Maximum likelihood

- Maximum likelihood
- ML is GMM on the scores
- When factors are returns, ML prescribes a time-series regression
- When factors are not excess returns, ML prescribes a cross-sectional regression

DGP

- We want to identify the Data generating process by a realization of it.
- Probability function is a main character of DGP.
- Different probability function can have the same realization.

Maximum likelihood

- The ML principle: pick the probability function(identified by parameters) that make the observed data most likely.

$$\hat{\theta} = \arg\max_{\{\theta\}} f\left(\{x_t\};\theta\right)$$

- *likelihood function-probability function*

Why conditional and log?(1)

- Finding the likelihood function isn't easy. So we turn to a easier one-*conditional likelihood function.* We do this because

$$p(abc) = p(a/bc)\,p(bc) = p(a/bc)\,p(b/c)\,p(c)$$

- In a time-series context,

$$f\left(\{x_t\};\theta\right) = f\left(\{x_t,\cdots x_1,x_0\};\theta\right)$$

$$= f\left(x_t \mid x_{t-1},x_{t-2},\cdots x_0;\theta\right) f\left(x_{t-1} \mid x_{t-2},x_{t-3},\cdots x_0;\theta\right)\cdots f\left(x_0;\theta\right)$$

Why conditional and log?(2)

- To make it simpler, we take log:

$$L(\{x_t\};\theta) = \ln\left(f(\{x_t\};\theta)\right)$$

$$= \ln\left(f(x_0;\theta)\right) + \sum_{t=1}^{T} \ln\left(f(x_t \mid x_{t-1}, x_{t-2}, \cdots x_0;\theta)\right)$$

- Remark: The first term is not a conditional probability.

estimation

- According to the definition, we should estimate the parameters by

$$\max\left\{L\left(\left\{x_t\right\};\theta\right)\right\}$$

- To do it we can take derivative but it always gets nonlinear function.
- Or do numerically maximization (Time series analysis; Hamilton 1994 p157)
- Or add normal assumptions

Adding normal assumptions(1)

- It is usually hard to evaluate the unconditional density or the first terms with only a few lagged observations.
- Suppose T is big enough, we can treat the first k observations as fixed rather than random variables, then

$$L\left(\{x_t\};\theta\right)=\sum_{t=k+1}^{T}\ln\left(f\left(x_t \mid x_{t-1},x_{t-2},\cdots x_0;\theta\right)\right)$$

Adding normal assumptions(2)

- Alternatively, one can treat k pre-sample values $x_0, x_{-1}, \ldots, x_{-k+1}$ as additional parameters, then the likelihood function becomes

$$L(\{x_t\};\theta) = \sum_{t=1}^{T} \ln\left(f\left(x_t, x_{t-1}, x_{t-2}, \cdots x_1;\theta\right)\right)$$

Adding normal assumptions(3)

- Define shocks:

$$\varepsilon_t = x_t - E\left(x_t \mid x_{t-1}, x_{t-2} \cdots x_0; \theta\right)$$

$$x_t = \left(x_{t1}, x_{t2} \cdots x_{tN}\right)'$$

- Assume the shocks of N assets are joint normal, then x is joint normal, so

$$f\left(x_t\right) = \frac{1}{\left(2\pi\right)^{N/2}\sqrt{\left|\Sigma\right|}} e^{-\left(x_t - u\right)' \Sigma^{-1}\left(x_t - u\right)}$$

$$u = E\left(x_{ti}\right) = \left(u_1, \cdots, u_N\right)', \Sigma = \text{cov}\left(\varepsilon_t, \varepsilon_t'\right)$$

Adding normal assumptions(4)

- Then $L\left(\{x_t\};\theta\right)=\sum_{t=1}^{T}\ln\left(f\left(x_t,x_{t-1},x_{t-2},\cdots x_1;\theta\right)\right)$

$$=\sum_{t=1}^{T}\ln\left(\frac{1}{(2\pi)^{N/2}\sqrt{|\Sigma|}}e^{-(x_t-u)'\Sigma^{-1}(x_t-u)}\right)$$

$$=-\frac{T}{2}\ln\left((2\pi)^{N}\sqrt{|\Sigma|}\right)-\frac{1}{2}\varepsilon_t'\Sigma^{-1}\varepsilon_t$$

- Using the first order condition, we can get the estimation

distribution of the estimates

$$\hat{\theta} \sim N\left(\theta, \left[-\frac{\partial^2 L}{\partial\theta\partial\theta'}\right]^{-1}\right)$$

- *asymptotically efficient* : no other estimator can produce a smaller covariance matrix.

Information matrix

- Information matrix is defined as the expected value of the second partial,which is estimated with the sample value.

$$\mathrm{I} = -\frac{1}{T}\left[\frac{\partial^2 L}{\partial\theta\partial\theta'}\right] = -\frac{1}{T}\sum_{1}^{T}\frac{\partial^2 \ln\left(f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots x_0; \theta\right)\right)}{\partial\theta\partial\theta'}$$

likelihood ratio test

- If the restriction is true, there should not exist big difference between unrestricted maximum value and restricted maximum value:

$$2\left(L_{\text{unrestriction}} - L_{\text{restriction}}\right) \sim \chi^2_{\text{no. of restriction}}$$

- H0: the restriction is true

ML is GMM on the scores(1)

- ML is a special case of GMM. Because the first order conditions for maximizing a likelihood function is

$$\frac{\partial\left(L\left(\{x_t\};\theta\right)\right)}{\partial\theta}=\sum_{t=1}^{T}\frac{\partial \ln f\left(x_t \mid x_{t-1},x_{t-2},\cdots;\theta\right)}{\partial\theta}=0$$

ML is GMM on the scores(2)

- In GMM language the model is

$$E\left[f\left(x_t, b\right)\right] = 0$$

- The estimation is

$$a_T g_T\left(\hat{b}\right) = a_T\left(\frac{1}{T}\sum_{t=1}^{T} f\left(x_t, b\right)\right) = 0$$

- In ML the estimation is

$$\frac{\partial\left(L\left(\left\{x_t\right\};\theta\right)\right)}{\partial\theta} = \sum_{t=1}^{T}\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots;\theta\right)}{\partial\theta} = 0$$

ML is GMM on the scores(3)

- So (multiply both sides by 1/T)

$$\frac{1}{T}\left(\sum_{t=1}^{T}\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta}\right) = 0$$

- here

$$a_T g_T\left(\hat{b}\right) = 1 \times g_T\left(\hat{\theta}\right)$$

$$= \frac{1}{T}\left(\sum_{t=1}^{T}\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta}\right) = 0$$

Score

- $$\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta}$$

Example(1)

- AR(1): $x_t = \rho x_{t-1} + \varepsilon_t$
- Suppose $\varepsilon_t \sim i.i.d, normal$
- Then

$$\ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \rho\right) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(x_t - u\right)^2 / 2\sigma^2}\right)$$

$$= const - \left(x_t - \rho x_{t-1}\right)^2 / 2\sigma^2$$

Example(2)

- The score is

$$\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots ; \rho\right)}{\partial \rho} = \frac{(x_t - \rho x_{t-1})x_{t-1}}{\sigma^2}$$

- The first order condition for ML is

$$\frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} (x_t - \rho x_{t-1})x_{t-1} = 0$$

Hint: The scores should be unforecastable

$$E_{t-1}\left[\frac{(x_t-\rho x_{t-1})x_{t-1}}{\sigma^2}\right]=E_{t-1}\left[\frac{\varepsilon_t x_{t-1}}{\sigma^2}\right]=0$$

Argument: The scores should be unforecastable

- $$1 = \int f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right) dx_t$$

$$0 = \int \frac{\partial f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta} dx_t$$

$$0 = \int \frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta} f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right) dx_t$$

$$so E_{t-1}\left[\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta}\right] = 0$$

The distribution of ML and GMM is the same(1)

- According to GMM

$$\sqrt{T}(\hat{b}-b) \rightarrow N[0,(ad)^{-1} aSa'(ad)^{-1'}]$$

- where

$$d \equiv E[\frac{\partial f}{\partial b'}(x_t,b)] = \frac{\partial g_T(b)}{\partial b'}$$

$$a = p\lim a_T$$

$$S \equiv \sum_{j=-\infty}^{\infty} E[f(x_t,b), f(x_{t-j},b)']$$

The distribution of ML and GMM is the same(2)

$$d = \frac{\partial g_T(\theta)}{\partial \theta'}$$

$$= \frac{\partial \left(\frac{1}{T} \left(\sum_{t=1}^{T} \frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta} \right) \right)}{\partial \theta'}$$

$$= \frac{1}{T} \left(\sum_{t=1}^{T} \frac{\partial^2 \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta \partial \theta'} \right)$$

$$= \text{information matrix}$$

The distribution of ML and GMM is the same(3)

- Because

$$E_{t-1}\left[\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta}\right] = 0$$

The distribution of ML and GMM is the same(4)

- so $$S \equiv \sum_{j=-\infty}^{\infty} E[f(x_t,b), f(x_{t-j},b)']$$

$$= E\left[f(x_t,b)f(x_{t-1},b)'\right] + E\left[f(x_t,b)f(x_{t-2},b)'\right] + \cdots$$

$$+E\left[f(x_t,b)f(x_{t+1},b)'\right] + E\left[f(x_t,b)f(x_{t+1},b)'\right] + \cdots$$

$$+E\left[f(x_t,b)f(x_t,b)'\right]$$

$$= E\left[E_{t-1}\left[f(x_t,b)f(x_{t-1},b)'\right]\right] + E\left[E_{t-1}\left[f(x_t,b)f(x_{t-2},b)'\right]\right] + \cdots$$

$$+E\left[E_t\left[f(x_t,b)f(x_{t+1},b)'\right]\right] + E\left[E_{t+1}\left[f(x_t,b)f(x_{t+2},b)'\right]\right] + \cdots$$

$$+E\left[f(x_t,b)f(x_t,b)'\right]$$

$$= E\left[f(x_t,b)f(x_t,b)'\right]$$

The distribution of ML and GMM is the same(5)

$$S = E\left[f(x_t, b) f(x_t, b)' \right]$$

$$= E\left[\frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)}{\partial \theta} \frac{\partial \ln f\left(x_t \mid x_{t-1}, x_{t-2}, \cdots; \theta\right)'}{\partial \theta} \right]$$

$$= \text{information matrix}$$

- $a = 1$ then GMM ML

$$\sqrt{T}(\hat{\theta} - \theta) \to N[0, (ad)^{-1} aSa'(ad)^{-1'}]$$

$$= N\left[0, I^{-1} I I^{-1}\right] = N\left[0, I^{-1}\right]$$

$$\hat{\theta} \sim N\left(\theta, \left[-\frac{\partial^2 L}{\partial \theta \partial \theta'} \right]^{-1} \right)$$

Empirical application: model to be tested(1)

- Suppose utility come from consumption.
- Maximizing utility we get p=E(mx) where

$$m = \beta \frac{U'(c_{t+1})}{U'(c_t)}$$

Empirical application: model to be tested(2)

- $$m = \beta \frac{U'(c_{t+1})}{U'(c_t)}$$

$$= g(f) \begin{cases} \text{g is a linear function} \\ \text{g is nolinear} \begin{bmatrix} \text{Talor approximations} \\ \text{nomal assumption} \\ \text{continouse time} \end{bmatrix} \end{cases}$$

$$\Rightarrow m = a + b'f \Leftrightarrow E\left(\mathrm{R}^{ei}\right) = \lambda'\beta$$

When factors are returns, ML is a time-series regression(1)

$$E\left(R^{ei}\right) = \beta_i'\lambda = \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \cdots$$

- When factors are excess returns, we can get

$$E\left(f^a\right) = \lambda_a$$

- So

$$E\left(R^{ei}\right) = \beta_{i,a}E\left(f^a\right) + \beta_{i,b}E\left(f^b\right) + \cdots$$

When factors are returns, ML is a time-series regression(2)

- Consider a single factor model, the economic model to be tested is

$$E\left(\mathrm{R}^{e}\right) = \beta E\left(f\right) \text{ where } f = \mathrm{R}^{em}$$

- Assume that the market return and regression errors are i.i.d. normal

$$R_t^e = \alpha + \beta f_t + \varepsilon_t \qquad \begin{bmatrix} \varepsilon_t \\ u_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_u^2 \end{bmatrix} \right)$$

$$f_t = E\left(f\right) + u_t$$

When factors are returns, ML is a time-series regression(3)

- CAPM implies that the intercepts α should all be zero.
- Under i.i.d. assumption

$$\ln\left(p\left(R_t^e f_t\right)\right)=\ln\left(p\left(R_t^e\right)p\left(f_t\right)\right)=\ln p\left(R_t^e\right)+\ln p\left(f_t\right)$$

- So the likelihood function is

$$L=const.-\frac{1}{2}\sum_{t=1}^{T}\left(R_t^e-\beta f_t\right)'\Sigma^{-1}\left(R_t^e-\beta f_t\right)-\frac{1}{2}\sum_{t=1}^{T}\left(f_t-E\left(f\right)\right)^2/\sigma_u^2$$

$$-\frac{T}{2}\ln\left(\left(2\pi\right)^N\sqrt{\left|\Sigma\right|}\right)-\frac{1}{2}\varepsilon_t'\Sigma^{-1}\varepsilon_t$$

When factors are returns, ML is a time-series regression(4)

- The estimates follow from the first order conditions

$$\frac{\partial L}{\partial \beta} = \Sigma^{-1} \sum_{t=1}^{T} \left(R_t^e - \beta f_t \right) f_t = 0 \Rightarrow \hat{\beta} = \left(\sum_{t=1}^{T} f_t^2 \right)^{-1} \sum_{t=1}^{T} R_t^e f_t$$

$$\frac{\partial L}{\partial E(f)} = \frac{1}{\sigma_u^2} \sum_{t=1}^{T} \left(f_t - E(f) \right) = 0 \Rightarrow \widehat{E(f)} = \hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} f_t$$

- The ML estimate of β is the time-series OLS regression *without* a constant.

distribution

-

$$\operatorname{cov}\left(\hat{\beta}\right)=\left[-\frac{\partial^2 L}{\partial\beta\partial\beta'}\right]^{-1}=\left[\Sigma^{-1}\sum_{t=1}^{T} f_t^{\,2}\right]^{-1}$$

$$=\frac{1}{T}\frac{1}{E\left(f^2\right)}\Sigma=\frac{1}{T}\frac{1}{E\left(f\right)^2+\sigma^2\left(f\right)}\Sigma$$

Estimation without the restriction $\alpha = 0$

- The unconstrained likelihood function is

$$L = const. - \frac{1}{2}\sum_{t=1}^{T}\left(R_t^e - \alpha - \beta f_t\right)' \Sigma^{-1}\left(R_t^e - \alpha - \beta f_t\right) + \cdots$$

- The estimates are now

$$\frac{\partial L}{\partial \alpha} = \Sigma^{-1}\sum_{t=1}^{T}\left(R_t^e - \alpha - \beta f_t\right) = 0 \Rightarrow \hat{\alpha} = E_T\left(R_t^e\right) - \hat{\beta} E_T\left(f_t\right)$$

$$\frac{\partial L}{\partial \beta} = \Sigma^{-1}\sum_{t=1}^{T}\left(R_t^e - \alpha - \beta f_t\right) f_t = 0 \Rightarrow \hat{\beta} = \frac{\mathrm{cov}_T\left(R_t^e, f_t\right)}{\sigma_T^2\left(f_t\right)}$$

distribution

$$\operatorname{cov}\left(\hat{\alpha}\right)=\frac{1}{T}\left[1+\left(\frac{E\left(f\right)}{\sigma\left(f\right)}\right)^{2}\right]\Sigma$$

$$\operatorname{cov}\left(\hat{\beta}\right)=\frac{1}{T}\frac{1}{\sigma^{2}\left(f\right)}\Sigma$$

Wald test

- $$T\left[1+\left(\frac{E(f)}{\sigma(f)}\right)^{2}\right]^{-1}\hat{\alpha}'\Sigma^{-1}\hat{\alpha} \sim \chi_{N}^{2}$$

warning

- ML will ruthlessly exploit the null hypothesis in order to get any small improvement in efficiency.
- Because ML estimator is *asymptotically efficient and*
- *25%*

$$\frac{\operatorname{cov}\left(\hat{\beta}\right)_{unconstrained}}{\operatorname{cov}\left(\hat{\beta}\right)_{constrained}} = 1 + \frac{E\left(f\right)^2}{\sigma^2\left(f\right)} > 1$$

When factors are not excess returns, ML is a cross-sectional regression(1)

$$E\left(R^{ei}\right) = \beta_i'\lambda = \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \cdots$$

- The betas are defined from time-series regressions

$$R_t^{ei} = a_i + \beta_i' f_t + \varepsilon_t^i$$

- Taking expectations

$$E\left(R^{ei}\right) = a_i + \beta_i' E\left(f_t\right) = \beta_i'\lambda \text{ so } a_i = \beta_i'\left(\lambda - E\left(f_t\right)\right)$$

$$\textit{and} \quad R_t^{ei} = \beta_i'\lambda + \beta_i'\left[f_t - E\left(f_t\right)\right] + \varepsilon_t^i$$

When factors are not excess returns, ML is a cross-sectional regression(2)

- denote B a $N \times K$ matrix of regression coefficients of the N assets on the K factors, then

$$R_t^e = B\lambda + B\left[f_t - E\left(f_t \right) \right] + \varepsilon_t \qquad \begin{bmatrix} \varepsilon_t \\ u_t \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma & 0 \\ 0 & V \end{bmatrix} \right)$$

$$f_t = E\left(f \right) + u_t$$

When factors are not excess returns, ML is a cross-sectional regression(3)

- The likelihood function is

$$L = const. - \frac{1}{2}\sum_{t=1}^{T}\varepsilon_t{}' \Sigma^{-1} \varepsilon_t - \frac{1}{2}\sum_{t=1}^{T}u_t{}'V^{-1}u_t$$

$$\varepsilon_t = R_t^e - B\left[\lambda + f_t - E\left(f_t\right)\right]; u_t = f_t - E\left(f_t\right)$$

- Maximizing the likelihood function

$$\frac{\partial \mathcal{L}}{\partial E(f)} \quad : \quad 0 = \sum_{t=1}^{T} B'\Sigma^{-1}\left(R_t^e - B\left[\lambda + f_t - E(f)\right]\right) + \sum_{t=1}^{T} V^{-1}\left(f_t - E(f)\right)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} \quad : \quad 0 = B'\sum_{t=1}^{T}\Sigma^{-1}\left(R_t^e - B\left[\lambda + f_t - E(f)\right]\right)$$

solution

$$\widehat{E(f)} = E_T(f_t)$$

$$\hat{\lambda} = \left(B' \Sigma^{-1} B\right)^{-1} B' \Sigma^{-1} E_T\left(R_t^e\right)$$

- *The ML estimate of the factor risk premium is a GLS cross-sectional regression of average returns on betas.*