Econ 2120: Section 9

MLE and Minimum Distance Estimation

Ashesh Rambachan

Fall 2018

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Maximum Likelihood Estimation

We'll now transition to another estimation strategy – maximum likelihood estimation. This is an example of an **M-estimator**, which is another class of **extremum estimators**.

Can think of maximum likelihood estimation as a maximum a posteori estimator with a flat prior.

If you believe that the likelihood is well-specified, then MLE has some extremely attractive properties.

Today: Just want to set-up the MLE estimator and next time, we'll discuss its properties.

Set Up

As always, assume random sampling $W_i \sim F$ i.i.d for $i = 1, \dots, n$. We specify a set of distributions

$$\{P_\theta : \theta \in \Theta\}.$$

Each distribution has a density that is well-defined. We denote it $f(w|\theta)$. For a given θ, $f(w|\theta)$ is the **likelihood function**.

We say the model is **well-specified** if for some $\theta^* \in \Theta$,

$$F = P_{\theta^*}.$$

If it is possible that there does not exist such a θ^*, we refer to f as a **pseudo-likelihood** or a **quasi-likelihood**.

Set Up

We could also set this up in terms of **conditional likelihoods.** That is, divide W_i into Y_i, Z_i and then we model the distribution of $Y_i|Z_i$. Our family of distributions

$$\{P_\theta : \theta \in \Theta\}$$

is now a set of conditional likelihood functions.

> **Example**: Normal linear model – we modeled $Y_i|X_i \sim N(X_i'\beta, \sigma^2)$. So, the set of distributions was the normal family with mean parameterized by β and variance σ^2.

MLE: Definition

We observe $W_1, \ldots, W_n$. Fix a particular value of θ. The joint density of $W_1, \ldots, W_n|\theta$ is given by

$$f(W_1, \ldots, W_n|\theta) = \Pi_{i=1}^{n} f(W_i|\theta).$$

The **maximum likelihood estimator** is the value of the parameter that maximizes the likelihood of the observed data.

$$\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} \Pi_{i=1}^{n} f(W_i|\theta).$$

Equivalently, we can take the log of the objective function because it is a monotone function. And define the **log-likelihood**

$$L_n(W_1, \ldots, W_n|\theta) = \sum_{i=1}^{n} \log(f(W_i|\theta))$$

Then,

$$\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} L_n(W_1, \ldots, W_n|\theta).$$

Outline

- Maximum Likelihood Estimation
  - Set up
  - Probit and Logit
  - Information Inequality
  - Neyman-Pearson Lemma and Optimal Tests
  - Best Approximation
  - Inference for MLE
  - Multivariate Normal Linear Model and Robustness
- Minimum Distance Estimation
  - Example
  - Set up
  - Delta Method

Probit and Logit applications

Not going to cover this in section – I assume this is material you have seen before. If not, make sure to read these sections of Note 10.

Next time: What are the properties of MLE? How can we derive its limit distribution?

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Entropy

We typically work with discrete random variables in this class, so all the definitions are presented for the discrete case. Of course, they generalize.

Let X be a random variable with values in $\mathcal{X}$ and pmf $p(x)$. The **entropy** $H(X)$ is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = E[\log \frac{1}{p(X)}].$$

We typically have the log be base 2. In this case, entropy is expressed in bits. When it is the natural log, it is measured in "nats."

Think of entropy as encoding the amount of "information" that can be learned from a random variable or how much uncertainty is present.

Entropy

You can arrive at this axiomatically. Suppose we wish to measure the amount of information that is generated from observing an event occur that has probability p.

Let $I(p)$ be the information function. We want it to satisfy:

(1) Information is non-negative, $I(p) \geq 0$.
(2) $I(1) = 0$ – events that always occur produce no information.
(3) If two events are independent with probabilities p_1, p_2, then the information produced by observing both events occur is additive: $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$.
(4) Information is monotone decreasing – more likely events produce less information.

You can show that $I(p) = -\log_b(p)$ for some base b.

Then, the entropy of an random variable $H(X)$ is the average information it produces.

The information inequality

Let F^* denote the true population distribution and suppose that F^* has an associated density f^*. Let E_F denote the expectation with respect to the true population distribution.

If the likelihood function is mis-specified, then there does NOT exist $\theta \in \Theta$ such that $f^*(w) = f(w|\theta)$.

Information Inequality: For all $\theta \in \Theta$,

$$E_F[\log f(W_i|\theta)] \leq E_F[\log f^*(W_i|\theta)].$$

This is also known as the **Shannon-Kolmogorov Information Inequality**.

Kullback Leibler distance or relative entropy

The **relative entropy** or **Kullback Leibler** distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p[\log \frac{p(X)}{q(X)}].$$

This measures the directed distance from p to q – it is always non-negative and zero if and only if $p = q$. It is a directed distance because it is not symmetric. In statistics, it measures the expected logarithm of the likelihood ratio between two distributions.

A larger Kullback Leibler distance or relative entropy $\implies$ more information is lost when we approximate p by q or similarly the worse an approximate q provides for p.

We could also write the **information inequality** as

$$D(f^*||f(\cdot|\theta)) \geq 0.$$

Proof of information inequality

Define $Q = \frac{f(W_i|\theta)}{f^*(W_i)}$. By Jensen's inequality, we have that

$$E_F[\log Q] \leq \log E_f[Q].$$

Note that

$$\begin{aligned}\log E_f[Q] &= \log \int \frac{f(w|\theta)}{f^*(w)} f^*(w) dw \\ &= 0.\end{aligned}$$

So, we have that

$$E_F[\log Q] = E_F[\log f(W_i|\theta) - \log f^*(W_i)] \leq 0.$$

The result follows.

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Optimal Tests

It will feel like we are completely changing gears here. We're now going to consider the problem of constructing optimal hypothesis tests. The solution to this problem will be deeply connected to **relative entropy**.

Suppose that the distribution of Y_i is discrete with finite support. We observe n observations $Y = (Y_1, \dots, Y_n)$ and we assume the data are i.i.d. with

$$P(Y_i = \alpha_j|\theta) = \theta_j \quad j = 1, \dots, J.$$

The parameter space is the unit simplex with $\Theta = \{\theta \in \mathbb{R}^J : \theta_j \geq 0, \sum_{j=1}^J \theta_j = 1\}$. The likelihood function is then

$$\begin{aligned} f(y|\theta) &= \Pi_{i=1}^n P(Y_i = y_i|\theta) \\ &= \Pi_{i=1}^n \Pi_{j=1}^J \theta_j^{1\{Y_i = \alpha_j\}} \\ &= \Pi_{j=1}^J \theta_j^{n_j}, \quad n_j = \sum_{j=1}^J 1\{Y_i = \alpha_j\}. \end{aligned}$$

Optimal Tests

We wish to test the following null against the alternative

$$H_0 : \theta = \theta^{(0)}, \quad H_a : \theta = \theta^{(1)}.$$

For a test procedure d, we define

Type 1 Error: $e_1(d) = P\{\text{Reject null}|\theta = \theta^0\}$.

Type 1 Error $\implies$ Shawshank Redemption.

The probability of Type 1 error is referred to as the **size** of a test.

Type 2 Error: $e_2(d) = P\{\text{Accept null} : \theta = \theta^1\}$.

Type 2 Error $\implies$ OJ Simpson.

The probability that the null is rejected given that the alternative is true is referred to as the **power** of a test.

Our test procedure will be based on realizations of the r.v. Y. They will specify a **critical region** W such that if $Y \in W$, we reject H_0 and otherwise we fail to reject H_0. **How do we choose the critical region?**

Optimal Tests – Neyman-Pearson Lemma

Classical approach to hypothesis testing: We choose the critical region to maximize power subject to a size constraint.

> Minimize the probability of type 2 error given a pre-specified rate of type 1 error.

Likelihood ratio tests: For some constant c,

> H_0 is accepted if $f(y|\theta^1) < cf(y|\theta^0)$.
>
> H_a is accepted if $f(y|\theta^1) > cf(y|\theta^0)$.

Theorem: Neyman-Pearson Lemma

> If d is a test procedure with $e_1(d) \leq e_1(d_{LR})$, then $e_2(d) \geq e_2(d_{LR})$. If $e_1(d) < e_1(d_{LR})$, then $e_2(d) > e_2(d_{LR})$.
>
> Equivalently, subject to a size constraint α, power is maximized by choosing a critical region based on the likelihood ratio $f(y|\theta^1)/f(y|\theta^0)$, where the constant c is chosen to satisfy the size constraint:
>
> $$P(f(y|\theta^1)/f(y|\theta^0)|\theta = \theta^0) = \alpha.$$

Neyman-Pearson Lemma

The Neyman-Pearson Lemma shows that the likelihood ratio is admissible – it cannot be domianted by some other test.

We can relate likelihood ratio tests to the Kullback Leibler distance. Let

$$f_0(\alpha_j) = \theta_j^0, \quad f_1(\alpha_j) = \theta_j^1, \quad \hat{f}(\alpha_j) = n_j/n.$$

Then, $D(\hat{f}||f_0), D(\hat{f}||f_1)$ is the distance from the empirical distribution to the distribution under the null and the distribution under the alternative. The likelihood ratio test simply compares these distances.

We have that (see the notes for the derivation):

$$\frac{1}{n}\log\frac{f(y|\theta^1)}{f(y|\theta^0)} = D(\hat{f}||f_0) - D(\hat{f}||f_1).$$

So, we have that

$$LR > c \iff D(\hat{f}||f_0) - D(\hat{f}||f_1) > \frac{1}{n}\log(c).$$

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Best Approximation

Define

$$\theta_F = \arg\max_{\theta\in\Theta} E_F[\log f(W_i|\theta)].$$

Or equivalently,

$$\theta_F = \arg\min_{\theta\in\Theta} E_F[\log \frac{f^*(W_i)}{f(W_i|\theta)}] = \arg\min_{\theta\in\Theta} K(f^*.f(\cdot|\theta)).$$

That is, θ_F minimizes the KL distance from the true population density over all possible densities in the model. If the model is well-specified such that $f^* = f(\cdot|\theta^*)$ for some $\theta^* \in \Theta$, then $\theta_F = \theta^*$.

We'll argue that the MLE is consistent for this θ_F, which we call the **best approximation**.

MLE is consistent for the best approximation

The MLE solves

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(W_i|a).$$

Under some regularity conditions (you'll see these in 2140), we can obtain a *uniform law of large numbers*:

$$\sup_{a \in \Theta} |\frac{1}{n} \sum_{i=1}^{n} \log f(W_i|a) - E_F[\log f(W_i|a)]| \xrightarrow{p} 0.$$

Then, it can be shown that

$$\hat{\theta} \xrightarrow{p} \theta_F.$$

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

How do we do inference on the MLE?

We can use our GMM results!

The best approximation satisfies

$$\theta_F = \arg\max_{\theta \in \Theta} E_F[\log f(W_i|\theta)].$$

The FOC is

$$E_F[\psi(W_i, \theta_F)] = 0, \quad \psi(W_i, \theta) = \frac{\partial \log f(W_i|\theta)}{\partial \theta}.$$

ψ is a moment function that equals zero at $\theta = \theta_F$. It is referred to as the **score function**.

Because $dim(\psi) = dim(\theta)$, we are just-identified and can set $\hat{D} = I$. The GMM estimator then satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i, \hat{\theta}) = 0.$$

MLE – asymptotic distribution

We have that

$$\sqrt{n}(\hat{\theta} - \theta_F) \xrightarrow{d} N(0, \Lambda),$$

where

$$\begin{aligned}
\Lambda &= H^{-1}\Sigma H^{-1}, \\
H &= E_F[\frac{\partial \psi(W_i, \theta_F)}{\partial \theta'}] = E_F[\frac{\partial^2 \log f(W_i|\theta_F)}{\partial \theta \partial \theta'}] \\
\Sigma &= E_F[\psi(W_i, \theta_f)\psi(W_i, \theta_F)'] = E_F[\frac{\partial \log f(W_i|\theta_F)}{\partial \theta}\frac{\partial \log f(W_i|\theta_F)}{\partial \theta'}].
\end{aligned}$$

If the model is well-specified, then $H = \Sigma$ and so, the asymptotic variance simplifies to be H^{-1}, which is the inverse of the **information matrix**. Well-specified MLE asymptotically achieves the Cramer-Rao lower bound.

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Multivariate Normal Linear Model

We begin by extending the normal linear model to the case in which Y_i is an $M \times 1$ vector. The parameter is now $\theta = (\Pi, \Sigma)$, where Π is a $K \times M$ vector and Σ is an $M \times M$ symmetric, positive definite matrix.

The likelihood function for a single observation y is

$$f(y|x,\theta) = (2\pi)^{-M/2} det(\Sigma)^{-1/2} \exp\{-\frac{1}{2}(y - \Pi' x)' \Sigma^{-1} (y - \Pi' x)\},$$

where y is $M \times 1$ and x is $K \times 1$.

We define the best approximation, which solves

$$\theta_F = (\Pi_F, \Sigma_F) = \arg\max_{\Pi,\Sigma} - \log det(\Sigma) - E_F[(Y_i - \Pi' X_i)' \Sigma^{-1} (Y_i - \Pi' X_i)].$$

Multivariate normal linear model and best approximation

Define $\Pi^* = (E_F[X_i X_i'])^{-1} E_F[X_i Y_i']$. It is simple to show that

$$\Pi_F = \Pi^*.$$

We can also show that

$$\Sigma_F = \Sigma^* = E_F[(Y_i - \Pi^{*\prime} X_i)(Y_i - \Pi^{*\prime} X_i)'].$$

See the notes for both arguments.

Robustness of Quasi-MLE

Suppose that there are additional restrictions placed on Σ, Π. We express these by specifying functions

$$\Pi(\theta), \Sigma(\theta) \text{ with } \theta \in \Theta.$$

We can analogously write the likelihood as

$$f(y|x,\theta) = (2\pi)^{-M/2} det(\Sigma(\theta))^{-1/2} \exp\{-\frac{1}{2}(y - \Pi(\theta)'x)'\Sigma(\theta)^{-1}(y - \Pi(\theta)'x$$

and the best approximation solves

$$\theta_F = \arg\max_{\theta \in \Theta} - \log det(\Sigma(\theta)) - E_F[(Y_i - \Pi(\theta)'X_i)'\Sigma(\theta)^{-1}(Y_i - \Pi(\theta)'X_i)].$$

The robustness property means that: even if the population distribution is not a multivariate normal provided that the mean, covariance matrix functions are well-specified i.e.

$$\Sigma^* = \Sigma(\theta^*), \Pi^* = \Pi(\theta^*) \text{ for some } \theta^* \in \Theta,$$

then

$$\Theta_F = \Theta^*.$$

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Minimum Distance Estimation

We now transition to the final tool for estimation and inference that we'll cover in this course. The set-up will be different but the asymptotic arguments will be familiar.

Our inputs are some sample statistics.

> Think of these as unrestricted estimates such as reduced-form least-squares estimates or unrestricted sample covariances.

Minimum distance will be a tool for imposing restrictions on these unrestricted sample statistics.

Before going to the general set-up, I want to offer a simple example.

Minimum Distance Estimation and IV

Suppose our model is

$$\begin{aligned} Y_i &= X_i\beta + U_i, \\ X_i &= Z_i\pi + V_i, \end{aligned}$$

where U_i, V_i are correlated and $E[Z_i V_i] = E[Z_i U_i] = 0$. The model implies restrictions on the reduced-form regressions. We have that

$$E^*[Y_i|Z_i] = \gamma Z_i, \text{ with } \gamma = \beta \cdot \pi, E^*[X_i|Z_i] = \lambda Z_i, \text{ with } \lambda = \pi$$

So, one way to estimate β, π would be with minimum distance. First, construct estimates of the reduced-form coefficients

$$\hat{\gamma} = \frac{\sum_{i=1}^n Y_i Z_i}{\sum_{i=1}^n Z_i^2}, \quad \hat{\lambda} = \frac{\sum_{i=1}^n X_i Z_i}{\sum_{i=1}^n X_i^2}.$$

Minimum Distance Estimation and IV

Then, we estimate π, β by solving

$$\arg\min_{b,p} \begin{pmatrix} \hat{\gamma} - b \cdot p \\ \hat{\lambda} - p \end{pmatrix}' C \begin{pmatrix} \hat{\gamma} - b \cdot p \\ \hat{\lambda} - p \end{pmatrix},$$

for some weight matrix 2×2 C. That is, we estimate β, π by finding the parameter values that best approximate the unrestricted reduced-form coefficients while imposing the restrictions of our IV model.

This is all that minimum distance is doing. Keep this example in mind as it's a good sanity check as we're moving along.

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Set up

We are given a statistic $\hat{\pi}$ with limit distribution

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} N(0, \Omega).$$

We are also given a distance functon $h(\cdot, \cdot)$, which is continuously differentiable. We assume that there is a unique parameter γ such that

$$h(\pi, \gamma) = 0.$$

Here, we have that h is $L \times 1$, γ is $K \times 1$ and $L \geq K$.

Set up

The **minimum distance estimator** solves

$$\hat{\gamma} = \arg\min_a h(\hat{\pi}, a)' \hat{C} h(\hat{\pi}, a),$$

where $\hat{C}$ converges in probability to a $L \times L$ non-random, positive definite, symmetric matrix C. The FOC is

$$(\partial h(\hat{\pi}, \hat{\gamma}) / \partial a)' \hat{C} h(\hat{\pi}, \hat{\gamma}) = 0.$$

So, we could also write the minimum distance estimator as satisfying

$$\hat{D} h(\hat{\pi}, \hat{\gamma}) = 0, \quad \hat{D} = (\partial h(\hat{\pi}, \hat{\gamma}) / \partial a)' \hat{C}.$$

Minimum distance – limit distribution

The derivation of the limiting distribution is very simple and follows our argument from GMM.

I'll leave this for you to review in the notes.

Outline

Maximum Likelihood Estimation

- Set up
- Probit and Logit
- Information Inequality
- Neyman-Pearson Lemma and Optimal Tests
- Best Approximation
- Inference for MLE
- Multivariate Normal Linear Model and Robustness

Minimum Distance Estimation

- Example
- Set up
- Delta Method

Delta Method

Suppose we wish to derive the asymptotic distribution of a non-linear function of parameters.

Consider $\gamma = g(\pi) \in \mathbb{R}^K$ for some continuously differentiable function g. Suppose that

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} N(0, \Omega).$$

Then, $\hat{\gamma} = g(\hat{\pi})$ is a consistent estimator. Suppose $\pi \in \mathbb{R}^L$ What is it's asymptotic distribution?

We use the **delta method** to derive it and an easy way to derive the delta method is through minimum distance estimation.

Delta Method and Minimum Distance

Define the distance function $h(\hat{\pi}, a) = g(\hat{\pi}) - a$. The minimum distance estimator is trivially $\hat{\gamma}) = g(\hat{\pi})$. So, we have that

$$h(\hat{\pi}, \hat{\gamma}) = 0.$$

So, we can just set $D = I$ and so, $\alpha = -I$. We then get that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \Sigma), \quad \Sigma = \frac{\partial g(\pi)}{\partial \pi} \Omega \frac{\partial g(\pi)}{\partial \pi'},$$

where this is just from our minimum distance results.