9. Nonlinear estimators

- introduction
- extremum estimators
- statistical inference
- maximum likelihood estimation
- nonlinear least-squares
Introduction

A nonlinear estimator is one that is a nonlinear function of the dependent variable

\[ \hat{\theta} = f(y), \quad f \text{ is nonlinear} \]

e.g., \( \hat{\theta} \) is the conditional mean

- Statistical results in small samples may be limited for nonlinear estimators
- The asymptotical theory has two major treatments derived from linear model:
  - Alternative methods of proof are needed since there is no direct formula for most nonlinear estimators
  - Asymptotic distribution is obtained under the weakest distributional assumptions possible
in a **nonlinear regression model** we have

- \( y \) (dependent variables)
- \( x \) (explanatory variables)
- \( y \) is a function of \( x \) and they have a joint distribution

**fact:** the best estimate of \( y \) given \( x \) is the conditional mean: \( \mathbb{E}[y|x] \)

**objective:** we would like to *model* \( \mathbb{E}[y|x] \) as a function of \( x \)

to this end, we define a **parametric model** for \( \mathbb{E}[y|x] \):

\[
m(x, \theta)
\]

- \( x \in \mathbb{R}^n \) is explanatory variable
- \( \theta \in \mathbb{R}^p \) is parameter vector (and \( p \) can be greater or less than \( n \))
examples of nonlinear regression functions:

- exponential regression function: useful model whenever $y \geq 0$

  \[ m(x, \theta) = \exp(x^T \theta) \]

- logistic function: when $y$ is restricted in $(0, 1)$

  \[ m(x, \theta) = \frac{e^{x^T \theta}}{1 + e^{x^T \theta}} \]

these examples are nonlinear functions in $\theta$

if we have a correctly specified model for $E[y|x]$, meaning

\[ \exists \theta^* \text{ such that } E[y|x] = m(x, \theta^*) \]

then we would like to estimate for $\theta$ given we know $y$
Examples of nonlinear estimators

a Poisson regression model for $y$ having nonnegative integer values $0, 1, \ldots$

aside: Poisson probability mass function:

$$f(y|\lambda) = e^{-\lambda} \frac{\lambda^y}{y!}, \quad y = 0, 1, \ldots, \quad \mathbb{E}[y] = \lambda, \quad \text{var}(y) = \lambda$$

objective: determine $\lambda$ from $y$

- assumption: $\lambda$ varies across regressors $x$ and parameter vector $\beta$
- propose to use the model $\lambda = e^{x^T \beta}$ to guarantee $\lambda > 0$
- based on one sample of $y, x$, the density of **Poisson regression model** is

$$f(y|x, \beta) = e^{-\exp(x^T \beta)} \exp(x^T \beta)^y / y!$$
suppose we have many independent samples: \((y_i, x_i), i = 1, 2, \ldots, N\)
each \(i\)th sample obeys the joint density (take the log )

\[
\log f(y_i | x_i, \beta) = -\exp(x_i^T \beta) + y_i x_i^T \beta - \log y_i!
\]

objective: choose \(\beta\) that maximizes the joint density

\[
\log f(y_1, \ldots, y_N | x_1, \ldots, x_N, \beta) = \frac{1}{N} \sum_{i=1}^{N} \left( -\exp(x_i^T \beta) + y_i x_i^T \beta - \log y_i \right)
\]

(where we apply that all samples are independent)

- choosing \(\beta\) this way is called maximum likelihood estimation
- no explicit solution for \(\hat{\beta}\), but requires numerical methods to solve
- once we obtain \(\beta\), we can determine \(\lambda\)
Estimate model of conditional expectation

A typical model for estimating conditional expectation is

\[ y = m(x, \theta) + u, \quad \mathbb{E}[u|x] = 0 \]

where \( u \) is an additive, unobservable error with a zero conditional mean.

- Define the error \( u = y - m(x, \theta) \)
- When \( y \) is restricted on some range, \( u \) and \( x \) cannot be independent, e.g.

\[ y \geq 0 \quad \Rightarrow \quad u \geq -m(x, \theta) \]

- It is too strong to assume that \( u_i \) and \( x_i \) are independent
Nonlinear least squares (NLS)

let $\Theta \subset \mathbb{R}^p$ be the parameter space

**assumptions:** for some $\theta^* \in \Theta$, $\mathbb{E}[y|x] = m(x, \theta^*)$

we seek for $\theta$ that solves the population problem

$$\minimize_{\theta \in \Theta} \mathbb{E}\{[y - m(x, \theta)]^2\}$$

where the expectation is taken over the joint distribution of $(x, y)$

we can show that

$$\mathbb{E}\{[y - m(x, \theta)]^2\} \geq \mathbb{E}\{[y - m(x, \theta^*)]^2\}, \quad \forall \theta \in \Theta$$

**conclusion:** $\theta^*$ indexing $\mathbb{E}[y|x]$ in fact minimizes the expected square error
the **nonlinear least-squares estimation** is the problem:

\[
\min_{\theta \in \Theta} \frac{1}{2N} \sum_{i=1}^{N} [y_i - m(x_i, \theta)]^2
\]

- it is the sample analogue problem, when samples of \(y_i\) and \(x_i\) are drawn from the population
- \(\hat{\theta}\) minimizes the **sum of squared residuals**
- the factor 1/2 simplifies the subsequent analysis
- can be solved by deriving the optimality condition: zero gradient condition
- no explicit solution
- the distribution of the NLS estimator depends on the dgp
more generally, we define an \( m \)-estimator \( \hat{\theta} \) of \( \theta \) as

\[
\hat{\theta} = \arg\max_{\theta} \quad Q_N(\theta) := \frac{1}{N} \sum_{i=1}^{N} q(y_i, x_i, \theta)
\]

where

- \( q(\cdot) \) is a scalar-valued function (but mapped from vector variables)
- \( Q_N \) is a sample average of \( q \) where \( N \) does not affect the minimization problem
- it is the sample analogue problem, as opposed to the population problem:

\[
\minimize_{\theta \in \Theta} \quad E[q(y, x, \theta)]
\]
examples:

- NLS is a special case of $m$-estimator where $q$ is the quadratic function:

  $$q(y, x, \theta) = (y - m(x, \theta))^2$$

- Poisson maximum likelihood estimation:

  $$q(y, x, \beta) = -e^{x^T \beta} + yx^T \beta - \log y!$$

- The term $m$-estimator stands for **maximum-likelihood estimation** where

  $$q(y, x, \theta) = -\log f(y|x, \theta)$$

  (-negative log of joint distribution of $y$ given $x$ and parameter $\theta$)
Properties of $m$-estimator

- identification
- consistency
- limit normal distribution

details in Cameron 2005, chapter 5.3
Identification of the true value

recall that if for some $\theta^* \in \Theta$

$$E[y|x] = m(x, \theta^*)$$

then we say we have a **correctly specified model** for the conditional mean

and often we say that $\theta^*$ is called the **true parameter value** of $\theta$

- when the model is correctly specified, $\theta^*$ is the unique solution to

$$\min_{\theta \in \Theta} E[q(y, x, \theta)]$$

- identification requires that $\theta^*$ be the unique solution:

$$E[q(y, x, \theta^*)] < E[q(y, x, \theta)], \quad \forall \theta \in \Theta, \quad \theta \neq \theta^*$$
Consistency of $m$-estimator

Consistency is established in the following manners:

- Suppose $Q_N(\theta) \xrightarrow{p} Q^*(\theta)$ as $N \to \infty$ (or other sense of convergence).
- Let $\theta^*$ be the solution that minimizes $Q^*(\theta)$.
- Let $\hat{\theta}$ be the solution that minimizes $Q_N(\theta)$.
- A consistency result is established to conclude if $\hat{\theta} \xrightarrow{p} \theta^*$.

Formal statements can be further read in Cameron 2005, chapter 5.3.
Limit normal distribution

we consider the behaviour of $\sqrt{N}(\hat{\theta} - \theta^*)$ as $N \to \infty$

under appropriate assumptions this yields the limit distribution of an $m$-estimator

$$\sqrt{N}(\hat{\theta} - \theta^*) \overset{d}{\to} \mathcal{N}(0, A^{-1}BA^{-1})$$

where

- $A$ is the probability limit of the term involving the Hessian of $q$
- $B$ is the probability limit of the term involving the gradient of $q$
Asymptotic Normality of $m$-estimators

define $z = (x, y)$ (or data samples), so $q(z, \theta)$ denote $q(y, x, \theta)$

notation: all derivatives here are w.r.t. $\theta$

assumptions:

- $\theta^*$ is in the interior of $\Theta$
- $\nabla q(z, \cdot)$ is continuously differentiable on the interior of $\Theta$
- each element of $\nabla^2 q(z, \theta)$ is bounded in absolute value by $b(z)$ where $\mathbb{E}[b(z)] < \infty$
- $A = \mathbb{E}[\nabla^2 q(z, \theta^*)]$ is positive definite
- $\mathbb{E}[\nabla q(z, \theta^*)] = 0$
- each element of $\nabla q(z, \theta^*)$ has finite second moment
under the given assumptions plus the conditions for consistency and identification, then we have

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$$

where

$$A = \mathbb{E}[\nabla^2 q(z, \theta^*)], \quad B = \mathbb{E}[\nabla q(z, \theta^*)\nabla q(z, \theta^*)^T] \triangleq \text{cov}(\nabla q(z, \theta^*))$$

thus the asymptotic covariance is given by

$$A_{\text{var}}(\hat{\theta}) = A^{-1}BA^{-1}/N$$
Maximum Likelihood (ML) Estimation

a special case of $m$-estimator

- likelihood function
- ML estimator
- examples
- distribution of ML estimator
Likelihood function

Let \( f(y, x | \theta) \) be the joint probability mass/density function

**log-likelihood function** is defined as

\[
\mathcal{L}_N(\theta) = \log f(y, x | \theta)
\]

- because \( f(y, x | \theta) \) can be viewed as a function of \( \theta \) given \( x, y \)
- \( y \) and \( x \) denote the data from \( N \) samples, hence \( \mathcal{L} \) depends on \( N \)

**the likelihood principle**: choose the value of \( \theta \) that maximize \( \mathcal{L}_N(\theta) \)

\[ e.g., \mathcal{L}_N(\theta_1) = 0.001, \quad \mathcal{L}_N(\theta_2) = 0.003 \]

\( \theta_2 \) gives a higher probability of the observed data occurring, hence is a better estimator
Conditional likelihood

A likelihood function can be rewritten as

\[ f(y, x | \theta) = f(y | x, \theta) f(x | \theta) \]

which requires both conditional density of \( y \) given \( x \) and the marginal of \( x \).

- The goal of regression is to model the behavior of \( y \) given \( x \).
- So estimation is usually based on the conditional likelihood function:

\[ \mathcal{L}_N(\theta) = \log f(y | x, \theta) \]

(Using that \( \log \) is an increasing function)

- We can view \( x \) as nonrandom vectors that are set ahead of time and appear in the unconditional distribution of \( y \).
if the observations \((y_i, x_i)\) are \textbf{independent} over \(i\) then the joint conditional density is

\[
f(y_1, y_2, \ldots, y_N | x_1, x_2, \ldots, x_N, \theta) = \prod_{i=1}^{N} f(y_i | x_i, \theta)
\]

this leads to the \textbf{conditional log-likelihood function}

\[
Q_N(\theta) = \frac{1}{N} \mathcal{L}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log f(y_i | x_i, \theta)
\]

where we divide by \(N\) so that the objective function is an average
example 1 (Bernoulli RVs): let $y_1, \ldots, y_N$ be random samples from a Bernoulli distribution

assume that the probability of success is given by $p$, a parameter to be estimated

the density function of Bernoulli distribution is

$$f(y_i|p) = p^{y_i}(1-p)^{1-y_i}$$

if we assume $y_i$’s are i.i.d. samples, the joint density function is

$$f(y_1, y_2, \ldots, y_N|p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}$$

the likelihood function is

$$Q_N(\theta)(1/N) \log f(y_1, y_2, \ldots, y_N|p) = (1/N) \sum_{i=1}^{N} y_i \log p + (1 - y_i) \log(1 - p)$$
**example 2 (Probit):** suppose the observation value of \( y \) is binary

\[
y = \text{sign}(x\theta + e), \quad e \sim \mathcal{N}(0, 1)
\]

where \( \text{sign}(\cdot) \) is the sign function, i.e., \( \text{sign}(y) = 1 \) if \( y \geq 0 \) and 0 otherwise

to derive the conditional density of \( y \), we first compute

\[
P(y = 1|x, \theta) = P(x\theta + e > 0|x, \theta) = P(e > -x\theta|x, \theta)
\]

\[
= 1 - \Phi(-x\theta) = \Phi(x\theta)
\]

\[
P(y = 0|x, \theta) = 1 - \Phi(x\theta)
\]

where \( \Phi(\cdot) \) denotes the standard normal CDF

therefore, the density of \( y \) given \( x \) and \( \theta \) is

\[
f(y|x, \theta) = [\Phi(x\theta)]^y[1 - \Phi(x\theta)]^{1-y}, \quad y = 0, 1
\]

and that \( f(y|x, \theta) = 0 \) when \( y \notin \{0, 1\} \)
suppose i.i.d. $N$ samples of observations are drawn: $y_1, y_2, \ldots, y_N$

the conditional density of $y_i$ given $x_i$ and $\theta$ is

$$f(y_i|x_i, \theta) = [\Phi(x_i\theta)]^{y_i}[1 - \Phi(x_i\theta)]^{1-y_i}, \quad y = 0, 1$$

hence, the joint conditional density function is

$$f(y_1, \ldots, y_N|x_1, \ldots, x_N, \theta) = \prod_{i=1}^{N} [\Phi(x_i\theta)]^{y_i}[1 - \Phi(x_i\theta)]^{1-y_i}$$

the conditional loglikelihood function is

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \{y_i \log(\Phi(x_i\theta)) + (1 - y_i) \log(1 - \Phi(x_i\theta))\}$$
example 3 (Poisson regression): from page 9-4

- determine $\lambda$, the mean of the poisson distribution from observations $y_i, x_i$
- propose to use the model $\lambda = e^{x^T \beta}$ to guarantee $\lambda > 0$
- based on one sample of $y, x$, the density of **Poisson regression model** is

$$f(y|x, \beta) = e^{-\exp(x^T \beta)} \exp(x^T \beta)^y/y!$$

- when all samples are i.i.d., the conditional loglikelihood function is

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log f(y_i|x_i, \beta) = \frac{1}{N} \sum_{i=1}^{N} -\exp(x_i^T \beta) + y_i x_i^T \beta - \log y_i!$$
example 4 (Gaussian vectors): estimate the mean and covariance matrix of Gaussian RVs

- observe a sequence of independent random vectors: \( y_1, y_2, \ldots, y_N \)
- each \( y_k \) is an \( n \)-dimensional Gaussian: \( y_k \sim \mathcal{N}(\mu, \Sigma) \), but \( \mu, \Sigma \) are unknown

the likelihood function of \( y_1, \ldots, y_N \) given \( \mu, \Sigma \) is

\[
    f(y_1, \ldots, y_N | \mu, \Sigma) = \frac{1}{(2\pi)^{Nn/2}} \cdot \frac{1}{|\Sigma|^{N/2}} \cdot \exp - \frac{1}{2} \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)
\]

the conditional log-likelihood function is

\[
    Q_N(\mu, \Sigma) = (1/N)\mathcal{L}(\mu, \Sigma)
\]

\[
    = (n/2) \log(2\pi) + (1/2) \log \det \Sigma^{-1} - (1/2N) \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)
\]
Maximum likelihood estimator (MLE)

the MLE is the estimator that maximizes the log-likelihood function

$$\hat{\theta} = \arg\max_{\theta} \log f(y, x|\theta)$$

or maximizes the conditional log-likelihood function

$$\hat{\theta} = \arg\max_{\theta} \log f(y|x, \theta)$$

• MLE is a special case of **extremum estimators** since it solves an optimization problem, which typically has no analytical solution

• usually MLE is a local maximum that solves the zero gradient condition:

$$\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta} = 0$$
the score of the loglikelihood for observation $i$ is defined as

$$s_i(\theta) = \frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta} = \frac{1}{f(y_i|x_i, \theta)} \nabla_\theta f(y_i|x_i, \theta)$$

- if $\theta \in \mathbb{R}^n$ then $s_i$ is the gradient vector of size $n \times 1$
- the zero gradient condition for solving MLE is then described as

$$\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta} = \sum_{i=1}^{N} s_i(\theta) = \sum_{i=1}^{N} \frac{1}{f(y_i|x_i, \theta)} \nabla_\theta f(y_i|x_i, \theta)$$

(the sum of the first derivatives of the log density)
- the gradient vector $\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta}$ is called the score vector
- when the score is evaluated at $\theta^*$, it is called the efficient score
Some ML estimators have closed-form expression

example 1 (Bernoulli): characterize the score likelihood

\[ s_i(p) = y_i \frac{1}{p} - (1 - y_i) \frac{1}{1 - p} \]

the zero gradient condition for solving MLE is

\[ 0 = \sum_{i=1}^{N} s_i(p) = \frac{1}{p} \sum_{i=1}^{N} y_i - \frac{1}{1 - p} \sum_{i=1}^{N} (1 - y_i) \]

with some algebra, we can solve that

\[ \hat{p} = \frac{1}{N} \sum_{i=1}^{N} y_i \]

MLE of probability of success is in fact the portion of success from \( N \) samples
example 4 (Gaussian): rewrite the relevant term in conditional likelihood

\[ Q_N(\Sigma, \mu) = \log \det \Sigma^{-1} - \frac{1}{N} \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1}(y_k - \mu) \]

two parameters to be estimated, but we can maximize over \( \mu \) first

the gradient w.r.t. \( \mu \) is set to zero

\[ \frac{\partial Q_N}{\partial \mu} = \sum_{k=1}^{N} \Sigma^{-1}(y_k - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} y_k \]

the likelihood function evaluated at \( \hat{\mu} \) can be expressed as

\[ Q_N(\Sigma, \hat{\mu}) = \log \det \Sigma^{-1} - \text{tr}(C \Sigma^{-1}) = \log \det X - \text{tr}(CX) \]

where \( C = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{\mu})(y_k - \hat{\mu})^T \) is the sample covariance matrix
taking the derivative w.r.t. $X$ gives

$$\frac{\partial Q_N}{\partial X} = X^{-1} - C \quad \Rightarrow \quad X = C^{-1}$$

in conclusion, the ML estimators of $\Sigma$ and $\mu$ are

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} y_k,$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{\mu})(y_k - \hat{\mu})^T$$

the sample mean and sample covariance matrix we already knew
Most ML estimations require numerical algorithms

eexample 2 (Probit): the zero gradient condition of the likelihood function is

$$\frac{\partial Q_N}{\partial \theta} = \sum_{i=1}^{N} \frac{x_i y_i f(x_i \theta)}{\Phi(x_i \theta)} + \frac{(1 - y_i)(-f_i(x_i \theta))x_i}{(1 - \Phi(x_i \theta))} = 0$$

(using $\Phi'(x) = f(x)$)

eexample 3 (Poisson): the zero gradient condition is

$$\frac{\partial Q_N}{\partial \beta} = \sum_{i=1}^{N} (-x_i e^{x_i^T \beta} + y_i x_i) = 0$$

- the zero gradient (or first-order) condition is a nonlinear equation in $\theta$
- numerically solving MLE involves nonlinear optimization such as Newton-Raphson method
Distribution of ML estimators

to derive asymptotic distribution of ML estimators, we discuss

- regularity condition
- Fisher information matrix
- theorem of asymptotic distribution
Regularity conditions

The ML regularity conditions are that

1. The score vector has expected value zero:

\[ E[\nabla_{\theta} \log f(y|x, \theta)] = \int \nabla_{\theta} \log f(y|x, \theta) f(y|x, \theta) dy = 0 \]

2. The expected Hessian is the expected outer product of the gradient

\[ -E[\nabla^2_{\theta} \log f(y|x, \theta)] = E[(\nabla_{\theta} \log f(y|x, \theta))(\nabla_{\theta} \log f(y|x, \theta))^T] \]

When evaluated at \( \theta = \theta^* \) it is known as the unconditional information matrix equality (UIME)

The regularity conditions hold when the expectation is w.r.t \( f(y|x, \theta) \)
Fisher information matrix

the **Fisher information matrix** for \( \theta \) contained in \( y \) (1 sample) is defined as

\[
I(\theta) = \mathbb{E} \left[ (\nabla_\theta \log f(y, |x, \theta)) (\nabla_\theta \log f(y|x, \theta))^T \right]
\]

the expectation of the **outer product of the score vector**

the Fisher information matrix for \( \theta \) contained in \( y_1, y_2, \ldots, y_N \) is

\[
I_N(\theta) = \mathbb{E} \left[ (\nabla_\theta L_N(\theta)) (\nabla_\theta L_N(\theta))^T \right]
\]

since \( y_1, y_2, \ldots, y_N \) are identical samples drawn from the same distribution

\[
I_N(\theta) = N I(\theta)
\]
• $\mathcal{I}(\theta)$ is a positive semidefinite matrix

• Since the score vector has mean zero, $\mathcal{I}_N(\theta)$ is the variance of $\nabla_\theta \mathcal{L}_N(\theta)$

• Large $\mathcal{I}_N(\theta)$ means small changes in $\theta$ lead to larger change in $\mathcal{L}_N$

• The second regularity condition implies that

$$\mathcal{I}(\theta) = -\mathbb{E} \left[ \nabla^2_\theta \log f(y|x, \theta) \right]$$

When evaluated at $\theta^*$ this is called the information matrix (IM) equality

• We will see later that $\mathcal{I}$ gives the quality of an estimator
Distribution of ML estimator

assumptions:

1. the dgp is the conditional density $f(y_i|x_i, \theta)$ used to defined the likelihood
2. the density $f(\cdot)$ satisfies $f(y, \theta) = f(y, \alpha)$ iff $\theta = \alpha$
3. the following matrix exists and is finite nonsingular

$$P = -E \left[ \frac{1}{N} \nabla^2 L_N(\theta^*) \right]$$

4. the order of differentiation and integration of $L$ can be reversed

then the ML estimator $\hat{\theta}_{ml}$ is consistent for $\theta^*$ and

$$\sqrt{N}(\hat{\theta}_{ml} - \theta^*) \xrightarrow{d} \mathcal{N}(0, P^{-1})$$
• condition 1: the conditional density is correctly specified
• condition 1&2: ensure that $\theta^*$ is identified
• condition 3: analogous to the assumption on $\text{plim} N^{-1} X^T X$ for OLS estimator
• condition 4: necessary for the regularity conditions to hold
• if $(y_i, x_i)$ are identical for all $i$, then

$$E[\nabla^2 \mathcal{L}_N(\theta^*)] = E[\sum_{i=1}^{N} \nabla^2 \log f(y_i|x_i, \theta^*)] = N E[\nabla^2 \log f(y|x, \theta^*)]$$

$P$ is replaced by evaluation based on one sample of $(y, x)$

$$P = -E[\nabla^2_{\theta} \log f(y|x, \theta^*)]$$

• asymptotic normality is obtained from the result on page 9-16 with $A = -B$
• $P$ is essentially the Fisher information matrix, $\mathcal{I}(\theta)$
Estimating the asymptotic covariance

asymptotic normality of ML:

\[ \hat{\theta}_{\text{ml}} \overset{d}{\to} N(\theta^*, P^{-1}/N) \]

where the asymptotic covariance can be also expressed as

\[ \text{Avar}(\hat{\theta}_{\text{ml}}) = P^{-1}/N = I(\theta)^{-1}/N = I_N(\theta)^{-1} \]

at least three possible estimators of \( I \) converges to \(-\mathbb{E}[\nabla^2 \log f(y|x, \theta^*)]\)

\[ -(1/N) \sum_{i=1}^{N} \nabla^2 \log f(y_i|\theta), \quad (1/N) \sum_{i=1}^{N} \nabla \log f(y_i|\theta) \nabla \log f(y_i|\theta)^T \]

\[ -(1/N) \sum_{i=1}^{N} \mathbb{E}_{y|x}[\nabla^2 \log f(y_i|x_i, \theta)] \]
thus $\widehat{\text{Avar}}(\hat{\theta}_{ml}) = \hat{\mathcal{I}}_N(\theta) = \frac{\hat{\mathcal{I}}(\theta)^{-1}}{N}$ can be taken to be any of the three matrices

$$\begin{bmatrix} -\sum_{i=1}^{N} \nabla^2 \log f(y_i|\hat{\theta}) \end{bmatrix}^{-1}, \begin{bmatrix} \sum_{i=1}^{N} \nabla \log f(y_i|\hat{\theta}) \nabla \log f(y_i|\hat{\theta})^T \end{bmatrix}^{-1}$$

$$\begin{bmatrix} -\sum_{i=1}^{N} E_{y|x}[\nabla^2 \log f(y_i|x_i, \hat{\theta})] \end{bmatrix}^{-1}$$
example 1 (Bernoulli): the loglikelihood based on one sample is

\[ \log f(y|p) = y \log p + (1 - y) \log(1 - p) \]

the gradient and the Hessian of the loglikelihood (w.r.t. \( p \)) is given by

\[ \nabla \log(y|p) = \frac{y}{p} - \frac{1 - y}{1 - p}, \quad \nabla^2 \log(y|p) = -\frac{y}{p^2} + \frac{1 - y}{(1 - p)^2} \]

the Fisher information matrix (based on 1 sample) is

\[ P = \mathcal{I}(\theta) = -\mathbb{E}[\nabla^2 \log(y|p)] = -\left( \frac{p}{p^2} + \frac{1 - p}{(1 - p)^2} \right) = \frac{1}{p(1 - p)} > 0 \]

hence, \( \mathcal{I}^{-1}(\theta) = p(1 - p) \) and the asymptotic distribution is

\[ \sqrt{N}(\hat{p}_{ml} - p^*) \xrightarrow{d} \mathcal{N}(0, p(1 - p)) \]
example 2 (Probit): consider the gradient of loglikelihood based on 1 sample

\[
\nabla \log f(y|x, \theta) = \frac{xyf(x\theta)}{\Phi(x\theta)} - \frac{(1-y)xf(x\theta)}{1-\Phi(x\theta)} = \frac{xf(x\theta)(y - \Phi(x\theta))}{\Phi(x\theta)(1 - \Phi(x\theta))}
\]

\[
\mathcal{I}(\theta) = -E[\nabla^2 \log f] = E[\nabla \log f \cdot \nabla \log f^T] = E_{y|x} \left[ \frac{x^2 f^2(x\theta)(y - \Phi(x\theta))^2}{\Phi^2(x\theta)(1 - \Phi(x\theta))^2} \right]
\]

\[
= \frac{x^2 f^2(x\theta)}{\Phi^2(x\theta)(1 - \Phi(x\theta))^2} E_{y|x}[(y - \Phi(x\theta))^2]
\]

note that \( y \) is Bernoulli with mean \( p = \Phi(x\theta) \) and variance \( \Phi(x\theta)(1 - \Phi(x\theta)) \)

\[
\mathcal{I}(\theta) = \frac{x^2 f^2(x\theta) \cdot \Phi(x\theta)(1 - \Phi(x\theta))}{\Phi^2(x\theta)(1 - \Phi(x\theta))^2} = \frac{x^2 f^2(x\theta)}{\Phi(x\theta)(1 - \Phi(x\theta))}
\]

\[
\widehat{\text{Avar}}(\hat{\theta}) = \left( \sum_{i=1}^{N} \frac{x_i^2 f^2(x_i\theta)}{\Phi(x_i\theta)(1 - \Phi(x_i\theta))} \right)^{-1}
\]
**example 3 (Poisson):** the gradient of loglikelihood based on 1 sample is

\[ \nabla \log f(y|x, \beta) = -xe^{x^T \beta} + yx \]

it follows that

\[ \nabla^2 \log f(y|x, \beta) = -xx^T e^{x^T \beta} \]

\[ I(\theta) = -E_{y|x}[\nabla^2 \log f(y|x, \beta)] = xx^T e^{x^T \beta} > 0 \]

the estimate of asymptotic covariance is

\[ \widehat{Avar}(\hat{\beta}) = \left[ \sum_{i=1}^{N} e^{x_i^T \hat{\beta}} x_i x_i^T \right]^{-1} \]
example 4 (scalar Gaussian): here $\theta = (d, \mu)$ where $d = \sigma^2 > 0$

$$\log f(y|\theta) = -(1/2) \log(d) - (1/2)(y - \mu)^2/d$$

$$\nabla \log f = (1/2) \begin{bmatrix} -1/d + (y - \mu)^2/d^2 \\ 2(y - \mu)/d \end{bmatrix}$$

$$\nabla^2 \log f = (1/2) \begin{bmatrix} 1/d^2 - 2(y - \mu)^2/d^3 & -2(y - \mu)/d^2 \\ -2(y - \mu)/d^2 & -2/d \end{bmatrix}$$

$$\mathcal{I}(\theta) = -\mathbb{E}[\nabla^2 \log f] = -(1/2) \begin{bmatrix} 1/d^2 - 2/d^3 & 0 \\ 0 & -2/d \end{bmatrix}$$

$$\mathcal{I}(\theta)^{-1} = \begin{bmatrix} 2d^2 & 0 \\ 0 & d \end{bmatrix} > 0$$

$$\widehat{\text{Avar}}(\hat{\sigma}^2) = 2\hat{\sigma}^4/N$$

$$\widehat{\text{Avar}}(\hat{\mu}) = \hat{\sigma}^2/N$$
Cramér-Rao inequality

for any unbiased estimator $\hat{\theta}$ with the covariance matrix of the error:

$$\text{cov}(\hat{\theta}) = E(\theta - \hat{\theta})(\theta - \hat{\theta})^T,$$

we always have a lower bound on $\text{cov}(\hat{\theta})$:

$$\text{cov}(\hat{\theta}) \succeq \mathcal{I}_N(\theta)^{-1}$$

- the RHS is called the Cramér-Rao lower bound, and also equal to $\mathcal{I}(\theta)^{-1}/N$
- provide the minimal covariance matrix over all possible estimators $\hat{\theta}$
• a consistent asymptotically normal estimator \( \hat{\theta} \) of \( \theta \) is said to be asymptotically efficient if

\[
A\text{var}(\hat{\theta}) = \mathcal{I}(\theta)^{-1}/N
\]

• ML estimator has the smallest asymptotic variance among root-\( N \) consistent estimators (requiring the correctly specified conditional density)
Example of CR bound

estimating $\lambda$ in exponential RVs: $f(x) = \lambda e^{-\lambda x}$

$$
\log f(x|\lambda) = \log \lambda - \lambda x, \quad \nabla \log f(x|\lambda) = \frac{1}{\lambda} - x, \quad \nabla^2 \log f(x|\lambda) = -\frac{1}{\lambda^2}
$$

therefore, $\mathcal{I}(\lambda) = 1/\lambda^2$ and CR bound is $\text{var}(\hat{\lambda}) \geq \lambda^2/N$

estimating $\theta$ in Bernoulli RVs: $p(x) = \theta^x (1 - \theta)^{1-x}$

$$
\log p(x|\theta) = x \log \theta + (1 - x) \log(1 - \theta), \quad \nabla \log p(x|\theta) = \frac{x}{\theta} - \frac{(1 - x)}{(1 - \theta)},
$$

$$
\nabla^2 \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{(1 - x)}{(1 - \theta)^2}, \quad \mathbf{E}[\nabla^2 \log p(x|\theta)] = -\frac{\theta}{\theta^2} - \frac{1 - \theta}{(1 - \theta)^2}
$$

therefore, $\mathcal{I}(\theta) = \frac{1}{\theta(1-\theta)}$ and CR bound is $\text{var}(\hat{\theta}) \geq \theta(1 - \theta)/N$
Important proofs

- derivation of regularity conditions
- proof of Cramér-Rao bound
Derivation of regularity conditions

- from $\int f(y|\theta)dy = 1$, differentiate both sides w.r.t $\theta$ gives $\nabla_\theta \int f(y|\theta)dy = 0$
- if the range of integration does not depend on $\theta$, by Leibniz integral rule
  
  $\int \nabla_\theta f(y|\theta)dy = 0$

- from the derivative of $\log(\cdot)$ function,
  
  $\nabla_\theta f(y|\theta) = \nabla_\theta \log f(y|\theta) \cdot f(y|\theta)$

- substitute into the previous equation
  
  $\int \nabla_\theta \log f(y|\theta) \cdot f(y|\theta)dy = 0 \Rightarrow \mathbf{E}[\nabla_\theta \log f(y|\theta)] = 0$

  this is the regularity condition (1) w.r.t. to the density $f(y|\theta)$
• from \( \int \nabla_\theta \log f(y|\theta) \cdot f(y|\theta) dy = 0 \), differentiate both sides w.r.t. \( \theta \)

\[
\int \left\{ \nabla_\theta^2 \log f(y|\theta) f(y|\theta) + (\nabla_\theta \log f(y|\theta))(\nabla_\theta f(y|\theta))^T \right\} dy = 0
\]

• substitute \( \nabla_\theta f(y|\theta) = \nabla_\theta \log f(y|\theta) \cdot f(y|\theta) \) to the previous equation

\[
\int \left\{ \nabla_\theta^2 \log f(y|\theta) f(y|\theta) + (\nabla_\theta \log f(y|\theta))(\nabla_\theta \log f(y|\theta))^T f(y|\theta) \right\} dy = 0
\]

• this is equivalent to

\[
\mathbf{E}[\nabla_\theta^2 \log f(y|\theta)] = -\mathbf{E}[(\nabla_\theta \log f(y|\theta))(\nabla_\theta \log f(y|\theta))^T]
\]

when the expectation is w.r.t. the density \( f(y|\theta) \)

this is the regularity condition (2)
Proof of the Cramér-Rao inequality

with abuse of notation, we mean $y = (y_1, y_2, \ldots, y_N)$ and $f(y|\theta)$ is a joint pdf

• since $\hat{\theta}$ is unbiased, we have $\theta = \int \hat{\theta}(y)f(y|\theta)dy$

• differentiate both sides w.r.t. $\theta$ and use $\nabla_\theta \log f(y|\theta) = \nabla f(y|\theta)/f(y|\theta)$

$$I = \int \hat{\theta}(y)\nabla \log f(y|\theta)f(y|\theta)dy = \mathbb{E}[\hat{\theta}(y)\nabla \log f(y|\theta)]$$

• from regularity condition (1), $\mathbb{E}[\nabla \log f(y|\theta)] = 0$ we have

$$\mathbb{E} \left[ (\hat{\theta}(y) - \theta)\nabla \log f(y|\theta) \right] = I$$

($\mathbb{E}$ is taken w.r.t $y$, and $\theta$ is fixed)
consider a positive semidefinite matrix

\[ E \left[ \begin{array}{c} \hat{\theta}(y) - \theta \\ \nabla_{\theta} \log f(y|\theta) \end{array} \right] \left[ \begin{array}{c} \hat{\theta}(y) - \theta \\ \nabla_{\theta} \log f(y|\theta) \end{array} \right]^T \succeq 0 \]

expand the product into the form

\[
\begin{bmatrix} A & I \\ I & D \end{bmatrix}
\]

where \( A = E(\hat{\theta}(y) - \theta)(\hat{\theta}(y) - \theta)^T \) and

\[
D = E[\nabla \log f(y|\theta) \cdot (\nabla \log f(y|\theta))^T] = \mathcal{I}_N(\theta)
\]

the Schur complement of the \((1, 1)\) block must be nonnegative:

\[
A - ID^{-1}I \succeq 0
\]

which implies the Cramér Rao inequality
Nonlinear Least Squares

- nonlinear least squares (NLS) estimator
- optimality condition
- examples
- distribution of NLS estimator
Nonlinear regression model

define the scalar dependent variable $y$ to have conditional mean

$$E[y|\mathbf{x}] = g(x, \beta)$$

- $g$ is a scalar-valued specified function
- $\mathbf{x}$ is a vector of explanatory variables
- $\beta$ is a parameter vector
- for linear case, $g(x, \beta) = \mathbf{x}^T \beta$
Exponential regression example

the nonlinear model is

\[ y = e^{x^T \beta} + u \]

to study household income with sociodemographic variables

- **y**: household income

- **x**: age, \( \text{age}^2 \), education, female, \( \text{female} \cdot \text{education} \), \( \text{age} \cdot \text{education} \)
The Box-Cox transformation

the Box-cox transformation for a fixed $\lambda$ is

$$z^{(\lambda)} = (z^\lambda - 1)/\lambda$$

- when $\lambda = 1$ the transformation is linear
- when $\lambda = 0$, it is a log transformation – by L'Hopital

a regression model can be generalized by using Box-cox transformation

$$y = \beta_0 + \sum_{k=1}^{n} \beta_k x_k^{(\lambda)} + u$$

- if $\lambda$ is fixed, the regression is linear in $\beta_k$'s
- if $\lambda$ is also a parameter, the regression is nonlinear
NLS estimator

The **nonlinear least-squares estimation** is the problem

\[
\text{minimize } Q_N(\beta) := \frac{1}{2N} \sum_{i=1}^{N} (y_i - g(x_i, \beta))^2
\]

- given the samples \((y_1, x_1), \ldots, (y_N, x_N)\) are available
- \(i\)th is the sample index
- \(\hat{\beta}_{\text{nls}}\) minimizes the sum of squared residuals
- the factor \(1/2\) is added for simplifying the analysis
Solving NLS

matrix notation: let

\[ y = (y_1, y_2, \ldots, y_N), \quad g(x, \beta) = (g(x_1, \beta), g(x_2, \beta), \ldots, g(x_N, \beta)) \]

the NLS problem can be written in a vector form as

\[
\min_{\beta} \frac{1}{2} \| y - g(x, \beta) \|^2
\]

so the optimality condition is

\[
\nabla_\beta Q_N(\beta) = Dg(x, \beta)^T(y - g(x, \beta)) = \sum_{i=1}^{N} \nabla_\beta g(x_i, \beta)(y_i - g(x_i, \beta)) = 0
\]

- no explicit solution for \( \hat{\beta}_{\text{nls}} \) satisfying the zero gradient condition
- one uses iterative methods (nonlinear optimization techniques) in solving NLS
Exponential regression example

suppose $y$ given $x$ has exponential conditional mean: $\mathbb{E}[y|x] = e^{x^T \beta}$

the model of nonlinear regression is

$$y = e^{x^T \beta} + u$$

- $u$ is the error term
- the conditional mean is nonlinear in $\beta$, parameter to be estimated
- the NLS estimator must satisfy the zero gradient condition:

$$\sum_{i=1}^{N} x_i e^{x_i^T \beta} (y_i - e^{x_i^T \beta}) = 0$$
the dgp can be written as
\[ y_i = g(x_i, \beta^*) + u_i \]

- \( u_i \) is additive error term
- \( \beta^* \) is the true value of parameter
- the conditional mean is correctly specified if
  \[ \mathbb{E}[y|x] = g(x, \beta^*) \]
  meaning the error must satisfy \( \mathbb{E}[u|x] = 0 \)
Distribution of NLS estimator

assumptions:

1. the model is \( y_i = g(x_i, \beta^*) + u_i \)

2. in the dgp \( E[u_i|x_i] = 0 \) and \( E[uu^T|x] = \Lambda \)

3. \( g(\cdot) \) satisfies \( g(x, \beta) = g(x, \alpha) \) iff \( \beta = \alpha \)

4. the following matrix exists and is finite nonsingular

\[
F(x, \beta) = (\nabla g(x_1, \beta)^T, \ldots, \nabla g(x_N, \beta)^T) \in \mathbb{R}^{N \times n}
\]

\[
A = \text{plim} \frac{1}{N} F(x, \beta^*)^T F(x, \beta^*)
\]

\[
= \text{plim} \frac{1}{N} \sum_{i=1}^{N} \nabla g(x_i, \beta^*) \nabla g(x_i, \beta^*)^T
\]
5. \( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla g(x_i, \beta^*) u_i \stackrel{d}{\rightarrow} \mathcal{N}(0, B) \) where

\[
B = \operatorname{plim} \frac{1}{N} F^T(x, \beta^*) F(x, \beta^*)
\]

\[
= \operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \nabla g(x_i, \beta^*) \nabla g(x_i, \beta^*)^T
\]

then the \textbf{NLS estimator} \( \hat{\beta}_{\text{nls}} \) defined to be a root of

\[
\nabla_{\beta} Q_N(\beta) = 0
\]

is consistent for \( \beta^* \) and

\[
\sqrt{N}(\hat{\beta}_{\text{nls}} - \beta^*) \stackrel{d}{\rightarrow} \mathcal{N}(0, A^{-1}BA^{-1})
\]
• condition 1-3: the regression is correctly specified and the regressors are uncorrelated with the errors and that $\beta^*$ is specified

• the errors can be heteroskedastic

• condition 4-5: assume the relevant limit results necessary for application of theorem on page 9-16

**special case:** spherical errors with $\Lambda = \sigma^2 I$

• this implies $B = \sigma^2 A$ and $A^{-1} B A^{-1} = \sigma^2 A^{-1}$

• nonlinear least-squares is then asymptotically efficient among LS estimators
Variance matrix estimation for NLS

from page 9-61, the **asymptotic distribution** of NLS estimators is

\[ \hat{\beta}_{\text{nls}} \sim \mathcal{N}(\beta^*, (F^T F)^{-1}F^T \Lambda F(F^T F)^{-1}) \]

where \( F := F(x, \beta^*) \) defined on page 9-61

- we consider independent errors with **heteroskedasticity of unknown functional form**
- we provide estimates of \( A, B \) and the asymptotic covariance matrix
let \( \hat{\beta} \) be a **consistent** estimate of \( \beta \) and define

\[
\hat{u} = y - g(x, \hat{\beta})
\]

- estimate of \( A \): \( \hat{A} = (1/N)F^T(x, \hat{\beta})F(x, \hat{\beta}) \)
- estimate of \( \Lambda \): \( \hat{\Lambda} = \text{diag}(\hat{u}^2) \) (squared element-wise)
- estimate of \( B \): \( \hat{B} = (1/N)F^T(x, \hat{\beta})\hat{\Lambda}F(x, \hat{\beta}) \)

these lead to the **heteroskedastic-consistent** estimate of the asymptotic variance matrix of the NLS estimator:

\[
\widehat{A\text{var}}(\hat{\beta}_{\text{nls}}) = (F^TF)^{-1}F^T\hat{\Lambda}F(F^TF)^{-1}
\]

(note that now \( F := F(x, \hat{\beta}) \); evaluated at \( \hat{\beta} \))
Exponential regression example

the model is

\[ y = e^{x^T \beta} + u \]

where \( u \) has \( \mathbb{E}[u|x] = 0 \) and \( u \) is potentially heteroskedastic

- \( g(x, \beta) = e^{x^T \beta} \), and \( \nabla g(x, \beta) = xe^{x^T \beta} \)
- \( F^T F := F^T(x, \hat{\beta})F(x, \hat{\beta}) = \sum_{i=1}^{N} x_i x_i^T e^{2x_i^T \hat{\beta}} \)
- \( \hat{\Lambda} = \text{diag}(\hat{u}^2) \) where \( \hat{u} = y - e^{x^T \hat{\beta}} \)
- the heteroskedastic-robust estimate is

\[
\widehat{\text{Avar}}(\hat{\beta}_{\text{nls}}) = \left( \sum_{i=1}^{N} x_i x_i^T e^{2x_i^T \hat{\beta}} \right)^{-1} \left( \sum_{i=1}^{N} \hat{u}_i^2 x_i x_i^T e^{2x_i^T \hat{\beta}} \right) \left( \sum_{i=1}^{N} x_i x_i^T e^{2x_i^T \hat{\beta}} \right)^{-1}
\]
References

Chapter 12-13 in

Chapter 7 in

Chapter 7,14 in