Nonlinear estimation and M-estimation

- Nonlinear estimation describes any problem in which estimators cannot be obtained in closed form

- M-estimators (M for maximisation) is a vary wide class of estimators. It includes

  – maximum likelihood
  – nonlinear least squares
  – least absolute deviation (LAD)
  – quasi-maximum likelihood
  – and many others

Nonlinear Least Squares (NLS)

- Nonlinear regression function. For some $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$

$$\mathrm{E}\left(y|\,\boldsymbol{x}\right) = m\left(\boldsymbol{x}, \boldsymbol{\theta}_o\right)$$

  where $m\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$ is a known function and $\boldsymbol{\theta}_o$ is a finite parameter vector

- Function $m\left(\boldsymbol{x}, \boldsymbol{\theta}_o\right)$ is describing the nonlinear depenence between the conditional expected values of the dependent variable and explanatory variables

Example 1. *Sometimes the linear model is not well suited to the data we have. If* $y > 0$ *(e.g. incomes) or* $y \in [0, 1]$ *(e.g. percentage of unempleyed) the*

linear model is problematic as for $\mathrm{E}\left(y|\,\boldsymbol{x}\right)=\boldsymbol{x}\boldsymbol{\beta}$ *the values of* $y\in(-\infty,\infty)$ *for* $\boldsymbol{x}\in R^{K}$*.*

For $y>0$ *exponential regression function can be used* $m\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\exp\left(\boldsymbol{x}\boldsymbol{\theta}\right)$*. For* $y\in[0,1]$ *the logistic model* $m\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\frac{\exp(\boldsymbol{x}\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}\boldsymbol{\theta})}$ *is often used.*

- We will denote true parameter vector as $\boldsymbol{\theta}_o$
- The model we consider is nonlinear regression model in error form

$$y=m\left(\boldsymbol{x},\boldsymbol{\theta}_o\right)+u$$

  and $\mathrm{E}\left(u|\,\boldsymbol{x}\right)=0$

- Let $\boldsymbol{w}_i\equiv(\boldsymbol{x}_i,y_i)$ be the vector of observation drawn from random sample

- Our estimation is based on the fact that $\boldsymbol{\theta}_o$ minimizes the squared error in the population

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \mathrm{E}\left[y - m\left(\boldsymbol{x}, \boldsymbol{\theta}_o \right) \right]^2 \right\}$$

- The estimation is based on the analogy principle we calculate estimator of $\boldsymbol{\theta}_o$ by minimizing the squared error in the sample

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \left[y_i - m\left(\boldsymbol{x}, \boldsymbol{\theta} \right) \right]^2$$

General case

- The M-estimator $\widehat{\boldsymbol{\theta}}$ solves the

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N} q\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)$$

  for some function $q\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)$

- The parameter vector $\boldsymbol{\theta}_o$ is assumed to solve uniquely the population problem (is identified)

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}\left[q\left(\boldsymbol{w},\boldsymbol{\theta}\right)\right]$$

- Because of the law of large numbers

$$N^{-1}\sum_{i=1}^{N} q\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right) \xrightarrow{p} \mathrm{E}\left[q\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]$$

and as $\widehat{\boldsymbol{\theta}}$ is minimizing the term on left and $\boldsymbol{\theta}_o$ the term on the right it seams plausible that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$

Identification

- The identification in the context of M-estimation means that $\boldsymbol{\theta}_o$ is a unique solution to the population problem

- For NLS it means that

$$\mathrm{E}\left\{\left[m\left(\boldsymbol{x},\boldsymbol{\theta}_o\right)-m\left(\boldsymbol{x},\boldsymbol{\theta}\right)\right]\right\}>0, \text{ for all } \boldsymbol{\theta}\in\boldsymbol{\Theta}, \boldsymbol{\theta}\neq\boldsymbol{\theta}_o$$

Example 2. *In regression function* $m\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\boldsymbol{\theta}_1+\boldsymbol{\theta}_2x_2+\boldsymbol{\theta}_3x_3^{\boldsymbol{\theta}_4}$ *this condition fails if* $\boldsymbol{\theta}_3=0$ *(poorly identified model)*

Asymptotic normality of M-estimators

- The estimator $\widehat{\boldsymbol{\theta}}$ minimizes the expected value of $q\left(\boldsymbol{w}, \boldsymbol{\theta}\right)$
- Then if $q\left(\boldsymbol{w}, \boldsymbol{\theta}\right)$ is continuously differentiable w.r.t. $\boldsymbol{\theta}$ than the first order derivative of $q\left(\boldsymbol{w}, \boldsymbol{\theta}\right)$ should be equal to zero at $\widehat{\boldsymbol{\theta}}$, as $\widehat{\boldsymbol{\theta}}$ solves the maximisation problem
- Denoted this first order derivative as $\boldsymbol{s}_i = \boldsymbol{s}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) = \frac{\partial q(\boldsymbol{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Then

$$\sum_{i-1}^{N} \boldsymbol{s}\left(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}\right) = 0$$

- Function $\boldsymbol{s}\left(\boldsymbol{w}, \boldsymbol{\theta}\right)$ is called score of the objective function

- Define $\ddot{\boldsymbol{H}}_i = \frac{\partial^2 q(\boldsymbol{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ as Hessian of the objective function for observation i

- Assume that $\ddot{\boldsymbol{H}}_i$ fulfills the conditions of the law of large numbers and so $\frac{1}{N} \sum_{i-1}^{N} \ddot{\boldsymbol{H}}_i \xrightarrow{p} \mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right)\right] = \boldsymbol{A}_o$ (and assume that $\boldsymbol{A}_o$ is nonsingular)

- Assume that $\boldsymbol{s}\left(\boldsymbol{w}_i, \boldsymbol{\theta}_o\right)$ satisfies the conditions of central limit theorem as a $i.i.d.$ random variable with mean zero and then $N^{-\frac{1}{2}} \sum_{i-1}^{N} \boldsymbol{s}\left(\boldsymbol{w}_i, \boldsymbol{\theta}_o\right) \xrightarrow{D} N\left(0, \boldsymbol{B}_o\right)$ where $\boldsymbol{B}_o = \mathrm{E}\left[\boldsymbol{s}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right) \boldsymbol{s}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right)'\right]$

- Under these conditions

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) \xrightarrow{D} N\left(0, \boldsymbol{A}_o^{-1} \boldsymbol{B}_o \boldsymbol{A}_o^{-1}\right)$$

- The asymptotic variance can then be estimated as

$$Avar\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{N}\boldsymbol{A}_o^{-1}\boldsymbol{B}_o\boldsymbol{A}_o^{-1}$$

Estimating the variance covariance matrix

- In practice it difficult to calculate directly $\boldsymbol{B}_o = \mathrm{E}\left[\boldsymbol{s}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right) \boldsymbol{s}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right)'\right]$ and $\boldsymbol{A}_o = \mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right)\right]$.

- We can however estimate $\boldsymbol{H}\left(\boldsymbol{w}, \boldsymbol{\theta}_o\right)$ using the Law of Large Numbers:

$$\widehat{\boldsymbol{A}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{H}\left(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}\right) = N^{-1} \sum_{i-1}^{N} \widehat{\boldsymbol{H}}_i \xrightarrow{p} \boldsymbol{A}_o$$

- Drawbacks of this estimator:
  - need to derive the second order derivatives
  - this estimator is not always positive definite

- In most economic application the inference is conditional on $\boldsymbol{x}_i$ then we use the conditional variance matrix

$$\boldsymbol{A}_i = \boldsymbol{A}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_o\right) = \mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) \middle| \boldsymbol{x}_i\right]$$

- But by the law of iterated expectations $\mathrm{E}\left[\boldsymbol{A}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_o\right)\right] = \mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] = \boldsymbol{A}_o$

- Again

$$\widehat{\boldsymbol{A}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{A}\left(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\right) = N^{-1} \sum_{i-1}^{N} \widehat{\boldsymbol{A}}_i \xrightarrow{p} \boldsymbol{A}_o$$

- This estimator is useful if $\mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) \middle| \boldsymbol{x}_i\right]$ can be obtained. In most cases it is possible to prove that this estimator is positive definite.

- We obtain the consistent estimator of $\boldsymbol{B}_o$ in the similar way

$$\widehat{\boldsymbol{B}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{s}\left(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}\right) \boldsymbol{s}\left(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}\right)' = N^{-1} \sum_{i-1}^{N} \widehat{\boldsymbol{s}}_i \widehat{\boldsymbol{s}}_i{}' \xrightarrow{p} \boldsymbol{B}_o$$

- Then the consistent estimator of the variance covariance matrix of $\sqrt{N}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right)$ has the form

$$\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{A}}^{-1} \widehat{\boldsymbol{B}} \widehat{\boldsymbol{A}}^{-1}$$

- And the estimate of the variance of $\widehat{\boldsymbol{\theta}}$

$$\widehat{Var}\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\boldsymbol{A}}^{-1} \widehat{\boldsymbol{B}} \widehat{\boldsymbol{A}}^{-1} \Big/ N$$

- In many contexts it is possible to simplify this formula under additional assumption
$$\mathrm{E}\left[\boldsymbol{s}\left(\boldsymbol{w},\boldsymbol{\theta}\right)\boldsymbol{s}\left(\boldsymbol{w},\boldsymbol{\theta}\right)'\right]=\boldsymbol{\sigma}_o^2\,\mathrm{E}\left[\boldsymbol{H}\left(\boldsymbol{w},\boldsymbol{\theta}_o\right)\right]$$
or simply $\boldsymbol{B}_o=\boldsymbol{\sigma}^2\boldsymbol{A}_o$

- This assumption if called generalized information matrix equality (GIME)

- Under this assumption the variance estimator has the form
$$\widehat{\boldsymbol{V}}=\widehat{\boldsymbol{\sigma}}^2\widehat{\boldsymbol{A}}^{-1}$$

Hypothesis Testing

- We have three test for testing nonlinear hypothesis $H_0 : \boldsymbol{c}(\boldsymbol{\theta}_o) = \mathbf{0}$
  - Wald
  - Score (Lagrange Multiplier test)
  - Tests based on the change of the objective function (Quasi Likelihood Ratio test)

- $\boldsymbol{C}(\boldsymbol{\theta}_0) = \frac{\partial \boldsymbol{c}(\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'}$ must have a full rank (models not identified under the H_0 excluded)

Wald test

- Wald test is of the form

$$W = \boldsymbol{c}\left(\widehat{\boldsymbol{\theta}}\right)' \left(\widehat{\boldsymbol{C}}'\widehat{\boldsymbol{V}}\widehat{\boldsymbol{C}}\right)^{-1} \boldsymbol{c}\left(\widehat{\boldsymbol{\theta}}\right) \xrightarrow{d} \boldsymbol{\chi}^2_Q$$

  where g is the number of restrictions, $\widehat{\boldsymbol{V}}$ is asymptotic variance of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{C}}$ is the Jacobian matrix of $\boldsymbol{c}\left(\widehat{\boldsymbol{\theta}}\right)$ (matrix of first order derivatives)

- The idea of the test is that we reject H_0 if $\boldsymbol{c}\left(\widehat{\boldsymbol{\theta}}\right)$ differs significantly from zero

- Most important advantage of Wald test: only requires calculation of the estimate $\widehat{\boldsymbol{\theta}}$ for *unrestricted* model

- Practical limitations of this test:
  - $\widehat{\boldsymbol{\theta}}_0$ must be in the interior of $\boldsymbol{\Theta}$ - not on the boundary
  - It is not invariant to how the nonlinear restrictions are imposed (in small samples)

Score test

- We impose the restriction directly on the model and estimate so called *restricted* model

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum q\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) \text{ s.t. } \boldsymbol{c}\left(\boldsymbol{\theta}\right) = 0$$

- It is possible to prove that

$$LM = \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i\right)' \widetilde{\boldsymbol{A}}^{-1} \widetilde{\boldsymbol{C}}' \left(\widetilde{\boldsymbol{C}} \widetilde{\boldsymbol{A}}^{-1} \widetilde{\boldsymbol{B}}^{-1} \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{C}}'\right)^{-1} \widetilde{\boldsymbol{C}} \widetilde{\boldsymbol{A}}^{-1} \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i\right) \Big/ N \xrightarrow{d} \boldsymbol{\chi}^2_Q$$

  where all the matrices are calculated for restricted estimates of $\boldsymbol{\theta}$

- This is the heteroscedasticity robust form
- If $GIME$ is true it is possible to simplify it to the form

$$LM = \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i \right)' \widetilde{\boldsymbol{M}}^{-1} \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i \right)$$

- Where $\widetilde{\boldsymbol{M}}$ can be estimated as $\sum_{i=1}^{N} \widetilde{\boldsymbol{A}}_i$, $\sum_{i=1}^{N} \widetilde{\boldsymbol{H}}_i$ or $\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i \widetilde{\boldsymbol{s}}_i'$.
- The last estimate results in so called outer product of the score LM statistics which can be calculated as NR_o^2 from the regression of 1 on $\widetilde{\boldsymbol{s}}_i$, where R_o^2 is uncentered R^2.
- Wooldrige gives a similar way how to calculate the hetoroscedasticity robust form of the test for NLS

- Vector $\sum_{i=1}^{N} \widetilde{\boldsymbol{s}}_i$ is called the score vector and is the first order derivative of the objective function calculated at *restricted* estimates

- The idea of the test is that if for the first order derivatives calculated for restricted estimates are far away from zero (f.o.c. for unrestricted model are far from being true), that we tend to reject H_0

- Advantages of LM test:
  - requires only the restricted estimates - ideal for specification testing
  - is invariant to specifications of the H_0
  - in some cases does not require that $\boldsymbol{\theta}$ is in interior of $\boldsymbol{\Theta}$, leading case $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $H_0 : \boldsymbol{\theta}_2 = 0$

Tests based on the change in objective function

- The QLR test has the form

$$QLR = 2\left[\sum_{i-1}^{N} q\left(\boldsymbol{w}_i, \widetilde{\boldsymbol{\theta}}\right) - \sum_{i-1}^{N} q\left(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}\right)\right] \xrightarrow{d} \boldsymbol{\chi}^2_Q$$

  where $\widetilde{\boldsymbol{\theta}}$ is the restricted estimate and $\widehat{\boldsymbol{\theta}}$ is unrestricted estimate

- Intuition of this test is that if imposing restriction causes the big reduction in the objective function at the restricted maximum we tend to reject the H_0

- Advantages of this statistics:

  - if the restricted and unrestricted model is easy to estimate then this statistic is very easy to calculate
  - invariant to the specifications of the H_0

- Disadvantages
  - cannot be made robust to heteroscedasticity

Optimization methods

- For many problem is impossible to find analytically the maximum of the objective function $\sum_{i=1}^{N} q_i\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$
- In this case we find the maximum numerically
- One of the simplest methods the bisection method
- However it is not efficient if $\boldsymbol{\theta}$ has many elements
- In this cases we use so called gradient methods
- One of the most popular is so called Newton-Raphson method

Maximum likelihood estimation

- In most economic application we specify conditional probability given $\boldsymbol{x}_i$ (we are not interested in distribution of exogenous variables)

- Then we use conditional maximum likelihood estimation $CMLE$

- We assume that the density is unknown up to the finite number of parameters (parametric estimation)

Example 3. *(Wooldridge) Suppose that latent variable follows*

$$y_i^* = \boldsymbol{x}_i\, \boldsymbol{\theta} + e_i$$

and $e_i \sim N(0,1)$. *We observe*

$$y_i = \left\{ \begin{array}{ll} 1 & \textit{if} \quad y_i^* > 0 \\ 0 & \textit{if} \quad y_i^* \leq 0 \end{array} \right|$$

Probability

$$\begin{aligned} \Pr(y_i = 1|\, \boldsymbol{x}_i) &= \Pr(\boldsymbol{x}_i\, \boldsymbol{\theta} + e_i > 0) = \Pr(e_i > -\boldsymbol{x}_i \boldsymbol{\theta}) \\ &= 1 - \Phi(-\boldsymbol{x}_i \boldsymbol{\theta}) = \Phi(\boldsymbol{x}_i \boldsymbol{\theta}) \end{aligned}$$

The probability of $\Pr(y_i = 0|\, \boldsymbol{x}_i) = 1 - \Phi(\boldsymbol{x}_i \boldsymbol{\theta})$. *The conditional density for an observation* i *can written as*

$$f(y_i|\, \boldsymbol{x}_i) = [\Phi(\boldsymbol{x}_i \boldsymbol{\theta})]^y [1 - \Phi(\boldsymbol{x}_i \boldsymbol{\theta})]^{1-y}$$

- The model is correctly specified model for conditional density if for some

$\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$

$$f\left(y|\,\boldsymbol{x};\boldsymbol{\theta}\right) = p_0\left(y|\,\boldsymbol{x}\right) \text{ for all } x \in \mathcal{X}$$

- Maximum likelihood estimation is special case of M-estimation as

$$\mathrm{E}\left[\ell_i\left(\boldsymbol{\theta}_0\right)|\,\boldsymbol{x}_i\right] \geq \mathrm{E}\left[\ell_i\left(\boldsymbol{\theta}\right)|\,\boldsymbol{x}_i\right] \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

where $\ell_i\left(\boldsymbol{\theta}\right) = \ell_i\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\theta}\right) = \log\left[f\left(y_i|\,\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right]$

- The M-estimator in the case of the Maximum Likelihood is found as the maximum of

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} \ell_i\left(\boldsymbol{\theta}\right)$$

Example 4. *For probit model conditional log likelihood have the form* $\ell_i\left(\boldsymbol{\theta}\right) = y_i \log\left[\Phi\left(\boldsymbol{x}_i\boldsymbol{\theta}\right)\right] + \left(1-y_i\right)\log\left[1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\theta}\right)\right]$

- Consistency and asymptotic normality of ML estimators are can be derived as a special case of the M-estimators

- The same is true about the test statistics (Wald, LM and LR)

Partial maximum likelihood methods

- Assume that we have the correct conditional density for each observation $f_t\left(y_t|\,\boldsymbol{x}_t,\boldsymbol{\theta}_o\right)$ - partial density
- However we do not assume that

$$p_o\left(\boldsymbol{y}|\,\boldsymbol{z}\right)=\prod_{i=1}^{N}f_t\left(y_t|\,\boldsymbol{x}_t,\boldsymbol{\theta}_o\right)$$

  is the correct density function for all the sample given some set of variables $\boldsymbol{z}$
- It is often much simpler to define $f_t\left(y_t|\,\boldsymbol{x}_t,\boldsymbol{\theta}_o\right)$ then the functions which would fulfill the above requirement

- With the standard assumptions concerning M-estimators, it possible to prove that the partial maximum likelihood estimator is consistent and asymptotically normal

- In order to obtain the correct variance matrix we have to use however the robust versions of variance estimators

Example 5. *(Probit with panel data). If the e_i are correlated than it is difficult to derive the conditional distribution of y_{it} given observation $\boldsymbol{x}_{it}$ and $\boldsymbol{x}_{it-1}, \ldots \boldsymbol{x}_{i1}, \boldsymbol{y}_{it-1}, \ldots, \boldsymbol{y}_{i1}$ (y_{it} depends on them indirectly because of the correlation of e_i). However the conditional distribution of the probit model conditional only on $\boldsymbol{x}_{it}$ is the same as original one. Then the consistent estimator can be found using the pooled ML estimator and using the robust variance covariance matrix.*

- For panel data we can use the estimators designed for clusters. We define the same units as clusters and estimated covariance matrix is adjusted

properly.