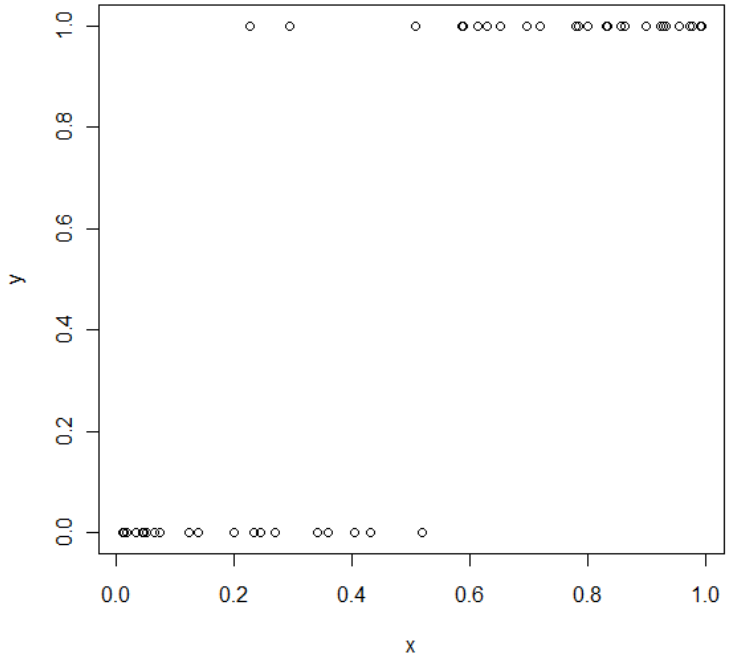
Maximum Likelihood Estimation

Walter Sosa-Escudero

Econ 507. Econometric Analysis. Spring 2009

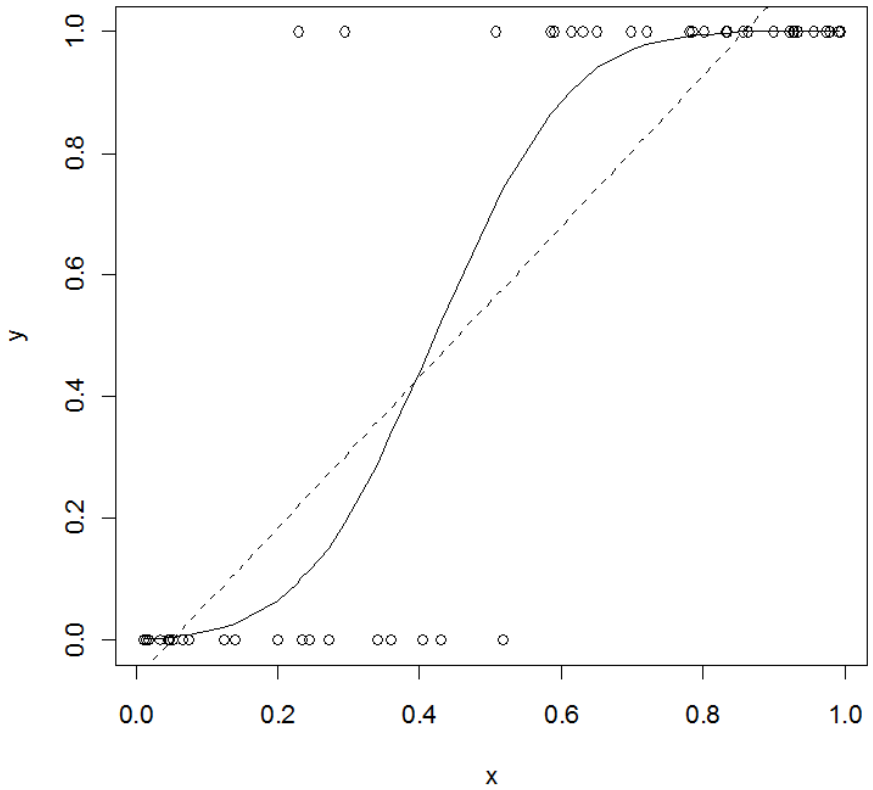
April 13, 2009

Consider the following data set



- The explained variable is binary.
- Admission to grad school depending on GRE score, Families send kids to school as a function of income, etc.

This case is a good candidate for a non-linear model



How to search for a model and estimate its parameters?

Consider the following non-linear model

$$E(y|x) = F(\beta_1 + \beta_2 x), \quad F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{s^2}{2}} ds$$

- For example, positive values for β_1 may produce a function that looks like the one in the previous graph (we well get $\hat{\beta}_1 = -2.8619$ and $\hat{\beta}_2 = 6.7612$.
- This is a truly non-linear model (in both, parameters and variables).
- This is the probit model.

In order to search for an estimation strategy for the parameters, we will exploit the following fact.

Since $y|x$ is a binary variable

$$E(y|x) = Pr(y = 1|x)$$

so, by being specific about the form of $E(y|x)$ we are being specific about $Pr(y = 1|x)$.

Likelihood and Basic Concepts

- $Z \sim f(z;\theta_0)$. $\theta_o \in \Re^K$. $f(z;\theta)$ is a member of a parametric class 'indexed' by θ.
- $\tilde{Z} = (Z_1, Z_2, \ldots, Z_n)'$ is an iid sample $\sim f(z;\theta_0)$.

The likelihood function for Z is

$$L(\theta; z) : \Re^K \rightarrow \Re : f(z;\theta)$$

In the density function θ is taken as given and z varies. In the likelihood function these roles are reversed

Note that due to the iid assumption:

$$L(\theta; \tilde{z}) = f(\tilde{z};\theta) = \Pi_{i=1}^n f(z_i;\theta) = \Pi_{i=1}^n L(\theta; z_i)$$

Example: $Z \sim N(\mu, \sigma^2)$

Here $\theta = (\mu, \sigma^2)'$, and $K = 2$.

Note:

- $f(z;\theta) : \Re \rightarrow \Re : \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]$
- $L(\theta;z) : \Re^2 \rightarrow \Re : \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]$

Intuitively, $L(\theta; z_0)$ quantifies how compatible is any choice of θ with the occurrence of z_0.

Maximum Likelihood

The maximum-likelihood estimator $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n \equiv \underset{\theta}{\operatorname{argmax}} \; L(\theta; \tilde{z})$$

It is kind of a 'reverse enginereeing' process: to generate random numbers for a certain distribution you first set parameter values and then get realizations. This is doing the reverse process: first set the realizations and try to get the parameters that are 'most likely' to have generated them.

Some normalizations

$$\hat{\theta}_n \equiv \underset{\theta}{\operatorname{argmax}} \; L(\theta; \tilde{z})$$

- Note

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \; \ln L(\theta; \tilde{z})$$

and $\sum_{i=1}^{n} \ln L(\theta; z_i) = \sum_{i=1}^{n} l(\theta; z_i)$

- Also

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \; \frac{1}{n} \sum_{i=1}^{n} l(\theta; z_i)$$

We will use whichever one is more convenient.

If $l(\theta; \tilde{z})$ is differentiable and has a local maximum in an interior point, then the FOC's for the problem are

$$\left.\frac{\partial l(\theta; \tilde{z})}{\partial \theta}\right|_{\theta=\hat{\theta}_n} = \sum_{i=1}^{n} \left.\frac{\partial l(\theta; z_i)}{\partial \theta}\right|_{\theta=\hat{\theta}_n} = 0.$$

- This is a system of K possibly non-linear equations with K unknowns, that define $\hat{\theta}_n$ implicitely.
- Even if we can guarantee that a solution to this problem exists, we do not have enough information to 'solve' for $\hat{\theta}$.

The discrete case

When Y is a discrete random variable, the likelihood function will be directly the probability function, that is

$$L(Y;\theta) = f(y;\theta)$$

where $f(y:\theta)$ is now $Pr(Y = y;\theta)$.

Conditional likelihood

Suppose $f(y, x, \eta)$ is the joint density function of two variables X and Y. Then, it can be decomposed as

$$f(y, x\ ; \eta) = f(y|X\ ; \theta) f(x; \phi)$$

Suppose we are interested in estimating θ: if θ and ϕ are functionally unrelated, then maximizing the joint likelihood is achivied through maximizing separately the conditional and the marginal likelihood: the MLE of θ also maximizes the conditional likelihood: we can obtain ML estimates by specifying the conditional likelihood only.

Three Examples

Poisson Distribution: $Y \sim \text{Poisson}(\mu)$ if it takes integer and positive values (including zero) and:

$$f(y) = Pr(Y = y) = \frac{e^{-\lambda_o} \lambda_o^y}{y!}$$

For an iid sample:

$$L(\lambda, \tilde{Y}) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

its log is:

$$\begin{aligned} l(\lambda, \tilde{Y}) &= \sum_{i=1}^{n} [-\lambda + y_i \ln \lambda - \ln y_i!] \\ &= -n\lambda + \ln \lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln y_i! \end{aligned}$$

FOC's are

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i = 0$$

so the MLE of λ is:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Probit Model:

$Y|X \sim$ Bernoulli(p), $p \equiv Pr(Y = 1|x) = F(x'\beta)$, and $F(z)$ is the normal CDF.

The sample (conditional) likelihood function will be:

$$L(\beta, \tilde{Y}) = \prod_{i/y_i=1} p_i \prod_{i/y_i=0} (1 - p_i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Then

$$l(\beta, \tilde{Y}) = \sum_{i=1}^{n} \left[y_i \ln F(x_i'\beta) + (1 - y_i) \ln \left(1 - F(x_i'\beta)\right) \right]$$

FOC's for a local maximum are:

$$\sum_{i=1}^{n} \frac{(y_i - F_i) f_i x_i}{F_i(1 - F_i)} = 0$$

,$F_i \equiv F(x_i'\hat{\beta})$, $f_i \equiv f(x_i'\hat{\beta})$. This is a system of K *non-linear* equations with K unknowns. Moreover, it is not possible to solve for $\hat{\beta}$ and obtain and explicit solution.

Gaussian regression model:

$$y = x'\beta_0 + u, \quad \text{with} \quad u|x \sim N(0, \sigma^2)$$

or, alternatively

$$y|x \sim N(x'\beta_0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y - x'\beta_0}{\sigma}\right)^2\right]$$

Then

$$l(\beta, \sigma^2; \tilde{Y}) = -n \ln \sigma - n \ln \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

Any idea what will be the MLE of β_0?

Asymptotic Properties

We will follow a similar path as we did with other estimation strategies

- Consistency
- Asymptotic Normality
- Estimation of the asymptotic variance
- Asymptotic efficiency
- Invariance (this is new)

But first we need to agree on some regularity conditions

Setup and Regularity Conditions

At this stage we will simply state them, and discuss them as we go along.

Some are purely technical, but some of them have important intuitive meaning.

1. $Z_i,\ i = 1, \ldots, n$, iid $\sim f(z_i; \theta_0)$
2. $\theta \neq \theta_0 \Rightarrow f(z_i; \theta) \neq f(z_i; \theta_0)$.
3. $\theta \in \Theta$, Θ a compact set.
4. $\ln f(z_i; \theta)$ is continuous at each $\theta \in \Theta$ w.p.1.
5. $E\left[\sup_{\theta \in \Theta} |\ln f(z; \theta)| \right] < \infty$.

These conditions will be used for consistency.

In addition, for asymptotic normality we will add the following:

6. θ_0 is an interior point of Θ.
7. $f(z;\theta)$ is twice continuously differentiable and strictly poisitive in a neighborhood $\mathcal{N}$ of θ_0.
8. $\int \sup_{\theta\in\mathcal{N}} \|\nabla_\theta f(z;\theta)\|\, dz < \infty$ and $\int \sup_{\theta\in\mathcal{N}} \|\nabla_{\theta\theta} f(z;\theta)\|\, dx < \infty$.
9. $J \equiv E\left[s(\theta_0, Z)s(\theta_0, Z)'\right]$ exists and is non-singular.
10. $E[\sup_{\theta\in\mathcal{N}} \|H(Z;\theta)\|] < \infty$.

Quick detour: on bounds.

Recall that:

$$|E(X)| < E(|X|)$$

Then by bounding $E(|X|)$ we are guaranteeing that $-\infty < E(X) < \infty$.

By considering something like $E\left[\sup_{\theta\in\Theta} |\ln f(z;\theta)|\right] < \infty$. we are bounding the 'worst case' scenario (the sup of the absolute value).

Consistency

Our starting point is the following normalized version of the MLE:

$$\hat{\theta}_n = \underset{\theta}{\text{argmax}} \; Q_n(\theta), \quad Q_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} l(\theta, Z_i)$$

For consistency we need to establish the following three results

1. $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta) \equiv E\left[l(\theta, Z)\right]$.
2. $Q_0(\theta)$ has a unique maximum at θ_0.
3.

$$Q_n(\theta) \xrightarrow{up} Q_0(\theta) \quad \Rightarrow \underbrace{\underset{\theta \in \Theta}{\text{argmax}} \; Q_n(\theta)}_{\hat{\theta}_n} \xrightarrow{p} \underbrace{\underset{\theta \in \Theta}{\text{argmax}} \; Q_0(\theta)}_{\theta_0}$$

Intuition

- $\hat{\theta}_n$ maximizes $Q_n(\theta)$ (definition of MLE).
- $Q_n(\theta) \to Q_0(\theta)$ (the MLE problem if well defined at ∞).
- θ_0 maximizes $Q_0(\theta)$ (the true value solves the problem at ∞).
- By maximizing $Q_n(\theta)$ we end up maximizing $Q_0(\theta)$, convergence of the sequence of functions guarantees converges of maximizers. (this is the difficult step).

If $\operatorname{argmax} Q_n(\theta)$ *is seen as function defined on functions, what property is implied by 3)?*

1) $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta) \equiv E\left[l(\theta, Z)\right]$

If we fix θ at any point θ^* then

$$Q_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} l(\theta^*, Z_i) \xrightarrow{p} E[l(\theta^*, Z)]$$

since by our assumptions (which ones?), $l(\theta^*, Z_i)$ is a sequence of rv's that satisfies Kolmogorov's LLN.

This establishes *pointwise* convergence of $Q_n(\theta)$ to $Q_0(\theta)$. But our strategy requires *uniform* convergence.

Uniform Convergence: a sequence of real valued functions f_n defined on a set $S \in \Re$ *converges uniformly* to a function f on S if for each $\epsilon > 0$, there is a number N such that

$$n > N \Rightarrow |f_n(x) - f(x)| < \epsilon \text{ for all } x \in S$$

Intuition: we are using the same ϵ for the whole domain, that is, eventually we can put f_n in the 'strip' $f \pm \epsilon$.

It can be shown that uniform convergence is equivalent to

$$\sup_{x \in S} |f(x) - f_n(x)| \to 0$$

(see Ross (1980, pp. 137))

Uniform convergence in probability: $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ means $\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$.

Uniform LLN: *if Z_i are iid, Θ is compact and $a(Z_i, \theta)$ is a continuous function at each $\theta \in \Theta$ wp1, and there is $d(Z)$ with $\|a(z, \theta)\| \leq d(Z)$ for all $\theta \in \Theta$ and $E[d(Z)] < \infty$, then $E[a(z, \theta)]$ is continuous and $n^{-1} \sum_{i=1}^{n} a(Z_i, \theta)$ converges uniformly in probability to $E[a(Z, \theta)]$ (Newey and West, 1994, pp. 2129).*

In our case, $a(Z_i, \theta) = l(\theta, Z_i)$, continuous on a compact set Θ, the required bound is provided by our assumption

$$E\Big[\sup_{\theta \in \Theta} |\ln f(z; \theta)|\Big] < \infty$$

and the desired result follows from the iid assumption.

2) $Q_0(\theta)$ has a unique maximum at θ_0.

Information inequality: *if* $\theta \neq \theta_0 \Rightarrow f(z_i'\theta) \neq f(z_i;\theta_0)$ *and* $E[|l(\theta;Z)|] < \infty$ *for all* θ*, then* $Q_0(\theta) = E[l(\theta,Z)]$ *has a unique maximum at* θ_0*.*

$$
\begin{aligned}
E[l(\theta_0;Z)] - E[l(\theta;Z)] &= E[l(\theta_0;Z) - l(\theta;Z)] \\
&= E\left[-\ln \frac{f(Z,\theta)}{f(Z,\theta_0)}\right] \\
&> -\ln E\left[\frac{f(Z;\theta)}{f(Z;\theta_0)}\right] \qquad \text{Jensen's inequality!} \\
&> -\ln \int \frac{f(z;\theta)}{f(z;\theta_0)} f(z;\theta_0)\, dz \\
&> -\ln 1 \\
&> 0
\end{aligned}
$$

Important: check where does the argument break down if assumptions do not hold.

3) 1) and 2) imply consistency.

Pick any $\epsilon > 0$. Let us get three inequalities wpa 1.

$\hat{\theta}_n$ maximizes $Q_n(\theta)$, so

$$\text{a) } Q_n(\hat{\theta}_n) > Q_n(\theta_0) - \epsilon/3$$

$Q_n(\theta)$ converges uniformly to $Q_0(\theta)$, so $Q_n(\hat{\theta}_n) - Q_0(\hat{\theta}_n) < \epsilon/3$, hence

$$\text{b) } Q_0(\hat{\theta}_n) > Q_n(\hat{\theta}_n) - \epsilon/3$$

and $Q_0(\theta_0) - Q_n(\theta_0) < \epsilon/3$ hence

$$\text{c) } Q_n(\theta_0) > Q_0(\theta_0) - \epsilon/3$$

Now start with b)

$$\begin{aligned} & Q_0(\hat{\theta}_n) > Q_n(\hat{\theta}_n) - \epsilon/3 \\ & > Q_n(\theta_0) - \epsilon\, 2/3 \qquad \text{Substract } \epsilon/3 \text{ in both sides of a)} \\ \text{d)} \quad & Q_0(\hat{\theta}_n) > Q_0(\theta_0) - \epsilon \qquad \text{Substract } \epsilon\, 2/3 \text{ in both sides of c)} \end{aligned}$$

Let $\mathcal{N}$ be an open subset of Θ containing θ_0. $\mathcal{N}$ open implies $\mathcal{N}^c \cap \Theta$ closed and bounded: compact in our case.

Since $Q_0(\theta)$ is continuous (the uniform limit of a continuous function is continuous) then:

$$\sup_{\theta \in \Theta \cap \mathcal{N}^c} Q_0(\theta) = Q_0(\theta^*)$$

for some $\theta^* \in \Theta \cap \mathcal{N}^c$ (continuous functions over compact sets achieve their maximum).

Now since θ_0 is the unique maximizer of $Q_0(\theta)$,

$$Q_0(\theta^*) < Q_0(\theta_0)$$

Now pick $\epsilon = Q_0(\theta_0) - Q_0(\theta^*)$, then by inequality d), wpa1

$$Q_0(\hat{\theta}) > Q_0(\theta^*)$$

so $\hat{\theta} \in \mathcal{N}$ wpa1. (if not $Q_0(\theta^*)$ would not be the sup).

Then using the definition of convergence in probability, since $\mathcal{N}$ was chosen arbitrarily

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

Asymptotic normalilty

Asymptotic normality is a bit easier to establish since we will follow a strategy very simimlar to what we did with all previous estimators.

But first we need to establish some notation and results.

- Score and hessian.
- Score equality
- Information matrix
- Information matrix equality

Score, Hessian and Information

- Score: $s(\theta; Z) \equiv \nabla_\theta \, l(\theta; Z)$, a $K \times 1$ vector.
- Sample score: $s(\theta; \tilde{Z}) = \nabla_\theta \, l(\theta; \tilde{Z}) = \sum_{i=1}^{n} s(\theta; Z_i)$
- Hessian: $H(\theta; Z) \equiv \nabla_{\theta\theta'} l(\theta; Z)$, a $K \times K$ matrix.
- Sample hessian: $H(\theta; \tilde{Z}) = \nabla_{\theta\theta'} l(\theta; \tilde{Z}) = \sum_{i=1}^{n} H(\theta; Z_i)$
- Information matrix: $J \equiv E\left[s(\theta_0; Z) s(\theta_0; Z)'\right]$, an $K \times K$ matrix.

Score equality: $E\left[s(\theta_0; Z)\right] = 0$ (*It is kind of a FOC of the likelihood inequality.*)

Information equality: $E\left[H(\theta_0; Z)\right] = -J$

Note that this imples $V\left[s(\theta_0; Z)\right] - J$

Proof of score equality (the continuous case):

For *any* θ

$$\int f(z;\theta)\, dz = 1$$

Taking derivatives in both sides

$$\frac{d[\int f(z;\theta)dz)]}{d\theta} = 0$$

If it is possible to interchange differentiation and integration:

$$\int \frac{df(z;\theta)}{d\theta} d\theta = 0$$

The score is a log-derivative, so

$$s(\theta;z) = \frac{d\ln f(z;\theta)}{d\theta} = \frac{df(z;\theta)/d\theta}{f(z;,\theta)}$$

hence

$$df(z;\theta)/d\theta = s(\theta;z)f(z,\theta)$$

Replacing above:

$$\int s(\theta;z)f(z;\theta)\ dz = 0$$

When $\theta = \theta_0$

$$\int s(\theta_0;z)f(z;\theta_0)\ dz = E\left[s(\theta_0;z)\right]$$

So

$$E\left[s(\theta_0;z)\right] = 0$$

Aside: Interchanging differentiation and integration

5.9 COROLLARY. *Suppose that for some* $t_0 \in [a, b]$, *the function* $x \to f(x, t_0)$ *is integrable on* X, *that* $\partial f/\partial t$ *exists on* $X \times [a, b]$, *and that there exists an integrable function* g *on* X *such that*

$$\left|\frac{\partial f}{\partial t}(x, t)\right| \leqslant g(x).$$

Then the function F *defined in Corollary* 5.8 *is differentiable on* $[a, b]$ *and*

$$\frac{dF}{dt}(t) = \frac{d}{dt}\int f(x, t)\, d\mu(x) = \int \frac{\partial f}{\partial t}(x, t)\, d\mu(x).$$

Source: Bartle, R., 1966, The Elements of Integration, Wiley, New York

Proof of information equality (the continuous case):

From the previous result:

$$\int s(\theta; z) f(z; \theta) dz = 0$$

Take derivatives in both sides, use the product rule and omit arguments in functions to simplify notation:

$$\begin{aligned} \int (sf' + s'f) dz &= 0 \\ \int sf' dz + \int s'f dz &= 0 \end{aligned}$$

From the score equality, $f' = sf$, replacing f' above

$$\int s(\theta; z) s(\theta; z)' f(z; \theta) dz + \int s(\theta; z)' f(z; \theta) dx = 0$$

When $\theta = \theta_0$

$$\begin{aligned} E(s(\theta_0, Z) s(\theta_0, Z)') + \int H(\theta_0; z) f(z; \theta_0) dz &= 0 \\ J + E(H(\theta_0; Z)) &= 0 \end{aligned}$$

which implies the desired result.

Asymptotic normality

Under our assumptions, wpa1 the MLE estimator satisfies the FOC's

$$s(\hat{\theta}_n; \tilde{Z}) = 0$$

Take a first order Taylor expansion around θ_0

$$s(\hat{\theta}_n; \tilde{Z}) = s(\theta_0; \tilde{Z}) + H(\bar{\theta}; \tilde{Z})(\hat{\theta}_n - \theta_0) = 0$$

where $\bar{\theta}$ is a 'mean value' located between $\hat{\theta}_n$ and θ_0. (Note that consistency implies $\theta_n \xrightarrow{p} \theta_0$).
Now solve

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \left(-\frac{H(\bar{\theta}; \tilde{Z})}{n}\right)^{-1} \left(\frac{s(\theta_0; \tilde{Z})}{\sqrt{n}}\right)$$

Now we are back in familiar territory: we will show that the first factor does not explode, and that the second is aymptotically normal

First we will show: $-\left(n^{-1}H(\bar{\theta};\tilde{Z})\right)^{-1} \xrightarrow{p} J^{-1}$

Preliminary result: if $g_n(\theta)$ is a sequence of random functions that converge uniformly in probability to $g_0(\theta)$ for all θ in a compact set Θ, and $g_0(\theta)$ is continuous, $\hat{\theta}_n \xrightarrow{p} \theta_0$ implies $g_n(\hat{\theta}_n) \xrightarrow{p} g_0(\theta_0)$ (see Ruud (2000, pp. 326).

According to our assumptions $n^{-1}H(\theta;\tilde{Z}) = n^{-1}\sum_{i=1}^{n} H(\theta;Z_i)$ converges uniformly in probability to $E\left[H(\theta;Z)\right]$, which by the previous results, it is continuous in θ.

Hence, by the previous result, since $\bar{\theta} \xrightarrow{p} \theta_0$

$$n^{-1}H(\bar{\theta};\tilde{Z}) \xrightarrow{p} E\left[H(\theta_0;Z)\right] = -J < \infty$$

by the information equality. Then the result follows by continuity of matrix inversion and existence of the information matrix.

Now we show:

$$\frac{s(\theta_0; \tilde{Z})}{\sqrt{n}} \xrightarrow{d} N(0, J)$$

Start with

$$\frac{s(\theta_0; \tilde{Z})}{\sqrt{n}} = \sqrt{n} \frac{s(\theta_0, \tilde{Z})}{n} = \sqrt{n} \frac{\sum_{i=1}^{n} s(\theta_0, Z_i)}{n}$$

In order to apply the CLT we check

- $E\left[s(\theta_0, Z_i)\right] = 0$, by the score equality.
- $V\left[s(\theta_0, Z_i)\right] = J < \infty$.

Then, using the Cramer Wold device (please fill details) we get the desired result.

Collecting results:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \underbrace{\left(-\frac{H(\theta_0,\tilde{X})}{n}\right)^{-1}}_{\xrightarrow{p} J^{-1}} \underbrace{\left(\frac{s(\theta_0,\tilde{X})}{\sqrt{n}}\right)}_{\xrightarrow{d} N\left(0,J\right)}$$

Then by Slutzky' theorem and linearity

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, J^{-1}JJ^{-1}\right) = N\left(0, J^{-1}\right)$$

Variance estimation

The asymptotic variance of $\hat{\theta}_n$ is J^{-1}, with $J = V(s(\theta_0, Z)$. We will propose three consistent estimators:

1. Inverse of empirical minus hessian:

$$\left[-\frac{1}{n} \sum_{i=1}^{n} H(\hat{\theta}_n; Z_i) \right]^{-1}$$

2. Inverse of empirical variance of score (OPG):

$$\left[\frac{1}{n} \sum_{i=1}^{n} s(\hat{\theta}_n; Z_i) s(\hat{\theta}_n; Z_i)' \right]^{-1}$$

3. Inverse of empirical information matrix:

$$\left[J(\hat{\theta}_n)^{-1} \right]$$

Invariance

Invariance: Let $\lambda = g(\theta)$, where $g(\theta)$ is a one-to-one function. Let θ_0 denote the true parameter, so $\lambda_0 = g(\theta_0)$ is the true parameter under the new reparametrizaton. Then, if $\hat{\theta}$ its MLE of θ_0, $\hat{\lambda} = g(\hat{\theta})$ is the MLE of λ_0

Example: if $\tilde{\theta}$ is the MLE of $\ln(\theta_0)$, how can we get the MLE of θ_0?

Proof:

Since $\hat{\theta}$ is the MLE

$$l(\hat{\theta}, \tilde{z}) \geq l(\theta, \tilde{z}),$$

for every $\theta \in \Theta$. Since $\lambda = g(\theta)$ is one-to-one:

$$l(g^{-1}(\hat{\lambda}), \tilde{z}) \geq l(g^{-1}(\lambda), \tilde{z})$$

then $\hat{\lambda} = g(\hat{\theta})$ maximizes the reparametrized log-likelihood.

MLE and unbiasedness

The invariance property makes the MLE estimator very likely to be biased in many relevant cases.

Consider the following intuition. Suppose $\tilde{\theta}$ is the MLE for θ_0 and suppose it is unbiased, so

$$E(\tilde{\theta}) = \theta_0$$

By invariance, $g(\tilde{\theta})$ is the MLE of $g(\theta_0)$. Is $g(\tilde{\theta})$ unbiased? In general

$$E(g(\hat{\theta})) \neq g(E(\hat{\theta})) = g(\theta_0$$

so if the MLE is unbiased for one parametrization, it is very likely to be biased for most other parametrizations.

MLE and Efficiency

Let θ^* be any unbiased estimator of θ_0. An important result is the following

Cramer-Rao Inequality: $V(\theta^*) - (nJ)^{-1}$ is psd.

This provides a *lower bound* for unbiased estimators.

Proof: the single parameter case ($K = 1$).

For any two random variables X and Y

$$Cov(X, Y)^2 \leq V(X)V(Y)$$

since the squared correlation is less than 1. Then

$$V(X) \geq \frac{Cov(X, Y)^2}{V(Y)}$$

We will take $X = \theta^*$ and $Y = s(\theta_0, \tilde{Z})$. It is immediate to check

$$V(s(\theta_0, \tilde{Z})) = V\left(\sum_{i=1}^{n} s(\theta_0, \tilde{Z}_i)\right) = n\ J$$

So...what do we need to show to finish the proof?

We have

$$V(\theta^*) \geq \frac{Cov\left(\theta^*, s(\theta_0, \tilde{Z})\right)^2}{nJ}$$

so we need to show $Cov\left(\theta^*, s(\theta_0, \tilde{Z})\right) = 1$.

(Sketch:) Since $E(s(\theta_0, \tilde{Z})) = 0$, $Cov\left(\theta^*, s(\theta_0, \tilde{Z})\right) = E\left(\theta^* s(\theta_0, \tilde{Z})\right)$

$$\begin{aligned}
E(\theta^* s) &= \int \theta^* s \; f(\tilde{z}) \; d\tilde{z} \\
&= \int \theta^* \frac{\dot{f}(\tilde{z})}{f(\tilde{z})} f(\tilde{z}) \; d\tilde{z} \\
&= \int \theta^* \dot{f}(\tilde{z}) \; d\tilde{z} \\
&= \frac{\partial}{\partial \theta} \int \theta^* f(\tilde{z}, \theta) \; d\tilde{z} \Big|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial \theta} \; E(\theta^*)|_{\theta=\theta_0} \\
&= 1
\end{aligned}$$

since $E(\theta^*) = \theta_0$.

MLE and efficiency

- The CR bound applies to unbiased estimators. MLE is likely to be biased.
- MLE estimators are asymptotically normal, centered around the true parameter with normalized variance equal to the CR lower bound for unbiased estimators.
- Problem: the class of consistent AN estimators includes some extreme (an highly unusual) cases that can improve upon the CR bound (the so called 'superefficient' estimator).
- Rao (1963): the MLE estimator is efficient (minimum variance) in the class of *consistent and uniformly asymptotically normal (CUAN)* estimators.
- CUAN estimators: $\tilde{\theta}$ is CUAN for θ_0 if it is consistent and $\sqrt{(}\tilde{\theta} - \theta_0)$ converges in distribution to a normal rv uniformly over compact subsets of Θ.