CHAPTER 9 MAXIMUM LIKELIHOOD ESTIMATION AND QUASI-MAXIMUM LIKELIHOOD ESTIMATION

Key words:

ARMA model, Censored data, Conditional probability distribution model, Discrete choice model, Dynamic information matrix test, GARCH model, Hessian matrix, Information matrix equality, Information matrix test, Lagrange multiplier test, Likelihood, Likelihood ratio test, Martingale, MLE, Pesudo likelihood function, QMLE, Score function, Truncated data, Wald test.

Abstract: Conditional distribution models have been widely used in economics and finance. In this chapter, we introduce two closely related popular methods to estimate conditional probability distribution models—Maximum Likelihood Estimation (MLE) and Quasi-MLE (QMLE). MLE is a parameter estimator that maximizes the model likelihood function of the random sample when the conditional probability distribution model is correctly specified, and QMLE is a parameter estimator that maximizes the model likelihood function of the random sample when the conditional probability distribution model is misspecified. Because the score function is an MDS process and the dynamic information matrix equality holds when a conditional distribution model is correctly specified, the asymptotic properties of the MLE is analogous to those of the OLS estimator when the regression disturbance is an MDS with conditional homoskedasticity, and we can use the Wald test, Lagrange Multiplier test and Likelihood Ratio test for hypothesis testing, where the Likelihood Ratio test is analogous to the $J \cdot F$ test statistic. On the other hand, when the conditional distributional model is misspecified, the score function has mean zero, but it may no longer be an MDS process and the dynamic information matrix equality may fail. As a result, the asymptotic properties of the QMLE are analogous to those of the OLS estimator when the regression disturbance displays serial correlation and conditional heteroskedasticity. Robust Wald tests and Lagrange Multiplier tests can be constructed for hypothesis testing, but the Likelihood ratio test can no longer be used, for a reason similar to the failure of the F-test statistic when the regression disturbance displays conditional heteroskedasticity and serial correlation. We discuss methods to test the MDS properties of the score function, and the dynamic information matrix equality, and correct specification of the entire conditional distribution model. Some empirical applications are considered.

9.1 Motivation

So far we have focused on the econometric models for conditional mean or conditional expectation, either linear or nonlinear. When do we need to model the conditional probability distribution of Y_t given X_t?

We first provide a number of economic examples which call for the use of a conditional probability distribution model.

Example 1 [Value at Risk, VaR]

In financial risk management, how to quantify extreme downside market risk has been an important issue. Let $I_{t-1} = (Y_{t-1}, Y_{t-2}, ..., Y_1)$ be the information set available at time $t-1$, where Y_t is the return on a portfolio in period t. Suppose

$$\begin{aligned} Y_t &= \mu_t(\beta^o) + \varepsilon_t \\ &= \mu_t(\beta^o) + \sigma_t(\beta^o) z_t, \end{aligned}$$

where $\mu_t(\beta^o) = E(Y_t|I_{t-1}), \sigma_t^2(\beta^o) = \text{var}(Y_t|I_{t-1})$, $\{z_t\}$ is an i.i.d. sequence with $E(z_t) = 0$, $\text{var}(z_t) = 1$, and pdf $f_z(\cdot|\beta^o)$. An example is that $\{z_t\} \sim i.i.d.N(0,1)$.

The value at risk (VaR), $V_t(\alpha) = V(\alpha, I_{t-1})$, at the significance level $\alpha \in (0,1)$, is defined as

$$P\left[Y_t < -V_t(\alpha)|I_{t-1}\right] = \alpha = 0.01 \text{ (say)}.$$

Intuitively, VaR is the threshold that the actual loss will exceed with probability α. Given that $Y_t = \mu_t + \sigma_t z_t$, where for simplicity we have put $\mu_t = \mu_t(\beta^o)$ and $\sigma_t = \sigma_t(\beta^o)$, we have

$$\begin{aligned} \alpha &= P\left(\mu_t + \sigma_t z_t < -V_t(\alpha)|I_{t-1}\right) \\ &= P\left[z_t < \left.\frac{-V_t(\alpha) - \mu_t}{\sigma_t}\right| I_{t-1}\right] \\ &= F_z\left[\frac{-V_t(\alpha) - \mu_t}{\sigma_t}\right], \end{aligned}$$

where the last equality follows by the independence assumption of $\{z_t\}$. It follows that

$$\frac{-V_t(\alpha) - \mu_t}{\sigma_t} = -C(\alpha).$$

$$V_t(\alpha) = -\mu_t + \sigma_t C(\alpha),$$

where $C(\alpha)$ is the left-tailed critical value of the distribution $F_z(\cdot)$ at level α, namely

$$P\left[z_t < -C(\alpha)\right] = \alpha$$

or

$$\int_{-\infty}^{-C(\alpha)} f_z(z|\beta^0)dz = \alpha.$$

For example, $C(0.05) = 1.65$ and $C(0.01) = 2.33$.

Obviously, we need to model the conditional distribution of Y_t given I_{t-1} in order to calculate $V_t(\alpha)$, which is a popular quantitative measure for downside market risk.

For example, J.P. Morgan's RiskMetrics uses a simple conditionally normal distribution model for asset returns:

$$\begin{aligned} Y_t &= \sigma_t z_t, \\ \sigma_t^2 &= (1-\lambda)\sum_{j=1}^{t-1} \lambda^j Y_{t-j}^2, \qquad 0 < \lambda < 1, \\ \{z_t\} &\sim i.i.d.N(0,1). \end{aligned}$$

Here, the conditional probability distribution of $Y_t|I_{t-1}$ is $N(0, \sigma_t^2)$, from which we can obtain

$$V_t(0.05) = 1.65\sigma_t.$$

Example 2 [Binary Probability Modelling] Suppose Y_t is a binary variable taking values 1 and 0 respectively. For example, a business turning point or a currency crisis may occur under certain circumstance; households may buy a fancy new product; and default risk may occur for some financial firms. In all these scenarios, the variables of interest can take only two possible values. Such variables are called binary.

We are interested in the probability that some economic event of interest occurs ($Y_t = 1$) and how it depends on some economic characteristics X_t. It may well be that the probability of $Y_t = 1$ differs among individuals or across different time periods. For example, the probability of students' success depends on their intelligence, motivation, effort, and the environment. The probability of buying a product may depend on income, age, and preference.

To capture such individual effects (denoted as X_t), we consider a model

$$P(Y_t = 1|X_t) = F(X_t'\beta^o),$$

where $F(\cdot)$ is a prespecified CDF. An example of $F(\cdot)$ is the logistic function, namely,

$$F(u) = \frac{1}{1+\exp(-u)}, \qquad -\infty < u < \infty.$$

This is the so-called logistic regression model. This model is useful for modeling (e.g.) credit default risk and currency crisis.

An economic interpretation for the binary outcome Y_t is a story of a latent variable process. Define

$$Y_t = \begin{cases} 1 & \text{if } Y_t^* \leq c, \\ 0 & \text{if } Y_t^* > c, \end{cases}$$

where c is a constant, the latent variable

$$Y_t^* = X_t' \beta^o + \varepsilon_t,$$

and $F(\cdot)$ is the CDF of the i.i.d. error term ε_t. If $\{\varepsilon_t\} \sim i.i.d. N(0, \sigma^2)$ and $c = 0$, the resulting model is called a probit model. If $\{\varepsilon_t\} \sim i.i.d.$ Logistic$(0, \sigma^2)$ and $c = 0$, the resulting model is called a logit model. The latent variable could be the actual economic decision process. For example, Y_t^* can be the credit score and c is the threshold with which a lending institute makes its decision on loan approvals.

This model can be extended to the multinomial model, where Y_t takes discrete multiple integers instead of only two values.

Example 3 [Duration Models]

Suppose we are interested in the time it takes for an unemployed person to find a job, the time that elapses between two trades or two price changes, the length of a strike, the length before a cancer patient dies, and the length before a financial crisis (e.g., credit default risk) comes out. Such analysis is called duration analysis or survival analysis.

In practice, the main interest often lies in the question of how long a duration of an economic event will continue, given that it has not finished yet. An important concept called the hazard rate measures the chance that the duration will end now, given that it has not ended before. This hazard rate therefore can be interpreted as the chance to find a job, to trade, to end a strike, etc.

Suppose Y_t is the duration from a population with the probability density function $f(y)$ and probability distribution function $F(y)$. Then the survival function is defined as

$$S(y) = P(Y_t > y) = 1 - F(y),$$

and the hazard rate is defined as

$$\begin{aligned}\lambda(y) &= \lim_{\delta\to 0^+}\frac{P(y<Y_t\le y+\delta|Y_t>y)}{\delta}\\ &= \lim_{\delta\to 0^+}\frac{P(y<Y_t\le y+\delta)/P(Y_t>y)}{\delta}\\ &= \frac{f(y)}{S(y)}\\ &= -\frac{d}{dy}\ln S(y).\end{aligned}$$

Hence, we have $f(y)=\lambda(y)S(y)$. The specification of $\lambda(y)$ is equivalent to a specification of $f(y)$. But $\lambda(y)$ is more interpretable in economics. For example, suppose we have $\lambda(y)=r$, a constant; that is, the hazard rate does not depend on the length of duration. Then

$$f(y)=r\exp(-ry)$$

is an exponential probability density.

The hazard rate may not be the same for all individuals (i.e., it may depend on individual characteristics X_t). To control heterogeneity across individuals, we assume a conditional hazard function

$$\lambda_t(y)=\exp(X_t'\beta)\lambda_0(y),$$

where $\lambda_0(y)$ is called the baseline hazard rate. This specification is called the proportional hazard model, proposed by Cox (1962). The parameter

$$\begin{aligned}\beta &= \frac{\partial}{\partial X_t}\ln\lambda_t(y)\\ &= \frac{1}{\lambda_t(y)}\frac{\partial}{\partial X_t}\lambda_t(y)\end{aligned}$$

is the marginal relative effect of X_t on the hazard rate of individual t. The survival function of the proportional hazard model is

$$S_t(t)=[S_o(t)]^{\exp(X_t'\beta)}$$

where $S_o(t)$ is the survival function of the baseline hazard rate $\lambda_0(t)$.

The probability density function of Y_t given X_t is

$$f(y|X_t)=\lambda_t(y)S_t(y).$$

To estimate parameter β, we need to use the maximum likelihood estimation (MLE) method, which will be introduced below.

Example 4 [Ultra-High Frequency Financial Econometrics and Engle and Russell's (1998) Autoregressive conditional duration model]

Suppose we have a sequence of tick-by-tick financial data $\{P_i, t_i\}$, where P_i is the price traded at time t_i, where i is the index for the i-th price change. Define the time interval between price changes

$$Y_i = t_i - t_{i-1}, \qquad i = 1, ..., n.$$

Question: How to model the serial dependence of the duration Y_i?

Engle and Russell (1998) propose a class of autoregressive conditional duration model:

$$\begin{cases} Y_i = \mu_i(\beta^o) z_i, \\ \mu_i(\beta^o) = E(Y_i | I_{i-1}), \\ \{z_i\} \sim i.i.d.\text{EXP}(1), \end{cases}$$

where I_{i-1} is the information set available at time t_{i-1}. Here, $\mu_i = \mu_i(\beta^o)$ is called the conditional expected duration given I_{i-1}. A model for μ_i is

$$\mu_i = \omega + \alpha \mu_{i-1} + \gamma Y_{i-1},$$

where $\beta = (\omega, \alpha, \gamma)'$.

From this model, we can write down the model-implied conditional probability density of Y_i given I_{i-1} :

$$f(y|I_{i-1}) = \frac{1}{\mu_i} \exp\left(-\frac{y}{\mu_i}\right), \qquad y > 0.$$

From this conditional density, we can compute the conditional intensity of Y_i (i.e., the instantaneous probability that the next price change will occur at time t_i), which is important for (e.g.) options pricing.

Example 5 [Continuous-time Diffusion models] The dynamics of the spot interest rate Y_t is fundamental to pricing fixed income securities. Consider a diffusion model for the spot interest rate

$$dY_t = \mu(Y_t, \beta^o)dt + \sigma(Y_t, \beta^o)dW_t,$$

where $\mu(Y_t, \beta^o)$ is the drift model, and $\sigma(Y_t, \beta^o)$ is the diffusion (or volatility) model, β^o is an unknown $K \times 1$ parameter vector, and W_t is the standard Brownian motion. Note that the time t is a continuous variable here.

Question: What is the Brownian motion?

Continuous-time models have been rather popular in mathematical finance and financial engineering. First, financial economists have the belief that informational flow into financial markets is continuous in time. Second, the mathematical treatment of derivative pricing is elegant when a continuous-time model is used.

The following are three well-known examples of the diffusion model:

- The random walk model with drift

$$dY_t = \mu dt + \sigma dW_t;$$

- Vasicek's (1977) model

$$dY_t = (\alpha + \beta Y_t)dt + \sigma dW_t;$$

Cox, Ingersoll, and Ross' (1985) model

$$dY_t = (\alpha + \beta Y_t)dt + \sigma Y_t^{1/2} dW_t.$$

These diffusion models are important for hedging, derivatives pricing and financial risk management.

Question: How to estimate model parameters of a diffusion model using a discretely sampled data $\{Y_t\}_{t=1}^n$?

Given $\mu(Y_t, \beta)$ and $\sigma(Y_t, \beta)$, we can determine the conditional probability density $f_{Y_t|I_{t-1}}(y_t|I_{t-1}, \beta)$ of Y_t given I_{t-1}. Thus, we can estimate β^o by the maximum likelihood estimation (MLE) or asymptotically equivalent methods using discretely observed data. For the random walk model, the conditional pdf of Y_t given I_{t-1} is

$$f(y|I_{t-1}, \beta) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left[-\frac{(y-\mu t)^2}{2\sigma^2 t}\right].$$

For Vasicek's (1977) model, the conditional pdf of Y_t given I_{t-1} is

$$f(y|I_{t-1}, \beta) = .$$

For the Cox, Ingersoll and Ross' (1985) model, the conditional pdf of Y_t given I_{t-1} is

$$f(y|I_{t-1}, \beta) = .$$

It may be noted that many continuous-time diffusion models do not have a closed form expression for their conditional pdf, which makes the MLE estimation infeasible. Methods have been proposed in the literature to obtain some accurate approximations to the conditional pdf so that MLE becomes feasible.

9.2 Maximum Likelihood Estimation (MLE) and Quasi-MLE

Recall a random sample of size n is a collection of random vectors $\{Z_1, \cdots, Z_n\}$, where $Z_t = (Y_t, X_t')'$. We denote the random sample as follows:

$$Z^n = (Z_1', \cdots, Z_n')'.$$

A realization of Z^n is a data set, denoted as $z^n = (z_1', \cdots, z_n')'$. A random sample Z^n can generate many realizations (i.e., data sets).

Question: How to characterize the random sample Z^n?

All information in Z^n is completely described by its joint probability density function (pdf) or probability mass function (pmf) $f_{Z^n}(z^n)$. [For discrete r.v.'s, we have $f_{Z^n}(z^n) = P(Z^n = z^n)$.] By sequential partitioning (repeatedly using the multiplication rule that $P(A \cap B) = P(A|B)P(B)$ for any two events A and B), we have

$$\begin{aligned} f_{Z^n}(z^n) &= f_{Z_n|Z^{n-1}}(z_n|z^{n-1}) f_{Z^{n-1}}(z^{n-1}) \\ &= \prod_{t=1}^{n} f_{Z_t|Z^{t-1}}(z_t|z^{t-1}). \end{aligned}$$

where $Z^{t-1} = (Z_{t-1}', Z_{t-2}', \cdots, Z_1')'$, and $f_{Z_t|Z^{t-1}}(z_t|z^{t-1})$ is the conditional pdf of Z_t given Z^{t-1}. Also, given $Z_t = (Y_t, X_t')'$ and using the formula that $P(A \cap B|C) = P(A|B \cap C)P(B|C)$ for any events A, B and C, we have

$$\begin{aligned} f_{Z_t|Z^{t-1}}(z_t|z^{t-1}) &= f_{Y_t|(X_t, Z^{t-1})}(y_t|x_t, z^{t-1}) f_{X_t|Z^{t-1}}(x_t|z^{t-1}) \\ &= f_{Y_t|\Psi_t}(y_t|\Psi_t) f_{X_t|Z^{t-1}}(x_t|z^{t-1}), \end{aligned}$$

where

$$\Psi_t = (X_t', Z^{t-1\prime})',$$

an extended information set which contains not only the past history Z^{t-1} but also the current X_t. It follows that

$$\begin{aligned} f_{Z^n}(z^n) &= \prod_{t=1}^{n} f_{Y_t|\Psi_t}(y_t|\Psi_t) f_{X_t|Z^{t-1}}(x_t|z^{t-1}) \\ &= \prod_{t=1}^{n} f_{Y_t|\Psi_t}(y_t|\Psi_t) \prod_{t=1}^{n} f_{X_t|Z^{t-1}}(x_t|z^{t-1}). \end{aligned}$$

Often, the interest is in modelling the conditional distribution of Y_t given $\Psi_t = (X_t, Z^{t-1})'$.

Some Important Special Cases

Case 1 [Cross-Sectional Observations]: Suppose $\{Z_t\}$ is i.i.d. Then $f_{Y_t|\Psi_t}(y_t|x_t, z^{t-1}) = f_{Y_t|X_t}(y_t|x_t)$ and $f_{X_t|Z^{t-1}}(x_t|z^{t-1}) = f_{X_t}(x_t)$. It follows that

$$f_{Z^n}(z^n) = \prod_{t=1}^{n} f_{Y_t|X_t}(y_t|x_t) \prod_{t=1}^{n} f_{X_t}(x_t),$$

where $f_{X_t}(x_t)$ is the marginal pdf/pmf of X_t.

Case 2: [Univariate Time Series Analysis] Suppose X_t does not exist, namely $Z_t = Y_t$. Then $\Psi_t = (X_t', Z^{t-1\prime})' = Z^{t-1} = (Y_{t-1}, ..., Y_1)'$, and as a consequence,

$$f_{Z^n}(z^n) = \prod_{t=1}^{n} f_{Y_t|Y^{t-1}}(y_t|y^{t-1}).$$

Variation-Free Parameters Assumption

We assume a parametric conditional probability model

$$f_{Z_t|Z^{t-1}}(z_t|z^{t-1}) = f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta) f_{X_t|Z^{t-1}}(x_t|z^{t-1}, \gamma),$$

where $f_{Y_t|\Psi_t}(\cdot|\Psi_t, \beta)$ is a known functional form up to some unknown $K \times 1$ parameter vector β^o, and $f_{X_t|Z^{t-1}}(\cdot|\Psi_t, \gamma)$ is a known or unknown parametric function with some unknown parameter γ. Note that $f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta)$ is a function of β rather than γ while $f_{X_t|Z^{t-1}}(x_t|z^{t-1}, \gamma)$ is a function of γ rather than β. This is called a variation free parameters assumption. It follows that the model log-likelihood function

$$\begin{aligned} \ln f_{Z^n}(z^n) &= \sum_{t=1}^{n} \ln f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta) \\ &\quad + \sum_{t=1}^{n} \ln f_{X_t|Z^{t-1}}(x_t|z^{t-1}, \gamma). \end{aligned}$$

If we are interested in using the extended information set $\Psi_t = (X_t', Z^{t-1\prime})'$ to predict the distribution of Y_t, then β is called the **parameter of interest**, and γ is called the **nuisance parameter**. In this case, to estimate β, we only need to focus on modelling the conditional pdf/pmf $f_{Y_t|\Psi_t}(y|\Psi_t, \beta)$. This follows because the second part of the likelihood function does not depend on β so that the maximization of $\ln f_{Z^n}(z^n)$ with respect to β is equivalent to the maximization of the first part of the likelihood with respect to β.

We now introduce various conditional distributional models. For simplicity, we only consider i.i.d. observations so that $f_{Y_t|\Psi_t}(y|\Psi_t, \beta) = f_{Y_t|X_t}(y|X_t, \beta)$.

Example 1 [Linear Regression Model with Normal Errors]: Suppose $Z_t = (Y_t, X_t')'$ is i.i.d., $Y_t = X_t'\alpha^o + \varepsilon_t$, where $\varepsilon_t|X_t \sim N(0, \sigma_o^2)$. Then the conditional pdf of $Y_t|X_t$ is

$$f_{Y_t|X_t}(y|x, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - x'\alpha)^2},$$

where $\beta = (\alpha', \sigma^2)'$. This is a classical linear regression model discussed in Chapter 3.

Example 2 [Logit Model]: Suppose $Z_t = (Y_t, X_t')'$ is i.i.d., Y_t is a binary random variable taking either value 1 or value 0, and

$$P(Y_t = y_t|X_t) = \begin{cases} \psi(X_t'\beta^o) & \text{if } y_t = 1, \\ 1 - \psi(X_t'\beta^o) & \text{if } y_t = 0, \end{cases}$$

where

$$\psi(u) = \frac{1}{1 + \exp(-u)}, \qquad -\infty < u < \infty,$$

is the CDF of the logistic distribution. We have

$$f_{Y_t|X_t}(y_t|X_t, \beta) = \psi(X_t'\beta)^{y_t}[1 - \psi(X_t'\beta)]^{1-y_t}.$$

Example 3 [Probit Model]: Suppose $Z_t = (Y_t, X_t')'$ is i.i.d., and Y_t is a binary random variable such that

$$P(Y_t = y_t|X_t) = \begin{cases} \Phi(X_t'\beta^o) & \text{if } y_t = 1 \\ 1 - \Phi(X_t'\beta^o) & \text{if } y_t = 0, \end{cases}$$

where $\Phi(\cdot)$ is the CDF of the N(0,1) distribution. We have

$$f_{Y_t|X_t}(y_t|X_t, \beta) = \Phi(X_t'\beta)^{y_t}[1 - \Phi(X_t'\beta)]^{1-y_t}.$$

There are wide applications of the logit and probit models. For example, a consumer chooses a particular brand of car; a student decides to go to PHD study, etc.

Example 4 [Censored regression (Tobit) Models]: A dependent variable Y_t is called censored when the response Y_t cannot take values below (left censored) or above (right censored) a certain threshold value. For example, the investment can only be zero or positive (when no borrowing is allowed). The censored data are mixed continuous-discrete. Suppose the data generating process is

$$Y_t^* = X_t'\alpha^o + \varepsilon_t,$$

where $\{\varepsilon_t\} \sim i.i.d.N(0, \sigma_o^2)$. When $Y_t^* > c$, we observe $Y_t = Y_t^*$. When $Y_t^* \leq c$, we only have the record $Y_t = c$. The parameter α^o should not be estimated by regressing Y_t on X_t based on the subsample with $Y_t > c$, because the data with $Y_t = c$ contain relevant information about α^o and σ_o^2. More importantly, in the subsample with $Y_t > c$, ε_t is a truncated distribution with nonzero mean (i.e., $E(\varepsilon_t|Y_t > c) \neq 0$ and $E(X_t\varepsilon_t|Y_t > c) \neq 0$). Therefore, OLS is not consistent for α^o if one only uses the subsample consisting of observations of $Y_t > c$ and throw away observations with $Y_t = c$.

Question: How to estimate α^o given an observed sample $\{Y_t, X_t'\}_{t=1}^n$ where some observations of Y_t are censored? Suppose $Z_t = (Y_t, X_t')'$ is i.i.d., with the observed dependent variable

$$Y_t = \begin{cases} Y_t^* & \text{if } Y_t^* > c \\ c & \text{if } Y_t^* \leq c, \end{cases}$$

where $Y_t^* = X_t'\alpha^o + \varepsilon_t$ and $\varepsilon_t|X_t \sim i.i.d.N(0, \sigma_o^2)$. We assume that the threshold c is known. Then we can write

$$\begin{aligned} Y_t &= \max(Y_t^*, c) \\ &= \max(X_t'\alpha^o + \varepsilon_t, c). \end{aligned}$$

Define a dummy variable indicating whether $Y_t^* > c$ or $Y_t^* \leq c$,

$$D_t = \begin{cases} 1 & \text{if } Y_t > c \text{ (i.e., if } Y_t^* > c) \\ 0 & \text{if } Y_t = c \text{ (i.e., if } Y_t^* \leq c). \end{cases}$$

Then the pdf of $Y_t|X_t$ is

$$\begin{aligned} & f_{Y_t|X_t}(y_t|x_t, \beta) \\ = & \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t - x_t'\alpha)^2}\right]^{D_t} \\ & \times \left[\Phi\left(\frac{c - x_t'\alpha}{\sigma}\right)\right]^{1-D_t}, \end{aligned}$$

where $\Phi(\cdot)$ is the $N(0,1)$ CDF, and the second part is the conditional probability

$$\begin{aligned} P(Y_t &= c|X_t) \\ &= P(Y_t^* \leq c|X_t) \\ &= P(\varepsilon_t \leq c - X_t'\alpha|X_t) \\ &= P\left(\frac{\varepsilon_t}{\sigma} \leq \frac{c - X_t'\alpha}{\sigma}|X_t\right) \\ &= \Phi\left(\frac{c - X_t'\alpha}{\sigma}\right), \end{aligned}$$

given $\frac{\varepsilon_t}{\sigma}|X_t \sim N(0,1)$.

Question: Can you give some examples where this model can be applied?

One example is a survey on unemployment spells. At the terminal date of the survey, the recorded time length of an unemployed worker is not the duration when his layoff will last. Another example is a survey on cancer patients. Those who have survived up to the ending date of the survey will usually live longer than the survival duration recorded.

Example 6 [Truncated Regression Models]: A random sample is called truncated if we know before hand that observations can come only from a restricted part of the underlying population distribution. The truncation can come from below, from above, or from both sides. We now consider an example where the truncation is from below with a known truncation point. More specifically, assume that the data generating process is

$$Y_t^* = X_t'\alpha^o + \varepsilon_t,$$

where $\varepsilon_t|X_t \sim i.i.d. N(0, \sigma_o^2)$. Suppose only those of Y_t^* whose values are larger than or equal to constant c are observed, where c is known. That is, we observe $Y_t = Y_t^*$ if and only if $Y_t^* = X_t'\alpha^o + \varepsilon_t \geq c$. The observations with $Y_t^* < c$ are not recorded. Assume the resulting sample is $\{Y_t, X_t\}_{t=1}^n$, where $\{Y_t, X_t\}$ is i.i.d. We now analyze the effect of truncation for this model. For the observed sample, $Y_t^* \geq c$ and so ε_t comes from the truncated version of the distribution $N(0, \sigma_o^2)$ with $\varepsilon_t \geq c - X_t'\alpha^o/\sigma_o$. It follows that $E(X_t\varepsilon_t|Y_t^* \geq c) \neq 0$ and therefore the OLS estimator based on the observed sample $\{Y_t, X_t'\}$ is not consistent.

Because the observation Y_t is recorded if and only if $Y_t^* \geq c$, the conditional probability distribution of Y_t given X_t is the same as the probability distribution of Y_t^* given X_t and $Y_t^* > c$.

Hence, for any observed sample point (y_t, x_t), we have

$$\begin{aligned}
f_{Y_t|X_t}(y_t|x_t, \beta) &= f_{Y_t^*|X_t,(Y_t^*>c)}(y_t|x_t, Y_t^* > c) \\
&= \frac{f_{Y_t^*|X_t,(Y_t^*>c)}(y_t|x_t, Y_t^* > c)P(Y_t^* > c|x_t)}{P(Y_t^* > c|x_t)} \\
&= \frac{f_{Y_t^*|X_t}(y_t|x_t)}{P(Y_t^* > c|x_t)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_t - x_t'\alpha)^2} \\
&\quad \times \frac{1}{1 - \Phi\left(\frac{c - x_t'\alpha}{\sigma}\right)},
\end{aligned}$$

where $\beta = (\alpha', \sigma^2)$,and the conditional probability

$$\begin{aligned}
P(Y_t^* &> c|X_t) = 1 - P(Y_t^* \leq c|X_t) \\
&= 1 - P\left(\frac{\varepsilon_t}{\sigma} \leq \frac{c - X_t'\alpha}{\sigma}|X_t\right) \\
&= 1 - \Phi\left(\frac{c - X_t'\alpha}{\sigma}\right).
\end{aligned}$$

Question: Can you give some examples where this model can be applied?

Example 1 [Loan applications]: Only those successful loan applications will be recorded.

Example 2 [Students and Examination Scores]:

Suppose we are interested in investigating how the examination scores of students depend on their effort, family support, and high schools, and we have a sample from those who have been admitted to colleges. This sample is obviously a truncated sample because we do not observe those who are not admitted to colleges because their scores are below certain minimum requirements.

Question: How to estimate β in a conditional distribution model $f_{Y_t|\Psi_t}(y|\Psi_t, \beta)$?

We first introduce the likelihood function.

Definition [Likelihood Function]: *The joint pdf/pmf of the random sample $Z^n = (Z_1, Z_2, ..., Z_n)$ as a function of (β, γ)*

$$L_n(\beta, \gamma; z^n) = f_{Z^n}(z^n, \beta, \gamma)$$

is called the likelihood function of Z^n when z^n is observed. Moreover, $\ln L_n(\beta, \gamma, z^n)$ is called the log-likelihood function of Z^n when z^n is observed.

Remarks:

The likelihood function $L_n(\beta, \gamma; z^n)$ is algebraically identical to the joint probability density function $f_{Z^n}(z^n, \beta, \gamma)$ of the random sample Z^n taking value z^n. Thus, given (β, γ), $L_n(\beta, \gamma; z^n)$ can be viewed as a measure of the probability or likelihood with which the observed sample z^n will occur.

Lemma [Variation-Free Parameter Spaces]: *Suppose β and γ are variation-free over parameter spaces $\Theta \times \Gamma$, in the sense that for all $(\beta, \gamma) \in \Theta \times \Gamma$, we have*

$$f_{Z_t|\Psi_t}(z_t|\Psi_t, \beta, \gamma) = f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta) f_{X_t|Z^{t-1}}(x_t|Z^{t-1}, \gamma),$$

where $\Psi_t = (X_t', Z^{t-1\prime})'$. Then the likelihood function of Z^n given $Z^n = z^n$ can be written as

$$L_n(\beta, \gamma; z^n) = \prod_{t=1}^n f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta) \prod_{t=1}^n f_{X_t|Z^{t-1}}(x_t|Z^{t-1}, \gamma),$$

and the log-likelihood function

$$\begin{aligned} \ln L_n(\beta, \gamma; z^n) &= \sum_{t=1}^n \ln f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta) \\ &\quad + \sum_{t=1}^n \ln f_{X_t|Z^{t-1}}(x_t|Z^{t-1}, \gamma). \end{aligned}$$

Suppose we are interested in predicting Y_t using the extended information set $\Psi_t = (X_t', Z^{t-1\prime})'$. Then only the first part of the log-likelihood is relevant, and β is called the parameter of interest. The other parameter γ, appearing in the second part of the log-likelihood function, is called the nuisance parameter.

We now define an estimation method based on maximizing the conditional log-likelihood function $\sum_{t=1}^n \ln f_{Y_t|\Psi_t}(y_t|\Psi_t, \beta)$.

Definition [(Quasi-)Maximum Likelihood Estimator for Parameters of Interest β; (Q)MLE]: *The MLE $\hat{\beta}$ for $\beta \in \Theta$ is defined as*

$$\begin{aligned} \hat{\beta} &= \arg\max_{\beta \in \Theta} \prod_{t=1}^n f_{Y_t|\Psi_t}(Y_t|\Psi_t, \beta) \\ &= \arg\max_{\beta \in \Theta} \sum_{t=1}^n \ln f_{Y_t|\Psi_t}(Y_t|\Psi_t, \beta), \end{aligned}$$

where Θ is a parameter space. When the conditional probability distribution model $f_{Y_t|\Psi_t}(y|\Psi_t, \beta)$ is correctly specified in the sense that there exists some parameter value $\beta \in \Theta$ such that

$f_{Y_t|\Psi_t}(y|\Psi_t,\beta)$ coincides with the true conditional distribution of Y_t given Ψ_t, then $\hat{\beta}$ is called the maximum likelihood estimator (MLE); when $f_{Y_t|\Psi_t}(y|\Psi_t,\beta)$ is misspecified in the sense that there exists no parameter value $\beta \in \Theta$ such that $f_{Y_t|\Psi_t}(y|\Psi_t,\beta)$ coincides with the true conditional distribution of Y_t given Ψ_t, $\hat{\beta}$ is called the quasi-maximum likelihood estimator (QMLE).

Remarks:

By the nature of the objective function, the MLE gives a parameter estimate which makes the observed sample z^n most likely to occur. By choosing a suitable parameter $\hat{\beta} \in \Theta$, MLE maximizes the probability that $Z^n = z^n$, that is, the probability that the random sample Z^n takes the value of the observed data z^n. Note that MLE and QMLE may not be unique.

The MLE is obtained over Θ, where Θ may be subject to some restriction. An example is the GARCH model where some parameters have to be restricted in order to ensure that the estimated conditional variance is nonnegative (e.g., Nelson and Cao 1992).

Under regularity conditions, we can characterize the MLE by a first order condition. Like the GMM estimator, However, there is usually no closed form for the MLE $\hat{\beta}$. The solution $\hat{\beta}$ has to be searched by computers. The most popular methods used in economics are BHHH, and Gauss-Newton.

Question: When does the MLE exist?

Suppose the likelihood function is continuous in $\beta \in \Theta$ and parameter space Θ is compact. Then a global maximizer $\hat{\beta} \in \Theta$ exists.

Theorem [Existence of MLE/QMLE] *Suppose for each $\beta \in \Theta$, where Θ is a compact parameter space, $f_{Y_t|\Psi_t}(Y_t|\Psi_t,\beta)$ is a measurable function of (Y_t,Ψ_t), and for each t, $f_{Y_t|\Psi_t}(Y_t|\Psi_t,\cdot)$ is continuous in $\beta \in \Theta$. Then MLE/QMLE $\hat{\beta}$ exists.*

This result is analogous to the Weierss Theorem in multivariate calculus that any continuous function over a compact support always has a maximum and a minimum.

9.3 Statistical Properties of MLE/QMLE

For notational simplicity, from now on we will write the conditional pdf/pmf of Y_t given Ψ_t as

$$f_{Y_t|\Psi_t}(y|\Psi_t,\beta) = f(y|\Psi_t,\beta).$$

We first provide a set of regularity conditions.

Assumptions

Assumption 9.1 [Parametric Distribution Model]: (i) $\{Z_t = (Y_t, X_t')'\}'$ is a stationary ergodic process, and (ii) $f(y_t|\Psi_t, \beta)$ is a conditional pdf/pmf model of Y_t given $\Psi_t = (X_t', Z^{t-1\prime})'$, where $Z^{t-1} = (Z_{t-1}', Z_{t-2}', \cdots, Z_1')'$. For each β, $\ln f(Y_t|\Psi_t, \beta)$ is measurable with respect to observations (Y_t, Ψ_t), and for each t, $\ln f(Y_t|\Psi_t, \cdot)$ is continuous in $\beta \in \Theta$, where Θ is a finite-dimensional parameter space.

Assumption 9.2 [Compactness]: Parameter space Θ is compact.

Assumption 9.3 [Uniform WLLN]: $\{\ln f(Y_t|\Psi_t, \beta) - E \ln f(Y_t|\Psi_t, \beta)\}$ obeys the uniform weak law of large numbers (UWLLN), i.e.,

$$\sup_{\beta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} \ln f(Y_t|\Psi_t, \beta) - l(\beta) \right| \to^p 0$$

where the population log-likelihood function

$$l(\beta) = E\left[\ln f(Y_t|\Psi_t, \beta)\right]$$

is continuous in $\beta \in \Theta$.

Assumption 9.4 [Identification]:

$$\beta^* = \arg\max_{\beta \in \Theta} l(\beta)$$

is the unique maximizer of $l(\beta)$ over Θ.

Question: What is the interpretation of β^*?

Assumption 9.4 is an identification condition which states that β^* is a unique solution that maximizes $l(\beta)$, the expected value of the logarithmic conditional likelihood function $\ln f(Y_t|\Psi_t, \beta)$. So far, there is no economic interpretation for β^*. This is analogous to the best linear least squares approximation coefficient $\beta^* = \arg\min_\beta E(Y - X'\beta)^2$ in Chapter 2.

Consistency

We first consider the consistency property of $\hat{\beta}$ for β^*. Because we assume that Θ is compact, $\hat{\beta}$ and β^* may be corner solutions. Thus, we have to use the extrema estimator lemma to prove the consistency of the MLE/QMLE $\hat{\beta}$.

Theorem [Consistency of MLE/QMLE]: *Suppose Assumptions 9.1–9.4 hold. Then as $n \to \infty$,*

$$\hat{\beta} - \beta^* \to^p 0.$$

Proof: Applying the extrema estimator lemma in Chapter 8, with

$$\hat{Q}(\beta) = n^{-1} \sum_{t=1}^{n} \ln f(Y_t|\Psi_t, \beta)$$

and

$$Q(\beta) = l(\beta) \equiv E[\ln f(Y_t|\Psi_t, \beta)].$$

Assumptions 9.1–9.4 ensure that all conditions for $\hat{Q}(\beta)$ and $Q(\beta)$ in the extrema estimator lemma are satisfied. It follows that $\hat{\beta} \to^p \beta^*$ as $n \to \infty$.

Model Specification and Interpretation of β^*

Definition [Correct Specification for Conditional Distribution] The model $f(y_t|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t if there exists some parameter value $\beta^o \in \Theta$ such that $f(y_t|\Psi_t, \beta^o)$ coincides with the true conditional pdf/pmf of Y_t given Ψ_t.

Under correct specification of $f(y|\Psi_t, \beta)$, the parameter value β^o is usually called the true model parameter value. It will usually have economic interpretation.

Question: What are the implications of correct specification of a conditional distributional model $f(y|\Psi_t, \beta)$?

Lemma: *Suppose Assumption 9.4 holds, and the model $f(y_t|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Then $f(y_t|\Psi_t, \beta^*)$ coincides with the true conditional pdf/pmf $f(y_t|\Psi_t, \beta^o)$ of Y_t given Ψ_t, where β^* is as given in Assumption 9.4. In other words, the population likelihood maximizer β^* coincides with the true parameter value β^o when the model $f(y_t|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t.*

Proof: Because $f(y|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t, there exists some $\beta^o \in \Theta$ such that

$$\begin{aligned} l(\beta) &= E[\ln f(Y_t|\Psi_t, \beta)] \\ &= E\{E[\ln f(Y_t|\Psi_t, \beta)|\Psi_t]\} \text{ by LIE} \\ &= E \int \ln[f(y|\Psi_t, \beta)] f(y|\Psi_t, \beta^o) dy, \end{aligned}$$

where the second equality follows by LIE and the expectation $E(\cdot)$ in the third equality is taken with respect to the true distribution of the random variables in Ψ_t.

By Assumption 9.4, we have $l(\beta) \leq l(\beta^*)$ for all $\beta \in \Theta$. By the law of iterated expectations, it follows that

$$\begin{aligned} & E\int \ln[f(y|\Psi_t, \beta)]f(y|\Psi_t, \beta^o)dy \\ \leq \; & E\int \ln[f(y|\Psi_t, \beta^*)]f(y|\Psi_t, \beta^o)dy, \end{aligned}$$

where $f(y_t|\Psi_t, \beta^o)$ is the true conditional pdf/pmf. Hence, by choosing $\beta = \beta^o$, we have

$$\begin{aligned} & E\int \ln[f(y|\Psi_t, \beta^o)]f(y|\Psi_t, \beta^o)dy \\ \leq \; & E\int \ln[f(y|\Psi_t, \beta^*)]f(y|\Psi_t, \beta^o)dy. \end{aligned}$$

On the other hand, by Jensen's inequality and the concavity of the logarithmic function, we have

$$\begin{aligned} & \int \ln\left[\frac{f(y|\Psi, \beta^*)}{f(y|\Psi_t, \beta^o)}\right] f(y|\Psi_t, \beta^o)dy \\ \leq \; & \ln\left\{\int \left[\frac{f(y|\Psi, \beta^*)}{f(y|\Psi_t, \beta^o)}\right] f(y|\Psi_t, \beta^o)dy\right\} \\ = \; & \ln\left\{\int f(y|\Psi, \beta^*)dy\right\} \\ = \; & \ln(1) \\ = \; & 0, \end{aligned}$$

where we have made use of the fact that $\int f(y|\Psi_t, \beta)dy = 1$ for all $\beta \in \Theta$. Therefore, we have

$$\begin{aligned} & \int \ln\left[f(y|\Psi_t, \beta^*)\right] f(y|\Psi_t, \beta^o)dy \\ \leq \; & \int \ln[f(y|\Psi_t, \beta^*)]f(y|\Psi_t, \beta^o)dy. \end{aligned}$$

Therefore, by taking the expectation with respect to the distribution of Ψ_t, we obtain

$$\begin{aligned} & E\int \ln[f(y|\Psi_t, \beta^*)]f(y|\Psi_t, \beta^o)dy \\ \leq \; & E\int \ln[f(y|\Psi_t, \beta^o)]f(y|\Psi_t, \beta^o)dy. \end{aligned}$$

It follows that we must have $\beta^* = \beta^o$; otherwise β^* cannot be the the maximizer of $l(\beta)$ over Θ. This completes the proof.

Remark:

This lemma provides an interpretation of β^* in Assumption 9.4. That is, the population likelihood maximizer β^* coincides with the true model parameter β^o when $f(y|\Psi_t, \beta)$ is correctly specified. Thus, by maximizing the population model log-likelihood function $l(\beta)$, we can obtain the true parameter value β^o.

Implication of Correct Model Specification

We now examine some important implications of correct model specification. For this purpose, we assume that β^o is an interior point of the parameter space Θ, so that we can impose differentiability condition on the log-likelihood function $\ln f(y|\Psi_t, \beta)$ at β^o:

Assumption 9.5: $\beta^o \in \text{int}(\Theta)$.

Question: Why do we need this assumption? This assumption is needed for the purpose of taking a Taylor series expansion.

We first state an important implication of a correctly specified conditional distribution model for Y_t given Ψ_t.

Lemma [The MDS Property of the Score Function of a Correctly Specified Conditional Distribution Model]: *Suppose that for each t, $\ln f(Y_t|\Psi_t, \cdot)$ is continuously differentiable with respect to $\beta \in \Theta$. Define a $K \times 1$ score function*

$$S_t(\beta) = \frac{\partial}{\partial \beta} \ln f(y_t|\Psi_t, \beta).$$

If $f(y|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t, then

$$E\left[S_t(\beta^o)|\Psi_t\right] = 0 \text{ a.s.},$$

where β^o is as in Assumption 9.4 and satisfies Assumption 9.5, and $E(\cdot|\Psi_t)$ is the expectation taken over the true conditional distribution of Y_t given Ψ_t.

Proof: Note that for any given $\beta \in \Theta$, $f(y|\Psi_t, \beta)$ is a valid pdf. Thus we have

$$\int_{-\infty}^{\infty} f(y|\Psi_t, \beta) dy = 1.$$

When $\beta \in \text{int}(\Theta)$, by differentiation, we have

$$\frac{\partial}{\partial \beta} \int_{-\infty}^{\infty} f(y|\Psi_t, \beta) dy = 0.$$

By exchanging differentiation and integration (assume that we can do so), we have

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \beta} f(y|\Psi_t, \beta) dy = 0,$$

which can be further written as

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t, \beta)}{\partial \beta} f(y|\Psi_t, \beta) dy = 0.$$

This relationship holds for all $\beta \in int(\Theta)$, including β^o. It follows that

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t, \beta^o)}{\partial \beta} f(y|\Psi_t, \beta^o) dy = 0,$$

where

$$\frac{\partial \ln f(y|\Psi_t, \beta^o)}{\partial \beta} = \frac{\partial \ln f(y|\Psi_t, \beta)}{\partial \beta}|_{\beta=\beta^o}.$$

Because $f(y|\Psi_t, \beta^o)$ is the true conditional pdf/pmf of Y_t given Ψ_t when $f(y|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t, we have

$$E[S_t(\beta^o)|\Psi_t] = 0.$$

This completes the proof.

Note that $E[S_t(\beta^o)|\Psi_t] = 0$ implies that $E[S_t(\beta^o)|Z^{t-1}] = 0$, namely $\{S_t(\beta^o)\}$ is an MDS.

Question: Suppose $E[S_t(\beta^o)|\Psi_t] = 0$ for some $\beta^o \in \Theta$. Can we claim that the conditional pdf/pmd model is correctly specified?

Answer: No. The MDS property is one of many implications of correct model specification. In certain sense, the MDS property is equivalent to correct specification of the conditional mean. Misspecification of $f(y|\Psi_t, \beta)$ may occur in higher order conditional moments of Y_t given Ψ_t. Below is an example in which $\{S_t(\beta^o)\}$ is MDS but the model $f(y_t|\Psi_t, \beta)$ is misspecified.

Example 1: Suppose $\{Y_t\}$ is a univariate time series process such that

$$Y_t = \mu_t(\beta) + \sigma_t(\beta) z_t,$$

where $\mu_t(\beta^o) = E(Y_t|I_{t-1})$ for some β^o and $I_{t-1} = (Y_{t-1}, Y_{t-2}, ..., Y_1)$ but $\sigma_t^2(\beta) \neq \text{var}(Y_t|I_{t-1})$ for all β. Then, correct model specification for the conditional mean $E(Y_t|I_{t-1})$ implies that

$E(z_t|I_{t-1})=0$. Assume that $\{z_t\}\sim$ i.i.d.$N(0,1)$. Then the conditional probability density model

$$f(y|\Psi_t,\beta)=\frac{1}{\sqrt{2\pi\sigma_t^2(\beta)}}\exp\left[-\frac{(Y_t-\mu_t(\beta))^2}{2\sigma_t^2(\beta)}\right],$$

where $\Psi_t=I_{t-1}$. It is straightforward to verify that

$$E\left[S_t(\beta^o)|\Psi_t\right]=E[S_t(\beta^o)|I_{t-1}]=0,$$

although the conditional variance $\sigma_t^2(\beta)$ is misspecified for $\text{var}(Y_t|I_{t-1})$.

Next, we state another important implication of a correctly specified conditional distribution model for Y_t given Ψ_t.

Lemma [Conditional Information Matrix Equality]: *Suppose Assumptions 9.1–9.5 hold, $f(y|\Psi_t,\beta)$ is twice continuously differentiable with respect to $\beta\in int(\Theta)$, and $f(y_t|\Psi_t,\beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Then*

$$E\left[S_t(\beta^o)S_t(\beta^o)'+H_t(\beta^o)|\Psi_t\right]=0,$$

where

$$\begin{aligned}H_t(\beta) &\equiv \frac{d}{d\beta}S_t(\beta)\\ &= \frac{\partial^2}{\partial\beta\partial\beta'}\ln f(Y_t|\Psi_t,\beta),\end{aligned}$$

or equivalently,

$$\begin{aligned}& E\left[\frac{\partial}{\partial\beta}\ln f(Y_t|\Psi_t,\beta^o)\frac{\partial}{\partial\beta'}\ln f(Y_t|\Psi_t,\beta^o)\right|\Psi_t\left.\vphantom{\frac{\partial}{\partial\beta}}\right]\\ =& -E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln f(Y_t|\Psi_t,\beta^o)\right|\Psi_t\left.\vphantom{\frac{\partial^2}{\partial\beta}}\right].\end{aligned}$$

Proof: For all $\beta\in\Theta$, we have

$$\int_{-\infty}^{\infty}f(y|\Psi_t,\beta)dy=1.$$

By differentiation with respect to $\beta\in int(\Theta)$, we obtain

$$\frac{\partial}{\partial\beta}\int_{-\infty}^{\infty}f(y|\Psi_t,\beta)dy=0.$$

Exchanging differentiation and integration, we have

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial f(y|\Psi_t,\beta)}{\partial \beta} dy &= 0, \\ \int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta} f(y|\Psi_t,\beta) dy &= 0.\end{aligned}$$

With further differentiation of the above equation again, we have

$$\begin{aligned}& \frac{\partial}{\partial \beta} \int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta} f(y|\Psi_t,\beta) dy \\ = & \int_{-\infty}^{\infty} \frac{\partial}{\partial \beta} \left[\frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta} f(y|\Psi_t,\beta) \right] dy \\ = & \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(y|\Psi_t,\beta)}{\partial \beta \partial \beta'} f(y|\Psi_t,\beta) dy \\ & + \int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta} \frac{\partial f(y|\Psi_t,\beta)}{\partial \beta'} dx \\ = & \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(y|\Psi_t,\beta)}{\partial \beta \partial \beta'} f(y|\Psi_t,\beta) dy \\ & + \int_{-\infty}^{\infty} \frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta} \frac{\partial \ln f(y|\Psi_t,\beta)}{\partial \beta'} f(y|\Psi_t,\beta) dy \\ = & \ 0.\end{aligned}$$

The above relation holds for all $\beta \in \Theta$, including β^o. This and the fact that $f(y|\Psi_t,\beta^o)$ is the true conditional pdf/pmf of Y_t given Ψ_t imply the desired conditional information matrix equality stated in the lemma. This completes the proof.

Remarks:

The $K \times K$ matrix

$$\begin{aligned}& E[S_t(\beta^o)S_t(\beta^o)'|\Psi_t] \\ = & \ E\left[\frac{\partial \ln f(Y_t|\Psi_t,\beta^o)}{\partial \beta} \frac{\partial \ln f(Y_t|\Psi_t,\beta^o)}{\partial \beta'} \middle| \Psi_t \right]\end{aligned}$$

is called the conditional Fisher's information matrix of Y_t given Ψ_t. It measures the content of the information contained in the random variable Y_t conditional on Ψ_t. The larger the expectation is, the more information Y_t contains.

Question: What is the implication of the conditional information matrix equality?

In certain sense, the IM equality could be viewed as equivalent to correct specification of conditional variance. It has important implications on the form of the asymptotic variance of the MLE. More specifically, the IM equality will simplify the asymptotic variance of the MLE in the same way as conditional homoskedasticity simplifies the asymptotic variance of the OLS estimator.

To investigate the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$, we need the following conditions.

Assumption 9.6: (i) For each t, $\ln f(y_t|\Psi_t, \cdot)$ is continuously twice differentiable with respect to $\beta \in \Theta$; (ii) $\{S_t(\beta^o)\}$ obeys a CLT, i.e.,

$$\sqrt{n}\hat{S}(\beta^o) \equiv n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o) \to^d N(0, V_o)$$

for some $K \times K$ matrix $V_o \equiv \text{avar}[n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)]$ which is symmetric, finite and positive definite; (iii) $\{H_t(\beta) \equiv \frac{\partial^2}{\partial\beta\partial\beta'} \ln f(y_t|\Psi_t, \beta)\}$ obeys a uniform weak law of large numbers (UWLLN) over Θ. That is, as $n \to \infty$,

$$\sup_{\beta\in\Theta}\left\| n^{-1}\sum_{t=1}^{n} H_t(\beta) - H(\beta)\right\| \to^p 0,$$

where the $K \times K$ Hessian matrix

$$\begin{aligned} H(\beta) &\equiv E\left[H_t(\beta)\right] \\ &= E\left[\frac{\partial^2 \ln f(Y_t|\Psi_t, \beta)}{\partial\beta\partial\beta'}\right] \end{aligned}$$

symmetric, finite and nonsingular, and is continuous in $\beta \in \Theta$.

Question: What is the form of the asymptotic variance V of $\sqrt{n}\hat{S}(\beta^o)$ when $f(y|\Psi_t, \beta)$ is correctly specified?

By the stationary MDS property of $S_t(\beta^o)$ with respect to Ψ_t, we have

$$\begin{aligned} V_o &\equiv \text{avar}\left[n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)\right] \\ &= E\left\{\left[n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)\right]\left[n^{-1/2}\sum_{\tau=1}^{n} S_\tau(\beta^o)\right]'\right\} \\ &= n^{-1}\sum_{t=1}^{n}\sum_{\tau=1}^{n} E[S_t(\beta^o)S_\tau(\beta^o)'] \\ &= E[S_t(\beta^o)S_t(\beta^o)'], \end{aligned}$$

where the expectations of cross-products, $E[S_t(\beta^o)S_\tau(\beta^o)']$, are identically zero for all $t \neq \tau$, as implied by the MDS property of $\{S_t(\beta^o)\}$ from the Lemma on the score function.

Furthermore, from the conditional information matrix equality, we have

$$\begin{aligned} V_o &= E[S_t(\beta^o)S_t(\beta^o)'] \\ &= -H_o. \end{aligned}$$

Note that H_o is a $K \times K$ symmetric negative definite matrix.

Asymptotic Distribution

Next, we derive the asymptotic normality of the MLE.

Theorem [Asymptotic Normality of MLE]: *Suppose Assumptions 9.1–9.6 hold, and $f(y_t|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, -H_o^{-1}).$$

Proof: Because β^o is an interior point in Θ and $\hat{\beta} - \beta^o \rightarrow^p 0$ as $n \rightarrow \infty$, we have $\hat{\beta} \in int(\Theta)$ for n sufficiently large. It follows that the FOC of maximizing the log-likelihood holds when n is sufficiently large:

$$\begin{aligned} \hat{S}(\hat{\beta}) &\equiv n^{-1}\sum_{t=1}^{n} \frac{\partial \ln f(Y_t|\Psi_t, \hat{\beta})}{\partial \beta} \\ &= n^{-1}\sum_{t=1}^{n} S_t(\hat{\beta}) \\ &= 0. \end{aligned}$$

The FOC provides a link between MLE and GMM: MLE can be viewed as a GMM estimation with the moment condition

$$E[m_t(\beta^o)] = E[S_t(\beta^o)] = 0 \text{ for some } \beta^o$$

in an exact identification case.

By the first order Taylor series expansion of $\hat{S}(\hat{\beta})$ around the true parameter β^o, we have

$$\begin{aligned} 0 &= \sqrt{n}\hat{S}(\hat{\beta}) \\ &= \sqrt{n}\hat{S}(\beta^o) + \hat{H}(\bar{\beta})\sqrt{n}(\hat{\beta} - \beta^o), \end{aligned}$$

where $\bar{\beta}$ lies between $\hat{\beta}$ and β^o, namely, $\bar{\beta} = a\hat{\beta} + (1-a)\beta^o$ for some $a \in [0,1]$, and

$$\begin{aligned}\hat{H}(\beta) &= n^{-1}\sum_{t=1}^{n} H_t(\beta) \\ &= n^{-1}\sum_{t=1}^{n} \frac{\partial^2 \ln f(Y_t|\Psi_t,\beta)}{\partial\beta\partial\beta'}\end{aligned}$$

is the derivative of $\hat{S}(\beta)$. Given that $\hat{\beta} - \beta^o \to^p 0$, we have

$$\begin{aligned}||\bar{\beta} - \beta^o|| &= ||a(\hat{\beta} - \beta^o)|| \le ||\hat{\beta} - \beta^o|| \\ &\to {}^p 0.\end{aligned}$$

Also, by the triangle inequality, the UWLLN for $\{H_t(\beta)\}$ over Θ and the continuity of $H(\beta)$, we obtain

$$\begin{aligned}&\left\|\hat{H}(\bar{\beta}) - H_0\right\| \\ =& \left\|\hat{H}(\bar{\beta}) - H(\bar{\beta}) + H(\bar{\beta}) - H(\beta^o)\right\| \\ \le& \sup_{\beta\in\Theta}\left\|\hat{H}(\beta) - H(\beta)\right\| + \left\|H(\bar{\beta}) - H(\beta^o)\right\| \\ \to& {}^p 0.\end{aligned}$$

Because H_0 is nonsingular, so is $\hat{H}(\bar{\beta})$ for n sufficiently large. Therefore, from FOC we have

$$\sqrt{n}(\hat{\beta} - \beta^o) = -\hat{H}^{-1}(\bar{\beta})\sqrt{n}\hat{S}(\beta^o)$$

for n sufficiently large. [Compare with the OLS estimator $\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}^{-1}\sqrt{n}\frac{X'\varepsilon}{n}$.]

Next, we consider $\sqrt{n}\hat{S}(\beta^o)$. By the CLT, we have

$$\sqrt{n}\hat{S}(\beta^o) \xrightarrow{d} N(0, V_o),$$

where, as we have shown above,

$$\begin{aligned}V_o &\equiv \text{avar}\left[\sqrt{n}\hat{S}(\beta^o)\right] \\ &= E[S_t(\beta^o)S_t(\beta^o)']\end{aligned}$$

given that $\{S_t(\beta^o)\}$ is an MDS with respect to Ψ_t.

It follows by the Slutsky theorem that

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta^o) &= -\hat{H}^{-1}(\bar{\beta})\sqrt{n}\hat{S}(\beta^o) \\ &\xrightarrow{d} N(0, H_o^{-1}V_oH_o^{-1}) \\ &\sim N(0, -H_o^{-1})\end{aligned}$$

or equivalently

$$\sqrt{n}(\hat{\beta}-\beta^o) \xrightarrow{d} N(0, H_o^{-1}V_oH_o^{-1}) \sim N(0, V_o^{-1})$$

using the information matrix equality $V_o = E[S_t(\beta^o)S_t(\beta^o)'] = -H_o$. This completes the proof. ■

Remarks:

Now it is easy to understand why $V_o = E[S_t(\beta^o)S_t(\beta^o)'] = -H_o$ is called the information matrix of Y_t given Ψ_t. The larger $-H_o$ is, the smaller the variance of $\hat{\beta}$ is (i.e., the more precise the estimator $\hat{\beta}$ is). Intuitively, as a measure of the curvature of the population log-likelihood function, the absolute value of the magnitude of H_o characterizes the sharpness of the peak of the population log-likelihood function at β^o.

The simplification of $H_o^{-1}V_oH_o^{-1}$ to $-H_o^{-1}$ by the information matrix equality is similar in spirit to the case of the asymptotic variance of the OLS estimator under conditional homoskedasticity.

Efficiency of MLE

From statistics theory, it is well-known that the asymptotic variance of MLE $\hat{\beta}$ achieves the Cramer-Rao lower bound. Therefore, the MLE $\hat{\beta}$ is asymptotically most efficient.

Question: What is the Cramer-Rao lower bound?

We now discuss consistent estimation of the asymptotic variance-covariance matrix of MLE.

Consistent Estimation of the Asymptotic Variance of the MLE

Because $\text{avar}(\sqrt{n}\hat{\beta}) = V_o^{-1} = -H_o^{-1}$, there are two methods to estimate $\text{avar}[\sqrt{n}(\hat{\beta}-\beta^o)]$.

Method 1: Use $\hat{\Omega} \equiv -\hat{H}^{-1}(\hat{\beta})$, where

$$\hat{H}(\beta) = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 \ln f(Y_t|\Psi_t,\beta)}{\partial\beta\partial\beta'}.$$

This requires taking second derivatives of the log-likelihood function. By Assumption 9.6(iii) and $\hat{\beta} \to^p \beta^o$, we have $\hat{\Omega} \to^p -H_o^{-1}$.

Method 2: Use $\hat{\Omega} \equiv \hat{V}^{-1}$, where

$$\hat{V} \equiv \frac{1}{n}\sum_{t=1}^{n} S_t(\hat{\beta})S_t(\hat{\beta})'.$$

This requires the computation of the first derivatives (i.e., score functions) of the log-likelihood function.

Suppose the $K \times K$ process $\{S_t(\beta)S_t(\beta)'\}$ follows the UWLLN, namely,

$$\sup_{\beta\in\Theta}\left\| n^{-1}\sum_{t=1}^{n} S_t(\beta)S_t(\beta)' - V(\beta)\right\| \to^p 0,$$

where

$$V(\beta) = E[S_t(\beta)S_t(\beta)']$$

is continuous in β. Then if $\hat{\beta} \to^p \beta^o$, we can show that $\hat{V} \to^p V_o$. Note that $V_o = V(\beta^o)$.

Question: Which asymptotic variance estimator (method 1 or method 2) is better in finite samples?

Hypothesis Testing

We now consider the hypothesis of interest

$$\mathbf{H}_0 : R(\beta^o) = r,$$

where $R(\cdot)$ is a $J \times 1$ continuously differentiable vector function with the $J \times K$ matrix $R'(\beta^o)$ being of full rank. We allow both linear and nonlinear restrictions on parameters. Note that in order for $R'(\beta^o)$ to be of full rank, we need the condition that $J \le K$, that is, the number of restrictions is smaller than or at most equal to the number of unknown parameters.

We will introduce three test procedures, namely the Wald test, the Likelihood Ratio (LR) test, and the Lagrange Multiplier (LM) test. We now derive these tests respectively.

Wald Test

By the Taylor series expansion, $\mathbf{H}_0$, and the Slustky theorem, we have

$$\begin{aligned}\sqrt{n}[R(\hat{\beta})-r] &= \sqrt{n}[R(\beta^o)-r] \\ &\quad +R'(\bar{\beta})\sqrt{n}(\hat{\beta}-\beta^o) \\ &= R'(\bar{\beta})\sqrt{n}(\hat{\beta}-\beta^o) \\ &\quad \xrightarrow{d} N[0,-R'(\beta^o)H_0^{-1}R'(\beta^o)'],\end{aligned}$$

where $\bar{\beta}=a\hat{\beta}+(1-a)\beta^o$ for some $a\in[0,1]$. It follows that the quadratic form

$$n[R(\hat{\beta})-r]'[-R'(\beta^o)H_0^{-1}R'(\beta^o)']^{-1}[R(\hat{\beta})-r]\to^d\chi^2_J.$$

By the Slutsky theorem, we have the Wald test statistic

$$W=n[R(\hat{\beta})-r]'[-R'(\hat{\beta})\hat{H}^{-1}(\hat{\beta})R'(\hat{\beta})']^{-1}[R(\hat{\beta})-r]\to^d\chi^2_J,$$

where again

$$\hat{H}(\beta)=n^{-1}\sum_{t=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta'}\ln f(Y_t|\Psi_t,\beta).$$

Note that only the unconstrained MLE $\hat{\beta}$ is needed in constructing the Wald test statistic.

Theorem [MLE-based Hypothesis Testing: Wald test] *Suppose Assumptions 9.1-9.6 hold, and the model $f(y_t|\Psi_t,\beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Then under $\mathbf{H}_0: R(\beta^o)=r$, we have as $n\to\infty$,*

$$\hat{W}\equiv n[R(\hat{\beta})-r]'[-R'(\hat{\beta})\hat{H}^{-1}(\hat{\beta})R'(\hat{\beta})']^{-1}[R(\hat{\beta})-r]\xrightarrow{d}\chi^2_J.$$

Question: Do we have the following result: Under $\mathbf{H}_0$

$$\begin{aligned}\tilde{W} &= n[R(\hat{\beta})-r]'[R'(\hat{\beta})\hat{V}^{-1}R'(\hat{\beta})']^{-1}[R(\hat{\beta})-r] \\ &= [R(\hat{\beta})-r]'[R'(\hat{\beta})[S(\hat{\beta})'S(\beta)]^{-1}R'(\hat{\beta})']^{-1}[R(\hat{\beta})-r] \\ &\to {}^d\chi^2_J \text{ as } n\to\infty,\end{aligned}$$

where

$$\hat{V}=n^{-1}\sum_{t=1}^{n}S_t(\hat{\beta})S_t(\hat{\beta})'=S(\hat{\beta})'S(\hat{\beta})/n,$$

and $S(\beta)=[S_1(\beta),S_2(\beta),...,S_n(\beta)]'$ is a $n\times K$ matrix.

Answer: Yes. But Why?

Likelihood Ratio Test

Theorem [Likelihood Ratio Test]: *Suppose Assumptions 9.1-9.6 hold, and $f(y|\Psi_t,\beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Define the average log-likelihoods*

$$\begin{aligned}\hat{l}(\hat{\beta}) &= n^{-1}\sum_{t=1}^{n}\ln f(Y_t|\Psi_t,\hat{\beta}),\\ \hat{l}(\tilde{\beta}) &= n^{-1}\sum_{t=1}^{n}\ln f(Y_t|\Psi_t,\tilde{\beta}),\end{aligned}$$

where $\hat{\beta}$ is the unconstrained MLE and $\tilde{\beta}$ is the constrained MLE subject to the constraint that $R(\tilde{\beta})=r$. Then under $\mathbf{H}_0 : R(\beta^o)=r$, we have

$$LR = 2n[\hat{l}(\hat{\beta})-\hat{l}(\tilde{\beta})] \xrightarrow{d} \chi^2_J \text{ as } n\to\infty.$$

Proof: We shall use the following strategy of proof:

(i) Use a second order Taylor series expansion to approximate $2n[\hat{l}(\hat{\beta})-\hat{l}(\tilde{\beta})]$ by a quadratic form in $\sqrt{n}(\tilde{\beta}-\hat{\beta})$.

(ii) Link $\sqrt{n}(\tilde{\beta}-\hat{\beta})$ with $\sqrt{n}\tilde{\lambda}$, where $\tilde{\lambda}$ is the Lagrange multiplier of the constrained MLE.

(iii) Derive the asymptotic distribution of $\sqrt{n}\tilde{\lambda}$.

Then combining (i)–(iii) will give an asymptotic χ^2_J distribution for the LR test statistic $LR = 2n[\hat{l}(\hat{\beta})-\tilde{l}(\tilde{\beta})]$.

The unconstrained MLE $\hat{\beta}$ solves for

$$\max_{\beta\in\Theta}\hat{l}(\beta).$$

The corresponding FOC is

$$\hat{S}(\hat{\beta})=0.$$

On the other hand, the constrained MLE $\tilde{\beta}$ solves the maximization problem

$$\max_{\beta\in\Theta}\left\{\hat{l}(\beta)+\lambda'[r-R(\beta)]\right\},$$

where λ is a $J\times 1$ Lagrange multiplier vector. The corresponding FOC are

$$\begin{aligned}\hat{S}(\tilde{\beta})-R'(\tilde{\beta})'\tilde{\lambda} &= 0,\\ (K\times 1)-(K\times J)\times(J\times 1) &= K\times 1\\ R(\tilde{\beta})-r &= 0.\end{aligned}$$

[Recall $R'(\beta)$ is a $K \times J$ matrix.] We now take a second order Taylor series expansion of $\hat{l}(\tilde{\beta})$ around the unconstrained MLE $\hat{\beta}$:

$$\begin{aligned} -LR &= 2n[\hat{l}(\tilde{\beta}) - \hat{l}(\hat{\beta})] \\ &= 2n[\hat{l}(\hat{\beta}) - \hat{l}(\hat{\beta})] + 2n\hat{S}(\hat{\beta})'(\tilde{\beta} - \hat{\beta}) \\ &\quad + \sqrt{n}(\tilde{\beta} - \hat{\beta})'\hat{H}(\bar{\beta}_a)\sqrt{n}(\tilde{\beta} - \hat{\beta}) \\ &= \sqrt{n}(\tilde{\beta} - \hat{\beta})'\hat{H}(\bar{\beta}_a)\sqrt{n}(\tilde{\beta} - \hat{\beta}) \end{aligned}$$

where $\bar{\beta}_a$ lies between $\tilde{\beta}$ and $\hat{\beta}$, namely $\bar{\beta}_a = a\tilde{\beta} + (1-a)\hat{\beta}$ for some $a \in [0,1]$. It follows that

$$2n[\hat{l}(\hat{\beta}) - \hat{l}(\tilde{\beta})] = \sqrt{n}(\tilde{\beta} - \hat{\beta})'[-\hat{H}(\bar{\beta}_a)]\sqrt{n}(\tilde{\beta} - \hat{\beta}). \tag{9.1}$$

This establishes the link between the LR test statistic and $\tilde{\beta} - \hat{\beta}$.

Next, we consider $\sqrt{n}(\tilde{\beta} - \hat{\beta})$. By a Taylor expansion for $\hat{S}(\tilde{\beta})$ around the unconstrained MLE $\hat{\beta}$ in the FOC $\hat{S}(\tilde{\beta}) - R'(\tilde{\beta})'\tilde{\lambda} = 0$, we have

$$\hat{S}(\hat{\beta}) + \hat{H}(\bar{\beta}_b)(\tilde{\beta} - \hat{\beta}) - R'(\tilde{\beta})'\tilde{\lambda} = 0,$$

where $\bar{\beta}_b = b\hat{\beta} + (1-b)\tilde{\beta}$ for some $b \in [0,1]$. Given $\hat{S}(\hat{\beta}) = 0$, we have

$$\hat{H}(\bar{\beta}_b)\sqrt{n}(\tilde{\beta} - \hat{\beta}) - R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} = 0$$

or

$$\sqrt{n}(\tilde{\beta} - \hat{\beta}) = \hat{H}^{-1}(\bar{\beta}_b)R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} \tag{9.2}$$

for n sufficiently large. This establishes the link between $\tilde{\lambda}$ and $\tilde{\beta} - \hat{\beta}$. In particular, it implies that the Lagrange multiplier $\tilde{\lambda}$ is an indicator for the magnitude of the difference $\tilde{\beta} - \hat{\beta}$.

Next, we derive the asymptotic distribution of $\sqrt{n}\tilde{\lambda}$. By a Taylor expansion of $\hat{S}(\tilde{\beta})$ around the true parameter β^o in the FOC $\sqrt{n}\hat{S}(\tilde{\beta}) - R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} = 0$, we have

$$\begin{aligned} R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} &= \sqrt{n}\hat{S}(\tilde{\beta}) \\ &= \sqrt{n}\hat{S}(\beta^o) + \hat{H}(\bar{\beta}_c)\sqrt{n}(\tilde{\beta} - \beta^o), \end{aligned}$$

where $\bar{\beta}_c$ lies between $\tilde{\beta}$ and β^o, namely, $\bar{\beta}_c = c\tilde{\beta} + (1-c)\beta^o$ for some $c \in [0,1]$. It follows that

$$\hat{H}^{-1}(\bar{\beta}_c)R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} = \hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o) + \sqrt{n}(\tilde{\beta} - \beta^o) \tag{9.3}$$

for n sufficiently large. Now, we consider a Taylor series expansion of $R(\tilde{\beta}) - r = 0$ around β^o :

$$\sqrt{n}[R(\beta^o) - r] + R'(\bar{\beta}_d)\sqrt{n}(\tilde{\beta} - \beta^o) = 0,$$

where $\bar{\beta}_d$ lies between $\tilde{\beta}$ and β^o. Given that $R(\beta^o) = r$ under $\mathbf{H}_0$, we have

$$R'(\bar{\beta}_d)\sqrt{n}(\tilde{\beta} - \beta^o) = 0. \tag{9.4}$$

It follows from Eq. (9.3) and Eq. (9.4) that

$$\begin{aligned} & R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} \\ = \; & R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o) \\ & +R'(\bar{\beta}_d)\sqrt{n}(\tilde{\beta} - \beta^o) \\ = \; & R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o) \\ \to \; & {}^d N(0, R'(\beta^o)H_o^{-1}V_oH_o^{-1}R'(\beta^o)') \end{aligned}$$

and therefore for n sufficiently large, we have

$$\begin{aligned} \sqrt{n}\tilde{\lambda} \; = \; & \left[R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)R'(\tilde{\beta})'\right]^{-1} R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o) \\ & \xrightarrow{d} N(0, [-R'(\beta^o)H_0^{-1}R'(\beta^o)']^{-1}) \end{aligned} \tag{9.5}$$

by the CLT for $\sqrt{n}\hat{S}(\beta^o)$, the MDS property of $\{S_t(\beta^o)\}$, the information matrix equality, and the Slutsky theorem.

Therefore, from Eq. (9.2) and Eq. (9.5), we have

$$\begin{aligned} & \hat{H}(\bar{\beta}_a)^{1/2}\sqrt{n}(\tilde{\beta} - \hat{\beta}) \\ = \; & \hat{H}(\bar{\beta}_a)^{1/2}\hat{H}^{-1}(\bar{\beta}_b)R'(\tilde{\beta})'\sqrt{n}\tilde{\lambda} \\ & \xrightarrow{d} N(0, \Pi) \\ \sim \; & \Pi^{1/2} \cdot N(0, I), \end{aligned} \tag{9.6}$$

where

$$\Pi = H_o^{-1/2}R'(\beta^o)'[-R'(\beta^o)H_o^{-1}R'(\beta^o)']^{-1}R'(\beta^o)H_o^{-1/2}$$

is a $K \times K$ symmetric and idempotent matrix ($\Pi^2 = \Pi$) with rank equal to J (using the formula that tr(ABC) =tr(BCA)).

Recall that if $v \sim N(0, \Pi)$, where Π is a symmetric and idempotent matrix with rank J, then the quadratic form $v'\Pi\nu \sim \chi^2_J$. It follows from Eq. (9.1) and Eq. (9.6) that

$$\begin{aligned} 2n[\hat{l}(\tilde{\beta}) - \hat{l}(\hat{\beta})] &= \sqrt{n}(\tilde{\beta} - \hat{\beta})'[-\hat{H}(\bar{\beta}_a)]^{1/2}[-\hat{H}(\bar{\beta}_a)]^{1/2}\sqrt{n}(\tilde{\beta} - \hat{\beta}) \\ &\xrightarrow{d} \chi^2_J. \end{aligned}$$

This completes the proof.

Remark:

The LR test is based on comparing the objective functions—the log likelihood functions under the null hypothesis $\mathbf{H}_0$ and the alternative to $\mathbf{H}_0$. Intuitively, when $\mathbf{H}_0$ holds, the likelihood $\hat{l}(\hat{\beta})$ of the unrestricted model is similar to the likelihood $\hat{l}(\tilde{\beta})$ of the restricted model, with the little difference subject to sampling variations. If the likelihood $\hat{l}(\hat{\beta})$ of the unrestricted model is sufficiently larger than the likelihood $\hat{l}(\tilde{\beta})$ of the restricted model, there exists evidence that $\mathbf{H}_0$ is false. How large a difference between $\hat{l}(\hat{\beta})$ and $\hat{l}(\tilde{\beta})$ is considered as sufficiently large to reject $\mathbf{H}_0$ is determined by the associated asymptotic χ^2_J distribution.

The likelihood ratio test statistic is similar in spirit to the F test statistic in the classical linear regression model, which compares the objective functions—the sum of squared residuals under the null hypothesis $\mathbf{H}_0$ and the alternative to $\mathbf{H}_0$ respectively. In other words, the negative log-likelihood is analogous to the sum of squared residuals. In fact, the LR test statistic and the $J \cdot F$ statistic are asymptotically equivalent under $\mathbf{H}_0$ for a linear regression model

$$Y_t = X_t'\alpha^o + \varepsilon_t,$$

where $\varepsilon_t|\Psi_t \sim N(0, \sigma_o^2)$. To see this, put $\beta = (\alpha', \sigma^2)'$ and note that

$$\begin{aligned} f(Y_t|\Psi_t, \beta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_t - X_t'\alpha)^2}, \\ \hat{l}(\beta) &= n^{-1}\sum_{t=1}^{n} \ln f(Y_t|\Psi_t, \beta) \\ &= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} n^{-1}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2. \end{aligned}$$

It is straightforward to show (please show it!) that

$$\begin{aligned} \hat{l}(\hat{\beta}) &= \frac{1}{2}\ln(e'e), \\ \hat{l}(\tilde{\beta}) &= \frac{1}{2}\ln(\tilde{e}'\tilde{e}), \end{aligned}$$

where e and $\tilde{e}$ are the $n \times 1$ unconstrained and constrained estimated residual vectors respectively. Therefore, under $\mathbf{H}_0$, we have

$$\begin{aligned} 2n[\hat{l}(\tilde{\beta}) - \hat{l}(\hat{\beta})] &= n \ln(\tilde{e}'\tilde{e}/e'e) \\ &= \frac{(\tilde{e}'\tilde{e} - e'e)}{e'e/n} + o_P(1) \\ &= J \cdot F + o_P(1), \end{aligned}$$

where we have used the inequality that $|\ln(1+z) - z| \leq z^2$ for small z, and the asymptotically negligible ($o_P(1)$) reminder term is contributed by the quadratic term in the expansion.

In the proof of the above theorem, we see that the asymptotic distribution of the LR test statistic depends on correct model specification of $f(y|\Psi_t, \beta)$, because it uses the MDS property of the score function and the IM equality. In other words, if the conditional distribution model $f(y|\Psi_t, \beta)$ is misspecified such that the MDS property of the score function or the IM equality does not hold, then the LR test statistic will not be asymptotically χ^2-distributed.

Lagrange Multiplier (LM) or Efficient Score Test

We can also use the Lagrange multiplier $\tilde{\lambda}$ to construct a Lagrange Multiplier (LM) test, which is also called Rao's efficient score test. Recall the Lagrange multiplier λ is introduced in the constrained MLE problem:

$$\max_{\beta \in \Theta} \hat{L}(\beta) + \lambda'[r - R(\beta)].$$

The $J \times 1$ Lagrange multipier vector $\tilde{\lambda}$ measures the effect of the restriction of H_0 on the

maximized value of the model likelihood. When $\mathbf{H}_0$ holds, the imposition of the restriction results in little change in the maximized likelihood. Thus the value of the Lagrange multiplier $\tilde{\lambda}$ for a correct restriction should be small. If a sufficiently large Lagrange mutiplier $\tilde{\lambda}$ is obtained, it implies that the maximized likelihood value of the restricted model is sufficiently smaller than that of the unrestricted model, thus leading to the rejection of $\mathbf{H}_0$.Therefore, we can use $\tilde{\lambda}$ to construct a test for $\mathbf{H}_0$.

In deriving the asymptotic distribution of the LR test statistic, we have obtained

$$\begin{aligned} \sqrt{n}\tilde{\lambda} &= \left[R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)R'(\tilde{\beta})'\right]^{-1} R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o) \\ &\xrightarrow{d} N(0, [-R'(\beta^o)H_o^{-1}R'(\beta^o)']^{-1}) \end{aligned}$$

It follows that the quadratic form

$$n\tilde{\lambda}'[-R'(\beta^o)H_o^{-1}R'(\beta^o)']\tilde{\lambda} \xrightarrow{d} \chi^2_J,$$

and so by the Slutsky theorem, we have

$$n\tilde{\lambda}'[-R'(\tilde{\beta})\hat{H}^{-1}(\tilde{\beta})R'(\tilde{\beta})']\tilde{\lambda} \xrightarrow{d} \chi^2_J.$$

We have actually proven the following theorem.

Theorem [LM/Efficient Score test] *Suppose Assumptions 9.1–9.6 hold, and the model $f(y|\Psi_t,\beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t. Then we have*

$$LM_0 \equiv n\tilde{\lambda}'R'(\tilde{\beta})[-\hat{H}^{-1}(\tilde{\beta})]R'(\tilde{\beta})'\tilde{\lambda} \xrightarrow{d} \chi^2_J$$

under $\mathbf{H}_0$.

The LM test statistic only involves estimation of the model $f(y_t|\Psi_t,\beta)$ under $\mathbf{H}_0$, its computation may be simpler than the computation of the Wald test statistic or the LR test statistic in many cases.

Question: Is it true that under $\mathbf{H}_0$,

$$\begin{aligned} &n\tilde{\lambda}'R'(\tilde{\beta})\tilde{V}^{-1}R'(\tilde{\beta})'\tilde{\lambda} = n^2\tilde{\lambda}'R'(\tilde{\beta})[S(\tilde{\beta})'S(\tilde{\beta})]^{-1}R'(\tilde{\beta})'\tilde{\lambda} \\ &\xrightarrow{d} \chi^2_J, \end{aligned}$$

where

$$\begin{aligned} \tilde{V} &= n^{-1}\sum_{t=1}^{n} S_t(\tilde{\beta})S_t(\tilde{\beta})' \\ &= S(\tilde{\beta})'S(\tilde{\beta})/n. \end{aligned}$$

Question: What is the advantage of the LM test?

Question: What is the relationship among the Wald, LR and LM test statistics?

9.4 Quasi-Maximum Likelihood Estimation

When $f(y_t|\Psi_t,\beta)$ is misspecified, for all $\beta\in\Theta$, $f(y|\Psi_t,\beta)$ is not equal to the true conditional pdf/pmf of Y_t given Ψ_t.

Question: What happens if $f(y_t|\Psi_t,\beta)$ is not correctly specified for the conditional pdf/pmf of Y_t given Ψ_t?

Question: What is the interpretation for β^o, where $\beta^o = \arg\max_{\beta\in\Theta} l(\beta)$ is as in Assumption 9.4 when $f(y|\Psi_t,\beta)$ is misspecified?

We can no longer interpret β^o as the true model parameter, because $f(y|\Psi_t, \beta^o)$ does not coincide with the true conditional probability distribution of Y_t given Ψ_t.

Below, we provide an alternative interpretation for β^o when $f(y|\Psi_t, \beta)$ is misspecified.

Lemma: *Suppose Assumption 9.4 holds. Define the conditional relative entropy*

$$I(f : p|\Psi) = \int \ln \left[\frac{p(y|\Psi)}{f(y|\Psi, \beta)} \right] p(y|\Psi)dy,$$

where $p(y|\Psi)$ is the true conditional pdf/pmf of Y on Ψ. Then $I(f : p|\Psi)$ is nonnegative almost surely for all β, and

$$\beta^o = \arg\min_{\beta \in \Theta} E[I(f : p|\Psi)],$$

where $E(\cdot)$ is taken over the probability distribution of Ψ.

Remark:

The parameter value β^o minimizes the "distance" of $f(\cdot|\cdot, \beta^o)$ from the true conditional density $p(\cdot|\cdot)$ in terms of conditional relative entropy. Relative entropy is a divergence measure for two alternative distributions. It is zero if and only if two distributions coincide with each other. There are many distance/divergence measures for two distributions. Relative entropy has the appealing information-theoretic interpretation and the invariance property with respect to data transformation. It has been widely used in economics and econometrics.

Question: Why is a misspecified pdf/pmf model $f(y_t|\Psi_t, \beta)$ still useful in economic applications?

In many applications, misspecification of higher order conditional moments does not render inconsistent the estimator for the parameters appearing in the lower order conditional moments. For example, suppose a conditional mean model is correctly specified but the conditional higher order moments are misspecified. We can still obtain a consistent estimator for the parameter β appearing in the conditional mean model. Of course, the parameters appearing in the higher order conditional moments cannot be consistently estimated.

We now consider a few illustrative examples.

Example 1 [Nonlinear Regression Model] Suppose $(Y_t, X_t')'$ is i.i.d.,

$$Y_t = g(X_t, \alpha^o) + \varepsilon_t,$$

where $E(\varepsilon_t|X_t) = 0$ a.s.

Here, the regression model $g(X_t, \alpha)$ is correctly specified for $E(Y_t|X_t)$ if and only if $E(\varepsilon_t|X_t) = 0$ a.s.. We need not know the distribution of $\varepsilon_t|X_t$.

Question: How to estimate the true parameter α^o when the conditional mean model $g(X_t, \alpha)$ is correctly specified for $E(Y_t|X_t)$?

In order to estimate α^o, we assume that $\varepsilon_t|X_t \sim i.i.d.N(0, \sigma^2)$, which is likely to be incorrect (and we know this). Then we can obtain the pesudo conditional likelihood function

$$f(y_t|x_t, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[y_t - g(x_t,\alpha)]^2},$$

where $\beta = (\alpha', \sigma^2)'$.

Define the Quasi-MLE

$$\hat{\beta} = (\hat{\alpha}', \hat{\sigma}^2)' = \arg\max_{\alpha,\sigma^2} \sum_{t=1}^{n} \ln f(Y_t|X_t, \beta).$$

Then $\hat{\alpha}$ is a consistent estimator for α^o. In this example, misspecification of i.i.d. $N(0, \sigma^2)$ for $\varepsilon_t|X_t$ does not render inconsistent the parameter for α^o. The QMLE $\hat{\alpha}$ is consistent for α^o as long as the conditional mean of Y_t is correctly specified by $f(y|X_t, \beta)$. Of course, the parameter estimator $\hat{\beta} = (\hat{\alpha}', \hat{\sigma}^2)'$ cannot consistently estimate the true conditional distribution of Y_t given Ψ_t if the conditional distribution of $\varepsilon_t|X_t$ is misspecified.

Suppose the true conditional distribution $\varepsilon_t|X_t \sim i.i.d.N(0, \sigma_t^2)$, where $\sigma_t^2 = \sigma^2(X_t)$ is a function of X_t but we assume $\varepsilon_t|X_t \sim i.i.d.N(0, \sigma^2)$. Then we still have $E[S_t(\beta^o)|X_t] = 0$ a.s. but the conditional informational matrix equality does not hold.

Example 2 [Capital Asset Pricing Model (CAPM):

Define Y_t as an $L \times 1$ vector of excess returns for L assets (or portfolios of assets). For these L assets, the excess returns can be described using the excess-return market model:

$$\begin{aligned} Y_t &= \alpha_0^o + \alpha_1^o Z_{mt} + \varepsilon_t \\ &= \alpha^{o\prime} X_t' + \varepsilon_t, \end{aligned}$$

where $X_t = (1, Z_{mt})'$ is a bivariate vector, Z_{mt} is the excess market return, α^o is a $2 \times L$ parameter matrix, and ε_t is an $L \times 1$ disturbance, with $E(\varepsilon_t|X_t) = 0$. With this orthogonality condition, CAPM is correctly specified for the expected excess return $E(Y_t|X_t)$.

To estimate unknown parameter matrix α^o, one can assume

$$\varepsilon_t|\Psi_t \sim N(0, \Sigma),$$

where $\Psi_t = \{X_t, Y_{t-1}, X_{t-1}, Y_{t-2}, ...\}$ and Σ is an $L \times L$ symmetric and positive definite matrix.

Then we can write the conditional pdf of Y_t given Ψ_t as follows:

$$\begin{aligned} f(Z_t|\Psi_t,\beta) &= (2\pi)^{-\frac{L}{2}}\left|\Sigma\right|^{-\frac{1}{2}} \\ &\quad\times\exp\left[-\frac{1}{2}(Y_t-\alpha' X_t)'\Sigma^{-1}(Y_t-\alpha' X_t)\right], \end{aligned}$$

where $\beta=(\alpha'$,vech$(\Sigma)')'$.

Although the i.i.d. normality assumption for $\{\varepsilon_t\}$ may not hold, the estimator based on the pesudo Gaussian likelihood function will be consistent for parameter matrix α^o appearing in the CAPM model.

Example 3 [Univariate ARMA(p,q) Model]: In Chapter 5, we introduced a class of time series models called ARMA(p,q). Suppose

$$Y_t=\alpha_0+\sum_{j=1}^{p}\alpha_j Y_{t-j}+\sum_{j=1}^{q}\gamma_j\varepsilon_{t-j}+\varepsilon_t,$$

where ε_t is an MDS with mean 0 and variance σ^2. Then this ARMA(p,q) model is correctly specified for $E(Y_t|I_{t-1})$, where $I_{t-1}=\{Y_{t-1},Y_{t-2},...,Y_1\}$ is the information set available at time $t-1$. Note that the distribution of ε_t is not specified. How can we estimate parameters $\alpha_0,\alpha_1,...,\alpha_p,\gamma_1,...,$ and γ_q?

Assuming that $\{\varepsilon_t\}\sim i.i.d.N(0,\sigma^2)$, then the conditional pdf of Y_t given $\Psi_t=I_{t-1}$ is

$$f(y|\Psi_t,\beta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu_t(\alpha,\gamma))^2}{2\sigma^2}\right],$$

where $\beta=(\alpha_0,\alpha_1,...,\alpha_p,\gamma_1,...,\gamma_q,\sigma^2)'$, and

$$\mu_t(\beta)=\alpha_0+\sum_{j=1}^{p}\alpha_j Y_{t-j}+\sum_{j=1}^{q}\gamma_j\varepsilon_{t-j}.$$

Although the i.i.d. normality assumption for $\{\varepsilon_t\}$ may be false, the estimator based on the above pesudo Gaussian likelihood function will be consistent for parameters (α^o,γ^o) appearing in the ARMA(p,q) model.

In practice, we have a random sample $\{Y_t\}_{t=1}^n$ of size n to estimate an ARMA(p,q) model and need to assume some initial values for $\{Y_t\}_{t=-p}^0$ and $\{\varepsilon_t\}_{t=-q}^0$. For example, we can set $Y_t=\bar{Y}$ for $-p\leq t\leq 0$ and $\varepsilon_t=0$ for $-q\leq t\leq 0$. When an ARMA(p,q) is a stationary process, these choice of initial values does not affect the asymptotic properties of the QMLE $\hat{\beta}$ under regularity conditions.

Example 4 [Vector Autoregression Model]: Suppose $Y_t = (Y_{1t}, ..., Y_{Lt})'$ is a $L \times 1$ stationary ergodic autoregressive process of order p :

$$Y_t = \alpha_0^o + \sum_{j=1}^{p} \alpha_j^{o\prime} Y_{t-j} + \varepsilon_t, \qquad t = p+1, ..., n,$$

where α_0^o is an $L \times 1$ parameter vector, α_j^o is a $L \times L$ parameter matrix for $j = \{1, ..., p\}$, and $\{\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{Lt})'\}$ is an $L \times 1$ MDS with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t \varepsilon_t') = \Sigma^o$, an $L \times L$ finite and positive definite matrix. When Σ^o is not a diagonal matrix, there exist contemporaneous correlations between different components of ε_t. This implies that a shock on ε_{1t} will be spilled over to other variables. With the MDS condition for $\{\varepsilon_t\}$, the VAR(p) model is correctly specified for $E(Y_t|I_{t-1})$, where $I_{t-1} = \{Y_{t-1}, Y_{t-2}, ..., Y_1\}$. Note that the VAR($p$) model can be equivalently represented as follows:

$$\begin{array}{lcl} Y_{1t} & = & \alpha_{10} + \sum_{j=1}^{p} \alpha_{11j} Y_{1t-j} + \cdots + \sum_{j=1}^{p} \alpha_{1Lj} Y_{Lt-j} + \varepsilon_{1t}, \\ Y_{2t} & = & \alpha_{20} + \sum_{j=1}^{p} \alpha_{21j} Y_{1t-j} + \cdots + \sum_{j=1}^{p} \alpha_{2Lj} Y_{Lt-j} + \varepsilon_{2t}, \\ \cdots & \cdots & \cdots \\ Y_{Lt} & = & \alpha_{L0} + \sum_{j=1}^{p} \alpha_{L1j} Y_{1t-j} + \cdots + \sum_{j=1}^{p} \alpha_{LLj} Y_{Lt-j} + \varepsilon_{Lt}. \end{array}$$

Let β^o denote a parameter vector containing all components of unknown parameters from $\alpha_0^o, \alpha_1^o, ..., \alpha_p^o$, and Σ^o. To estimate β^o, one can assume

$$\varepsilon_t | I_{t-1} \sim N(0, \Sigma).$$

Then $Y_t|I_{t-1} \sim N(\alpha_0 + \sum_{j=1}^{p} \alpha_j' Y_{t-j}, \Sigma)$, and the pesudo conditional pdf of Y_t given $\Psi_t = Y^{t-1}$ is

$$\begin{array}{rcl} f(Y_t|\Psi_t, \beta) & = & \dfrac{1}{\sqrt{(2\pi)^L \det(\Sigma)}} \times \\ & & \exp\left\{ -\dfrac{1}{2} \left[Y_t - \mu_t(\alpha) \right]' \Sigma^{-1} Y_t - \mu_t(\alpha) \right\}, \end{array}$$

where $\mu_t(\alpha) = \alpha_0 + \sum_{j=1}^{p} \alpha_j' Y_{t-j}$.

Example 5 [GARCH Model]: Time-varying volatility is an important empirical stylized facts for many economic and financial time series. For example, it has been well-known that there exists volatility clustering in financial markets, that is, a large volatility today tends to be followed by another large volatility tomorrow; a small volatility today tends to be followed by another small volatility tomorrow, and the patterns alternate over time. In financial econometrics, the following GARCH model has been used to capture volatility clustering or more generally time-varying volatility. Suppose (Y_t, X_t) is a strictly stationary process with

$$\begin{aligned} Y_t &= \mu(\Psi_t, \beta^o) + \sigma(\Psi_t, \beta^o) z_t, \\ E(z_t|\Psi_t) &= 0 \text{ a.s.}, \\ E(z_t^2|\Psi_t) &= 1 \text{ a.s..} \end{aligned}$$

The models $\mu(\Psi_t, \beta)$ and $\sigma^2(\Psi_t, \beta)$ are correctly specified for $E(Y_t|\Psi_t)$ and $\text{var}(Y_t|\Psi_t)$ if and only if $E(z_t|\Psi_t) = 0$ a.s. and $\text{var}(z_t|\Psi_t) = 1$ a.s. We need not know the conditional distribution of $z_t|\Psi_t$ (in particular, we need not know the higher order conditional moments of z_t given Ψ_t).

An example for $\mu(\Psi_t, \beta)$ is the ARMA(p, q) in Example 2. We now give some popular models for $\sigma^2(\Psi_t, \beta)$. For notational simplicity, we put $\sigma_t^2 = \sigma^2(\Psi_t, \beta)$.

- Engle's (1982) ARCH(q) model

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}^2,$$

where $\varepsilon_t = \sigma_t z_t$.

- Bollerslev's (1986) GARCH(p, q) model

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j \sigma_{t-j}^2 + \sum_{j=1}^{q} \gamma_j \varepsilon_{t-j}^2;$$

- Nelson's (1990) EGARCH(p, q) model

$$\ln \sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j \ln \sigma_{t-j}^2 + \sum_{j=0}^{q} \gamma_j g(z_{t-j}),$$

where $g(z_t)$ is a nonlinear function defined as

$$g(z_t) = \theta_1(|z_t| - E|z_t|) + \theta_2 z_t.$$

- Threshold GARCH(p, q) model:

$$\begin{aligned} \sigma_t^2 &= \omega + \sum_{j=1}^{p} \alpha_j \sigma_{t-j}^2 + \sum_{j=1}^{q} \gamma_j \varepsilon_{t-j}^2 \mathbf{1}(z_{t-j} > 0) \\ + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}^2 \mathbf{1}(z_{t-j} &\leq 0), \end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Question: How to estimate β^o, the parameters appearing in the first two conditional moments?

A most popular approach is to assume that $z_t|\Psi_t \sim$ i.i.d.N$(0,1)$. Then $Y_t|\Psi_t \sim N(\mu_t(\Psi_t,\beta^o), \sigma^2(\Psi_t,\beta^o))$, and the pseudo conditional pdf of Y_t given Ψ_t is

$$f(y|\Psi_t,\beta) = \frac{1}{\sqrt{2\pi}\sigma(\Psi_t,\beta)} e^{-\frac{1}{2\sigma^2(\Psi_t,\beta)}[y-\mu(\Psi_t,\beta)]^2}.$$

It follows that the log-likelihood function

$$\begin{aligned} & \sum_{t=1}^n \ln f(Y_t|\Psi_t,\beta) \\ = & -\frac{n}{2}\ln 2\pi - \sum_{t=1}^n \ln \sigma_t(\Psi_t,\beta) \\ & -\frac{1}{2}\sum_{t=1}^n \frac{[Y_t-\mu(\Psi_t,\beta)]^2}{\sigma^2(\Psi_t,\beta)}. \end{aligned}$$

The i.i.d. N(0,1) innovation assumption does not affect the specification of the conditional mean $\mu(\Psi_t,\beta)$ and conditional variance $\sigma^2(\Psi_t,\beta)$, so it does not affect the consistency of the QMLE $\hat{\beta}$ for the true parameter value β^o appearing in the conditional mean and conditional variance specifications. In other words, ε_t may not be i.i.d. N(0,1) but this does not affect the consistency of the Gaussian QMLE $\hat{\beta}$.

In addition to the i.i.d.N(0,1) assumption, the following two error distributions have also been popularly used in practice:

- Standardized Student$\sqrt{(\nu-2)/\nu} \cdot t(\nu)$ Distribution

  The scale factor $\sqrt{(\nu-2)/\nu}$ ensures that z_t has unit variance. The pdf of z_t is

$$f(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad -\infty < z < \infty.$$

- Generalized Error Distribution

$$f(z_t) = \frac{b}{2a\Gamma\left(\frac{1}{b}\right)} \exp\left[-\left(\frac{|z-\mu|}{a}\right)^b\right], \qquad -\infty < z < \infty$$

where μ, a and b are location, scale and shape parameters respectively. Note that both standardized t-distribution and generalized error distribution include N(0,1) as a special case.

Like estimation of an ARMA(p, q) model, we may have to choose initial values for some variables in estimating GARCH models. For example, in estimating GARCH(1,1) models, we will encounter the initial value problem for the conditional variance σ_0^2 and ε_0. One can set h_0 to be the unconditional variance $E(\sigma_t^2) = \omega/(1 - \alpha_1 - \gamma_1)$, and set $\varepsilon_0 = 0$.

We note that the ARMA model in Example 2 can be estimated via QMLE as a special case of the GARCH model by setting $\sigma^2(\Psi_t, \beta) = \sigma^2$.

Question: What is the implication of a misspecified probability distribution model?

Although misspecification of $f(y_t|\Psi_t, \beta)$ may not affect the consistency of the QMLE (or the consistency of a subset of parameters) under suitable regularity conditions, it does affect the asymptotic variance (and so efficiency) of the QMLE $\hat{\beta}$.

Remark: The parameter β^o is not always consistently estimable by QMLE when the likelihood function is misspecified. In some cases, β^o cannot be consistently estimated when the likelihood model is misspecified.

We first investigate the implication of a misspecifed conditional distribution model $f(y|\Psi_t, \beta)$ on the score function and the IM equality.

Lemma: *Suppose Assumptions 9.4–9.6(i) hold. Then*

$$E\left[S_t(\beta^o)\right] = 0,$$

where $E(\cdot)$ is taken over the true distribution of the data generating process.

Proof: Because β^o maximizes $l(\beta)$ and is an interior point in Θ, the FOC holds: at $\beta = \beta^o$:

$$\frac{dl(\beta^o)}{d\beta} = 0.$$

By differentiating, we have

$$\frac{d}{d\beta} E[\ln f(Y_t|\Psi_t, \beta^o)] = 0.$$

Exchanging differentiation and integration yields the desired result:

$$E\left[\frac{\partial \ln f(Y|\Psi_t, \beta^o)}{\partial \beta}\right] = 0.$$

This completes the proof. ■

Remarks:

No matter whether the conditional distributional model $f(y|\Psi_t, \beta)$ is correctly specified, the score function $S_t(\beta^o)$ evaluated at β^o always has mean zero. This is due to the consequence of the FOC of the maximization of $l(\beta)$. This is analogous to the FOC of the best linear least squares approximation where one always has $E(X_t u_t) = 0$ with $u_t = Y_t - X_t'\beta^*$ and $\beta^* = [E(X_t X_t')]^{-1}E(X_t Y_t)$.

When $\{Z_t = (Y_t, X_t')'\}$ is i.i.d., or $\{Z_t\}$ is not independent but $\{S_t(\beta^o)\}$ is MDS (we note that $S_t(\beta^o)$ could still be MDS when $f(Y_t|\Psi_t, \beta)$ is misspecified for the conditional distribution of Y_t given Ψ_t), we have

$$\begin{aligned}
V_o &= \text{avar}\left(n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)\right) \\
&= \lim_{n\to\infty} E\left[\left(n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)\right)\left(n^{-1/2}\sum_{\tau=1}^{n} S_\tau(\beta^o)\right)'\right] \\
&= E[S_t(\beta^o)S_t(\beta^o)'].
\end{aligned}$$

Thus, even when $f(y|\Psi_t, \beta)$ is a misspecified conditional distribution model, we do not have to estimate a long-run variance-covariance matrix for V_o as long as $\{S_t(\beta^o\}$ is an MDS process.

Question: Can you give a time series example in which $f(y_t|\Psi_t, \beta)$ is misspecified but $\{S_t(\beta^o)\}$ is MDS?

Answer: Consider a conditional distribution model which correctly specifies the conditional mean of Y_t but misspecifies the higher order conditional moments (e.g., conditional variance).

Question: Is $\{S_t(\beta^o)\}$ always MDS, when $\{S_t(\beta^o)\}$ is stationary ergodic?

Answer: In the time series context, when the conditional pdf/pmf $f(y_t|\Psi_t, \beta)$ is misspecified, then $S_t(\beta^o)$ may not be MDS. In this case, we have

$$\begin{aligned}
V_o &\equiv \text{avar}\left[\sqrt{n}\hat{S}(\beta^o)\right] \\
&= \lim_{n\to\infty} n^{-1}\sum_{t=1}^{n}\sum_{\tau=1}^{n} E[S_t(\beta^o)S_\tau(\beta^o)'] \\
&= \sum_{j=-\infty}^{\infty} E[S_t(\beta^o)S_{t-j}(\beta^o)'] \\
&= \sum_{j=-\infty}^{\infty} \Gamma(j),
\end{aligned}$$

where

$$\Gamma(j) = E[S_t(\beta^o)S_{t-j}(\beta^o)'].$$

In other words, we have to estimate the long-run variance-covariance matrix for V when $\{S_t(\beta^o)\}$

is not an MDS.

Question: If the model $f(y|\Psi_t, \beta)$ is misspecified for the conditional distribution of Y_t given Ψ_t, do we have the conditional information matrix equality?

Generally, no. That is, we generally have neither $E\left[S_t\left(\beta^o\right)|I_{t-1}\right] = 0$ nor

$$E\left[S_t(\beta^o)S_t(\beta^o)'|\Psi_t\right] + E\left[\frac{\partial^2 \ln f(Y_t|\Psi_t, \beta^o)}{\partial\beta\partial\beta'}|\Psi_t\right] = 0,$$

where $E(\cdot|\Psi_t)$ is taken under the true conditional distribution which differs from the model $f(y_t|\Psi_t, \beta^o)$ when $f(y_t|\Psi_t, \beta)$ is misspecified. Please check.

Question: What is the impact of the failure of the MDS property for the score function and the failure of the conditional information matrix equality?

Theorem [Asymptotic Normality of QMLE]: *Suppose Assumptions 9.1–9.6 hold. Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \rightarrow^d N(0, H_o^{-1}V_oH_o^{-1}),$$

where $V_o \equiv avar[\sqrt{n}\hat{S}(\beta^o)]$.

Remarks:
Without the MDS property of the score function, we have to estimate $V_o \equiv \text{avar}[\sqrt{n}\hat{S}(\beta^o)]$ by (e.g.) the Newey-West (1987, 1994) type estimator in the time series context. Without the conditional information matrix equality (even if the MDS holds), we cannot simplify the asymptotic variance of the QMLE from $H_o^{-1}V_oH_o^{-1}$ to $-H_o^{-1}$ even if the score function is i.i.d. or MDS. In certain sense, the MDS property of the score function is analogous to serial uncorrelatedness in a regression disturbance, and the information matrix equality is analogous to conditional homoskedasticity.

Compared with the asymptotic variance $-H_o^{-1}$ of MLE, the asymptotic variance $-H_o^{-1}V_oH_o^{-1}$ of QMLE is more complicated than that of MLE, because we cannot use the information matrix equality to simplify the asymptotic variance. In addition, V_o has to be estimated using a kernel-based method when $\{S_t(\beta^o)\}$ is not an MDS.

In the literature, the variance $H_o^{-1}V_oH_o^{-1}$ is usually called the robust asymptotic variance-covariance matrix of QMLE $\hat{\beta}$. It is robust to misspecification of model $f(y_t|\Psi_t, \beta)$. That is, no matter whether $f(y_t|\Psi_t\beta)$ is correctly specified, $H_o^{-1}V_oH_o^{-1}$ is always the correct asymptotic variance of $\sqrt{n}\hat{\beta}$.

Question: Is QMLE asymptotically less efficient than MLE?

Yes. The asymptotic variance of the MLE, equal to H_o^{-1}, the inverse of the negative Hessian matrix, achieves the Cramer-Rao lower bound, and therefore is asymptotically most efficient. On the other hand, the asymptotic variance $-H_o^{-1}V_oH_o^{-1}$ of the QMLE is not the same as the asymptotic variance $-H_o^{-1}$ of the MLE and thus does not achieve the Cramer-Rao lower bound. It is asymptotically less efficient than the MLE. This is the price one has to pay with use of a misspecified pdf/pmf model, although some model parameters still can be consistently estimated.

Asymptotic Variance Estimation

Question: How to estimate the asymptotic variance $H_o^{-1}V_oH_o^{-1}$ of the QMLE?

First, it is straightforward to estimate H_0 :

$$\hat{H}(\hat{\beta}) = n^{-1}\sum_{t=1}^{n}\frac{\partial^2 \ln f(Y_t|\Psi_t,\hat{\beta})}{\partial\beta\partial\beta'}.$$

The UWLLN for $\{H_t(\beta)\}$ and the continuity of $H(\beta)$ ensure that $\hat{H}(\hat{\beta}) \to^p H_o$.

Next, how to estimate $V_o = \text{avar}[n^{-1/2}\Sigma_{t=1}^n S_t(\beta^o)]$?

We consider two cases, depending on whether $\{S_t(\beta^o)\}$ is MDS:

Case 1: $\{Z_t = (Y_t, X_t')'\}$ is i.i.d. or $\{Z_t\}$ is not independent but $\{S_t(\beta^o)\}$ is MDS.

In this case,

$$V_o = E[S_t(\beta^o)S_t(\beta^o)']$$

so we can use

$$\hat{V} = n^{-1}\sum_{t=1}^{n} S_t(\hat{\beta})S_t(\hat{\beta})'$$

which is consistent for V.

Case 2: When $\{Z_t\}$ is not independent, $\{S_t(\beta^o)\}$ may not be MDS.

In this case, we can use the kernel method

$$\hat{V} = \sum_{j=1-n}^{n-1} k(j/p)\hat{\Gamma}(j),$$

where

$$\hat{\Gamma}(j) = n^{-1} \sum_{t=j+1}^{n} S_t(\hat{\beta}) S_{t-j}(\hat{\beta})' \text{ if } j \geq 0$$

and $\hat{\Gamma}(j) = \hat{\Gamma}(-j)'$ for $j < 0$.

We directly assume that $\hat{V}$ is consistent for V_o.

Assumption 9.7: $\hat{V} \to^p V_o$, where V_o is finite and nonsingular.

Lemma [Asymptotic Variance Estimator for QMLE]: *Suppose Assumptions 9.1–9.7 hold. Then as $n \to \infty$,*

$$\hat{H}^{-1}(\hat{\beta}) \hat{V} \hat{H}^{-1}(\hat{\beta}) \to^p H_o^{-1} V_o H_o^{-1}.$$

Hypothesis Testing

With the consistent asymptotic variance estimator, we can now construct suitable hypothesis tests under a misspecified conditional distributional model.

Again, we consider the null hypothesis

$$\mathbf{H}_0 : R(\beta^o) = r,$$

where $R(\beta)$ is a $J \times 1$ continuously differentiable vector function with the $J \times K$ matrix $R'(\beta^o)$ being of full rank, and r is a $J \times 1$ vector.

Wald Test Under Model Misspecification

We first consider a Wald test.

Theorem [QMLE-based Hypothesis Testing, Wald Test]: *Suppose Assumptions 9.1–9.7 hold. Then under $\mathbf{H}_0 : R(\beta^o) = r$, we have*

$$\begin{aligned} \hat{W} &= n[R(\hat{\beta}) - r]' \\ &\quad \times [R'(\hat{\beta})[\hat{H}^{-1}(\hat{\beta}) \hat{V} \hat{H}^{-1}(\hat{\beta})]^{-1} R'(\hat{\beta})']^{-1} \\ &\quad \times [R(\hat{\beta}) - r] \\ &\quad \xrightarrow{d} \chi^2_J \end{aligned}$$

Proof: By the first order Taylor series expansion, we obtain

$$\begin{aligned} \sqrt{n}[R(\hat{\beta}) - r] &= \sqrt{n}[R(\beta^o) - r] + R'(\bar{\beta})\sqrt{n}(\hat{\beta} - \beta^o) \\ &= R'(\bar{\beta})\sqrt{n}(\hat{\beta} - \beta^o) \\ &\to {}^d N(0, R'(\beta^o) H_o^{-1} V_o H_o^{-1} R'(\beta^o)') \end{aligned}$$

where we have made use of the fact that $\sqrt{n}(\hat{\beta}-\beta^o) \xrightarrow{d} N(0, H_o^{-1}V_oH_o^{-1})$, and the Slutsky theorem. The desired result for $\hat{W}$ follows immediately.

Remarks:

Only the unconstrained QMLE $\hat{\beta}$ is used in constructing the robust Wald test statistic. The Wald test statistic under model misspecification is similar in structure to the Wald test in linear regression modeling that is robust to conditional heteroskedasticity (under the i.i.d. or MDS assumption) or that is robust to conditional heteroskedasticity and autocorrelation (under the non-MDS assumption).

LM/Score Test Under Model Misspecification

Question: Can we use the LM test principle for $\mathbf{H}_0$ when $f(y|\Psi_t, \beta)$ is misspecified?

Yes, we can still derive the asymptotic distribution of $\sqrt{n}\tilde{\lambda}$, with a suitable (i.e., robust) asymptotic variance, which of course will be generally different from that under correct model specification.

Recall that from the FOC of the constrained MLE $\tilde{\beta}$,

$$\begin{aligned} \hat{S}(\tilde{\beta}) - R'(\tilde{\beta})'\tilde{\lambda} &= 0, \\ R(\tilde{\beta}) - r &= 0, \end{aligned}$$

In deriving the asymptotic distribution of the LR test statistic, we have obtained

$$\sqrt{n}\tilde{\lambda} = \left[R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)R'(\tilde{\beta})'\right]^{-1} R'(\bar{\beta}_d)\hat{H}^{-1}(\bar{\beta}_c)\sqrt{n}\hat{S}(\beta^o)$$

for n sufficiently large. By the CLT, we have $\sqrt{n}\hat{S}(\beta^o) \to^d N(0, V_o)$, where $V_o = \text{avar}[\sqrt{n}\hat{S}(\beta^o)]$. Using the Slutsky theorem, we can obtain

$$\sqrt{n}\tilde{\lambda} \xrightarrow{d} N(0, \Omega),$$

where

$$\begin{aligned} \Omega &= [R'(\beta^o)H_0^{-1}R'(\beta^o)']^{-1} \\ &\quad \times R'(\beta^o)H_o^{-1}V_oH_o^{-1}R'(\beta^o)' \\ &\quad \times [R'(\beta^o)H_o^{-1}R'(\beta^o)']^{-1}. \end{aligned}$$

Then a robust LM test statistic

$$LM \equiv n\tilde{\lambda}'\tilde{\Omega}^{-1}\tilde{\lambda} \to^d \chi^2_J$$

by the Slutsky theorem, where the asymptotic variance estimator

$$\begin{aligned}\tilde{\Omega} &= [R'(\tilde{\beta})\hat{H}^{-1}(\tilde{\beta})R'(\tilde{\beta})']^{-1} \\ &\quad \times[R'(\tilde{\beta})\hat{H}^{-1}(\tilde{\beta})\tilde{V}\hat{H}^{-1}(\tilde{\beta})R'(\tilde{\beta})'] \\ &\quad \times[R'(\tilde{\beta})\hat{H}^{-1}(\tilde{\beta})R'(\tilde{\beta})']^{-1},\end{aligned}$$

and $\tilde{V}$ satisfies the following condition:

Assumption 9.8: $\tilde{V} \to^p V_o$, where $\tilde{V}$ is defined as $\hat{V}$ in Assumption 9.7 with $\hat{\beta}$ replaced with $\tilde{\beta}$.

With this assumption, the LM test statistic will only involves estimation of the conditional pdf/pmf model $f(y|\Psi_t, \beta)$ under the null hypothesis $\mathbf{H}_0$.

Theorem [QMLE-based LM Test]: *Suppose Assumptions 9.1–9.6 and 9.8 and* $\mathbf{H}_0 : R(\beta^o) = r$ *holds. Then as* $n \to \infty$,

$$LM \equiv n\tilde{\lambda}'\tilde{\Omega}^{-1}\tilde{\lambda} \to^d \chi^2_J.$$

Remark:

The LM$_0$ test statistic under MLE and the LM test statistic under QMLE differ in the sense that they use different asymptotic variance estimators. The LM test statistic here is robust to misspecification of the conditional pdf/pmf model $f(y|\Psi_t, \beta)$.

Question: Could we use the likelihood ratio (LR) test under model specification?

$$LR = 2n[\hat{l}(\hat{\beta}) - \hat{l}(\tilde{\beta})].$$

No. This is because in deriving the asymptotic distribution of the LR test statistic, we have used the MDS property of the score function $\{S_t(\beta^o)\}$ and the information matrix equality ($V_o = -H_o$),

which may not hold when the conditional distribution model $f(y|\Psi_t, \beta)$ is misspecified. If the MDS property of the score function or the information matrix equality fails, the LR statistic is not asymptotically χ^2_J under $\mathbf{H}_0$. This is similar to the fact that J times the F-test statistic does not converge to χ^2_J when there exists serial correlation in $\{\varepsilon_t\}$ or when there exists conditional heteroskedasticity.

In many applications (e.g., estimating CAPM models), both GMM and QMLE can be used to estimate the same parameter vector. In general, by making fewer assumptions on the DGP, GMM will be less efficient than QMLE if the pesudo-model likelihood function is close to the true conditional distribution of Y_t given Ψ_t.

9.5 Model Specification Testing

It is important to check whether a conditional probability distribution $f(y|\Psi_t, \beta)$ is correctly specified. There are various reasons:

(i) A misspecified pdf/pmf model $f(y|\Psi_t, \beta)$ implies suboptimal forecasts of the true probability distribution of the underlying process.

(ii) The QMLE based on a misspecified pdf/pmf model $f(y|\Psi_t, \beta)$ is less efficient than the MLE based on a correctly specified pdf/pmf model.

(iii) A misspecified pdf/pmf model $f(y|\Psi_t, \beta)$ implies that we have to use a robust version of the asymptotic variance of QMLE, because the conditional information matrix equality no longer holds among other things. As a consequence, the resulting statistical inference procedures are more tedious.

Question: How to check whether a conditional distribution model $f(y|\Psi_t, \beta)$ is correctly specified?

We now introduce a number of specification tests for conditional distributional model $f(y|\Psi_t, \beta)$.

Case 1: When $\{Z_t = (Y_t, X_t')'\} \sim$ i.i.d.

When the data generating process is an i.i.d. sequence, we have

$$\sqrt{n}(\hat{\beta} - \beta^o) \to N(0, H_o^{-1} V_o H_o^{-1}),$$

where

$$V_o = E[S_t(\beta^o) S_t(\beta^o)'].$$

White's (1982) Information Matrix Test

In the i.i.d. random sample context, White (1982) proposes a specification test for $f(y|\Psi_t, \beta) = f(y|X_t, \beta)$ by checking whether the information matrix equality holds:

$$E\left[S_t(\beta^o) S_t(\beta^o)'\right] + E[H_t(\beta^o)] = 0.$$

This is implied by correct model specification. If the information matrix equality does not hold, then there is evidence of model misspecification for the conditional distribution of Y given X.

Define the $\frac{K(K+1)}{2} \times 1$ sample average

$$\hat{m}(\beta) = \frac{1}{n} \sum_{t=1}^{n} m_t(\beta),$$

where

$$m_t(\beta) = \text{vech}\left[S_t(\beta)S_t(\beta)' + H_t(\beta)\right].$$

Then one can check whether the sample average $\hat{m}(\hat{\beta})$ is close to zero (the population moment).

How large the magnitude of $\hat{m}(\hat{\beta})$ should be in order to be considered as significantly larger than zero can be determined by the asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$.

Question: How to derive the asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$?

White (1982) proposes an information matrix test using a suitable quadratic form of $\sqrt{n}\hat{m}(\hat{\beta})$ that is asymptotically $\chi^2_{K(K+1)/2}$ under correct model specification. Specifically, White (1982) shows that

$$\begin{aligned} n^{1/2}\hat{m}(\hat{\beta}) &= n^{-1/2}\sum_{t=1}^{n}[m_t(\beta^o) - D_0H_0^{-1}S_t(\beta^o)] \\ &\to {}^d N(0, W), \end{aligned}$$

where $D_o \equiv D(\beta^o) = E\left[\frac{\partial m_t(\beta^o)}{\partial \beta}\right]$, and the asymptotic variance

$$W = \text{var}\left[m_t(\beta^o) - D_oH_o^{-1}S_t(\beta^o)\right].$$

It follows that a test statistic can be constructed by using the quadratic form

$$M = n\hat{m}(\hat{\beta})'\hat{W}^{-1}\hat{m}(\hat{\beta}) \to^d \chi^2_{K(K+1)/2}$$

for some consistent variance estimator $\hat{W}$ for W. Putting $\hat{W}_t = m_t(\hat{\beta}) - \hat{D}(\hat{\beta})\hat{H}^{-1}(\hat{\beta})S_t(\hat{\beta})$, we can use the variance estimator

$$\hat{W} = \frac{1}{n}\sum_{t=1}^{n}\hat{W}_t\hat{W}_t'.$$

Question: If the information matrix equality holds, is the model $f(y|X_t, \beta)$ correctly specified for the conditional distribution of Y_t given X_t?

Answer: No. Correct model specification implies the information matrix equality but the converse may not be true. The information matrix equality is only one of many (infinite) implications of the correct specification for $f(y|\Psi_t, \beta)$.

Although White (1982) considers i.i.d. random samples only, his IM test is applicable for both cross-sectional and time series models as long as the score function $\{S_t(\beta^o)\}$ is an MDS.

Case 2: $\{Z_t = (Y_t, X_t')'\}$ is a serially dependent process.

White's (1994) Dynamic Information Matrix Test:

In a time series context, White (1994) proposes a dynamic information matrix test that essentially checks the MDS property of the score function $\{S_t(\beta^o)\}$:

$$E[S_t(\beta^o)|\Psi_t] = 0,$$

which is implied by correct model specification for $f(y|\Psi_t, \beta)$.

Let

$$m_t(\beta) = \text{vech}[S_t(\beta) \otimes W_t(\beta)],$$

where $W_t(\beta) = [S_{t-1}(\beta)', S_{t-2}(\beta)', ..., S_{t-p}(\beta)']'$ and $\otimes$ is the Kronecker product. Then the MDS property implies

$$E[m_t(\beta^o)] = 0.$$

This test is essentially checking whether $\{S_t(\beta^o)\}$ is a white noise process up to lag order p. If $E[m_t(\beta^o)] \neq 0$, i.e., if there exists serial correlations in $\{S_t(\beta^o)\}$, then there is evidence of model misspecification.

White (1994) considers the sample average

$$\hat{m} = n^{-1} \sum_{t=1}^{n} m_t(\hat{\beta})$$

and checks if this is close to zero. White (1994) develops a so-called dynamic information matrix test by using a suitable quadratic form of $\sqrt{n}\hat{m}$ that is asymptotically chi-squares distributed under correct dynamic model specification.

Question: If $\{S_t(\beta^o)\}$ is MDS, is $f(y|\Psi_t, \beta)$ correctly specified for the conditional distribution of Y_t given Ψ_t?

No. Correct model specification implies that $\{S_t(\beta^o)\}$ is a MDS but the converse may not be true. It is possible that $S_t(\beta^o)$ is an MDS even when the model $f(y|\Psi_t, \beta)$ is misspecified for the conditional distribution of Y_t given Ψ_t. A better approach is to test the conditional density model itself, rather than the properties of its derivatives (e.g., the MDS of the score function or the information matrix equality).

Next, we consider a test that directly checks the conditional distribution of Y_t given Ψ_t.

Hong and Li's (2005) Nonparametric Test for Time Series Conditional Distribution Models

Suppose Y_t is a univariate continuous random variable, and $f(y|\Psi_t, \beta)$ is a conditional distribution model of Y_t given Ψ_t. Define the dynamic probability integral transform

$$U_t(\beta) = \int_{-\infty}^{Y_t} f(y|\Psi_t, \beta)dy.$$

Lemma: *If $f(y|\Psi_t, \beta^o)$ coincides with the true conditional pdf of Y_t given Ψ_t, then*

$$\{U_t(\beta^o)\} \sim \ i.i.d.\text{U[0,1]}.$$

Thus, one can test whether $\{U_t(\beta^o)\}$ is i.i.d.U[0,1]. If it is not, there exists evidence of model misspecification.

Question: Suppose $\{U_t(\beta^o)\}$ is i.i.d.U[0,1], is the model $f(y|\Psi_t, \beta)$ correctly specified for the conditional distribution of Y_t given Ψ_t?

For univariate time series (so that $\Psi_t = \{Y_{t-1}, Y_{t-2}, ...\}$), the i.i.d.U[0,1] property holds if and only if the conditional pdf model $f(y_t|\Psi_t, \beta)$ is correctly specified.

Hong and Li (2005) use a nonparametric kernel estimator for the joint density of $\{U_t(\beta^o), U_{t-j}(\beta^o)\}$ and compare the joint density estimator with $1 = 1 \cdot 1$, the product of the marginal densities of $U_t(\beta^o)$ and $U_{t-j}(\beta^o)$ under correct model specification. The test statistic follows an asymptotical N(0,1) distribution. See Hong and Li (2005) for more discussion.

9.6 Empirical Applications

Empirical Application I: China's Evolving Managerial Labor Market

Groves, Hong, McMillan and Naughton (1995, *Journal of Political Economy*)

Question: How does the industrial bureau decide to use the competitive auction to select firm managers?

We define a binary variable as follows: $Y_t = 1$ if the current manager of firm t selected by competitive auction, and $Y_t = 0$ otherwise. We shall use the past performance of a firm and the size of a firm to predict the probability of $Y_t = 1$. Thus, we put $X_t = (1, X_{1t}, X_{2t})'$, where X_{1t} = past performance of firm t (the average output per worker in the past 3-year relative to the industry average), X_{2t} = the size of firm t (the number of employees of firm t relative to the industry)

We specify a probit model:

$$P(Y_t = 1|X_t) = \Phi(X_t'\beta),$$

where $\Phi(\cdot)$ is the N(0,1) CDF.

Estimation Results:

X_{1t}	X_{2t}	n
-0.2769^{**}	-0.2467^{**}	645 ,
(-7.485)	(-7.584)	

where ** indicates significance at the 5% level. These results suggest that the poor-performing and/or smaller firms are more likely to have their managers selected by competitive auction.

Empirical Application II: Full Dynamics of the Short-Term Interest Rates

Data: Daily series of 7-day Eurodollar rates $\{r_t\}$ from June 1, 1973 to February 25, 1995. The sample size $T = 5050$.

We are interested in modelling the conditional probability distribution of the short-term interest rate. There are two popular discrete-time models for the spot interest rate: one is the GARCH model, and the other is the Markov chain regime-switching model.

Model 1: GARCH(1,1)-Level Effect with an i.i.d. N(0,1) innovation:

$$\begin{cases} \Delta r_t & = \alpha_{-1} r_{t-1}^{-1} + \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-2}^2 + \sigma r_{t-1}^{\rho} h_t^{1/2} z_t, \\ h_t & = \beta_0 + \beta_1 h_{t-1} + \beta_2 h_{t-1} z_{t-1}^2, \\ \{z_t\} & \sim i.i.d. N(0,1). \end{cases}$$

Here, the conditional mean of the interest rate change is a nonlinear function of the interest rate level:

$$\mu_t = E(\Delta r_t | I_{t-1}) = \alpha_{-1} r_{t-1}^{-1} + \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-2}^2.$$

This specification can capture nonlinear dynamics in the interest rate movement.

The conditional variance model of the interest rate change is

$$\sigma_t^2 = \text{var}(\Delta r_t | I_{t-1}) = \sigma^2 r_{t-1}^{2\rho} h_t,$$

where r_{t-1}^{ρ} captures the so-called "level effect" in the sense that when $\rho > 0$, volatility will increase when the interest rate level is high. On the other hand, the GARCH component h_t captures volatility clustering.

Estimation Results

Parameter Estimates for the GARCH Model (with nonlinear drift and level effect)

Parameters	Estimates (GARCH)	Std. Error (GARCH)
α_{-1}	-0.0984	0.1249
α_0 (1e-02)	5.0494	6.3231
α_1 (1e-03)	-4.4132	9.2876
α_2	0.0000	0.0004
ρ	1.0883	0.0408
β_0 (1e-03)	0.0738	0.0119
β_2 (1e-01)	6.4117	0.1359
β_1 (1e-01)	3.5260	0.2181
Log-Likelihood	654.13	

Model 2: Regime-Switching Model with GARCH and Level Effects

$$\begin{aligned}
\Delta r_t &= \alpha\left(S_{t-1}\right) + \beta\left(S_{t-1}\right) r_{t-1} + \sigma\left(S_{t-1}\right) r_{t-1t}^{\rho(S_{t-1})} h_t^{1/2} z_{t-1}, \\
h_t &= \beta_0 + h_{t-1}\left(\beta_1 + \beta_2 z_{t-1}^2\right), \\
\{z_t\} &\sim i.i.d.N(0,1),
\end{aligned}$$

where the state variable S_t is a latent process that is assumed to follow a two-state Markov chain with time-varying transition matrix, as specified in Ang and Bekaert (1998):

$$\begin{aligned}
P\left(S_t = 1 | S_{t-1} = 1\right) &= \left[1 + \exp(-a_{01} - a_{11} r_{t-1})\right]^{-1}, \\
P\left(S_{t-1} = 0 | S_{t-1} = 0\right) &= \left[1 + \exp(-a_{00} - a_{10} r_{t-1})\right]^{-1}.
\end{aligned}$$

Question: What is the model likelihood function? That is, what is the conditional density of Δr_t given $I_{t-1} = \{r_{t-1}, r_{t-2}, ...\}$, the observed information set available at time $t-1$?

The difficulty arises because the state variable S_t is not observable. See Hamilton (1994, Chapter 22) for treatment.

Estimation Results

Parameter estimates for the Regime Switching Model (with GARCH and level effect)

Parameters	Estimates (RS)	Std. Error (RS)
α_0	1.5378	1.5378
β_0	-1.0646	0.4207
α_1	-0.0013	0.0351
β_1	-0.0076	0.0484
σ_1	0.3355	0.0483
ρ_0	0.3566	0.0693
ρ_1	0.0064	0.0512
b_0 (1e-03)	6.5126	1.9898
b_1	0.0224	0.0034
b_2	0.7810	0.0254
a_{00}	0.2350	0.2192
a_{01}	4.5398	0.2691
a_{10}	0.0208	0.0184
a_{11}	-0.2800	0.0296
Log-Likelihood	2712.97	

Empirical III: Volatility Models of Foreign Exchange Returns

Hong (2001, *Journal of Econometrics*)

Suppose one is interested in studying volatility spillover between two exchange rates—German Deutschmark and Japanese Yen. A first step is to specify a univariate volatility for German Deutschmark and Japanese yen respectively. Hong fits an AR(3)-GARCH(1,1) model for weekly German Deutschmark exchange rate changes and Japanese Yen exchange rate changes:

Model: AR(3)-GARCH(1,1)-i.i.d.N(0,1)

$$\begin{cases} X_t = \mu_t + \varepsilon_t, \\ \mu_t = b_0 + \sum_{j=1}^3 b_j X_{t-j}, \\ \varepsilon_t = h_t^{1/2} z_t, \\ h_t = \omega + \alpha\varepsilon_{t-1}^2 + \gamma h_{t-1}, \\ \beta = (b_0, b_1, b_2, b_3, \omega, \alpha, \gamma)'. \end{cases}$$

Assuming that $\{z_t\} \sim i.i.d.N(0,1)$, we obtain the following QMLE.

Data: First week of 1976:1 to last week of 1995:11, with totally 1039 observations.

Estimation results

	DM		*YEN*	
Parameter	Estimate	s.d.	Estimate	s.d.
b_0	-0.073	0.041	-0.097	0.042
b_1	0.049	0.033	0.051	0.034
b_2	0.067	0.033	0.093	0.034
b_3	-0.028	0.033	0.066	0.033
ω	0.051	0.030	0.116	0.068
α	0.114	0.027	0.084	0.026
γ	0.873	0.033	0.863	0.055
Sample Size	1038		1038	
Log-Likelihood	-1862.307		-1813.625	

The standard errors reported here are robust standard errors.

9.7 Summary and Conclusion

Conditional probability distribution models have wide applications in economics and finance. For some applications, one is required to specify the entire distribution of the underlying process. If the distribution model is correct, the resulting estimator $\hat{\beta}$ which maximizes the likelihood function is called MLE.

For some other applications, on the other hand, one is only required to specify certain aspects (e.g., conditional mean and conditional variance) of the distribution. One important example is volatility modeling for financial time series. To estimate model parameters, one usually makes some auxiliary assumptions on the distribution that may be incorrect so that one can estimate β by maximizing the pseudo likelihood function. This is called QMLE. MLE is asymptotically more efficient than QMLE, because the asymptotic variance of MLE attains the Cramer-Rao lower bound.

The likelihood function of a correctly specified conditional distributional model has different properties from that of a misspecified conditional distributional model. In particular, for a correctly specified distributional model, the score function is an MDS and the conditional information matrix equality holds. As a consequence, the asymptotic distributions of MLE and QMLE are different (more precisely, their asymptotic variances are different). In particular, the asymptotic variance of MLE is analogous to that of the OLS estimator under MDS regression errors with conditional homoskedasticity; and the asymptotic variance of QMLE is analogous to that of the OLS estimator under possibly non-MDS with conditional heteroskedasticity.

Hypothesis tests can be developed using MLE or QMLE. For hypothesis testing under a correct specified conditional distributional models, the Wald test, Lagrange Multiplier test, and Likelihood Ratio tests can be used. When a conditional distributional model is misspecified, robust Wald tests and LM tests can be constructed. Like the F-test in the regression context, Likelihood ratio tests are valid only when the distribution model is correctly specified. The reasons are that they exploit the MDS property of the score function and the information matrix equality which may not hold under model misspecification.

It is important to test correct specification of a conditional distributional model. We introduce some specification tests for conditional distributional models under i.i.d. observations and time series observations respectively. In particular, White (1982) proposes an Information Matrix test for i.i.d. observations and White (1994) proposes a dynamic information matrix test that essentially checks the MDS property of the score function of a correctly specified conditional distribution model with time series observations.

References

Hong, Y. and H. Li (2005), Review of Financial Studies

Rosenblatt, M. (1952), Annals of Mathematical Statistics

White, H. (1982), Econometrica

White, H. (1994), Estimation, Inference and Specification Analysis

EXERCISES

9.1. For the probit model $P(Y_t = y|X_t) = \Phi(X_t'\beta^o)^y[1 - \Phi(X_t'\beta^o)]^{1-y}$, where $y = 0, 1$. Show that
(a) $E(Y_t|X_t) = \Phi(X_t'\beta^o)$;
(b) $\mathrm{var}(Y_t|X_t) = \Phi(X_t'\beta^o)[1 - \Phi(X_t'\beta^o)]$.

9.2. For a censored regression model, show that $E(X_t\varepsilon_t|Y_t > c) \neq 0$. Thus, the OLS estimator based on a censored random sample cannot be consistent for the true model parameter β^o.

9.3. Suppose $f(y|\Psi, \beta)$ is a conditional pdf model for Y given Ψ, where $\beta \in \Theta$, a parameter space. Show that for all $\beta, \dot{\beta} \in \Theta$ and all ψ,

$$\int \ln[f(y|\psi, \beta)]f(y|\psi, \dot{\beta})dy \leq \int \ln[f(y|\psi, \dot{\beta})]f(y|\psi, \dot{\beta})dy.$$

9.4. **(a)** Suppose $f(y|\psi, \beta)$, $\beta \in \Theta$, is a correctly specified model for the conditional probability density of Y given Ψ, such that $f(y|\psi, \beta^o)$ coincides with the true conditional probability density of Y given Ψ. We assume that $f(Y|\Psi, \beta)$ is continuously differentiable with respect to β and β^o is an interior point in Θ. Please show that

$$E\left[\left.\frac{\partial \ln f(Y|\Psi, \beta^o)}{\partial \beta}\right| \Psi\right] = 0.$$

(b) Suppose Part (a) is true. Can we conclude that $f(y|\Psi, \beta)$ is correctly specified for the conditional distribution of Y given Ψ? If yes, give your reasoning. If not, give a counter example.

9.5. Suppose $f(y|x, \beta)$, $\beta \in \Theta \subset R^K$, is a correctly specified model for the conditional probability density of Y given X, such that for some parameter value β^o, $f(y|x, \beta^o)$ coincides with the true conditional probability density of Y given X. We assume that $f(Y|x, \beta)$ is continuously differentiable with respect to β and β^o is an interior point in Θ. Please show that

$$E\left[\left.\frac{\partial \ln f(Y|X, \beta^o)}{\partial \beta}\frac{\partial \ln f(Y|X, \beta^o)}{\partial \beta'}\right| X\right] + E\left[\left.\frac{\partial^2 \ln f(Y|X, \beta^o)}{\partial \beta \partial \beta'}\right| X\right] = 0,$$

where $\frac{\partial \ln f}{\partial \beta}$ is a $K \times 1$ vector, $\frac{\partial \ln f}{\partial \beta'}$ is the transpose of $\frac{\partial \ln f}{\partial \beta}$, $\frac{\partial^2 \ln f}{\partial \beta \partial \beta'}$ is a $K \times K$ matrix, and the expectation $E(\cdot)$ is taken under the true conditional distribution of Y given X.

9.6. Put $V_o = E[S_t(\beta^o)S_t(\beta^o)']$ and $H_o = E[\frac{\partial}{\partial \beta}S_t(\beta^o)] = E[\frac{\partial^2}{\partial \beta \partial \beta'} \ln f_{Y_t|\Psi_t}(y|\Psi_t, \beta^o)]$, where $S_t(\beta) = \frac{\partial}{\partial \beta} \ln f(Y_t|\Psi_t, \beta)$, and $\beta^o = \arg\min_{\beta \in \Theta} l(\beta) = E[\ln f_{Y_t|\Psi_t}(Y_t|\Psi_t, \beta)]$. Is $H_o^{-1}V_oH_o^{-1} -$

$(-H_o^{-1})$ always positive semi-definite? Give your reasoning and any necessary regularity conditions. Note that the first term $H_o^{-1}V_oH_o^{-1}$ is the formula for the asymptotic variance of $\sqrt{n}\hat{\beta}_{QMLE}$ and the second term $-H_o^{-1}$ is the formula for the asymptotic variance of $\sqrt{n}\hat{\beta}_{MLE}$.

9.7. Suppose a conditional pdf/pmf model $f(y|x,\beta)$ is misspecified for the conditional distribution of Y given X, namely, there exists no $\beta \in \Theta$ such that $f(y|x,\beta)$ coincides with the true conditional distribution of Y given X. Show that generally,

$$E\left[\frac{\partial \ln f(Y|X,\beta^o)}{\partial \beta}\frac{\partial \ln f(Y|X,\beta^o)}{\partial \beta'}\middle| X\right] + E\left[\frac{\partial^2 \ln f(Y|X,\beta^o)}{\partial \beta \partial \beta'}\middle| X\right] = 0,$$

does not hold, where β^o satisfies Assumptions 9.4 and 9.5. In other words, the conditional information matrix equality generally does not hold when the conditional pdf/pmf model $f(y|x,\beta)$ is misspecified for the conditional distribution of Y given X.

9.8. Consider the following maximum likelihood estimation problem:

Assumption 7.1: $\{Y_t, X_t'\}'$ is a stationary ergodic process, and $f(Y_t|\Psi_t,\beta)$ is a *correctly specified* conditional probability density model of Y_t given $\Psi_t = (X_t', Z^{t-1\prime})'$, where $Z^{t-1} = (Z_{t-1}', Z_{t-2}', \cdots, Z_1')'$ and $Z_t = (Y_t, X_t')'$. For each β, $\ln f(Y_t|\Psi_t,\beta)$ is measurable of the data, and for each t, $\ln f(Y_t|\Psi_t,\cdot)$ is twice continuously differentiable with respect to $\beta \in \Theta$, where Θ is a compact set.

Assumption 7.2: $l(\beta) = E[\ln f(Y_t|\Psi_t,\beta)]$ is continuous in $\beta \in \Theta$.

Assumption 7.3: (i) $\beta^o = \arg\max_{\beta\in\Theta} l(\beta)$ is the unique maximizer of $l(\beta)$ over Θ, and (ii) β^o is an interior point of Θ.

Assumption 7.4: (i) $\{S_t(\beta^o) \equiv \frac{\partial}{\partial\beta}\ln f(Y_t|\Psi_t,\beta)\}$ obeys a CLT, i.e.,

$$\sqrt{n}\hat{S}(\beta^o) = n^{-1/2}\sum_{t=1}^{n} S_t(\beta^o)$$

converges to a multivariate normal distribution with some $K \times K$ variance-covariance matrix; (ii) $\{H_t(\beta) \equiv \frac{\partial^2}{\partial\beta\partial\beta'}\ln f(Y_t|\Psi_t,\beta)\}$ obeys a uniform weak law of large numbers (UWLLN) over Θ. That is,

$$\lim_{n\to\infty}\sup_{\beta\in\Theta}\left\|n^{-1}\sum_{t=1}^{n} H_t(\beta) - H(\beta)\right\| = 0 \text{ a.s.},$$

where the $K \times K$ Hessian matrix $H(\beta) \equiv E[H_t(\beta)]$ is symmetric, finite and nonsingular, and is continuous in $\beta \in \Theta$.

The maximum likelihood estimator is defined as $\hat{\beta} = \arg\max_{\beta\in\Theta}\hat{l}_n(\beta)$, where $\hat{l}_n(\beta) \equiv n^{-1}\sum_{t=1}^{n}\ln f(Y_t|\Psi_t,\beta)$. Suppose we have had $\hat{\beta} \to \beta^o$ almost surely, and this consistency result can be used in answering the following questions in parts (a)–(d). Show your reasoning in *each* step.

(a) Find the first order condition of the MLE.

(b) Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}-\beta^o)$. Note that the asymptotic variance of $\sqrt{n}(\hat{\beta}-\beta^o)$ should be expressed as the Hessian matrix $H(\beta^o)$.

(c) Find a consistent estimator for the asymptotic variance of $\sqrt{n}(\hat{\beta}-\beta^o)$ and justify why it is consistent.

(d) Construct a Wald test statistic for the null hypothesis $\mathbf{H}_0 : R(\beta^o) = r$, where r is a $J \times 1$ constant vector, and $R(\cdot)$ is a $J \times 1$ vector with the derivative $R'(\beta)$ is continuous in β and $R'(\beta^o)$ is of full rank. Derive the asymptotic distribution of the Wald test under $\mathbf{H}_0$.

9.9. In a linear regression model $Y_t = X_t'\alpha^o + \varepsilon_t$, where $\varepsilon_t|\Psi_t \sim N(0, \sigma_o^2)$. Put $\beta = (\alpha', \sigma^2)'$ and note that

$$\begin{aligned}
f(Y_t|X_t, \beta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_t - X_t'\alpha)^2}, \\
\hat{l}(\beta) &= n^{-1}\sum_{t=1}^{n} \ln f(Y_t|X_t, \beta) \\
&= -\frac{1}{2\sigma^2}\ln(2\pi) - \frac{1}{2\sigma^2} n^{-1}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2.
\end{aligned}$$

Suppose $\mathbf{H}_0 : R\beta^o = r$ is the hypothesis of interest.

(a) Show

$$\begin{aligned}
\hat{l}(\hat{\beta}) &= \frac{1}{2}\ln(e'e), \\
\hat{l}(\tilde{\beta}) &= \frac{1}{2}\ln(\tilde{e}'\tilde{e}),
\end{aligned}$$

where $\tilde{\beta}$ is the MLE under $\mathbf{H}_0$.

(b) Show that under $\mathbf{H}_0$,

$$\begin{aligned}
2n[\hat{l}(\tilde{\beta}) - \hat{l}(\hat{\beta})] &= n\ln(\tilde{e}'\tilde{e}/e'e) \\
&= J \cdot \frac{(\tilde{e}\tilde{e} - e'e)/J}{e'e/n} + o_P(1) \\
&= J \cdot F + o_P(1).
\end{aligned}$$

9.10. Show the dynamic probability integral transforms $\{U_t(\beta^o)\}$ is i.i.d.U[0,1] if the conditional probability density model $f(y|\Psi_t, \beta)$ is correctly specified for the conditional distribution of Y_t given Ψ_t.