as the sum of the OLS estimator and a random component which has mean zero and is uncorrelated with the OLS estimator.

This asymptotic efficiency result is really an asymptotic version of the Cramer-Rao lower bound: that the covariance matrix of an unbiased estimator can never be smaller than $\mathbf{I}^{-1}$, which we have seen is asymptotically equal to the covariance matrix of the MLE.

Problem Set

In all questions the data is IID, and a sample size of n is observed. For a random variable, the notation $w_i \in \mathbb{R}^{d_w}$ should be interpreted as a shorthand for 'w_i is a d_w-dimensional random vector'. For conciseness, I generally drop the '0' subscript from true parameter values.

Q1 Convergence in probability and asymptotic distributions

Let $\{x_i\}_{i=1}^n$ be an IID sample with mean 0 and variance σ^2. Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $s_n = \frac{1}{n}\sum_{i=1}^n x_i^2$, and consider the following statistics:

1. $\bar{x}_n$;
2. s_n;
3. $\log s_n$;
4. $\bar{x}_n / s_n$;
5. $\log(1 + \bar{x}_n)$;
6. $\bar{x}_n^2 / s_n$.

For each of these statistics, find the probability limit, and then the asymptotic distribution of the rescaled statistic around that limit. For example, in 1) we have $\bar{x}_n \xrightarrow{p} 0$, and $\sqrt{n}(\bar{x}_n - 0) = \sqrt{n}\bar{x}_n \overset{a}{\sim} N(0, \sigma^2)$. State all results used, and any additional assumptions that are needed to apply those results.

2) For s_n, we have

$$\begin{aligned}\operatorname*{plim}_{n\to\infty} s_n &= \mathbb{E}[x_i^2] \\ &= \operatorname{Var}(x_i) = \sigma^2,\end{aligned}$$

and a scaled statistic around the plim:

$$\sqrt{n}(s_n - \sigma^2) \overset{a}{\sim} N(0, V),$$

where $V = \operatorname{Var}(x_i^2) = \mathbb{E}\left[(x_i^2 - \mathbb{E}[x_i^2])^2\right] = \mathbb{E}\left[x_i^4\right] - \sigma^4$. So we have

$$\sqrt{n}(s_n - \sigma^2) \xrightarrow{d} N(0, \mathbb{E}\left[x_i^4\right] - \sigma^4). \tag{47}$$

3) For $\log s_n$, we have

$$\operatorname*{plim}_{n\to\infty} \log s_n = \log \mathbb{E}[x_i^2] = \log \sigma^2,$$

via the Continuous Mapping Theorem (CMT), provided that $\sigma^2 \neq 0$. To find the asymptotic distribution, we need to use the delta method. To start, we know that

$$\sqrt{n}(s_n - \sigma^2) \overset{a}{\sim} N(0, \mathbb{E}[x_i^4] - \sigma^4),$$

and so, taking a first-order Taylor expansion about $\mathbb{E}[x_i^2] = \sigma^2$, and applying the right power of n, we get:

$$\begin{aligned} \log s_n &\approx \log \sigma^2 + \frac{1}{2\sigma}(s_n - \sigma^2) \\ \therefore \sqrt{n}(\log s_n - \log \sigma^2) &\overset{a}{=} \frac{2}{\sigma}\sqrt{n}(s_n - \sigma^2). \end{aligned}$$

Combining with (47), we have:

$$\sqrt{n}(\log s_n - \log \sigma^2) \overset{a}{\sim} N\left(0, \frac{4}{\sigma^2}\left[\mathbb{E}[x_i^4] - \sigma^4\right]\right). \tag{48}$$

4) For $\bar{x}_n/s_n$ we can use our previous results and apply CMT/Slutsky's Theorem:

$$\operatorname*{plim}_{n\to\infty}\left(\frac{\bar{x}_n}{s_n}\right) = \mathbb{E}\left[\frac{x_i}{x_i^2}\right] = \frac{0}{\sigma^2} = 0.$$

Likewise, for the asymptotic distribution, use CMT/Slutsky's Theorem:

$$\sqrt{n}\left(\frac{\bar{x}_n}{s_n} - 0\right) = \sqrt{n}\frac{\bar{x}_n}{s_n} \overset{a}{\sim} N\left(0, \frac{\sigma^2}{\sigma^4}\right). \tag{49}$$

5) For $\log(1 + \bar{x}_n)$, we have

$$\operatorname*{plim}_{n\to\infty} \log(1 + \bar{x}_n) = \log(1 + \mathbb{E}[x_i]) = 0,$$

via the CMT. Then, for the asymptotic distribution, we use the delta method. Take a first-order Taylor expansion about 0:

$$\log(1 + \bar{x}_n) \approx 0 + \frac{1}{1+0}(\bar{x}_n - 0),$$

then multiply by the appropriate powers of n to get:

$$\begin{aligned} \sqrt{n}\log(1 + \bar{x}_n) &\overset{a}{=} \sqrt{n}(\bar{x}_n - 0) \\ \therefore \sqrt{n}\log(1 + \bar{x}_n) &\overset{a}{\sim} N(0, \sigma^2). \end{aligned} \tag{50}$$

6) For $\bar{x}_n^2/s_n$, we have

$$\operatorname*{plim}_{n\to\infty}\left(\frac{\bar{x}_n^2}{s_n}\right) = \frac{\mathbb{E}[x_i]^2}{\mathbb{E}[x_i^2]} = \frac{0^2}{\sigma^2} = 0,$$

by CMT/Slutsky's Theorem. For the asymptotic distribution, start by rewriting the quantity as

$$\frac{\bar{x}_n \bar{x}_n}{s_n},$$

and then distribute $\sqrt{n}$ for each of the mean terms:

$$\frac{\sqrt{n}\bar{x}_n\sqrt{n}\bar{x}_n}{s_n} = n\frac{\bar{x}_n\bar{x}_n}{s_n} = \frac{(\sqrt{n}\bar{x}_n)^2}{s_n}.$$

This may seem a bit obtuse, but it's to illustrate the importance of allocating $\sqrt{n}$ when we use the CLT. Now, we can deduce that the numerator has the following asymptotic distribution:

$$\begin{aligned}(\sqrt{n}\bar{x}_n)^2 &\overset{a}{\sim} \big(N(0,\sigma^2)\big)^2 \\ &= (\sigma N(0,1))^2 \\ &= \sigma^2 N(0,1)^2 \\ &= \sigma^2\chi^2(1),\end{aligned}$$

if you recall that a squared standard normal distribution was nothing but a chi-square distribution (with 1 degree of freedom). Now, putting it all together, we get:

$$\frac{(\sqrt{n}\bar{x}_n)^2}{s_n} \overset{a}{\sim} \frac{\sigma^2\chi^2(1)}{\sigma^2} = \chi^2(1). \tag{51}$$

Q2 Censored Gaussian distribution

Suppose that we observe

$$w_i = \begin{cases} 0 & \text{if } w_i^* \leq 0, \\ w_i^* & \text{if } w_i^* \in (0, c), \\ c & \text{if } w_i^* \geq c, \end{cases}$$

where $w_i^* \sim N(\mu, \sigma^2)$, and $c > 0$ is a known right hand censor point. Given the sample $\{w_i\}_{i=1}^n$, construct the log-likelihood for the parameters $\boldsymbol{\theta} = [\mu, \sigma]^\top$.

The density is given by

$$f(\mathbf{w}; \mu, \sigma) = \mathbf{1}(w_i = 0)\Phi\left(-\frac{\mu}{\sigma}\right) + \mathbf{1}(w_i = w_i^* \in (0, c))\frac{\phi}{\sigma}\left(\frac{w - \mu}{\sigma}\right) + \mathbf{1}(w_i = c)\left[1 - \Phi\left(\frac{c - \mu}{\sigma}\right)\right],$$

and the log-likelihood is given by

$$\begin{aligned} l(\mathbf{w}; \mu, \sigma) = \log f(\mathbf{y}; \mu, \sigma) &= \sum_{i=1}^n l_i(w_i; \mu, \sigma) \\ &= \sum_{i=1}^n \log \left\{ \begin{array}{c} \mathbf{1}(w_i = 0) \cdot \Phi\left(-\frac{\mu}{\sigma}\right) + \mathbf{1}(w_i = w_i^* \in (0, c))\frac{\phi}{\sigma}\left(\frac{w-\mu}{\sigma}\right) \\ +\mathbf{1}(w_i = c)\left[1 - \Phi\left(\frac{c-\mu}{\sigma}\right)\right] \end{array} \right\}. \end{aligned}$$

Q2.1)

Now specialise to the case where $c = \infty$, i.e. where there is only censoring at zero. Show that if $w_i = 0$ for at least one (but not all) i, then the MLE for μ lies strictly to the left of the sample mean. [Hint: Show that $\partial l_n(\mu, \sigma)/\partial\mu < 0$ for all $\mu \geq n^{-1}\sum_{i=1}^n w_i$ and $\sigma^2 > 0$. You may use the fact that:

$$\nu[1 - \Phi(\nu)] - \phi(\nu) < 0,$$

for all $\nu \geq 0$.]

The sample log-likelihood for when $c = \infty$ is given by

$$\begin{aligned} l_n(\mathbf{w}; \mu, \sigma) &= \sum_{i=1}^n \log\left\{\mathbf{1}(w_i = 0)\Phi\left(-\frac{\mu}{\sigma}\right) + \mathbf{1}(w_i > 0)\frac{\phi}{\sigma}\left(\frac{w - \mu}{\sigma}\right)\right\} \\ &= n_0 \log\Phi\left(-\frac{\mu}{\sigma}\right) - \frac{n_1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i|w_i>0}^{n_1}(w_i - \mu)^2, \end{aligned}$$

where n_0 and n_1 are the number of censored and uncensored observations, respectively. Taking the derivative with respect to μ gives:

$$\begin{aligned} \frac{\partial l_n}{\partial\mu} &= -n_0\frac{\phi\left(-\frac{\mu}{\sigma}\right)}{\sigma\Phi\left(-\frac{\mu}{\sigma}\right)} - \frac{1}{\sigma^2}\sum_{i|w_i>0}^{n_1}(w_i - \mu)(-1), \\ &= -n_0\frac{\phi\left(-\frac{\mu}{\sigma}\right)}{\Phi\left(-\frac{\mu}{\sigma}\right)} - \frac{1}{\sigma}\sum_{i|w_i>0}^{n_1}(w_i - \mu) \\ \implies n_0\frac{\phi\left(-\frac{\mu}{\sigma}\right)}{\Phi\left(-\frac{\mu}{\sigma}\right)} &= -\frac{1}{\sigma}(n\bar{w} - n\mu). \end{aligned} \tag{52}$$

We can see that if $n_0 = 0$, then the RHS of the above equation must be equal to zero, which implies that μ lies in the middle of the sample mean. Thus for any values $n_0 > 0$, it must be that μ lies to the left of the sample mean.

Q3 Binary response model

Consider the binary response model:

$$y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \sigma u_i,$$

where y_i^* is the unobserved latent variable, $u_i \sim F$ is independent of $\mathbf{X}_i$, and

$$y_i = \mathbf{1}(y_i^* \geq 0)$$

is the unobserved binary outcome. Supposing that the first element of $\mathbf{X}_i$ is continuously distributed, derive the marginal effect of a change in $X_{1,i}$ on the conditional probability of a success $\Pr(y_i = 1|\mathbf{X}_i = \mathbf{X})$. Show, in particular, that this depends only on the ratio $\boldsymbol{\beta}/\sigma$, and not on the parameters $\boldsymbol{\beta}$ and σ separately. You may assume F is symmetric.

Begin by writing the CDF

$$\begin{aligned}\Pr(y_i = 1|\mathbf{X}_i = \mathbf{X}) &= \Pr(\mathbf{X}_i\boldsymbol{\beta} + \sigma\mu \geq 0|\mathbf{X}_i = \mathbf{X}) \\ &= \Pr(\mathbf{X}_i\boldsymbol{\beta} \geq -\sigma\mu|\mathbf{X}_i = \mathbf{X}) \\ &= \Pr\left(-\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma} \geq \mu|\mathbf{X}_i = \mathbf{X}\right) \\ &= F\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right),\end{aligned}$$

where the last line is attained since F is symmetric. To get the marginal effect, begin by partitioning $\mathbf{X}_i = (X_{1,i}, X_{-1,i})$. Then differentiate:

$$\begin{aligned}\frac{\partial \Pr(y_i = 1|\mathbf{X}_i = \bar{\mathbf{X}})}{\partial \bar{x}_1} &= \frac{\partial F\left(\frac{\bar{X}_1\beta_1}{\sigma} + \frac{\bar{\mathbf{X}}_{-1}\boldsymbol{\beta}_{-1}}{\sigma}\right)}{\partial \bar{X}_1} \\ &= \frac{\beta_1}{\sigma} f\left(\frac{\bar{\mathbf{X}}\boldsymbol{\beta}}{\sigma}\right). \qquad (53)\end{aligned}$$

Thus, the marginal effect of a change in $X_{1,i}$ depends on the ratio β/σ.

Q4 Average log-likelihood

Let $\{w_i\}_{i=1}^n$ be an IID sample with joint density:

$$f(\mathbf{w};\mu,\sigma)=\frac{1}{2\sigma}\exp\left(-\frac{|\mathbf{w}-\mu|}{\sigma}\right).$$

Show that the MLE of μ is the sample median of w_i. If it helps, you may assume that n is odd.

Since the sample observations are IID, the joint density is

$$\prod_{i=1}^n f(w_i;\mu,\sigma). \tag{54}$$

This implies that the likelihood function is

$$L(\mu,\sigma)=\prod_{i=1}^n \frac{1}{2\sigma}\exp\left(-\frac{|w_i-\mu|}{\sigma}\right), \tag{55}$$

and that the average log-likelihood is

$$\begin{aligned}
\bar{l}_n(\mu,\sigma) &= \frac{1}{n}\log L(\mu,\sigma)=\frac{1}{n}\log\left(\prod_{i=1}^n \frac{1}{2\sigma}\exp\left(-\frac{|w_i-\mu|}{\sigma}\right)\right)\\
&= \frac{1}{n}\left\{\sum_{i=1}^n \log\left(\frac{1}{2\sigma}\exp\left(-\frac{|w_i-\mu|}{\sigma}\right)\right)\right\}\\
&= -\frac{1}{n}n\log 2\sigma-\frac{1}{n}\sum_{i=1}^n\frac{|w_i-\mu|}{\sigma}\\
&= -\log 2\sigma-\frac{1}{n}\sum_{i=1}^n\frac{|w_i-\mu|}{\sigma} \qquad (56)
\end{aligned}$$

Thus, the MLE for μ is nothing but an M-estimator with the following objective function that is a sample average:

$$Q_n(\mu,\sigma)=\frac{1}{n}\sum_{i=1}^n\underbrace{\frac{|w_i-\mu|}{\sigma}}_{m(w_t;\mu,\sigma)},$$

where m is a real-valued function of (w_i,μ,σ) (recall (2)). Since $\hat{\mu}$ is an M-estimator which satisfies $Q_n(\mu,\sigma)=0$, and since σcannot equal 0, it must be that $\hat{\mu}$ is the sample median of $\mathbf{w}$.

Q5 The uniform distribution and misspecification

Let $w_i \overset{IID}{\sim} F$ for some unknown distribution F. We model $\{w_i\}_{i=1}^n$ as having been drawn from a uniform distribution $U[-\theta, \theta]$ for some $\theta \in \mathbb{R}_+$. Derive the likelihood and a closed-form expression for the MLE $\hat{\theta}_n$. Suppose that $F = U[-\theta_0, \theta_0]$ for some $\theta_0 \in \mathbb{R}_+$. Show $\hat{\theta}_n \xrightarrow{p} \theta_0$.

The density for w_i is:

$$f(w_i, \theta) = \begin{cases} 0 & \text{if } w_i < -\theta, \\ \frac{1}{2\theta} & \text{if } -\theta \leq w_i \leq \theta, \\ 0 & \text{if } w_i > \theta. \end{cases}$$

provided that $-\theta \leq w_i \leq \theta$ for all observations, the likelihood function is equal to:

$$f(w_i, \theta) = \frac{1}{(2\theta)^n},$$

and thus the log-likelihood function is therefore

$$l(\mathbf{w}, \theta) = -n \log(2\theta). \tag{57}$$

This function cannot be maximised by differentiating it with respect to θ and setting the derivative equal to 0. Instead, the way to maximise $l(\mathbf{w}, \theta)$ is to make 2θ as small as possible. But we cannot make θ smaller than the largest observed w_i, otherwise the likelihood function would be equal to 0. It follows that the MLE is:

$$\hat{\theta} = \max |w_i|. \tag{58}$$

The MLE $\hat{\theta} \leq \theta_0$ since w_i must lie between $-\theta$ and θ. Despite this, the MLE is consistent. This is because as the sample size n gets large, the observed values of w_i fill up the space between $-\theta_0$ and θ_0.

To see this, let $\epsilon \in (0, \theta_0)$. Since $\hat{\theta} = \max |w_i|$, we have:

$$\begin{aligned} \Pr(|\hat{\theta} - \theta_0| > \epsilon) &= \Pr(\hat{\theta} > \epsilon + \theta_0) \\ &= \Pr(\hat{\theta} < \theta_0 - \epsilon) \\ &= \prod_{i=1}^n \Pr(|w_i| < \theta_0 - \epsilon) \\ &= \left(1 - \frac{\epsilon}{\theta_0}\right)^n \\ &\to 0. \end{aligned}$$

Q5.1)

Now suppose the model is misspecified: In particular, $F = N(0, 1)$. Let ϕ denote the density of a $N(0, 1)$ distribution, and u_θ the density of a $U[-\theta, \theta]$ distribution. Show that

$d_{KL}(\phi||u_\theta) = \infty$ for all $\theta \in \mathbb{R}_+$. Show $\hat{\theta}_n \xrightarrow{p} \infty$. (For a random sequence $\{x_n\}$, $x_n \xrightarrow{p} \infty$ means: for every $M < \infty, \Pr(x_n \geq M) \to 1$ as $n \to \infty$.)

The Kullback-Leibler divergence between ϕ and u_θ is:

$$\begin{aligned} d_{KL}(\phi||u_\theta) &= \mathbb{E}_\phi \log \frac{\phi(w_i)}{u(w_i)} \\ &= \int_{\{\mathbf{w} \in \mathbb{R} | \phi(\mathbf{w}) > 0} \phi(\mathbf{w}) \log \left(\frac{\phi(\mathbf{w})}{u(\mathbf{w}, \theta)} \right) d\mathbf{w} \\ &= \int_{\{\mathbf{w} \in [-\theta, \theta]\}} \phi(\mathbf{w}) \log \left(2\theta\phi(\mathbf{w}) \right) d\mathbf{w} + \int_{\{\mathbf{w} \notin [-\theta, \theta]\}} \phi(\mathbf{w}) \infty d\mathbf{w} \\ &= \infty, \end{aligned}$$

since for the uniform distribution, if you try to integrate outside the interval of $[-\theta, \theta]$, you get a value of infinity.

Now, let $M < \infty$. Then:

$$\begin{aligned} \Pr(\hat{\theta}_n \leq M) &= \Pr(\max |w_i| \leq M) \\ &= \prod_{i=1}^{n} \Pr(|w_i| \leq M) \\ &= (1 - 2\Phi(-M))^n \\ &\to 0, \end{aligned}$$

hence $\hat{\theta}_n \xrightarrow{p} \infty$.