5c. MLE

A. Colin Cameron Pravin K. Trivedi

These slides were prepared in 1999.
They cover material similar to Sections 5.6-5.7 and 5.2.4 of our subsequent book *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.

INTRODUCTION

- The likelihood principle, due to R.A. Fisher, is to choose as estimator of the parameter vector θ_0 that value of θ that maximizes the probability of observing the actual sample.
- For discrete random variables this probability is simply the probability mass function and for continous the joint density.

- Example: If one value of θ gives a probability of 0.0012 of the observed data $\mathbf{y}, \mathbf{X}$ occurring, while a second value of θ gives a probability of 0.0014 data, then the second value of θ is a better estimator.
- The MLE maximizes the joint density.
  This is called the likelihood function in this context, because it is being viewed as a function of θ given data, to distinguish it from the joint density which is the probability of data given θ.

- The MLE holds special place amongst estimators.
- The small sample result that the MLE is the most efficient unbiased estimator and attains the Cramer-Rao lower bound carries over asymptotically.
- The MLE is also important pedagogically.
  Many nonlinear regression methods such as extremum estimation can be viewed as extensions and adaptations of results first obtained for ML estimation.

LIKELIHOOD FUNCTION

- We consider data obtained by exogenous sampling, in which we can condition on $\mathbf{X}$ and consider the likelihood function

$$\mathsf{L}_n(\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}),$$

  where $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is the conditional density of $\mathbf{y}$ given $\mathbf{X}$.

- We also consider its natural logarithm

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln \mathsf{L}_n(\boldsymbol{\theta}).$$

- The MLE is the extremum estimator that maximizes the log-likelihood function.

- For cross-section data y_i are assumed to be independent but not necessarily identically distributed (inid) with conditional density function $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$.
- The conditional density $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ by independence, so

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

EXAMPLES

- ML estimation for the logit model has already been presented.
  Then we assume y is Bernoulli with parameter p, where $p = \Lambda(\mathbf{x}'\boldsymbol{\theta})$, leading to density given earlier.
- A second example is the linear regresson model under normality.
- Then we assume $y \sim \mathtt{N}\left[\mu, \sigma^2\right]$ where $\mu = \mathbf{x}'\boldsymbol{\beta}$ and σ^2 is not modelled.

- Across a wide range of data types this same method is used to generate fully parametric cross-section regression models.
  - Choose the one-parameter or two-parameter (or in some rare cases three-parameter) distribution that would be used for the dependent variable y in the iid case studied in a basic statistics course.
  - Then parameterize the one or two underlying parameters in terms of regressors $\mathbf{x}$ and parameters $\boldsymbol{\theta}$.

Commonly-used distributions include

- *Normal* for data continuous on $(-\infty, \infty)$.
- *Exponential*, *Weibull* or *lognormal* for positive data continuous on $(0, \infty)$.
- Censored normal (*tobit* model) for data on on $[0, \infty)$ where there is a mass at 0 but otherwise the data are continuous.
- Bernoulli for discrete binary data taking values 0 or 1, with different parameterizations of p leading to the *logit* model and the *probit* model.

- *Poisson* or *negative binomial* for count data taking discrete values $0, 1, 2, ...$

Many of these distributions are analyzed in detail later.

DISTRIBUTION OF THE MLE

- We consider cross-section data.
- The general theory of extremum estimation is directly applicable.
- We make the assumptions that
  (*) the range of y does not depend on $\boldsymbol{\theta}$.
  (Then the order of differentiation and integration of the log-density can be reversed).
  (**) the density $f(y|\mathbf{x}, \boldsymbol{\theta})$ is correctly specified

- Assumption (*) implies

$$\mathsf{E}\left[\frac{\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0},$$

and

$$\mathsf{E}\left[\frac{\partial^2 \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = -\mathsf{E}\left[\frac{\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right],$$

where

expectation is taken with respect to the density $f(y|\mathbf{x},\boldsymbol{\theta})$.

- Proof is by diffentiating $\mathsf{E}\left[\ln f(y|\mathbf{x},\boldsymbol{\theta})\right] = 0$ and manipulating.
- These are called the regularity conditions.

- Now

$$\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \sum_{i=1}^{n} \partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta},$$

  and

$$\mathsf{E}\left[\partial \ln f(y|\mathbf{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}\right] = \mathbf{0},$$

  implies

$$\mathsf{E}\left[\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\big|_{\boldsymbol{\theta}_0}\right] = \mathbf{0},$$

  if the dgp is $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$, i.e. assumption (**).
- But this is the informal condition for consistency .
- So MLE is consistent if dgp correctly specified and regularity conditions hold.

- Also
$$\mathsf{E}\left[\frac{\partial^2 \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = -\mathsf{E}\left[\frac{\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln f(y|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right],$$
  implies the information matrix equality,
$$\mathsf{E}\left[\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0}\right] = -\mathsf{E}\left[\frac{\partial \mathcal{L}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0}\right],$$
  if the dgp is $f(y|\mathbf{x},\boldsymbol{\theta}_0)$.
- Then $\mathbf{A}(\boldsymbol{\theta}_0) = -\mathbf{B}(\boldsymbol{\theta}_0)$, where $\mathbf{A}(\boldsymbol{\theta}_0)$ and $\mathbf{B}(\boldsymbol{\theta}_0)$.
- It follows that $\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{'-1} = -\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$.

Proposition: *Distribution of ML Estimator. Make the assumptions:*
(i) The dgp has conditional density $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)$*;*
(ii) The density function $f(\cdot)$ *satisfies* $f(y, \boldsymbol{\theta}^{(1)}) = f(y, \boldsymbol{\theta}^{(2)})$ *iff* $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)}$*;*
(iii) The following matrix exists and is finite nonsingular

$$\mathbf{A}(\boldsymbol{\theta}_0) = \lim \frac{1}{n} \mathsf{E} \left[\left. \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}_0} \right];$$

(iv) The order of differentiation and integration of the likelihood can be reversed.

Then the MLE $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}}$*, defined to be a solution of the first-order conditions* $\partial \frac{1}{n} \mathcal{L}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$*, is consistent for* $\boldsymbol{\theta}_0$*, and*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathsf{N}\left[\mathbf{0}, -\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\right].$$

CRAMER RAO LOWER BOUND

- Then asymptotically

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} \overset{a}{\sim} \mathsf{N}\left[\boldsymbol{\theta}_0, -\mathsf{E}\left[\left.\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0}\right]^{-1}\right].$$

- It follows that the ML estimator attains the Cramer-Rao lower bound (CRLB).
- From basic statistics courses, the CRLB is the lower bound for unbiased estimators in small samples.
- For large samples, which we consider here, the CRLB is the lower bound for the variance matrix of consistent

asymptotically normal (CAN) estimators with convergence to normality of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ uniform in compact intervals of $\boldsymbol{\theta}_0$, see Rao (1973, pp.344-351).

VARIANCE MATRIX ESTIMATION

- As already noted there are several ways to consistently estimate the variance matrix of the estimator.
- In principle one can use the more general sandwich estimate $\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}'^{-1}$ rather than $-\widehat{\mathbf{A}}^{-1}$ or $\widehat{\mathbf{B}}^{-1}$.
  This is called the Huber estimate or White estimate after Huber (1965) and White (1982).
- The sandwich estimate is in theory more robust. The cause of failure of the information matrix equality may, however, additionally lead to the more fundamental complication of inconsistency of $\widehat{\boldsymbol{\theta}}$.

ML TERMINOLOGY

- A special terminology has developed for ML estimation.
- The gradient vector $\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is called the score.
- When evaluated at θ_0, $\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$is called the efficient score.

- The expectation of the outer product of the first derivative of the log-likelihood function, $\mathsf{E}[\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \times \partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}']$ is called the information matrix .
- This is because it is the variance of $\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, with large values meaning that small changes in $\boldsymbol{\theta}$ lead to large changes in the log-likelihood which accordingly contains a lot of information about $\boldsymbol{\theta}$. By the information matrix equality the information matrix also equals $-\mathsf{E}[\partial^2 \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}']$.

- The information matrix equality is a special case of the generalized information matrix equality

$$\mathsf{E}\left[\frac{\partial \mathbf{m}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = -\mathsf{E}\left[\mathbf{m}(y,\boldsymbol{\theta})\frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right],$$

where $\mathbf{m}(\cdot)$ is a vector function and the expectation is whith respect to the density $f(y|\boldsymbol{\theta})$.

ALTERNATIVE SAMPLING SCHEMES

- If sampling is instead *endogenous* or *choice-based* then we need to instead use the joint density $f(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta})$, as the MLE based on the conditional density $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is inconsistent.

- For time series data y_t with *strongly exogenous* regressor variables $\mathbf{x}_t$, the conditional density function

$$f(\mathbf{y}|\mathbf{X},y_0,\boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t|y_{t-1},...,y_0,x_t,x_{t-1},...,x_1,\boldsymbol{\theta})$$

  upon repeated conditioning on past y_t and using the strong exogeneity assumption.

- Then

$$Q_T(\boldsymbol{\theta}) = \frac{1}{T}\mathcal{L}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \ln f(y_t|y_{t-1},...,y_0,x_t,x_{t-1},...,x_1,\boldsymbol{\theta}).$$

- Cross-section data results can be adapted to time series data by replacing $f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$ by $f(y_t|y_{t-1},...,y_0,x_t,x_{t-1},...,x_1,\boldsymbol{\theta})$

QUASI-MLE

- The MLE in a model with misspecified density is called the quasi-MLE.
- This is investigated by Huber (1965) and White (1982).
- In general any misspecification leads to inconsistency, as then the expectation in $\mathsf{E}\left[\partial \mathcal{L}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}\right]$ is no longer with respect to $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$.

- The quasi-MLE $\widehat{\boldsymbol{\theta}}$ instead converges in probability to the pseudo-true value $\boldsymbol{\theta}^*$ which maximizes $\mathsf{E}[n^{-1}\mathcal{L}_n(\boldsymbol{\theta})]$, where the expectation is taken with respect to the true dgp which is no longer $f(y|\mathbf{x}, \boldsymbol{\theta}_0)$.
- The variance matrix of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is then of the more general form $\mathbf{A}(\boldsymbol{\theta}^*)^{-1}\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)'^{-1}$, where again expectation is taken with respect to the true dgp.

GENERALIZED LINEAR MODELS

- In some special cases the MLE may be consistent when the density is partially misspecified.
- For example, in the linear regression model with normality the quasi-MLE may be consistent even if the errors are non-normal. The key condition in this example is that the conditional mean of the error equal zero.

- Similar robustness to misspecification is enjoyed by other models based on densities in the linear exponential family (LEF), in which case the density can be expressed as

$$f(y) = \exp\{a(\mu) + b(y) + c(\mu)y\},$$

where different functions $a(\cdot)$ and $b(\cdot)$ lead to different densities in the family.

- For regression the parameter $\mu = \mathsf{E}[y|\mathbf{x}]$ is modelled as $\mu = g(\mathbf{x}, \boldsymbol{\theta})$ for some specified function $g(\cdot)$.

- Gourieroux and Monfort (1984a) proved that the quasi-MLE $\widehat{\boldsymbol{\theta}}$ which maximizes the LEF log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} a(g(\mathbf{x}_i, \boldsymbol{\theta})) + b(y_i) + c(g(\mathbf{x}_i, \boldsymbol{\theta}))y_i\},$$

  is consistent for $\boldsymbol{\theta}_0$, even if the dgp is not an LEF density, provided that the conditional mean of y given $\mathbf{x}$ is correctly specified.
- This result holds because for this class of densities $\partial\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be shown to be a weighted sum of $y_i - g(\mathbf{x}_i, \boldsymbol{\theta}_0)$, which has expected value zero if $\mathsf{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i, \boldsymbol{\theta}_0)$ in the true dgp.

- Models based on the LEF are called generalized linear models in the statistics literature, see the book with this title by McCullagh and Nelder (1989).
- The Poisson, probit, logit, gamma and exponential models are special cases.
- Generalized linear models are widely used in applied statistics.

- While the quasi-MLE in these cases will be consistent provided only that the conditional mean is correctly specified, adjustment will have to be made to the usual MLE output for variance, standard errors, and t-statistics, since $\mathbf{A}(\boldsymbol{\theta}_0) \neq -\mathbf{B}(\boldsymbol{\theta}_0)$ for the LEF unless the conditional variance is also correctly specified.
- See the preceding references and Cameron and Trivedi (1986, 1998) for further details.
- Aside from this special case one should be aware that in general misspecification of any aspect of the density leads to inconsistency of the MLE.

COEFFICIENT INTERPRETATION

- Consider the impact on the expected value of y of a one unit change in a regressor.
- For linear regression model $\mathsf{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ implies $\partial \mathsf{E}[y|\mathbf{x}]/\partial \mathbf{x} = \boldsymbol{\beta}$.
  So the coefficient $\boldsymbol{\beta}$ has a direct interpretation as this impact.

- For nonlinear regression models this interpretation is no longer possible.
- Can again consider $\partial \mathsf{E}[y|\mathbf{x}]/\partial \mathbf{x}$, which in general will be a function of both parameters and regressors.
- For example, for the logit model $\mathsf{E}[y|\mathbf{x}] = \Lambda(\mathbf{x}'\boldsymbol{\beta})$.
  So $\partial \mathsf{E}[y|\mathbf{x}]/\partial \mathbf{x} = \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\boldsymbol{\beta}$.
- This may then be evaluated at representative values of $\mathbf{x}$, such as $\bar{\mathbf{x}}$, or evaluated for each $\mathbf{x}_i$, $i = 1, ..., n$, and averaged.

- A useful result is that in many cases one can directly interpret the relative impact of different regressors.
- Consider base density with a scalar parameter γ.
- This is allowed to depend on regressors through the *single-index* form $\gamma = g(\mathbf{x}'\boldsymbol{\beta})$, i.e. a nonlinear transformation of the linear combination $\mathbf{x}'\boldsymbol{\beta}$.
- For example, in the logit model $\gamma = \Lambda(\mathbf{x}'\boldsymbol{\beta})$.
- Then

$$\partial\gamma/\partial\mathbf{x} = [\partial g(\mathbf{x}'\boldsymbol{\beta})/\partial\mathbf{x}'\boldsymbol{\beta}] \times \boldsymbol{\beta}.$$

- Now consider the relative effect of the j^{th} and k^{th} regressors.
- This is given by $(\partial\gamma/\partial x_j)/(\partial\gamma/\partial x_k)$ which using above result simplifies to β_j / β_k.
- This is constant regardless of the value of the regressors.
- Thus if, for example β_j is two times β_k then the impact on $\gamma = g(\mathbf{x}'\boldsymbol{\beta})$ of a one unit change in the j^{th} regressor is twice that of a one unit change in the k^{th} regressor.

- If additionally the function $g(\cdot)$ is monotonic, $\partial\gamma/\partial x_j$ has the same sign (determined by the sign of β_j) for all $\mathbf{x}$.
- Usually $g(\mathbf{x}'\boldsymbol{\beta})$ is chosen to be monotonically increasing, so that $\beta_j > 0$ means that an increase in the j^{th} regressor leads to an increase in γ.