Introduction to Estimation Methods for Time Series models
Lecture 2

Fulvio Corsi

SNS Pisa

Estimators: Large Sample Properties

- Purposes:
  - study the behavior of $\hat{\theta}_n$ when $n \to \infty$
  - approximate unknown finite sample distributions of $\hat{\theta}_n$
- Being $\hat{\theta}_n$ a distribution $\forall n$, how to define $\hat{\theta}_n \to \theta_0$?
- Convergence in Probability (plim): The random variable $\hat{\theta}_n$ converges in probability to a constant θ_0 if $\forall \epsilon > 0$

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta_0| < \epsilon) = 1$$

- If plim $\hat{\theta}_n = \theta_0$ the estimator is Consistent
- Convergence in Quadratic Mean:

$$\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n - \theta_0)^2 = \lim_{n\to\infty} MSE[\hat{\theta}_n] = 0$$

- Being

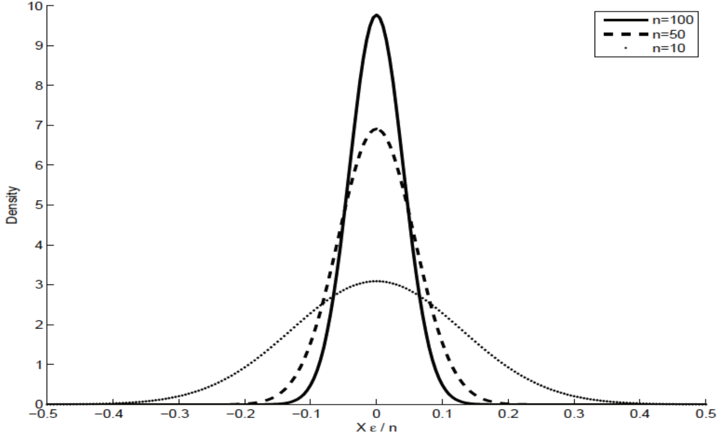
$$MSE[\hat{\theta}_n] = Var[\hat{\theta}_n] + Bias[\hat{\theta}_n]^2$$

  we have Convergence in Quadratic Mean $\Leftrightarrow$

$$\lim_{n\to\infty} Bias[\hat{\theta}_n] = 0 \quad \text{and} \quad \lim_{n\to\infty} Var[\hat{\theta}_n] = 0$$

- Convergence in Quadratic Mean $\Rightarrow$ Convergence in Probability

Consistency



OLS without Normality: Large Sample Theory

When H.4 of Normality is violated, OLS is still unbiased and BLUE, however confidence intervals and test statistics are not valid

Assumptions for the large sample theory:

- A.1 $\mathbb{E}[x_i \epsilon_i] = 0$ regressors uncorrelated with errors (weaker than strict exogeneity)
- A.2 $\lim_{n\to\infty} \frac{1}{n} X'X = Q$ definite positive

Being

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i \epsilon_i\right)$$

- Consistency (convergence in Probability): plim$\widehat{\beta}_n = \beta$

$$\widehat{\beta}_n = \beta + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}}_{\to Q^{-1} \text{(A.2)}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i\right)}_{\to 0 \quad \text{(A.1+LLN)}} \;\to\; \beta$$

- Asymptotic Normality

$$\sqrt{n}\left(\widehat{\beta}_n - \beta\right) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}}_{\to Q^{-1} \text{(A.2)}} \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \epsilon_i\right)}_{\to N(0,\sigma^2 Q) \quad \text{(A.2+CLT)}} \;\to\; N(0, Q^{-1}(\sigma^2 Q)Q^{-1}) = N(0, \sigma^2 Q^{-1})$$

Hence $\widehat{\beta}_n \to N(\beta, \sigma^2 (X'X)^{-1})$ as in the Normal case (H.4) but only asymptotically.

NLS (the idea)

- The Nonlinear Regression model is

$$y_i = h(x_i, \beta) + \epsilon_i \quad \text{with} \quad \mathbb{E}[\epsilon_i | x_i] = 0$$

- Nonlinear Least Square (NLS) estimator:

$$\widehat{\beta}_{NLS} = \arg\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - h(x_i, \beta))^2$$

- FOC: NLS estimator $\widehat{\beta}_{NLS}$ satisfy

$$\sum_{i=1}^{n} \left[y_i - h(x_i, \widehat{\beta}_{NLS}) \right] \frac{\partial h(x_i, \widehat{\beta}_{NLS})}{\partial \beta} = 0$$

- In general, no close form solutions $\Rightarrow$ numerical minimization

Maximum Likelihood

- Basic, strong assumption: distribution of the data known up to θ.
- Likelihood function: The joint density of an $i.i.d$ random sample $(x_1, x_2, \ldots, x_n)$ from $f(x;\theta_0)$

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \ldots f(x_n;\theta)$$

a different perspective: see the joint density as a function of the parameters θ (as opposed to the sample)

$$L(\theta|X) \equiv f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

it is usually simpler to work with the log of the likelihood

$$l(\theta|x_1, x_2, \ldots, x_n) \equiv \ln L(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

- ML Estimator: Given sample data generated from parametric model, find parameters that maximize probability of observing that sample.

$$\widehat{\theta}_{ML} = \arg\max_{\theta} L(\theta|x_1, x_2, \ldots, x_n) = \arg\max_{\theta} l(\theta|x_1, x_2, \ldots, x_n)$$

F.0.C. $\Rightarrow$ the Score:

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

Maximum Likelihood: Example

Consider a Univariate Normal model:

$$f(y,\theta) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right)$$

The log-Likelihood is

$$l(\mu,\sigma^2) = \sum_{i=1}^{n} f(y_i;\theta) = -\frac{n}{2}\ln(2\pi) - -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}$$

and then the score of the μ and σ^2 parameters are

$$\frac{\partial l(\mu,\sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = 0$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

Therefore, by first solving for $\widehat{\mu}$ and inserting it in the score of $\widehat{\sigma}^2$ we get the ML estimators

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu}_{ML})^2$$

Maximum Likelihood: Properties

- Consistency:

$$\text{plim } \widehat{\theta}_{ML} = \theta_0$$

- Asymptotic normality:

$$\widehat{\theta}_{ML} \overset{a}{\sim} N\left(\theta_0, [I(\theta_0)]^{-1}\right) \quad \text{where} \quad I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right]$$

  $I(\theta)$ is the Fisher Information matrix

- Asymptotic efficiency: has the smallest asymptotic variance being $I(\theta)^{-1}$ the Cramér-Rao lower bound

- Invariance: if $\widehat{\theta}$ is the MLE for θ_0, and if $g(\theta)$ is any (invertible) transformation of θ, then the MLE for $g(\theta_0) = g(\widehat{\theta})$.

  Ex: precision parameter $\gamma^2 = 1/\sigma^2 \Rightarrow \gamma^2_{ML} = 1/\sigma^2_{ML}$

Bottom line: MLE makes “best use” of information (asymptotically)

Maximum Likelihood: Properties of regular density

Under some regularity conditions (whose goal is to use Taylor approximation and interchange differentiation and expectation)

- D1: $\ln f(y_i;\theta)$, $g_i = \frac{\partial \ln f(y_i;\theta)}{\partial\theta}$, $H_i = \frac{\partial^2 \ln f(y_i;\theta)}{\partial\theta\theta'}$ are random sample
- D2: $\mathbb{E}_0[g_i(\theta_0)] = 0$
- D3: $Var_0[g_i(\theta_0)] = -\mathbb{E}_0[H_i(\theta_0)]$

D1 implied by assumption: $y_i, i = 1, \ldots, n$ is random sample

D2 is a consequence of $\int \ln f(y_i;\theta_0)dy_i = 1$ since by differencing both sides by θ_0

$$0 = \int \frac{\partial f(y_i;\theta_0)}{\partial\theta_0}dy_i = \int \frac{\partial \ln f(y_i;\theta_0)}{\partial\theta_0} f(y_i;\theta_0)dy_i = \mathbb{E}_0[g_i(\theta_0)]$$

D3 is obtained by differencing once more w.r.t θ_0

D1 (random sample) $\Rightarrow Var_0\left[\sum_{i=1}^n g_i(\theta_0)\right] = \sum_{i=1}^n Var_0[g_i(\theta_0)]$, thus

$$\underbrace{J(\theta_0) \equiv Var_0\left[\frac{\partial \ln L(\theta_0;y_i)}{\partial\theta_0}\right] = -\mathbb{E}_0\left[\frac{\partial^2 \ln L(\theta_0;y_i)}{\partial\theta_0\theta_0'}\right] \equiv I(\theta_0)}_{\text{Information matrix equality}}$$

Asymptotic normality of MLE

being the max of $\ln L$, MLE satisfy by construction the likelihood equation $g(\hat{\theta}) = \sum_{i=1}^{n} g_i(\hat{\theta}) = 0$

define $H(\theta_0) = \frac{\partial^2 \ln L(\theta_0;y_i)}{\partial\theta_0\theta_0'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i;\theta_0)}{\partial\theta_0\theta_0'} = \sum_{i=1}^{n} H_i(\theta_0)$

1. take first order Taylor expansion of the score $g(\hat{\theta})$ around θ_0

$$g(\hat{\theta}) = g(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) + R_1 = 0,$$

2. rearrange and scale by $\sqrt{n}$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \underbrace{\left(-\frac{1}{n}\sum_{i=1}^{n} H_i(\theta_0)\right)^{-1}}_{\to -\mathbb{E}_0\left[\frac{1}{n}H(\theta_0)\right]} \times \underbrace{\sqrt{n}\ \frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)}_{\to N\left(0, Var_0\left[\frac{1}{n}g(\theta_0)\right]\right)} + \underbrace{R_1}_{\to 0}$$

3. Use LLN (on first term) and CLT (on second term)

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) \quad &\to \quad \mathbb{E}_0\left[\frac{1}{n}H(\theta_0)\right]^{-1} \times N\left(0, Var_0\left[\frac{1}{n}g(\theta_0)\right]\right) = \left(\frac{1}{n}I(\theta_0)\right)^{-1} \times N\left(0, \frac{1}{n}J(\theta_0)\right) \\ &\to \quad N\left(0,\ n\ I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1}\right)\end{aligned}$$

if information matrix equality $J(\theta_0) = I(\theta_0)$ holds then

$$\hat{\theta} \overset{a}{\sim} N\left(\theta_0, I(\theta_0)^{-1}\right)$$

Estimating asymptotic covariance matrix of MLE

Three asymptotically equivalent estimators of the Asy.Var[$\hat{\theta}$]:

1. Calculate $\mathbb{E}_0[H(\theta_0)]$ (very difficult) and evaluate it at $\hat{\theta}$ to estimate

$$\{I(\hat{\theta})\}^{-1} = \left\{ -\mathbb{E}_0 \left[\frac{\partial^2 \ln L(\hat{\theta}; y_i)}{\partial \hat{\theta} \hat{\theta}'} \right] \right\}^{-1}$$

2. Calculate $H(\theta_0)$ (still quite difficult) and evaluate it at $\hat{\theta}$ to get

$$\{\hat{I}(\hat{\theta})\}^{-1} = \left\{ \frac{\partial^2 \ln L(\hat{\theta}; y_i)}{\partial \hat{\theta} \hat{\theta}'} \right\}^{-1}$$

3. BHHH or OPG estimator (easy): use information matrix equality $I(\theta_0) = J(\theta_0)$

$$\{\tilde{I}(\hat{\theta})\}^{-1} = \left\{ Var \left[\frac{\partial \ln L(\hat{\theta}; y_i)}{\partial \hat{\theta}} \right] \right\}^{-1} = \left\{ \sum_{i=1}^{n} g_i(\hat{\theta}) g_i(\hat{\theta})' \right\}^{-1}$$

Hypothesis testing (idea)

Test of hypothesis $H_0 : c(\theta) = 0$

Three tests, asymptotically equivalent (not in finite sample):

- Likelihood ratio test : If $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference,

  $$2(\ln L - \ln L_R) \sim \chi^2_{df}$$

  Both unrestricted $\ln L$ and restricted $\ln L_R$ ML estimators are required

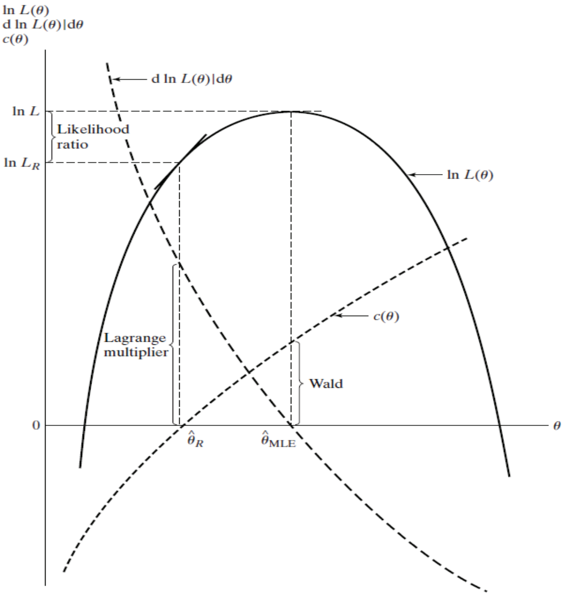
- Wald test: If $c(\theta) = 0$ is valid, then $c(\theta_{ML}) \approx 0$

  Only unrestricted (ML) estimator is required

- Lagrange multiplier test: If $c(\theta) = 0$ is valid, then the restricted estimator should be near the point that maximizes the $\ln L$. Therefore, the slope of $\ln L$ should be near zero at the restricted estimator

  Only restricted estimator is required

Hypothesis testing



Application of MLE: Linear regression model

Model: $y_i = x_i'\beta + \epsilon_i$ and $y_i|x_i \sim N(x_i'\beta, \sigma^2)$

Log-likelihood based on n conditionally independent observations:

$$\begin{aligned} \ln L &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)}{\sigma^2} \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \end{aligned}$$

Likelihood equations

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= \frac{X'(y - X\beta)}{\sigma^2} = 0 \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4} = 0 \end{aligned}$$

solving likelihood equations

$$\hat{\beta}_{ML} = (X'X)^{-1}X'Y \qquad \text{and} \qquad \hat{\sigma}^2_{ML} = \frac{e'e}{n}$$

$\hat{\beta}_{ML} = \hat{\beta}_{OLS} \Rightarrow$ OLS has all desirable asymptotic properties of MLE

Maximum Likelihood in time series: AR model

- In a time series y_t, the innovations ϵ_t are usually not i.i.d.

  $\Rightarrow$ It is then very convenient to use the **"prediction–error"** decomposition of the likelihood:

$$L(y_T, y_{T-1}, \ldots, y_1; \theta) = f(y_T|\Omega_{T-1}; \theta)\ f(y_{T-1}|\Omega_{T-2}; \theta)\ \ldots\ f(y_1|\Omega_0; \theta)$$

- For example for the AR(1)

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

  the full log-Likelihood can be written as

$$l(\phi) = \underbrace{f_{Y_1}(y_1;\phi)}_{\text{marginal } 1^{st} \text{ obs}} + \underbrace{\sum_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\phi)}_{\substack{\text{conditional likelihood} \\ \text{under normality OLS=MLE}}} = f_{Y_1}(y_1;\phi) - \frac{T}{2}\log(2\pi) - \sum_{t=1}^{T}\log\sigma^2 - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - \phi y_{t-1})^2}{\sigma^2}$$

  Hence, maximizing the conditional likelihood for ϕ is equivalent to minimize

$$\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2$$

  which is the OLS criteria.

- In general for AR(p) process OLS are consistent and, under gaussianity, asymptotically equivalent to MLE $\Rightarrow$ asymptotically efficient

Maximum Likelihood in time series: ARMA model

- For a general ARMA(p,q)

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

Y_{t-1} is correlated with $\epsilon_{t-1}, ..., \epsilon_{t-q}$ $\Rightarrow$ OLS not consistent.

$\rightarrow$ MLE with numerical optimization procedures.

Maximum Likelihood in time series: GARCH model

- A GARCH process with gaussian innovation:

$$r_t|\Omega_{t-1} \sim N(\mu_t(\theta), \sigma_t^2(\theta))$$

- has conditional densities:

$$f(r_t|\Omega_{t-1};\theta) = \frac{1}{\sqrt{2\pi}}\sigma_t^{-1}(\theta)\exp\left(-\frac{1}{2}\frac{(r_t-\mu_t(\theta))^2}{\sigma_t^2(\theta)}\right)$$

- using the prediction–error decomposition the log-likelihood becomes:

$$\log L(r_T, r_{T-1}, ..., r_1;\theta) = -\frac{T}{2}\log(2\pi) - \sum_{t=1}^{T}\log\sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^{T}\frac{(r_t-\mu_t(\theta))^2}{\sigma_t^2(\theta)}$$

- Non–linear function in $\theta \Rightarrow$ Numerical optimization techniques.

Quasi Maximum Likelihood

- ML requires complete specification of $f(y_i|x_i;\theta)$, usually Normality is assumed.

  Nevertheless, even if the true distribution is not Normal, assuming Normality gives consistency and asymptotic normality provided that the conditional mean and variance processes are correctly specified.

- However, the information matrix equality does not hold anymore i.e. $J(\theta_0) \neq I(\theta_0)$ hence, the covariance matrix of $\widehat{\theta}_{ML}$ is not $I(\theta_0)^{-1} = -\mathbb{E}\left[\frac{\partial^2 l(\theta_0)}{\partial\theta_0 \partial\theta_0'}\right]^{-1}$ but

  $$\widehat{\theta}_{QML} \overset{a}{\sim} N\left(\theta_0, [I(\theta_0)]^{-1} J(\theta_0) [I(\theta_0)]^{-1}\right)$$

  where

  $$J(\theta_0) = \lim_{n\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\frac{\partial l_t(\theta_0)}{\partial\theta_0} \frac{\partial l_t(\theta_0)}{\partial\theta_0}'\right]$$

  the std error provided by $\left[\widehat{I(\theta_0)}\right]^{-1} \widehat{J(\theta_0)} \left[\widehat{I(\theta_0)}\right]^{-1}$ are called the robust standard errors.

GMM (idea)

- Idea: Sample moments $\xrightarrow{p}$ Population moments = function(parameters)

- A vector function $g(\theta, w_t)$ which under the true value θ_0 satisfies

$$\mathbb{E}_0\left[g(\theta_0, w_t)\right] = 0$$

  is called a set of orthogonality or moment conditions

- Goal: estimate θ_0 from the informational content of the moment conditions
  $\Rightarrow$ semiparametric approach i.e. no distributional assumptions!

- replace population moment conditions with sample moments

$$\widehat{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(\theta, w_t)$$

  and minimize, w.r.t. θ, a quadratic form of $\widehat{g}_n(\theta)$ with a certain weighting matrix W

$$\widehat{\theta}_{GMM} = \arg\min_{\theta}\left(\frac{1}{T}\sum_{t=1}^{T} g(\theta, w_t)\right)' W \left(\frac{1}{T}\sum_{t=1}^{T} g(\theta, w_t)\right)$$

GMM: optimal weighting matrix

- exactly identified (Method of Moment, MM):

  # orthogonality conditions = # parameters $\Rightarrow$ moment equations satisfied exactly, i.e.

  $$\frac{1}{T}\sum_{t=1}^{T} g(\widehat{\theta}, w_t) = 0 \qquad \Rightarrow \qquad W \text{ irrelevant}$$

- overidentified (Generalize MM):

  # orthogonality conditions $>$ # parameters $\Rightarrow W$ is relevant.

  The optimal weighting matrix W^* is the inverse of the asymptotic var-cov of $g(\theta_0, w_t)$

  $$W^* = Var\left[g(\theta_0, w_t)\right]^{-1}$$

  but it depends on the unknonw θ_0

- Feasible two-step procedure:

  Step 1. Use $W = I$ to obtain a consistent estimator, $\hat{\theta}_1$, then estimate

  $$\hat{\Phi} = \frac{1}{T}\sum_{t=1}^{T} g(\hat{\theta}_1, w_t) g(\hat{\theta}_1, w_t)'$$

  Step 2. Compute second step GMM estimator using the weighting matrix $\hat{\Phi}^{-1}$

  $$\widehat{\theta}_{GMM} = \arg\min_{\theta} \left(\frac{1}{T}\sum_{t=1}^{T} g(\theta, w_t)\right)' \hat{\Phi}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} g(\theta, w_t)\right)$$

  The two-step estimator $\widehat{\theta}_{GMM}$ is asymptotically efficient in the GMM class

GMM, OLS and ML

- Many estimation methods can be seen as GMM

  Examples:

  - 1) OLS is a GMM with ortoghonality condition

$$\mathbb{E}\left[x_i \epsilon_i\right] = \mathbb{E}\left[x_i(y_i - x_i'\theta)\right] = 0$$

  - 2) ML is a GMM on the score

$$\mathbb{E}\left[\frac{\partial \log f(y_t;\theta)}{\partial \theta}\right] = \mathbb{E}\left[g(\theta, w_t)\right] = 0$$