Maximum Estimation

Paolo Zacchia

Econometric Theory

Lecture 11

Overview of Maximum Estimation

- This Lecture introduces Maximum Estimation: a family of estimation frameworks for linear and non-linear models.
- All these frameworks are based upon the optimization (that is, maximization or minimization) of an objective function.
- Hence, a more general name for the family of frameworks is that of **Extremum Estimation**.
- However, **Maximum Estimation** is more commonly used.
- **Maximum Likelihood Estimation** (MLE) is one special case of Maximum Estimation, which is emphasized here.
- Lecture 12 focuses on another special case: the Generalized Method of Moments.

Criteria and M-Estimators

- Suppose that the probabilistic assumptions of a **structural** econometric model with variables $(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$ make $\boldsymbol{\theta}_0$ *in the population* the *only* solution to a maximization problem as:

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{Q}_0(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta})\right]$$

  where $q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta})$ is a **criterion** for $i = 1, \ldots, N$.

- Write the latter hereinafter as $q(\boldsymbol{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$.

- With i.i.d. data, it is simply $\mathcal{Q}_0(\boldsymbol{\theta}) = \mathbb{E}\left[q(\boldsymbol{x}_i; \boldsymbol{\theta})\right]$.

- An **M-Estimator** $\widehat{\boldsymbol{\theta}}_M$ for $\boldsymbol{\theta}_0$ maximizes the sample analog of $\mathcal{Q}_0(\boldsymbol{\theta})$, written as $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$.

$$\widehat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} q(\mathbf{x}_i; \boldsymbol{\theta})$$

Example: OLS as an M-Estimator

- Let the CEF of some endogenous variable Y_i, given some K exogenous variables $\boldsymbol{x}_i$, be linear.

$$\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$$

- In such a case, the "true" parameter vector $\boldsymbol{\beta}_0$ is shown to minimize the "limiting" mean squared error (MSE).

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\max} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} - \mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2\right]$$

- The OLS estimator is the minimizer of the sample analog.

$$\widehat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\max} - \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2$$

- This is an M-Estimator for $q\left(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}\right) = - \left(Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2$.

Example: NLLS as an M-Estimator

- Let the CEF be now non-linear, governed by some function denoted as $h(\boldsymbol{x}_i; \boldsymbol{\theta})$.

$$\mathbb{E}[Y_i | \boldsymbol{x}_i] = h(\boldsymbol{x}_i; \boldsymbol{\theta}_0)$$

- The result linking here $\boldsymbol{\beta}_0$ with the MSE is here as follows.

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} - \mathbb{E}\left\{ [Y_i - h(\boldsymbol{x}_i; \boldsymbol{\theta})]^2 \right\}$$

- This lets express the **Non-Linear Least Squares** (NLLS) estimator as the solution to the sample analog of the above problem, so long as $h(\boldsymbol{x}_i; \boldsymbol{\theta})$ is invertible in $\boldsymbol{\theta}$.

$$\widehat{\boldsymbol{\theta}}_{NLLS} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} -\frac{1}{N} \sum_{i=1}^{N} [y_i - h(\mathbf{x}_i; \boldsymbol{\theta})]^2$$

- In this case it is $q(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}) = -[Y_i - h(\boldsymbol{x}_i; \boldsymbol{\theta})]^2$.

Score and Hessian

It is useful to define the **score** as the criterion's gradient for $\boldsymbol{\theta}$:

$$\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta}\right)=\frac{\partial q\left(\mathbf{x}_i;\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}=\begin{bmatrix}\frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_1}\\ \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_2}\\ \vdots\\ \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_K}\end{bmatrix}$$

as well as the **Hessian** as... the criterion's Hessian for $\boldsymbol{\theta}$.

$$\mathbf{H}_i\left(\mathbf{x}_i;\boldsymbol{\theta}\right)=\frac{\partial\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}^{\mathrm{T}}}=\frac{\partial^2 q\left(\mathbf{x}_i;\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}$$

$$=\begin{bmatrix}\frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_1\partial\theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_1\partial\theta_K}\\ \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial^2\theta_2\partial\theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_2\partial\theta_K}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_K\partial\theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_K\partial\theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_K\partial\theta_K}\end{bmatrix}$$

Both objects are observation-specific (for $i=1,\ldots,N$).

Identification of M-Estimators

Theorem 1

Identification of M-Estimators. *In a M-Estimation environment, the "true" parameter set* $\boldsymbol{\theta}_0$ *is locally point identified if the following limiting average Hessian matrix evaluated at* $\boldsymbol{\theta}_0$ *has full* K *rank.*

$$\mathbf{Q}_0 \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{H}_i\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right]$$

Proof.

This is indeed a simple application of the Implicit Function Theorem. In a well-defined M-Estimator, the true parameter vector $\boldsymbol{\theta}_0$ sets the K First Order Conditions of the empirical criterion function $\widehat{\mathcal{Q}}_N\left(\boldsymbol{\theta}_0\right)$ equal to zero *at the probability limit*, at some *limiting average score*.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \widehat{\mathcal{Q}}_N\left(\boldsymbol{\theta}_0\right) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\partial}{\partial \boldsymbol{\theta}} q\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right)\right] \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{s}_i\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] = \mathbf{0}$$

If the Jacobian $\mathbf{Q}_0$ has full rank, the local solution $\boldsymbol{\theta}_0$ is unique. $\square$

Examples: Identification of OLS and NLLS (1/2)

- In OLS, the score and Hessian are:

$$\mathbf{s}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\beta}\right) = 2\mathbf{x}_i\varepsilon_i$$
$$\mathbf{H}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\beta}\right) = -2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$$

  hence, identification requires full rank of a familiar matrix.

$$\mathbf{Q}_0 = -2\mathbf{K}_0 = \lim_{N\to\infty} -\frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$$

- In NLLS, define the *error term* by $\varepsilon_i \equiv y_i - h\left(\mathbf{x}_i; \boldsymbol{\theta}\right)$; thus the score and the Hessian are as follows.

$$\mathbf{s}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = 2\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\mathbf{x}_i; \boldsymbol{\theta}\right)\cdot\varepsilon_i$$
$$\mathbf{H}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = -2\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\mathbf{x}_i; \boldsymbol{\theta}\right)\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\mathbf{x}_i; \boldsymbol{\theta}\right) + 2\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\mathbf{x}_i; \boldsymbol{\theta}\right)\cdot\varepsilon_i$$

Examples: Identification of OLS and NLLS (2/2)

- Note that, by the Law of Iterated Expectations:

$$\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\cdot\varepsilon_i\right]=\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\cdot\mathbb{E}\left[\varepsilon_i\middle|\,\boldsymbol{x}_i\right]\right]$$
$$=\mathbf{0}$$

  as $\mathbb{E}\left[\varepsilon_i\middle|\,\boldsymbol{x}_i\right]=\mathbb{E}\left[Y_i\middle|\,\boldsymbol{x}_i\right]-h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)=0$.

- Thus, any expected Hessian matrix of NLLS is just:

$$\mathbb{E}\left[\mathbf{H}_i\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=-2\,\mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

  where:

$$\boldsymbol{h}_{0i}\equiv\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)$$

  is the derivative of the CEF evaluated at $\mathbf{x}_i$ and at the true parameters $\boldsymbol{\theta}_0$.

- The identification of NLLS is evaluated through $\mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$.

Remarks on identification of M-Estimators

- In practical applications, it is customary to verify that the *sample mean* of the Hessian has full rank, as an indication that the model is identified ($N^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ for OLS).

- Furthermore, it is useful to check that the rows-columns of the Hessian's sample mean are not *too correlated*; otherwise, identification is said to be *weak*, and the estimates are thus very imprecise with large standard errors.

- This problem is called **quasi-multicollinearity**, and it is intuitively due to the statistical difficulty of distinguishing between two "factors" (like different explanatory variables, columns in $\mathbf{X}$) if they are very similar.

- In the IV and 2SLS cases, this corresponds to the problem of *weak instruments* which occurs if the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}$ is too large.

Maximum (Likelihood) Estimation (1/3)

- Consider one special case of M-Estimation: MLE.

- Start from P **structural** relationships $\boldsymbol{y}_i = \boldsymbol{s}\left(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta}\right)$, as well as a joint distribution $F_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i\right)$ of the exogenous variables (both observable and unobservable).

- One can write the distribution of $\boldsymbol{x}_i = \left(\boldsymbol{y}_i, \boldsymbol{z}_i\right)$ as:

$$f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta}\right) = f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i; \boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}\right); \boldsymbol{\theta}\right)$$

  where: $f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\cdot\right)$ is the probability mass/density function that is associated with $F_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\cdot\right)$, whereas $\boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}\left(\cdot\right)$ is the solution of $\boldsymbol{s}\left(\cdot\right)$ with respect to the unobservables $\boldsymbol{\varepsilon}_i$ (if it exists).

- Hence, MLE is the M-Estimator with criterion equaling the logarithm of $f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\cdot\right)$; write it as $\ell\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)$ where $\boldsymbol{x}_i = \left(\boldsymbol{y}_i, \boldsymbol{z}_i\right)$.

$$q\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}\right) = \log f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i; \boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}\right); \boldsymbol{\theta}\right) \equiv \ell\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)$$

Maximum (Likelihood) Estimation (2/3)

- This complies as follows with the definition of M-Estimator.

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\ell\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right]$$

- This holds since for $i = 1, \ldots, N$ one can maintain that:

$$\begin{aligned}
\mathbb{E}\left[\ell\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right] - \mathbb{E}\left[\ell\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] &= \mathbb{E}\left[\log \frac{f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)}{f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)}\right] \\
&\leq \log \mathbb{E}\left[\frac{f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)}{f_{\boldsymbol{x}}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)}\right] \\
&= \log \int_{\mathbb{X}} \frac{f_{\boldsymbol{x}}\left(\mathbf{x}_i; \boldsymbol{\theta}\right)}{f_{\boldsymbol{x}}\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right)} f_{\boldsymbol{x}}\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right) \mathrm{d}\mathbf{x}_i \\
&= 0
\end{aligned}$$

per an application of Jensen's inequality in the second line; $\mathbb{X}$ is here the joint support of $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$.

Maximum (Likelihood) Estimation (3/3)

- It follows that $\mathbb{E}\left[\ell\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right] \geq \mathbb{E}\left[\ell\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right]$ for $i = 1,\ldots,N$: this shows the argued compliance with M-Estimators.

- A variant of such an approach is to specify the **conditional** distribution of the unobserved factors $\boldsymbol{\varepsilon}_i$ **only**, while taking the realizations $\mathbf{y}_i$ of the exogenous variables **as given**.

- The criterion function would then be as follows.

$$q\left(\boldsymbol{y}_i,\boldsymbol{z}_i;\boldsymbol{\theta}\right) = \log f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}\left(\boldsymbol{y}_i,\boldsymbol{z}_i;\boldsymbol{\theta}\right)\middle|\,\boldsymbol{z}_i\right)$$

- This is referred as: **Conditional Maximum Likelihood Estimation** (CMLE).

- Due to its flexibility, CMLE is prevalent in econometrics – as the next example helps clarify.

Example: MLE and linear regression (1/4)

- Consider a linear regression model like $y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$.
- Suppose that the error term is homoscedastic, independent across observations and **normally distributed**:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

  where σ^2 is an unknown parameter of the model.
- Also assume that the right-hand side variables $\boldsymbol{x}_i$ are **fixed** (*non-stochastic*): one specific realization $\mathbf{x}_i$ occurs *always*.
- The probability density function of ε_i is then as follows.

$$\begin{aligned} f_\varepsilon\left(\varepsilon_i \middle| \boldsymbol{\beta}, \sigma^2\right) &= f_\varepsilon\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \middle| \boldsymbol{\beta}, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2}{2\sigma^2}\right) \end{aligned}$$

Example: MLE and linear regression (2/4)

- The likelihood function for the parameters $\theta = (\beta; \sigma^2)$ is:

$$\mathcal{L}\left(\theta | \{y_i, \mathbf{x}_i\}_{i=1}^N\right) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^T\beta\right)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^T\beta\right)^2}{2\sigma^2}\right)$$

- . . . and the log-likelihood function is as follows.

$$\log\mathcal{L}\left(\theta | \{y_i, \mathbf{x}_i\}_{i=1}^N\right) = -\frac{N}{2}\left(\log 2\pi + \log\sigma^2\right) - \frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^T\beta\right)^2}{2\sigma^2}$$

- Both functions have a unique maximum.

Example: MLE and linear regression (3/4)

- The First Order Conditions are:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}\left(\boldsymbol{\theta} | \left\{y_i, \mathbf{x}_i\right\}_{i=1}^N\right) =$$

$$= \begin{bmatrix} \sum_{i=1}^N \mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \sigma^{-2} \\ -\frac{N}{2}\sigma^{-2} + \frac{1}{2}\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2 \sigma^{-4} \end{bmatrix} = \mathbf{0}$$

- ...and the ML estimator $\widehat{\boldsymbol{\theta}}_{MLE} = \left(\widehat{\boldsymbol{\beta}}_{MLE}; \widehat{\sigma}^2_{MLE}\right)$ is:

$$\widehat{\boldsymbol{\beta}}_{MLE} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)^2}{N}$$

- ...and it is shown to satisfy the Second Order Conditions for a maximum.

Example: MLE and linear regression (4/4)

- The ML estimator for $\boldsymbol{\beta}$ is identical to OLS, while the ML estimator for σ^2 is smaller than the unbiased estimator $\widehat{\sigma}^2$ from the small sample analysis of OLS.
- The latter is larger by a factor $\frac{N}{N-K}$: hence $\widehat{\sigma}^2_{MLE}$ is more efficient even if biased.
- Suppose that $\boldsymbol{x}_i$ is not fixed; only assume that *conditional on any realization* $\mathbf{x}_i$, the error term is normal and that it has constant variance: $\varepsilon_i | \, \mathbf{x}_i \sim \mathcal{N}\left(0, \sigma^2\right)$.
- As $\varepsilon_i = y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$, the following conditional density applies.

$$f_{Y|\boldsymbol{x}}\left(y_i | \, \mathbf{x}_i; \boldsymbol{\beta}, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

- Hence, the more flexible CMLE has the same solution as in the unrealistic "fixed regressors" case!

Consistency of M-Estimators (1/5)

- It is necessary to study the asymptotics of M-Estimators.
- To prove consistency, one must show that the maximizer of the sample average criterion converges in probability to the maximizer of the *population expected* criterion.

$$\widehat{\boldsymbol{\theta}}_M = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) \xrightarrow{p} \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{Q}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0$$

- Intuitively, **pointwise convergence** of $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$ to $\mathcal{Q}_0(\boldsymbol{\theta})$ – for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ – is necessary.

$$\left| \frac{1}{N} \sum_{i=1}^{N} \{ q(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbb{E}[q(\boldsymbol{x}_i; \boldsymbol{\theta})] \} \right| \xrightarrow{p} 0$$

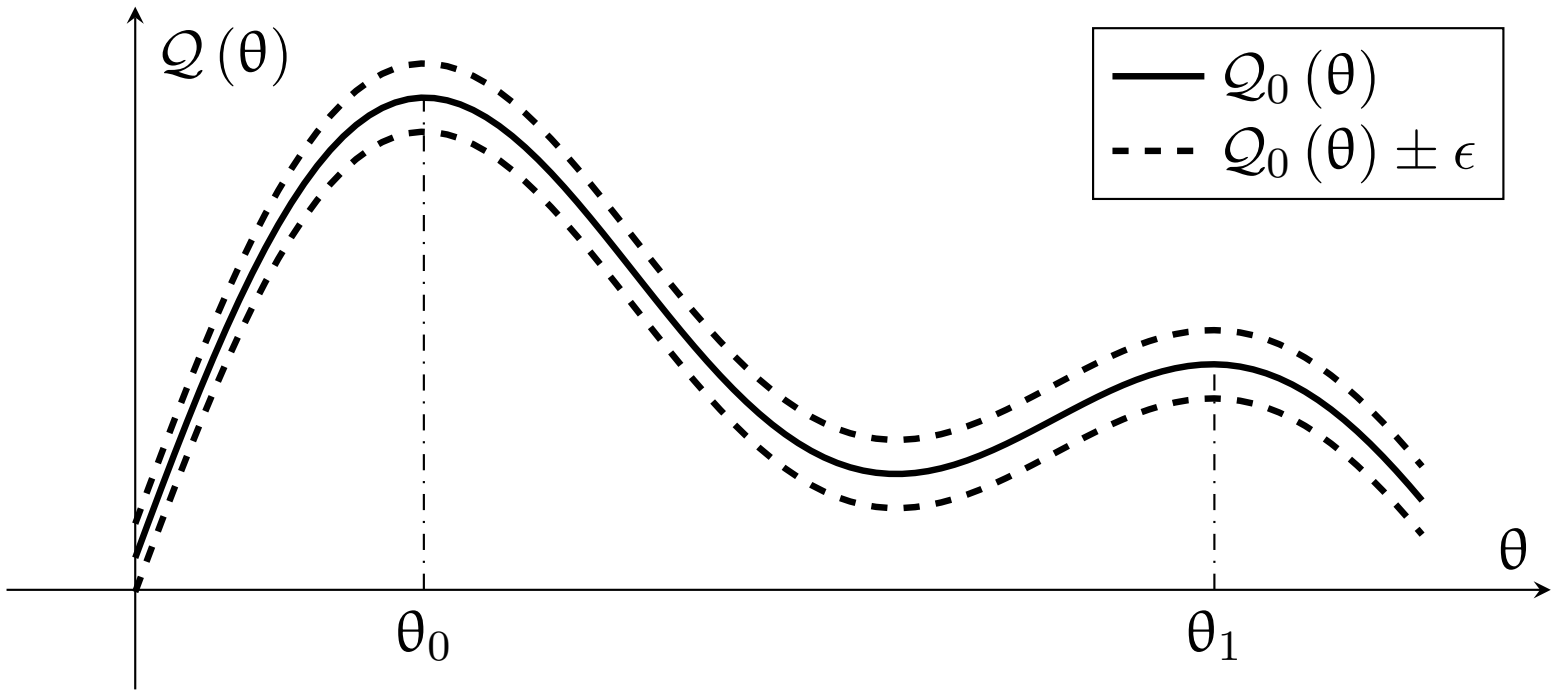
- However, it is not sufficient – unlike the stronger **uniform convergence** condition.

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{N} \sum_{i=1}^{N} \{ q(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbb{E}[q(\boldsymbol{x}_i; \boldsymbol{\theta})] \} \right| \xrightarrow{p} 0$$

Consistency of M-Estimators (2/5)

- Uniform convergence requires in fact that $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$ converges in probability towards $\mathcal{Q}_0(\boldsymbol{\theta})$ "at the same speed" over the entire parameter space $\boldsymbol{\Theta}$.

- The intuition is graphically represented in the figure below: $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) \in [\mathcal{Q}_0(\boldsymbol{\theta}) - \epsilon, \mathcal{Q}_0(\boldsymbol{\theta}) + \epsilon]$ for any $\epsilon > 0$ and $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ ($\widehat{\mathcal{Q}}_N$ always stays within a "sleeve" of $\mathcal{Q}(\theta)$).



Consistency of M-Estimators (3/5)

Uniform convergence is ensured if these four conditions hold:

i. $q(\mathbf{x}_i; \boldsymbol{\theta})$ is continuous;

ii. $\boldsymbol{\Theta}$ is a compact set;

iii. $\mathbb{E}[|q(\boldsymbol{x}_i; \boldsymbol{\theta})|] < \infty$: that is, $q(\mathbf{x}_i; \boldsymbol{\theta})$ has a bounded first absolute moment;

iv. $q(\mathbf{x}_i; \boldsymbol{\theta})$ is Borel-measureable on its support.

These conditions – together – allow to invoke a result known as **Uniform Weak Law of Large Numbers**, which then implies uniform convergence.

These conditions are technical; note that *i.* and *ii.* relate to the fact that M-Estimators are maxima; furthermore *iii.* and *iv.* are analogous to conditions from other Laws of Large Numbers.

The full-fledged proof by Newey and McFadden (1994) follows.

Consistency of M-Estimators (4/5)

Theorem 2

Consistency of M-Estimators. *If i.* $\mathcal{Q}_0(\theta)$ *is uniquely maximized at* θ_0*, ii.* $\boldsymbol{\Theta}$ *is a compact set, iii.* $\mathcal{Q}_0(\theta)$ *is a continuous function, and iv.* $\widehat{\mathcal{Q}}_N(\theta)$ *uniformly converges in probability to* $\mathcal{Q}_0(\theta)$*, it follows that M-Estimators are consistent.*

Proof.

For any $\epsilon > 0$, with probability approaching 1 (w.p.a. 1);

$$\text{by } i.: \qquad \widehat{\mathcal{Q}}_N\left(\widehat{\theta}_M\right) > \widehat{\mathcal{Q}}_N(\theta_0) - \frac{\epsilon}{3} \qquad (a)$$

$$\text{by } iv.: \qquad \mathcal{Q}_0\left(\widehat{\theta}_M\right) > \widehat{\mathcal{Q}}_N\left(\widehat{\theta}_M\right) - \frac{\epsilon}{3} \qquad (b)$$

$$\text{by } iv.: \qquad \widehat{\mathcal{Q}}_N(\theta_0) > \mathcal{Q}_0(\theta_0) - \frac{\epsilon}{3} \qquad (c)$$

which allows to show that w.p.a. 1, $\mathcal{Q}_0\left(\widehat{\theta}_M\right) > \mathcal{Q}_0(\theta_0) - \epsilon$.

(Continues...)

Consistency of M-Estimators (5/5)

Theorem 2

Proof.

(Continued.) Such a property is shown via a chain of inequalities.

$$\mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) \overset{(b)}{>} \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \frac{\epsilon}{3} \overset{(a)}{>} \widehat{\mathcal{Q}}_N\left(\boldsymbol{\theta}_0\right) - \frac{2\epsilon}{3} \overset{(c)}{>} \mathcal{Q}_0\left(\boldsymbol{\theta}_0\right) - \epsilon$$

Now, denote by $\mathbb{U}$ any given open neighborhood of $\boldsymbol{\theta}_0$ and by $\mathbb{U}^c$ its complement in $\boldsymbol{\Theta}$. Also define,

$$\mathcal{Q}_0\left(\boldsymbol{\theta}^*\right) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \cap \mathbb{U}^c} \mathcal{Q}_0\left(\boldsymbol{\theta}\right)$$

for some $\boldsymbol{\theta}^*$, and notice that $\mathcal{Q}_0\left(\boldsymbol{\theta}^*\right) < \mathcal{Q}_0\left(\boldsymbol{\theta}_0\right)$ by *i.*-*ii.*-*iii.*: thus, by setting $\epsilon = \mathcal{Q}_0\left(\boldsymbol{\theta}_0\right) - \mathcal{Q}_0\left(\boldsymbol{\theta}^*\right)$ it follows that:

$$\mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \mathcal{Q}_0\left(\boldsymbol{\theta}_0\right) - \epsilon \Rightarrow \mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \mathcal{Q}_0\left(\boldsymbol{\theta}^*\right)$$

implying $\widehat{\boldsymbol{\theta}}_M \in \mathbb{U}$ for any open neighborhood $\mathbb{U}$, and so $\widehat{\boldsymbol{\theta}}_M \xrightarrow{p} \boldsymbol{\theta}_0$. $\square$

Asymptotic normality of M-Estimators (1/5)

The next step is the analysis of M-Estimators' limiting distribution.

Theorem 3

Asymptotic Normality of M-Estimators. *A given M-Estimator* $\widehat{\theta}_M$ *follows an asymptotically normal distribution if the following five conditions hold simultaneously:*

i. $\widehat{\theta}_M$ *is a consistent estimator of* θ_0*;*

ii. $q(\mathbf{x}_i; \theta)$ *is a concave and twice continuously differentiable function in an open neighborhood of* θ_0*;*

iii. $\frac{\partial}{\partial \theta} \mathbb{E}[q(\boldsymbol{x}_i; \theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta} q(\boldsymbol{x}_i; \theta)\right]$*: the derivative can pass through the expectation integral;*

iv. the data meet the requirements for the application of a Central Limit Theorem (the data are "well behaved");

v. the Hessian matrix is nonsingular, it is continuous in θ *and it has a bounded absolute first moment.*

(Continues. . .)

Asymptotic normality of M-Estimators (2/5)

Theorem 3

(Continued.) *The limiting distribution is:*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_0^{-1}\boldsymbol{\Upsilon}_0\mathbf{Q}_0^{-1}\right)$$

where $\mathbf{Q}_0$ *and* $\boldsymbol{\Upsilon}_0$ *are defined as the following probability limits:*

$$\lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right] \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{H}_i\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right] \xrightarrow{p} \mathbf{Q}_0$$

which implies the following asymptotic distribution, for a fixed N*.*

$$\widehat{\boldsymbol{\theta}}_M \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{N}\mathbf{Q}_0^{-1}\boldsymbol{\Upsilon}_0\mathbf{Q}_0^{-1}\right)$$

Asymptotic normality of M-Estimators (3/5)

Theorem 3

Proof.

The derivation is analogous to those given in Lecture 6. By the mean value theorem:

$$\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_M\right)=\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta}_0\right)+\mathbf{H}_i\left(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\left(\widehat{\boldsymbol{\theta}}_M-\boldsymbol{\theta}_0\right)$$

where $\widetilde{\boldsymbol{\theta}}_N$ is some convex combination of $\widehat{\boldsymbol{\theta}}_M$ and $\boldsymbol{\theta}_0$. Summing over the N observations and dividing by $\sqrt{N}$, one gets:

$$\begin{aligned}\mathbf{0}&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_M\right)\\&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta}_0\right)+\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i\left(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_M-\boldsymbol{\theta}_0\right)\end{aligned}$$

by recalling that the sample score evaluated at the solution is equal to zero by definition of M-Estimators. **(Continues...)**

Asymptotic normality of M-Estimators (4/5)

Theorem 3

Proof.

(Continued.) The expression above can be rewritten as:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i\left(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta}_0\right)$$

as condition *v.* lets invert the average Hessian matrix. Consider that:

1. By *i.* and *v.* one can apply some suitable Law of Large Numbers to the "sample-averaged" Hessian matrix, showing that:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i\left(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}_N\right) \xrightarrow{p} \mathbf{Q}_0$$

which follows from the Continuous Mapping Theorem given that, at the same time, $\widetilde{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$. **(Continues...)**

Asymptotic normality of M-Estimators (5/5)

Theorem 3

Proof.
(Continued.)

2. condition *iii.* yields $\frac{\partial \mathcal{Q}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{s}_i\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right)\right] = \mathbf{0}$, thus:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right) \xrightarrow{p} \mathbf{0}$$

so by condition *iv.* and the Continuous Mapping Theorem, it is as follows.

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i\left(\mathbf{x}_i; \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Upsilon}_0\right)$$

All the intermediate results are – as usual – recombined via Slutskij's Theorem and the Cramér-Wold Device to show the desired result. □

Inference for M-Estimators (1/2)

- To perform inference on M-Estimators, $\boldsymbol{\Upsilon}_0$ and $\mathbf{Q}_0$ must be estimated: as usual, the analogy principle is applied here.

- The "bread" matrix $\mathbf{Q}_0$ is estimated easily.

$$\widehat{\mathbf{Q}}_N \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i \left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M \right) \xrightarrow{p} \mathbf{Q}_0$$

- With the "meat" matrix $\boldsymbol{\Upsilon}_0$, distributional assumptions on the sample matter. If the observations are independent but possibly not identically distributed, it is:

$$\boldsymbol{\Upsilon}_0 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\mathbf{s}_i \left(\boldsymbol{x}_i; \boldsymbol{\theta}_0 \right) \mathbf{s}_i^{\mathrm{T}} \left(\boldsymbol{x}_i, \boldsymbol{\theta}_0 \right) \right]$$

with $\boldsymbol{\Upsilon}_0 = \mathbb{E} \left[\mathbf{s}_i \left(\boldsymbol{x}_i; \boldsymbol{\theta}_0 \right) \mathbf{s}_i^{\mathrm{T}} \left(\boldsymbol{x}_i, \boldsymbol{\theta}_0 \right) \right]$ in the i.i.d. case.

Inference for M-Estimators (2/2)

- With independent observations $\boldsymbol{\Upsilon}_0$ is estimated as follows.

$$\widehat{\boldsymbol{\Upsilon}}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_i\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M\right)\mathbf{s}_i^{\mathrm{T}}\left(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

- Dependent observations complicate matters. The HAC case is feasible, but often complicated. The CCE case yields:

$$\widehat{\boldsymbol{\Upsilon}}_{CCE} = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c} \mathbf{s}_{ic}\left(\mathbf{x}_{ic}; \widehat{\boldsymbol{\theta}}_M\right)\mathbf{s}_{jc}^{\mathrm{T}}\left(\mathbf{x}_{jc}; \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

  where both observations and scores are also indexed by the group or cluster $c = 1, \ldots, C$ they belong to.

- For a suitable estimator $\widehat{\boldsymbol{\Upsilon}}_N \xrightarrow{p} \boldsymbol{\Upsilon}_0$, the variance-covariance matrix of M-Estimators is ultimately estimated as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_M\right) = \frac{1}{N}\widehat{\mathbf{Q}}_N^{-1}\widehat{\boldsymbol{\Upsilon}}_N\widehat{\mathbf{Q}}_N^{-1}$$

Examples: Asymptotics of OLS and NLLS (1/3)

- This analysis is reconciled with that about OLS (Lecture 8) by noting that there consistency follows from $\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0$, and that in that case it is $\boldsymbol{\Upsilon}_0 = 4\boldsymbol{\Xi}_0$ and $\mathbf{Q}_0 = -2\mathbf{K}_0$.

- To discuss NNLS, it is useful to define a $K \times 1$ vector: that is, the derivative of the CEF evaluated at $\mathbf{x}_i$ and at $\widehat{\boldsymbol{\theta}}_{NLLS}$.

$$\widehat{\mathbf{h}}_i \equiv \frac{\partial}{\partial \boldsymbol{\theta}} h\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{NLLS}\right)$$

- Recall that by construction, NLLS sets the average score at zero for every value of N.

$$\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{NLLS}\right) = \frac{1}{N}\sum_{i=1}^{N} 2\widehat{\mathbf{h}}_i \varepsilon_i \xrightarrow{p} \mathbf{0}$$

- Refer to it as the "condition about the probability limit of the score."

Examples: Asymptotics of OLS and NLLS (2/3)

- Also recall the CEF motivation of NLLS.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = \mathbb{E}\left[y_i \middle| \boldsymbol{x}_i\right] - h\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) = 0$$

- The immediate implication of such a CEF condition is that:

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} h\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right) \varepsilon_i\right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{h}_{0i} \varepsilon_i\right] \\ &= \mathbf{0}\end{aligned}$$

  which can be reconciled with the above condition about the probability limit of the score if $\widehat{\mathbf{h}}_i \xrightarrow{p} \boldsymbol{h}_{0i}$ for $i = 1, \ldots, N$ as per some applicable Law of Large Numbers.

- The Continuous Mapping Theorem finally implies that:

$$\widehat{\boldsymbol{\theta}}_{NLLS} \xrightarrow{p} \boldsymbol{\theta}_0$$

  that is, the NLLS estimator is indeed consistent.

Examples: Asymptotics of OLS and NLLS (3/3)

- Regarding the asymptotic distribution, note that under the hypothesis of *independent observations* the following holds.

$$\mathbf{Q}_0 = \lim_{N\to\infty} -\frac{1}{N}\sum_{i=1}^{N} 2\cdot\mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

$$\boldsymbol{\Upsilon}_0 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} 4\cdot\mathbb{E}\left[\varepsilon_i^2\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

- Therefore, the residual $e_i \equiv y_i - h\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_{NLLS}\right)$ delivers:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_{NLLS}\right) = \left[\sum_{i=1}^{N}\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]^{-1}\left[\sum_{i=1}^{N}e_i^2\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]\left[\sum_{i=1}^{N}\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]^{-1}$$

  the **heteroscedasticity-consistent** estimator of the NLLS variance-covariance, akin to the OLS "robust" formula.

- The homoscedastic, CCE and HAC versions are akin too.

The Trinity of Asymptotic Tests

After estimating an econometric model, a researcher is typically interested into performing some **tests of hypothesis**; these are possibly non-linear:

$$\mathrm{H}_0 : \ \boldsymbol{v}\left(\boldsymbol{\theta}_0\right) = \mathbf{0} \qquad\qquad \mathrm{H}_1 : \ \boldsymbol{v}\left(\boldsymbol{\theta}_0\right) \neq \mathbf{0}$$

where $\boldsymbol{v}\left(\cdot\right)$, a vector-valued function, has length L (for multiple hypotheses).

There are three alternative methods to perform such tests; they are known together as the "Trinity."

1. Generalized Wald Statistics.
2. Distance, or Likelihood Ratio test.
3. Score – or Lagrange multiplier – test.

These methods are briefly review under the unifying framework of M-Estimation.

The Generalized Wald Statistics (1/2)

The **generalized** Wald Statistics is:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} = \boldsymbol{v}^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_M\right) \cdot \left[\widehat{\mathbf{V}} \cdot \widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_M\right) \cdot \widehat{\mathbf{V}}^{\mathrm{T}}\right]^{-1} \cdot \boldsymbol{v}\left(\widehat{\boldsymbol{\theta}}_M\right)$$

where $\widehat{\mathbf{V}} \equiv \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{v}\left(\widehat{\boldsymbol{\theta}}_M\right)$. Its limiting distribution is:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} \xrightarrow{d} \chi^2_L$$

that is, **under the null hypothesis** H_0 the test statistic has a limiting χ^2_L distribution with L degrees of freedom.

To gain intuition recall the Wald Statistic from the linear model (where $\boldsymbol{v}(\boldsymbol{\beta}) = \mathbf{R}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$ is some linear function). There, the Wald Statistic is a special case of Hotelling's t-squared statistic, which is asymptotically chi-squared distributed. This non-linear case is analogously derived by exploiting the Delta Method.

The Generalized Wald Statistics (2/2)

Example. Suppose that interest lies in one specific hypothesis about the linear model:

$$\mathrm{H}_0 : \sum_{k=1}^{K} \beta_k = 1 \qquad\qquad \mathrm{H}_1 : \sum_{k=1}^{K} \beta_k \neq 1$$

like say the hypothesis of constant return to scale in production functions (with the constant parameter being β_0). In this case:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} = \frac{\left(\sum_{k=1}^{K} \widehat{\beta}_{k,OLS} - 1\right)^2}{\sum_{k=1}^{K}\sum_{q=1}^{K} \widehat{\sigma}_{\beta kq}} \xrightarrow{d} \chi_1^2$$

where the kq-th element of the estimated variance-covariance of the OLS estimates is denoted as $\widehat{\sigma}_{\beta kq}$.

The Wald Test is very easy to implement but it suffers from two issues. First, it performs poorly in small samples; and second, it is not transformation-invariant – for example, different statistics are calculated whether $\mathrm{H}_0 : \beta_k = 0$ or $\mathrm{H}_0 : \exp(\beta_k) = 1$.

The Distance, or Likelihood Ratio test (1/2)

The "Distance Test" or "Likelihood Ratio Test,"originates with MLE. With respect to the simpler Generalized Wald Test, it is transformation-invariant and deals non-linear hypotheses quite well. However, it is more computationally demanding.

In fact, to perform it one must compute the main estimate of $\boldsymbol{\theta}$ as well as an additional "restricted" estimate

$$\widehat{\boldsymbol{\theta}}_V = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_V}{\arg\max}\, \widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$$

with a "restricted parameter space" $\boldsymbol{\Theta}_V = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{v}(\boldsymbol{\theta}) = \mathbf{0}\}$. Then, the "Distance" Statistic in all cases but MLE is:

$$\mathrm{D}_{\mathrm{H}_0} = N\left[\widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_V\right)\right] \xrightarrow{d} \chi^2_L$$

while:

$$\mathrm{LR}_{\mathrm{H}_0} = 2\left[\log \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \log \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_V\right)\right] \xrightarrow{d} \chi^2_L$$

is the "Likelihood Ratio" from MLE, where $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) = \widehat{\mathcal{L}}_N(\boldsymbol{\theta})$.

The Distance, or Likelihood Ratio test (2/2)

Intuitively, the test compares the advantage in terms of **fitting** when letting the model be estimated without the restriction.

Example. Expanding on the toy test that used the Generalized Wald Statistic, one can estimate a "restricted" model, such as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{K-1} X_{(K-1)i} + \left(1 - \sum_{k=1}^{K-1} \beta_k\right) X_{Ki} + \epsilon_i$$

which can also be written, for $\ddot{X}_{ki} \equiv X_{ki} - X_{Ki}$ $(k = 1, \ldots, K)$:

$$Y_i - X_{Ki} = \beta_0 + \beta_1 \ddot{X}_{1i} + \beta_2 \ddot{X}_{2i} + \cdots + \beta_{K-1} \ddot{X}_{(K-1)i} + \varepsilon_i$$

restricting β_K or possibly another coefficient (except β_0). Thus:

$$\mathrm{D}_{\mathrm{H}_0} = \left[\sum_{i=1}^{N} \left(y_i - x_{Ki} - \ddot{\mathbf{x}}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_V\right)^2 - \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{OLS}\right)^2\right] \xrightarrow{d} \chi^2$$

is the Distance Test, and $\widehat{\boldsymbol{\beta}}_V$ are the "restricted" estimates.

The score, or Lagrange multiplier, test (1/2)

This last test has the same advantages as the previous one, but it does not require the "unrestricted" model to be estimated: it is computationally more parsimonious. The test is based on the properties of the sample average score function evaluated at **one specific parameter value $\boldsymbol{\theta_v}$** implied by the null hypothesis.

By construction, in all M-Estimators the average score is always equal to zero when evaluated at the unrestricted estimate $\widehat{\boldsymbol{\theta}}_M$.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_M\right)=\mathbf{0}$$

Consider a *restricted* parameter value $\boldsymbol{\theta_v}$ such that $\boldsymbol{v}\left(\boldsymbol{\theta_v}\right)=\mathbf{0}$. It follows that:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{s}_i\left(\mathbf{x}_i;\boldsymbol{\theta_v}\right)\right|>\mathbf{0}$$

the K-dimensional sample score vector deviates from zero when evaluated at any "suboptimal" parameter choice.

The score, or Lagrange multiplier, test. (2/2)

The Lagrange Multiplier statistic is:

$$\mathrm{LM}_{\mathrm{H}_0} = N \left[\sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}}\right)\right]^{\mathrm{T}} \widehat{\boldsymbol{\Upsilon}}_{\mathbf{v}}^{-1} \left[\sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}}\right)\right] \xrightarrow{d} \chi^2_K$$

where:

$$\widehat{\boldsymbol{\Upsilon}}_{\mathbf{v}} = \widehat{\mathbb{A}\mathrm{var}}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}}\right)\right].$$

It indicates how "statistically relevant" are the deviations of the restricted score from the zero benchmark.

Example. Return again to the running examples: consider the "restricted" estimates $\boldsymbol{\beta}_{\mathbf{v}}$ obtained by forcing all key coefficients to sum up to one. The Lagrange Multiplier Statistic is:

$$\mathrm{LM}_{\mathrm{H}_0} = \left[\sum_{i=1}^{N} \mathbf{x}_i e_i \left(\boldsymbol{\beta}_{\mathbf{v}}\right)\right]^{\mathrm{T}} \left[\sum_{i=1}^{N} e_i^2 \left(\boldsymbol{\beta}_{\mathbf{v}}\right) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{x}_i e_i \left(\boldsymbol{\beta}_{\mathbf{v}}\right)\right]$$

with $\mathrm{LM}_{\mathrm{H}_0} \xrightarrow{d} \chi^2_K$, given the new residuals $e_i \left(\boldsymbol{\beta}_{\mathbf{v}}\right) = y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{\mathbf{v}}$.

The Maximum Likelihood special case

- As seen in Lectures 5 and 6, with MLE if the data are i.i.d. the Information Matrix Equality applies.

$$\boldsymbol{\Upsilon}_0 = -\mathbf{Q}_0$$

- Hence, the asymptotic distribution is:

$$\widehat{\theta}_{MLE} \overset{A}{\sim} \mathcal{N}\left(\theta_0, \left[\boldsymbol{I}_N\left(\theta_0\right)\right]^{-1}\right)$$

  where $\boldsymbol{I}_N\left(\theta_0\right)$ is the information matrix.

- Importantly, the information matrix hits the **Cramér-Rao bound**, making MLE the most efficient estimator under its distributional assumptions.

- The information matrix can be **estimated** either by $\widehat{\boldsymbol{\Upsilon}}_N$ or via $-\widehat{\mathbf{Q}}_N$; the former approach is called **outer product of the gradients**, typically more computationally convenient.

Example: MLE linear regression expanded (1/2)

- Return to the MLE framework for linear regression. There, the score is as follows.

$$\mathbf{s}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = \begin{bmatrix} \mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \sigma^{-2} \\ -\frac{1}{2}\sigma^{-2} + \frac{1}{2}\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2 \sigma^{-4} \end{bmatrix}$$

- The (symmetric) Hessian matrix is instead as follows.

$$\mathbf{H}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = \\ = \begin{bmatrix} -\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \sigma^{-2} & -\mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \sigma^{-4} \\ -\mathbf{x}_i^{\mathrm{T}} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \sigma^{-4} & \frac{1}{2}\sigma^{-4} - \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2 \sigma^{-6} \end{bmatrix}$$

- By plugging the MLE estimators $\widehat{\boldsymbol{\beta}}_{MLE}$ and $\widehat{\sigma}^2_{MLE}$ into the above expressions one can construct **consistent** estimators for the information matrix.

Example: MLE linear regression expanded (2/2)

- The estimator based on the Hessian is as follows.

$$\begin{aligned}\frac{\widehat{\mathbf{Q}}_N^{-1}}{N} &= \left[\sum_{i=1}^{N} \mathbf{H}_i \left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1} \\ &= -\begin{bmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \widehat{\sigma}^2_{MLE} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \frac{2}{N}\widehat{\sigma}^4_{MLE} \end{bmatrix}\end{aligned}$$

- The next is based on the outer product of the gradients.

$$\begin{aligned}\frac{\widehat{\boldsymbol{\Upsilon}}_N^{-1}}{N} &= \left[\sum_{i=1}^{N} \mathbf{s}_i \left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right) \mathbf{s}_i^{\mathrm{T}} \left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1} \\ &= \begin{bmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \widehat{\sigma}^2_{MLE} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \frac{2}{N}\widehat{\sigma}^4_{MLE} \end{bmatrix}\end{aligned}$$

- Note the different sign, the zero vectors at the border, and the bottom-right element for the variance of $\widehat{\sigma}^2_{MLE}$.

Quasi-Maximum Likelihood (1/2)

- The nice properties of MLE break apart when one deviates from the i.i.d. benchmark. For example, in case of clustered dependence the likelihood function reads:

$$\mathcal{L}\left(\boldsymbol{\theta}|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right)=\prod_{c=1}^{C} f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{N_c}}\left(\mathbf{x}_{1c},\ldots,\mathbf{x}_{N_c c};\boldsymbol{\theta}\right)$$

  but it cannot be factorized further. Hence, the Information Matrix Equality no longer holds and $\widehat{\boldsymbol{\Upsilon}}_{CCE}$ must be used.

- Things are even worse when MLE is built around incorrect distributional assumptions (even if the data are i.i.d.).

- In *misspecification* cases, the estimator in question is called **Quasi-Maximum Likelihood Estimator** $\widehat{\boldsymbol{\theta}}_{QMLE}$ and it has a probability limit called **pseudo-true value** $\boldsymbol{\theta}^*$.

$$\widehat{\boldsymbol{\theta}}_{QMLE}=\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\,\mathcal{L}\left(\boldsymbol{\theta}|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\xrightarrow{p}\boldsymbol{\theta}^*$$

Quasi-Maximum Likelihood (2/2)

- A key question is: when is the Quasi-Maximum Likelihood Estimator (QMLE) **consistent**, that is $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$?

- In the MLE **linear** regression example, $\widehat{\boldsymbol{\beta}}_{MLE}$ is consistent under a linear CEF even if the errors are not normal!

- If the assumed distribution is $f_{Y,\boldsymbol{x}}(y_i, \mathbf{x}_i; \boldsymbol{\theta})$ for some scalar endogenous variable Y_i, and it belongs to the **exponential** macro-**family** – it can be decomposed in terms of primitive scalar functions $a[\cdot]$, $b[\cdot]$ and $c[\cdot]$ as:

$$f_{Y,\boldsymbol{x}}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \\ = \exp\left\{a\left[\mu_{Y|\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta})\right] + b[y_i] + c\left[\mu_{Y|\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta})\right] y_i\right\}$$

where: $\mu_{Y|\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta}) \equiv \mathbb{E}[Y_i | \boldsymbol{x}_i; \boldsymbol{\theta}]$ is **a correctly specified** parametric expression of the CEF of Y_i given $\boldsymbol{x}_i$, in general (Q)MLE consistently estimates the function $\mu_{Y|\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta})$.

Example: Poisson regression (1/4)

- A **count data model** is a model for explaining variables Y_i that assume non-negative integer values $Y_i = 0, 1, 2, \ldots$ *only* and where smaller values occur more frequently.

- An example is the **Poisson regression**, that is:

$$\mathbb{P}\left(Y_i \middle| \boldsymbol{x}_i\right) = \frac{\lambda_i\left(\boldsymbol{x}_i\right)^{Y_i} \exp\left(-\lambda_i\left(\boldsymbol{x}_i\right)\right)}{Y_i!}$$

  where the Poisson parameter $\lambda_i\left(\boldsymbol{x}_i\right)$ is treated as a function of the individual characteristics $\boldsymbol{x}_i$.

- By the properties of the Poisson distribution:

$$\lambda_i\left(\boldsymbol{x}_i\right) = \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = \mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right]$$

  that is, $\lambda_i\left(\boldsymbol{x}_i\right)$ equals both the **conditional mean** and the **conditional variance** of Y_i given $\boldsymbol{x}_i$.

Example: Poisson regression (2/4)

- The most common choice is $\lambda_i(\boldsymbol{x}_i) = \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$, yielding:

$$\mathbb{P}(Y_i|\,\boldsymbol{x}_i) = \frac{\exp\left(Y_i \cdot \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\exp\left[-\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right]}{Y_i!}$$

- ...an implication of which is:

$$\frac{\partial \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)}{\partial \boldsymbol{x}_i} = \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\boldsymbol{\beta}_0$$

$$\Rightarrow\ \boldsymbol{\beta}_0 = \frac{\partial\,\mathbb{E}[Y_i|\,\boldsymbol{x}_i]}{\partial \boldsymbol{x}_i}\frac{1}{\mathbb{E}[Y_i|\,\boldsymbol{x}_i]}$$

  hence, $\boldsymbol{\beta}_0$ can be interpreted as a **semi-elasticiy** like in a *log-lin* model that can be estimated via OLS.

- Unlike a *log-lin* model the Poisson regression allows $Y_i = 0$, which can be often useful or convenient!

Example: Poisson regression (3/4)

- The **log-likelihood** function, given a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, writes as follows for this model.

$$\log \mathcal{L}\left(\boldsymbol{\beta} | \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \\ = \sum_{i=1}^N \left[y_i \cdot \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \exp\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) - \log y_i!\right]$$

- The First Order Conditions of the MLE problem, expressed as the sum of the individual scores, consequently are:

$$\sum_{i=1}^N \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{MLE}\right) = \sum_{i=1}^N \mathbf{x}_i \left[y_i - \exp\left(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)\right] = \mathbf{0}$$

- . . . and they lack a closed form solution; therefore, the MLE estimator must be obtained by numerical methods.

Example: Poisson regression (4/4)

- The empirical Hessian matrix is:

$$\begin{aligned}\widehat{\mathbf{Q}}_N &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i\left(y_i,\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_{MLE}\right)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\exp\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)\cdot\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\end{aligned}$$

  the opposite of which – if divided by N – is an appropriate estimator of the information matrix.

- The Poisson distribution belongs to the exponential family: therefore, even if the likelihood is misspecified, the **CEF** of Y_i given $\boldsymbol{x}_i$ is consistently estimated if it is well specified.

- With the "exponential" specification for $\lambda_i(\boldsymbol{x}_i)$ this implies that the MLE is interpretable in terms of semi-elasticities.

Quasi-Maximum Likelihood: more remarks (1/3)

- The Poisson regression is an extreme case where QMLE is safe, whereas in the linear model $\widehat{\sigma}^2_{MLE}$ has little meaning without the normality assumption.

- This bears implications for inference. In general:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{QMLE} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\mathbf{Q}^*\right]^{-1} \boldsymbol{\Upsilon}^* \left[\mathbf{Q}^*\right]^{-1}\right)$$

  where $\mathbf{Q}^*$ and $\boldsymbol{\Upsilon}^*$ are analogues of $\mathbf{Q}_0$ and $\boldsymbol{\Upsilon}_0$ respectively, but evaluated at $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_0$.

- It is thus safer to conduct inference using the "sandwiched" formula of M-Estimation – the extensive one!

- What if QMLE is actually inconsistent: $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$? Not all is lost: QMLE can still be given an interpretation in terms of "best approximation," like OLS with a non-linear CEF.

Quasi-Maximum Likelihood: more remarks (2/3)

- To see this, consider the **Kullback-Leibler Information Criterion** (KLIC).

$$\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}) \equiv \mathbb{E}_g\left[\log\left(\frac{g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}_0)}{f_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta})}\right)\right] =$$

$$= \int_{\mathbb{X}} \log\left(\frac{g_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}_0)}{f_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta})}\right) g_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}_0)\,\mathrm{d}\mathbf{x}_1\ldots\mathrm{d}\mathbf{x}_N$$

- Here, $f_{\boldsymbol{x}}(\cdot)$ is the **assumed** joint mass or density function that is thought to generate the data.

- Instead $g_{\boldsymbol{x}}(\cdot)$ is the **true** function, which is taken as given; the expectation is taken with respect to $g_{\boldsymbol{x}}(\cdot)$.

- Note that by construction, $\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- If the distribution is correctly specified, it is $f_{\boldsymbol{x}}(\cdot) = g_{\boldsymbol{x}}(\cdot)$, and $\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}_0) = 0$: the KLIC would attain its minimum.

Quasi-Maximum Likelihood: more remarks (3/3)

- In addition:

$$\begin{aligned}\mathcal{K}_{\boldsymbol{x}}(\theta) &= \mathbb{E}_g\left[\log g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\theta_0) - \log f_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\theta)\right] \\ &= \mathbb{E}_g\left[\log g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\theta_0)\right] - \log \mathcal{L}_0^g(\theta|\,\mathbf{x}_1,\ldots,\mathbf{x}_N)\end{aligned}$$

  where $\log \mathcal{L}_0^g(\theta|\,\mathbf{x}_1,\ldots,\mathbf{x}_N)$ is here a **pseudo-population likelihood** that results if the assumed distribution is $f_{\boldsymbol{x}}(\cdot)$, but the true one is $g_{\boldsymbol{x}}(\cdot)$.

- So, under general assumptions:

$$\widehat{\theta}_{QMLE} \xrightarrow{p} \theta^* = \max_{\theta\in\boldsymbol{\Theta}} \log \mathcal{L}_0^g(\theta|\,\mathbf{x}_1,\ldots,\mathbf{x}_N) = \min_{\theta\in\boldsymbol{\Theta}} \mathcal{K}_{\boldsymbol{x}}(\theta)$$

  the pseudo-true value θ^*, to which the QMLE converges in probability, is both the maximizer of the pseudo-population likelihood and the minimizer of the KLIC!

- Trying to "augment" an MLE model, however misspecified, aids this KLIC-motivated "approximation" interpretation.

Introduction to Binary Outcome Models (1/4)

A subclass of econometric models is characterized by a **limited dependent variable** (LDV): Y_i has a *discrete, finite* support.

Many questions in economics revolve in fact around a **binary** dependent variable: $Y_i \in \{0, 1\}$. Here are some examples.

- What are the determinants of enrollment in college?
- Which factors influence a firm's probability of default?
- What are the causes of civil wars in a country?

Other LDV models take multiple outcomes.

- What means of transportations do individuals choose?
- What determines a country's political regime?
- What type of insurance contract is preferred by people?
- Which individual characteristics predict survey responses?

Introduction to Binary Outcome Models (2/4)

It is useful to overview models for **binary outcomes** $Y_i \in \{0, 1\}$ so as to appreciate the usefulness of MLE in this setting.

- If one treats Y_i as random (Bernoulli) event, it is natural to think about its realization *probability* as **conditional** upon some observable variables $\boldsymbol{x}_i$:

$$\begin{aligned}\mathbb{P}\left(Y_i = 1 \middle| \boldsymbol{x}_i\right) &= G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right) \\ \mathbb{P}\left(Y_i = 0 \middle| \boldsymbol{x}_i\right) &= 1 - G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\end{aligned}$$

  where $G(\cdot)$ is some function of $\boldsymbol{x}_i$ parametrized by $\boldsymbol{\beta}_0$.

- As the problem is binary, the probability of either outcome can be treated residually with respect to the other's. Also, one can write the CEF of Y_i given $\boldsymbol{x}_i$ as follows.

$$\begin{aligned}\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] &= 1 \cdot \left[G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\right] + 0 \cdot \left[1 - G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\right] \\ &= G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\end{aligned}$$

Introduction to Binary Outcome Models (3/4)

- Can such a model be estimated via linear regression? This is called the **linear probability model** (LPM):

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \epsilon_i, \quad y_i \in \{0, 1\}$$

  which is equivalent to assuming $G(\boldsymbol{x}_i, \boldsymbol{\beta}_0) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$.

- By definition of regression: $\epsilon_i = Y_i - \mathbb{E}[Y_i | \boldsymbol{x}_i] = Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$; this has a "natural heteroscedasticity" implication.

$$\mathbb{P}\left(\epsilon_i = 1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \middle| \boldsymbol{x}_i\right) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$$
$$\mathbb{P}\left(\epsilon_i = -\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \middle| \boldsymbol{x}_i\right) = 1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$$

- A consequent implication is:

$$\mathbb{E}[\epsilon_i | \boldsymbol{x}_i] = \left(1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \left(1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) = 0$$

  hence, OLS still delivers unbiased and consistent estimates of $\boldsymbol{\beta}_0$ even if the problem is naturally heteroscedastic.

Introduction to Binary Outcome Models (4/4)

"Natural heteroscedasticity" is no issue for the LPM: it can be addressed via "robust" estimation, or FGLS in small samples.

The main issue is that the linear conditional expectation $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ cannot be constrained to lie within the $(0,1)$ interval. Hence:

1. the conditional variance of the error term ϵ_i takes negative values;
$$\mathbb{V}\mathrm{ar}\left[\epsilon_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\left(1 - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) \gtreqless 0$$

2. the predicted probabilities $\widehat{\mathbb{E}}\left[Y_i \middle| \boldsymbol{x}_i = \mathbf{x}_i\right] = \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{LPM} = \widehat{y}_i$ take values outside the $[0,1]$ interval.

Both facts that make **no probabilistic sense** and call for more parametric models estimated by MLE. Yet, the LPM is however used especially when interest falls on transparent **causal** effects of $\boldsymbol{x}_i$ on Y_i, as it is easier to implement.

MLE for Binary Outcome Models (1/8)

All MLE approaches to binary outcome models are based on an assumption of the kind that, for each realization $\mathbf{x}_i$ of $\boldsymbol{x}_i$:

$$G\left(\mathbf{x}_i, \boldsymbol{\beta}_0\right) = F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) = F_{\boldsymbol{x}}\left(\lambda_i\right)$$

where $F_{\boldsymbol{x}}(\cdot)$ is some **probability distribution function** with one "free" parameter (usually a location parameter) $\lambda_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$.

Most applications use a **symmetric distribution** function, so that $F\left(\lambda_i\right) = 1 - F\left(-\lambda_i\right)$. This solves the problem of predicted probabilities, since conditionally on any realization $\mathbf{x}_i$:

$$\lim_{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \to +\infty} \mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) = \lim_{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \to +\infty} F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) = 1$$

$$\lim_{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \to -\infty} \mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) = \lim_{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \to -\infty} F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) = 0$$

and symmetrically for $Y_i = 0$.

MLE for Binary Outcome Models (2/8)

Another advantage is a model of this sort can be based upon a **structural** interpretation about "choices." Specifically, write:

$$y_i^* = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \alpha_0 \\ 0 & \text{if } y_i^* \leq \alpha_0 \end{cases}$$

where Y_i^* is a **latent variable** that represents the **cost-benefit** evaluation of the binary choice by the i-th individual.

- The latent variable Y_i^* is unobserved, and it is a theoretical abstraction used for interpretation's sake (like "utility").

In such a model, conditionally on any realization $\mathbf{x}_i$:

$$\mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) = \mathbb{P}\left(Y_i^* > \alpha_0 \middle| \mathbf{x}_i\right) = \mathbb{P}\left(\varepsilon_i > -\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \alpha_0 \middle| \mathbf{x}_i\right)$$

and intuitively, if $\mathbf{x}_i$ contains a constant element the associated "intercept" parameter and α_0 are not separately identified. The **normalization** $\alpha_0 = 0$ is thus typically imposed.

MLE for Binary Outcome Models (3/8)

If $F_{\boldsymbol{x}}(\cdot)$ is a symmetric distribution, the model reshapes as:

$$\begin{aligned}
\mathbb{P}(Y_i = 1 | \mathbf{x}_i) &= \mathbb{P}(Y_i^* > 0 | \mathbf{x}_i) \\
&= \mathbb{P}\left(\varepsilon_i > -\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \middle| \mathbf{x}_i\right) \\
&= 1 - \mathbb{P}\left(\varepsilon_i \leq -\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \middle| \mathbf{x}_i\right) \\
&= \mathbb{P}\left(\varepsilon_i < \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \middle| \mathbf{x}_i\right) \\
&= F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)
\end{aligned}$$

where the fourth line exploits the symmetry of $F_{\boldsymbol{x}}(\cdot)$. This fact reconciles the latent variable model with the specification of the conditional probability for the outcome Y_i.

- Latent variable models are not specific of binary outcomes: multinomial LDV models are usually motivated by complex variations of this approach. Latent variable models are also used in the structural analysis of empirical strategic games.

MLE for Binary Outcome Models (4/8)

Note that the distribution $F_{\boldsymbol{x}}(\cdot)$ should *not* contain a **variable** scale parameter; if it is say normal, its variance must be known or normalized, e.g. $\sigma^2 = 1$); or else the K parameters in $\boldsymbol{\beta}_0$ and the scale parameter would not be separately identified.

To see intuitively why consider the case where $\alpha_0 = 0$ and $F_{\boldsymbol{x}}(\cdot)$ features a scale parameter σ. Here, the two equations:

$$y_i^* = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$$
$$\sigma y_i^* = \sigma \left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i \right)$$

are observationally equivalent: $F_{\boldsymbol{x}} \left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \right) = F_{\boldsymbol{x}} \left(\sigma \cdot \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \right)$.

Intuitively, one can only observe whether Y_i^* lies above ($Y_i = 1$) or below ($Y_i = 0$) the implied threshold, and not its variation as a function of the variation of $\boldsymbol{x}_i$. For a similar reason σ *could* be identified *if* both $\alpha_0 = 0$ *and* $\boldsymbol{x}_i$ do not include a constant term.

MLE for Binary Outcome Models (5/8)

In light of all these considerations it is easier to characterize the MLE problem for binary outcome models.

Given a symmetric probability distribution $F_{\boldsymbol{x}}(\cdot)$ and a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, the likelihood function is expressed as:

$$\mathcal{L}\left(\boldsymbol{\beta} | \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \prod_{i=1}^N \left[F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{1-y_i}$$

which generalizes the likelihood function for a Bernoulli sample.

The corresponding log-likelihood function is as follows.

$$\begin{aligned}
&\log \mathcal{L}\left(\boldsymbol{\beta} | \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \\
&\qquad = \sum_{i=1}^N \left\{ y_i \log F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + (1 - y_i) \log\left[1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right] \right\}
\end{aligned}$$

MLE for Binary Outcome Models (6/8)

The First Order Conditions are:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}\left(\widehat{\boldsymbol{\beta}}_{MLE} \middle| \{(y_i, \mathbf{x}_i)\}_{i=1}^{N}\right) = \sum_{i=1}^{N} \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{MLE}\right) =$$

$$= \sum_{i=1}^{N} \left[\frac{y_i f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)}{F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)} - \frac{(1-y_i) f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)}{1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{MLE}\right)} \right] \mathbf{x}_i = \mathbf{0}$$

where $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)$ is the probability **density function** associated with the – implicitly continuous – distribution $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)$.

- Like in the Poisson model, there is no closed form solution and the estimator must be calculated numerically.
- The variance-covariance is more conveniently estimated via the outer product of the gradients.

MLE for Binary Outcome Models (7/8)

The most common choices for $F_{\boldsymbol{x}}(\cdot)$ are:

- the **probit** model, where $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \Phi\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$ and $\Phi(\cdot)$ is a **cumulative standard normal distribution**:

$$\mathbb{P}(Y_i = 1|\,\mathbf{x}_i) = \Phi\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \int_{-\infty}^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{t^2}{2}\right) dt$$

- the **logit** model, where $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \Lambda\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$ and $\Lambda(\cdot)$ is a **scale-normalized cumulative logistic distribution**.

$$\mathbb{P}(Y_i = 1|\,\mathbf{x}_i) = \Lambda\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \frac{\exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)}{1 + \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)}$$

The two models typically yield quite similar results; of the two, the logit is often more convenient to manipulate and compute.

MLE for Binary Outcome Models (8/8)

- How to interpret the estimates $\widehat{\boldsymbol{\beta}}_{MLE}$ in a binary outcome model? **Not** in terms of causal effects, or probabilities.

- They must interpreted in terms of **marginal effects**:

$$\frac{\partial \mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right)}{\partial \mathbf{x}_i} = \frac{\partial \mathbb{E}\left[Y_i \middle| \mathbf{x}_i\right]}{\partial \mathbf{x}_i} = \frac{\partial F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)}{\partial \mathbf{x}_i} = f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}$$

  which, however, vary as a function of the realizations $\mathbf{x}_i$.

- There are two **asymptotically equivalent** ways to derive marginal effects that aid the interpretation of the estimates.

- In one approach one evaluates $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}_{MLE}$ and at $\mathbf{x} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$, the average values of $\mathbf{x}_i$ in the data.

- In the other, one evaluates $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}_{MLE}$ and at $\mathbf{x}_i$ for $i = 1, \ldots, N$; the resulting quantities are averaged out.

Introduction to Simulated Maximum Estimation

- Sometimes, the **numerical** evaluation of the M-Estimation criterion function $q(\boldsymbol{x}_i; \boldsymbol{\theta})$ is so **complicated** that practical applications of the theory discussed thus far is unfeasible.

- In such cases, the typical solutions are estimators based on **simulation methods** that help circumvent the problem.

- The leading techniques based on simulations lie in the MLE domain, and are especially applied in some of LDV models.

- In econometrics this applies *especially but not exclusively* to models of Industrial Organization, and related ones.

- For convenience, the overview that follows next starts from simulation-based estimators that extend MLE.

- It then moves to general simulation-based M-Estimators.

Unsolvable criterion functions

- Suppose that the probability mass or density function of all observable variables $\boldsymbol{x}_i$ is expressed given the parameters $\boldsymbol{\theta}$ as follows.

$$f_{\boldsymbol{x}}\left(\mathbf{x}_i \middle| \boldsymbol{\theta}\right) = \int_{\mathbb{U}} f_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i \middle| \mathbf{u}_i; \boldsymbol{\theta}\right) dH_{\boldsymbol{u}}\left(\mathbf{u}_i\right)$$

- Here, $\boldsymbol{u}_i$ is a **random vector** with cumulative distribution $H_{\boldsymbol{u}}\left(\mathbf{u}_i\right)$. In the above expression, $\boldsymbol{u}_i$ is integrated out over its support $\mathbb{U}$.

- Consider the case where there is **no closed form solution** for the above integral, even if by itself the conditional mass or density function $f_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i \middle| \mathbf{u}_i; \boldsymbol{\theta}\right)$ is tractable.

- It is obvious that the MLE problem associated to $f_{\boldsymbol{x}}\left(\mathbf{x}_i \middle| \boldsymbol{\theta}\right)$ **cannot be solved**, at least not easily.

Example: Random coefficients logit (1/2)

- Consider the following **bivariate** logit model:

$$\mathbb{P}\left[Y_i = 1 \middle| X_i\right] = \Lambda\left(\beta_0 + \beta_{1i} X_i\right)$$

  a LDV with one *binary* dependent variable $Y_i \in \{0, 1\}$ and one *possibly* continuous independent variable X_i.

- Note that the parameter β_{1i} is **observation specific**.

- As there are only N observations, one cannot estimate **all** the N parameters β_{1i} for $i = 1, \ldots, N$ *and* β_0.

- Yet there are actual settings where it is important to assess **heterogeneity** in the individual response of Y_i to X_i.

- These settings are framed as **random coefficients** models – as the logit model expressed above; they are more difficult to handle in LDV cases than in, say, linear environments.

Example: Random coefficients logit (2/2)

- One could make some **distributional assumptions** about the individual coefficients, such as normality.

$$\beta_{1i} \sim \mathcal{N}\left(\beta_1, \sigma^2\right)$$

- The parameter set of interest is thus $\theta = (\beta_0, \beta_1, \sigma^2)$.

- Define $u_i \equiv (\beta_{1i} - \beta_1)/\sigma$ and recall that $\phi(\cdot)$ is the density function of the standard normal distribution.

- Under these hypotheses the conditional mass function of Y_i is expressed as an integral without closed form solution.

$$\begin{aligned} f_{Y_i|X_i}\left(y_i|\, x_i; \beta_0, \beta_1, \sigma^2\right) = \\ = \int_{\mathbb{R}} \Lambda\left[\beta_0 + (\beta_1 + \sigma u_i)\, x_i\right]^{y_i} \cdot \\ \cdot \left\{1 - \Lambda\left[\beta_0 + (\beta_1 + \sigma u_i)\, x_i\right]\right\}^{1-y_i} \phi(u_i)\, du_i \end{aligned}$$

Simulation-based likelihood evaluations

- How to solve this problem and similar ones? A *brute force* approach to the integral is generally impractical.

- Common approaches are based upon the **simulation** of the likelihood components $f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$.

- The **Direct Monte Carlo Sampling** method is based on a sample $\{\mathbf{u}_s\}_{s=1}^{S}$ of S random draws of $\boldsymbol{u}_i$ from $H_{\boldsymbol{u}}\left(\mathbf{u}_i\right)$ – these are used to calculate a *Monte Carlo estimate*:

$$\widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)=\frac{1}{S}\sum_{s=1}^{S}\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}\right)$$

  where function $\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}\right)$ is called a **subsimulator**.

- The subsimulator is typically an unbiased predictor:

$$\mathbb{E}\left[\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}\right)\right]=f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$$

  where the expectation is taken over the support of $\boldsymbol{u}_i$.

Maximum Simulated Likelihood

- With an unbiased subsimulator, by standard asymptotics:

$$\widehat{f}_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right) \xrightarrow{p} f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$$

  that is, the simulator is **consistent** insofar as $S \to \infty$.

- Therefore the **Maximum Simulated Likelihood** (MSL) can be easily motivated and constructed as follows.

$$\widehat{\boldsymbol{\theta}}_{MSL} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \log \widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$$

- MSL simulators need to be differentiable with respect to $\boldsymbol{\theta}$. **Example:** this is verifiable in the random coefficient logit.

$$\widetilde{f}_{Y_i,X_i|U_s}\left(y_i, x_i|\,u_s; \beta_0, \beta_1, \sigma^2\right) =$$
$$= \Lambda\left[\beta_0 + \left(\beta_1 + \sigma u_s\right) x_i\right]^{y_i} \left\{1 - \Lambda\left[\beta_0 + \left(\beta_1 + \sigma u_s\right) x_i\right]\right\}^{1-y_i}$$

- Elements $\widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$ are conveniently calculated using the same draw $\{\mathbf{u}_s\}_{s=1}^{S}$ for $i = 1, \ldots, N$.

Maximum Simulated Likelihood: Asymptotics

Theorem 4

Asymptotic Efficiency of Maximum Simulated Likelihood. *Suppose that the mass or density function* $f_{\boldsymbol{x}}(\mathbf{x}_i|\,\theta)$ *that describes the model's data generation process meets the requirements of Theorem 18 of Lecture 6, hence the corresponding "theoretical" MLE has a limiting distribution as per the statement of that Theorem. A SML estimator based on an unbiased subsimulator is asymptotically equivalent to the "theoretical" MLE and it has the same limiting distribution:*

$$\sqrt{N}\left(\widehat{\theta}_{SML}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\left[\boldsymbol{I}\left(\theta_0\right)\right]^{-1}\right)$$

if $S,N\to\infty$ *(which is sufficient for consistency) and* $\sqrt{N}/S\to 0$.

Proof.

(Outline.) Gourieroux and Monfort (1991) develop the standard Taylor expansion of the First Order Conditions and elaborate how it depends on two sources of noise: the one coming from the data $\{\mathbf{x}_i\}_{i=1}^N$ and the one due to the simulation draws $\{\mathbf{u}_s\}_{s=1}^S$. The latter vanishes asymptotically if S grows at a rate higher than that of N. □

Asymptotic bias-corrected MSL

- This result, while powerful, requires S to be very large, or else MSL is inconsistent: the problem arieses as the log of the subsimulator is *not* unbiased, even if the level is.

- A second-order Taylor expansion of $\log \widehat{f}_{\boldsymbol{x}|S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$ around $\log f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$ suggested by Gourieroux and Monfort (1991):

$$\log f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right) \simeq \mathbb{E}\left[\log \widehat{f}_{\boldsymbol{x}|S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right] + \frac{\mathbb{V}\mathrm{ar}\left[\left(\widehat{f}_{\boldsymbol{x}|S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right) - f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right)^2\right]}{2\left[f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right]^2}$$

- . . . motivates the **asymptotic bias-corrected MSL**.

$$\widehat{\boldsymbol{\theta}}_{BCMSL} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N} \left[\log \widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right) + \sum_{s=1}^{S} \frac{\left[\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}\right) - \widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right]^2}{2S\left[\widehat{f}_{\boldsymbol{x},S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right]^2} \right]$$

Simulated Maximum Estimation

- The analysis further extends to all M-Estimators where the criterion $q(\boldsymbol{x}_i; \boldsymbol{\theta})$ cannot be expressed in closed form.

- A **Simulated M-Estimator** (SM), of which MSL is but a special case, is defined as:

$$\widehat{\boldsymbol{\theta}}_{SM} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \widehat{q}_S(\mathbf{x}_i; \boldsymbol{\theta})$$

  and $\widehat{q}_S(\mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} \widetilde{q}_s(\mathbf{x}_i, \mathbf{u}_s; \boldsymbol{\theta})$ is typically an average of subsimulators that are written as $\widetilde{q}_s(\mathbf{x}_i, \mathbf{u}_s; \boldsymbol{\theta})$; these are based upon pseudo-random draws of $\boldsymbol{u}_i$ like in SML case.

- The SM estimator is **consistent** if $S, N \to \infty$; furthermore it is as efficient as the paired non-simulated M-Estimator if $\sqrt{N}/S \to 0$; if S is too small, like in SML a bias correction may be necessary.

Inference for Simulated Maximum Estimators

- In SM, a consistent estimator for $\boldsymbol{\Upsilon}_0$ – if the observations are independent (extensions to CCE and HAC exist) – is:

$$\widehat{\boldsymbol{\Upsilon}}_{M,S} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{s}}_{Si}\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M\right) \widehat{\mathbf{s}}_{Si}^{\mathrm{T}}\left(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

  where $\widehat{\mathbf{s}}_{Si}\left(\mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{\partial \widehat{q_S}(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

- That for $\mathbf{Q}_0$ is based on $\widehat{\mathbf{H}}_{Si}\left(\mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{\widehat{\mathbf{s}}_{Si}(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{\partial \widehat{q_S}(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}$.

$$\widehat{\mathbf{Q}}_N \equiv \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{H}}_{Si}\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \mathbf{Q}_0$$

- With i.i.d. data, the MSL information matrix is estimated as $\widehat{\boldsymbol{\Upsilon}}_{M,S} \xrightarrow{p} \left[\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)\right]^{-1}$, and the score specializes as follows.

$$\widehat{\mathbf{s}}_{Si}\left(\mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{\sum_{s=1}^{S} \frac{\partial \widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\mathbf{u}_s;\widehat{\boldsymbol{\theta}}_{SML}\right)}{\partial \boldsymbol{\theta}}}{\sum_{s=1}^{S} \widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s; \widehat{\boldsymbol{\theta}}_{SML}\right)}$$

Applications of Maximum Estimation: overview

This Lecture is concluded by showcasing three applications of M-Estimation in actual econometric problems.

The following applications are covered.

1. The Constant Elasticity of Substitution (CES) production functions, which naturally call for the use of NLLS.

2. Sample selection models: peculiar binary outcome models with a clear structural, economic interpretation.

3. The Bresnahan model for detecting collusion in a market: a classical application of MLE in Industrial Organization.

CES production functions and NLLS (1/2)

- **Constant Elasticity of Substitution** (CES) production functions are key ingredients of many economic models.

- In its simplest form a CES production function writes as:

$$Y_i = \left[\alpha_K K_i + \alpha_L L_i\right]^{\frac{1}{\rho}} + \varepsilon_i$$

  where Y_i, K_i, L_i are output, capital, labor; while $\alpha_K > 0$ and $\alpha_L > 0$ are the **saliency** parameters that determine the relative importance of each input. Furthermore, ε_i is an econometric error term.

- Instead, $\rho > 0$ is a parameter related to the **elasticity of substitution** between inputs, which is clearly a constant and writes as $\sigma = (1+\rho)^{-1} \in (0,1)$.

- The CES production function becomes a Cobb-Douglas as in Lecture 7, with $\alpha_K = \beta_K$ and $\alpha_L = \beta_L$, for $\rho \to 0$.

CES production functions and NLLS (2/2)

- Clearly, this model must be estimated via NLLS. No closed form solution exists, and numerical methods must be used.

- A typical estimation algorithm splits the problem as:

$$(\widehat{\rho}, \widehat{\alpha}_K, \widehat{\alpha}_L)_{NLLS} \in$$
$$\underset{\rho \in \mathbb{R}_{++}}{\arg\min} \left[\underset{(\alpha_K, \alpha_L) \in \mathbb{R}^2_{++}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \left[\alpha_K k_i + \alpha_L l_i \right]^{\frac{1}{\rho}} \right)^2 \right]$$

  where (y_i, k_i, l_i) denote observations of (Y_i, K_i, L_i).

- Hence, numerical algorithms feature an *inner maximizer* of (α_K, α_L) given a value of ρ, and an *outer maximizer* for ρ.

- Unfortunately, the solution is typically unstable: especially with more inputs. Hence, practitioners usually prefer linear approximations or alternative production functions.

The Heckit Sample Selection Model (1/4)

- Labor economists are interested in **labor supply** choices, especially those by women (who traditionally participated less than men in the labor market).

- Consider a labor supply **intensity** equation such as:

$$h_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$$

  where h_i is the amount of time spent by woman i in wage labor, and $\mathbf{x}_i$ are her characteristics.

- However, h_i is only observed for women who *do work*:

$$z_i^* = \mathbf{w}_i^{\mathrm{T}} \boldsymbol{\gamma} + \upsilon_i$$

$$h_i \begin{cases} > 0 & \text{if } z_i^* > 0 \\ = 0 & \text{if } z_i^* \leq 0 \end{cases}$$

  as governed by some latent variable z_i^*, a function of some set of characteristics $\mathbf{w}_i$, with a structural interpretation.

The Heckit Sample Selection Model (2/4)

- Although the binary outcome model for **participation** can be estimated (say via MLE) interest falls on the **intensity** equation for h_i, or other labor market outcomes like wages.

- OLS cannot estimate the intensity equation consistently: if H_i is the random variable whence h_i is drawn, it is:

$$\begin{aligned}\mathbb{E}\left[H_i \middle| \mathbf{x}_i, h_i > 0\right] &= \mathbb{E}\left[H_i \middle| \mathbf{x}_i, z_i^* > 0\right] \\ &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_i \middle| \mathbf{x}_i, z_i^* > 0\right] \\ &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_i \middle| \mathbf{x}_i, \upsilon_i > -\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma}\right]\end{aligned}$$

  and $\lambda\left(\mathbf{w}_i\right) \equiv \mathbb{E}\left[\varepsilon_i \middle| \upsilon_i > -\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\gamma}\right] \neq 0$ if the two error terms ε_i and υ_i are correlated (which is likely).

- The **omitted variable** $\lambda\left(\mathbf{w}_i\right)$ thus leads to inconsistency. For example, wealthy women may be **both** inclined not to participate, and to participate with low h_i if they do.

The Heckit Sample Selection Model (3/4)

- Heckman (1979) devised a solution to this problem, which awarded him the Nobel prize in economics.

- His **"heckit"** model posits a parametric assumption about the two error terms $(\varepsilon_i, \upsilon_i)$, like the bivariate normal.

$$\begin{pmatrix} \varepsilon_i \\ \upsilon_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

  where ρ is the correlation coefficient between ε_i and υ_i, σ^2 is the variance of ε_i, while the variance of υ_i is *normalized* to 1 like in probit and logit.

- The resulting model could be estimated jointly by MLE.

- Yet the full-fledged likelihood function may be complicated to handle. An alternative two-step procedure, which is also consistent, is more popular.

The Heckit Sample Selection Model (4/4)

The two-step procedure is as follows.

1. Run a probit on the participation equation; obtain $\widehat{\boldsymbol{\gamma}}_{MLE}$.

2. For each observation, calculate the **inverse Mills ratio**:

$$\lambda_i = \left[\frac{\phi\left(\mathbf{w}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{MLE}\right)}{\Phi\left(\mathbf{w}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{MLE}\right)} \right]$$

   where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the p.d.f. and c.d.f. of the standard normal distribution.

3. Run OLS on a *modified* intensity equation:

$$h_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \rho \lambda_i + \varepsilon_i$$

   where ρ is now also the OLS coefficient for λ_i.

While consistent, this procedure delivers higher standard errors.

Detecting collusion in oligopolies (1/6)

- Industrial Organization is currently the "structural" field of economics *par excellance*.
- Decades ago however it saw little empirical work. Classical questions, like the one regarding the sudden 45% increase in the US automobile production and sales in 1955 (that some attributed to a temporary price war), were left unanswered.
- Paul Samuelson himself had allegedly once said that he... *would flunk any econometrics paper that claimed to provide an explanation of 1955 auto sales.*
- Yet Bresnahan (1987) defied Samuelson as he developed an original model estimated via MLE that can be used to *test for collusion*. The results suggest that a price war occurred.
- Although this model is now obsolete, it is still a classic and has great instruction value.

Detecting collusion in oligopolies (2/6)

- Consider N types of cars each with **quality** $X_i = X(\boldsymbol{z}_i, \boldsymbol{\beta})$ being a function of one car's characteristics $\boldsymbol{z}_i$ (given some parameters $\boldsymbol{\beta}$). Qualities can be ordered from best to worst – without loss of generality, $X_i > X_h$ if $i > h$.

- Also consider some well-microfounded **demand functions** of each car for each year $t = 1, \ldots, T$:

$$Q_{it}^D = D(P_{ht}, P_{it}, P_{jt}, X_{ht}, X_{it}, X_{jt}, \boldsymbol{\gamma})$$

  where Q_{it} is the **quantity** of product i, P_{it} its **price**, h, i, j are three **consecutive** products in the **order** of "qualities" and $\boldsymbol{\gamma}$ are some parameters.

- This specification makes prices and quantities dependent in equilibrium only on those of the "neighbors" of one product in the product space – it results from a unique specification of consumers' utility as given by Bresnahan.

Detecting collusion in oligopolies (3/6)

Supply is standard: the profits from the sale of product i are:

$$\pi_{it} = P_{it}Q_{it} - c\left(X_{it}\right)Q_{it}$$

with $c\left(X_{it}\right) = \mu \exp\left(X_{it}\right)$. Bresnahan analyzes two scenarios.

1. **Competition**: each firm sets its own price P_{it} taking the price of neighbors h and j as given, with FOCs:

$$\frac{\partial \pi_{it}}{\partial P_{it}} = Q_{it} + \left(P_{it} - c\left(X_{it}\right)\right)\frac{\partial Q_{it}\left(\cdot\right)}{\partial P_{it}} = 0$$

   where Q_{it} is a function of P_{it} as in the demand function.

2. **Cooperation**: the firm(s) selling two products i and j set prices P_{it} and P_{jt} so as to maximize the joint profits, with FOCs for P_{it} price as follows (and symmetrically for P_{jt}).

$$\frac{\partial\left[\pi_{it} + \pi_{jt}\right]}{\partial P_{it}} = Q_{it} + \left(P_{it} - c\left(X_{it}\right)\right)\frac{\partial Q_{it}\left(\cdot\right)}{\partial P_{it}} + \left(P_{jt} - c\left(X_{jt}\right)\right)\frac{\partial Q_{jt}\left(\cdot\right)}{\partial P_{it}} = 0$$

Detecting collusion in oligopolies (4/6)

- Bresnahan then defines several **matrices $\mathbf{H}_t$** such that, in each year,

$$h_{(ij)t} = \begin{cases} 1 & \text{cooperation between products } i \text{ and } j \\ 0 & \text{competition between products } i \text{ and } j \end{cases}$$

  and he characterizes several hypothetical scenarios for 1955 and surrounding years with corresponding matrices $\mathbf{H}_t$.

- Thus, for a **given choice** of matrix $\mathbf{H}_t$ the supply function can be written as

$$q_{it}^S = S\left(P_{ht}, P_{it}, P_{jt}, X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\gamma}, \mu\right)$$

  where the demand parameters $\boldsymbol{\gamma}$ enter via the derivative of the demand functions implied in the supply FOCs.

- This sets up alternative **counterfactuals** about collusion.

Detecting collusion in oligopolies (5/6)

- By setting the equilibrium condition $Q_{it}^D = Q_{it}^S = Q_{it}^*$ (and similarly for prices) for each product $i = 1, \ldots, N$ in every year $t = 1, \ldots, T$, the **reduced form** is derived.

$$P_{it} = P^* \left(X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mu \right)$$
$$Q_{it} = Q^* \left(X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mu \right)$$

- This requires to develop the full-fledged solution given the assumptions about demand and supply, and a matrix $\mathbf{H}_t$.

- Introduce some error terms that make the model stochastic, and endow them of distributional assumptions as follows.

$$\begin{pmatrix} P_{it} - P^* \\ Q_{it} - Q^* \end{pmatrix} = \begin{pmatrix} \xi_{it}^P \\ \xi_{it}^Q \end{pmatrix} = \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_P^2 & 0 \\ 0 & \sigma_Q^2 \end{pmatrix} \right)$$

- All these error terms are implicit functions of the **reduced form** "predictions" of the model.

Detecting collusion in oligopolies (6/6)

- For any **given** matrix $\mathbf{H}_t$, the model is estimated via MLE using the following likelihood function.

$$\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mu | \, \mathbf{H}_t, \{p_{it}, q_{it}, \mathbf{z}_i\}_{i=1}^{N}\right) =$$

$$= \prod_{t=1}^{\mathrm{T}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_P^2}} \exp\left(-\frac{\left(\xi_{it}^P\right)^2}{2\sigma_P^2}\right)$$

$$\times \prod_{t=1}^{\mathrm{T}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_Q^2}} \exp\left(-\frac{\left(\xi_{it}^Q\right)^2}{2\sigma_Q^2}\right)$$

- Tests about alternative matrices $\mathbf{H}_{t0}$ and $\mathbf{H}_{t1}$ (say, one for "competition" and one for "collusion") are performed with the likelihood ratio method, as follows.

$$\mathrm{C}_{\mathbf{H}} = 2\Big[\log\mathcal{L}\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\mu}\Big| \, \mathbf{H}_{t1}, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)$$

$$- \log\mathcal{L}\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\mu}\Big| \, \mathbf{H}_{t0}, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)\Big]$$