Financial Econometrics

Yanchu Wang

Shanghai University of Finance and Economics

September 2024

Introduction

- Determination of prices and returns in stocks, bonds, and other financial assets.
- Empirical analysis of prices and returns $\Rightarrow$ econometric methods.
- The empirical analysis is based on economic and finance theories (asset pricing theory etc.).
- Use of econometric techniques to test and evaluate asset pricing theories.
- Programming skill is required.
- Exercises using actual data from financial markets.

Introduction

- Main focus on:
  - Introduction of basic theories in econometrics.
  - The test of asset pricing theory in equity markets.
  - The term structure of interest rates and bond pricing.
  - Nonlinearity and nonparametric models.

Examples of Topics

- How to measure "return" and "risk"?
- What basically determine stock returns - over time and across assets?
- Are stock returns predictable? If so, is it consistent with the "efficient market hypothesis", or can we use the predictability to earn abnormal riskfree profits?
- How can we measure portfolio managers performance?
- Are derivatives are really "redundant"?
- What causes different interest rates between long-term and short-term bonds?

The course addresses all these issues in a econometric framework.

Outline of the course

- Basic concepts in econometrics: Estimation, Inference, and Hypothesis Testing, etc.
- Time series analysis
  1. Review of statistical distributions and their moments.
  2. Stationary process, cointegration, unit root, etc.
  3. AR, MA, and ARMA models.
- The predictability of stock returns.
- The CAPM.
- The multifactor asset pricing models.
- Expectation Hypothesis and its test.
- Interest rate models.
- The GARCH model
- Nonparametric models.
- State-space models and the Kalman Filter.

Literature

- Campbell, J. Y., Lo, A. W., and Mackinlay, A. C., 1997. *The Econometrics of Financial Markets*, Princeton University Press.
- Tsay, R. S., 2010. *Analysis of Financial Time Series*, John Wiley & Sons, Inc.
- Kim, C. J., and Nelson, C. R., 1999. State-Space Models with Regime Switching. The MIT Press.
-
- Selected Articles

Other Information

- The form of evaluation: hand-in exercises + final exam.
- My email: wang.yanchu@mail.shufe.edu.cn.
- Office: 319, Tongde Lou.

Estimation, Inference, and Hypothesis Testing

- We first take an overview on estimation, distribution theory, inference, and hypothesis testing.
- Three steps must be completed to test the implications of a theory in economics or finance after a model is set up:
  1. Estimate unknown parameters.
  2. Determine the distribution of estimators.
  3. Conduct hypothesis tests to examine whether the data support the theoretical model.
- We start with cases where the data is independent and identically distributed (I.I.D.), later we will release the restriction.

Estimation

- Parametric methods:
  $y_t = \beta x_t + \varepsilon_t$. Linear.
  $y_t = e^{x_t} + x_t^2 + \varepsilon_t$, nonlinear.
- Nonparametric methods.
  For example, $y_t = f(x_t)$, but we don't know the functional form of $f(\cdot)$.
- Semi-parametric methods.
  Assume parametric relationships between variables but estimate the underlying distribution of errors flexibly.
- Semi-nonparametric methods.
  Estimators that take a stand on the distribution of the errors but allow for flexible relationships between variables.

Estimation

- The M-, L-, and R-estimator classes.
- M-estimators (extremum estimators) involve maximize or minimize some objective function.
  The class contains ordinary least square (OLS), maximum likelihood, classical minimum distance, and both the classical and generalized method of moments.
- L-estimators (linear estimators) are a class where the estimator can be expressed as a linear function of ordered data.
  It takes the form
  $$\sum_{i=1}^{n} w_i y_i,$$
  for some set of weights $\{w_i\}$, and $\{y_i\}$ are ordered such that $y_{j-1} \leqslant y_j$ for $j = 2, 3, \cdots, n$.
- R-estimators: explore the rank of the data. Examples including the minimum, maximum, and Spearman's rank correlation.
  Rank statistics are often robust to outliers and non-linearilities.

Statistical Distribution

- For a scalar random variable X, $F_X(a;\boldsymbol{\theta}) = P(X \leqslant a;\boldsymbol{\theta})$, which is called the cumulative distribution function (CDF) of X.
- The CDF of a random variable is nondecreasing, i.e., $F_X(x_1) \leqslant F_X(x_2)$ if $x_1 \leqslant x_2$.
- $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.
- For a given probability, pth quantile of x is given by:

$$a_p = \inf_a \{a | p \leqslant F_X(a)\}. \tag{1}$$

- Conditional distribution:

$$f_{x|y}(\mathbf{x};\boldsymbol{\theta}) = \frac{f_{x,y}(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{f_y(\mathbf{y};\boldsymbol{\theta})} \tag{2}$$

- If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $f_{x|y}(\mathbf{x};\boldsymbol{\theta}) = f_x(\mathbf{x};\boldsymbol{\theta})$, we have $f_{x,y}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = f_x(\mathbf{x};\boldsymbol{\theta}) \times f_y(\mathbf{y};\boldsymbol{\theta})$.

Statistical Distribution

- Let R^k and R^q be the $k-$dimensional and the $q-$dimensional space, respectively. Consider two random vectors, $\mathbf{X} = (X_1, ..., X_k)'$ and $\mathbf{Y} = (Y_1, ..., Y_q)'$. Let $P(\mathbf{X} \in A, \mathbf{Y} \in B)$ be the probability that $\mathbf{X}$ is in the subspace $A \subset R^k$ and $\mathbf{Y}$ is in the subspace $B \subset R^q$.
- Joint Distribution: $F_{X,Y}(\mathbf{a}, \mathbf{b}; \boldsymbol{\theta}) = P(\mathbf{X} \leqslant \mathbf{a}, \mathbf{Y} \leqslant \mathbf{b}; \boldsymbol{\theta})$. If the joint density function $fx, y(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ exists, then

$$F_{X,Y}(\mathbf{a}, \mathbf{b}; \boldsymbol{\theta}) = \int_{-\infty}^{\mathbf{a}} \int_{-\infty}^{\mathbf{b}} fx, y(\mathbf{w}, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} d\mathbf{w}. \tag{3}$$

- Marginal Distribution of $\mathbf{X}$: $F_X(\mathbf{a}; \boldsymbol{\theta}) = F_{X,Y}(\mathbf{a}, \infty, ..., \infty; \boldsymbol{\theta})$, i.e., the marginal distribution of $\mathbf{X}$ is obtained by integrating out $\mathbf{Y}$.

Method of Moments

- Noncentral Moments
$$\mu'_r = E[x^r], \tag{4}$$
for $r = 1, 2, \cdots$. For estimation
$$\widehat{\mu}'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r .$$

- Central Moments
$$\mu_r = E[(x - \mu_1)^r], \tag{5}$$
for $r = 2, 3, \cdots$. For estimation
$$\widehat{\mu}_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_1)^r .$$

Method of Moments

- The first moment is called mean of x, and the second central moment is called variance of x.

$$\begin{aligned}\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2.\end{aligned}$$

- Each noncentral moment can be shown as an expression of central moments.
- For example, it can also be shown that

$$\sigma^2 = \mu_2' - \mu_1^2.$$

- We can also show that

$$\mu_3' = \mu_3 + 3\mu_2\mu_1 + \mu_1^3.$$

Method of Moments

- The method of moments is to use the moment conditions to estimate unknown parameters.
- It is in the class of M-estimators. To understand this, note that the central moments are actually the first order condition for the following quadratic form:

$$\arg \min_{\mu,\mu_2,\ldots\mu_k} (n^{-1} \sum_{i=1}^{n} x_i - \mu)^2 + \sum_{j=2}^{k} (n^{-1} \sum_{i=1}^{n} (x_i - \mu)^j - \mu_j)^2.$$

- For example, for a given data set $\{y_i\}_{i=1}^n$, we can estimate the mean $\mu = E[y_i]$ and variance $\sigma^2 = E[(y_i - \mu)^2]$ as follows

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} y_i, \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \widehat{\mu}^2, \text{ or} \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\mu})^2. \end{aligned}$$

Classical Minimum Distance

- Classical minimum distance (CMD) method can be used when we need to add some constraints (by economic means) to the model.
- For example, we can use MLE or method of moments without constraints to obtain the initial estimator, $\widehat{\boldsymbol{\psi}}$, of a model, and then add the restrictions, $g(\boldsymbol{\theta})$. CMD is used to minimize the quadratic form:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (\widehat{\boldsymbol{\psi}} - g(\boldsymbol{\theta}))' \mathbf{W} (\widehat{\boldsymbol{\psi}} - g(\boldsymbol{\theta})), \tag{6}$$

where $\mathbf{W}$ is a positive definite weighting matrix.

- Bias and consistency

1. The bias of an estimator is defined as

$$B(\widehat{\theta}) = E[\widehat{\theta}] - \theta_0, \tag{7}$$

where θ_0 is the true value of the parameter and $\widehat{\theta}$ is an estimator of θ_0.

2. An estimator $\widehat{\theta}_n$ is said to be consistent if

$$\lim_{n\to\infty} \Pr(\widehat{\theta}_n - \theta_0 < \varepsilon) = 1, \forall \varepsilon > 0. \tag{8}$$

Consistency requires two features of an estimator:

2.1. When the sample size increases, any bias must be shrinking.

2.2. The distribution of $\widehat{\theta}_n$ around θ_0 must be shrinking, i.e. close to θ_0.

The Law of Large Numbers

- Let $\{y_i\}$ be a sequence of i.i.d. random variables with $\mu = E[y_i]$, and define $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as the sample mean, then
$$\lim_{n\to\infty} \overline{y} = \mu,$$
if and only if $E[|y_i|] < \infty$.
- Usually we can also denote $\widehat{\mu} = \overline{y}$ as the sample mean. Define
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2,$$
as the sample variance.
- We can use the law of large numbers (LLN) to show that
$$\begin{aligned} E[\widehat{\mu}] &= \mu, \\ E[\widehat{\sigma}^2] &= \frac{n-1}{n}\sigma^2, \end{aligned}$$
that is, the mean estimator is unbiased, but the variance estimator is biased. However, as $n \to \infty$, $E[\widehat{\sigma}^2] \to \sigma^2$.

Asymptotic Normality

- Central limit theory (CLT).
- Let $\{y_i\}$ be a sequence of i.i.d. random scalars with $E[y_i] = \mu$ and $Var[y_i] = \sigma^2 < \infty$, then

$$\frac{\overline{y}_n - \mu}{\overline{\sigma}_n} = \sqrt{n}\frac{\overline{y}_n - \mu}{\sigma} \xrightarrow{d} N(0, 1), \tag{9}$$

  where $\overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ and $\overline{\sigma}^2 = \sqrt{\sigma^2/n}$.
- CLT is the basis of most inference in econometrics, although the formal justification is only asymptotic.

Efficiency

- Let $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ are two $\sqrt{n}-$consistent asymptotically normal estimators for θ_0. Denote the asymptotic variance as $avar(\cdot)$, if

$$avar(\widehat{\theta}_n) < avar(\widetilde{\theta}_n), \tag{10}$$

  for any choice of $\widetilde{\theta}_n$, then $\widehat{\theta}_n$ is said to be the efficient estimator of θ_0.
- Note θ could be a vector, thus $avar(\widehat{\theta}_n)$ is a covariance matrix. Denote $avar(\widehat{\theta}_n) = A$ and $avar(\widetilde{\theta}_n) = B$. $A < B$ means $B - A$ is positive definite, so every element in B is larger than the corresponding element in A.

Distribution Theory

- In order to derive the inference of an estimator, we need to determine its distribution.
- Distribution theory for classical method of moments.
  Define the following estimators from method of moments:

$$
\begin{aligned}
\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\widehat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2, \\
&\vdots \\
\widehat{\mu}_k &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^k.
\end{aligned}
$$

Distribution Theory

- Define μ_i as the corresponding population values. Let $g_{1i} = x_i - \mu$, and $g_{ji} = (x_i - \mu)^j - \mu_j$, $j = 2, 3, \cdots, k$. Define the vector $\mathbf{g}_i$ as

$$\mathbf{g}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ki} \end{bmatrix} \tag{11}$$

- The method of moments estimator can be seen as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i = \mathbf{0}.$$

- From the mean value theory, we know that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \quad (12)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)+\mathbf{G}_n(\overline{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),
\end{aligned}
$$

  where $\overline{\boldsymbol{\theta}}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.
- Note $\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\widehat{\boldsymbol{\theta}})=\mathbf{0}$ by construction, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)+\mathbf{G}_n(\overline{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=0. \quad (13)
$$

Distribution Theory

- From eq. (13), we get

$$
\begin{aligned}
(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\mathbf{G}_n(\overline{\boldsymbol{\theta}})^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}_0) \Rightarrow \\
\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\mathbf{G}_n(\overline{\boldsymbol{\theta}})^{-1} \sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0),
\end{aligned}
$$

  where $\mathbf{g}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}_0)$ is the average of moment conditions.
- Assume that $\sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0) \sim N(0, \Sigma)$, we have

$$
\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, \mathbf{G}^{-1}\Sigma(\mathbf{G}')^{-1}), \tag{14}
$$

  where $\mathbf{G}n(\overline{\boldsymbol{\theta}})$ is replaced with its limit

$$
\mathbf{G} = p \lim_{n\to\infty} \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = p \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.
$$

- This form of covariance is known as the "sandwich" covariance.

- Example 1: Derive the inference on the mean and variance.
- To estimate the mean and variance by the method of moments, we need two moment conditions:

$$\begin{aligned}\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2,\end{aligned}$$

thus $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}^2)'$.

- Define

$$\mathbf{g}_i = \left[\begin{array}{c} x_i-\mu \\ (x_i-\mu)^2-\sigma^2 \end{array}\right].$$

Distribution Theory

- The covariance matrix of $\mathbf{g}_i$ is given by

$$\begin{aligned}\Sigma &= E[\mathbf{g}_i\mathbf{g}_i'] \\ &= \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4-\sigma^4 \end{bmatrix}.\end{aligned}$$

- The Jacobian is

$$\begin{aligned}\mathbf{G} &= p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} -1 & 0 \\ -2(x_i-\mu) & -1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.\end{aligned}$$

Distribution Theory

- Apply the sandwich variance, eq. (14), we have

$$
\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \sqrt{n}\left(\begin{bmatrix} \widehat{\mu} \\ \widehat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \right) \\
&\sim N(\mathbf{0}, \mathbf{G}^{-1}\Sigma(\mathbf{G}')^{-1}) \\
&\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \right)
\end{aligned}
$$

Maximum Likelihood Estimation

- The use of M-estimators is pervasive in financial econometrics.
- Maximum likelihood.
  Use the distribution of the data to estimate unknown parameters by maximizing the likelihood.
  For example, suppose we randomly select 10 stocks and observe their prices in one day, 9 stocks have prices increased, and 1 decreases. Now we want to estimate the probability of price decrease for all stocks on that day. Suppose the probability for price decrease is p, thus for price increase is $1-p$. The probability for 9 increases and 1 decrease is

$$f(p) = (1-p)^9 p.$$

  We maximize this probability with respect to p, i.e. let $df(p)/dp = 0 \Rightarrow p = 0.1$.

Maximum Likelihood Estimation

- More generally, maximum likelihood estimation begins by specifying the joint distribution $f(\mathbf{y}|\boldsymbol{\theta})$ of observed data $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$ given the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)'$. Usually we denote $L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$ as the likelihood function, to emphasize the parameter vector $\boldsymbol{\theta}$ depends on observed data $\mathbf{y}$. The maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}$ is defined as the solution to

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}). \tag{15}$$

- Since $L(\cdot)$ is strictly positive, we can use the log of $L(\cdot)$ to estimate $\boldsymbol{\theta}$. Define $\ell(\cdot) = \ln L(\cdot)$. The maximum likelihood estimator can be found by solving the score function

$$\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{16}$$

- Eq. (16) sometimes cannot be solved if $\boldsymbol{\theta}$ is constrained.

Maximum Likelihood Estimation

- Maximizing $\ell(\cdot)$ allows us to estimate covariance matrix $Cov(\widehat{\boldsymbol{\theta}}_{ML})$ of the maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{ML}$.
- Define the information matrix $I(\boldsymbol{\theta})$ as follows

$$I(\boldsymbol{\theta}) = -E[\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}]. \tag{17}$$

- The inverse of $I(\boldsymbol{\theta})$ is the Cramer-Rao lower bound, i.e., if $\widetilde{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, then we have

$$Cov(\widetilde{\boldsymbol{\theta}}) - I(\boldsymbol{\theta})^{-1} \geqslant 0.$$

Maximum Likelihood Estimation

- In addition, we have

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \sim N(0, H^{-1}),$$

  where

$$H = \lim_{T\to\infty} \frac{1}{T} I(\boldsymbol{\theta}).$$

- This property also tells us how to estimate $Cov(\widehat{\boldsymbol{\theta}}_{ML})$

$$Cov(\widehat{\boldsymbol{\theta}}_{ML}) = -\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}{}_{|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{ML}}\right]^{-1}.$$

Maximum Likelihood Estimation

- Maximum likelihood estimation of a normal (Gaussian) distributed model.
- Assume y_i is i.i.d. normally distributed with mean μ and variance σ^2, i.e. $y \sim N(\mu, \sigma^2)$. Thus the p.d.f. of y_i is

$$f(y_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mu)^2}{2\sigma^2}), \tag{18}$$

  where $\boldsymbol{\theta} = (\mu, \sigma^2)'$.
- The joint likelihood is the product of the n individual likelihoods (since y_i is i.i.d.):

$$f(\mathbf{y}; \boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mu)^2}{2\sigma^2}). \tag{19}$$

- Taking log, we obtain

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}. \tag{20}$$

- Taking the derivative with respect to the parameter $\boldsymbol{\theta} = (\mu, \sigma^2)'$,

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma^2}, \tag{21}$$

$$\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2} = -\frac{2n}{\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^4}. \tag{22}$$

Setting eqs. (21) and (22) to zero, we obtain

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

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

which is the sample mean and sample variance, as we expected.

Maximum Likelihood Estimation

- In general, for a dependent series $\{y_t\}$ and parameter vector $\boldsymbol{\theta}$, we have

$$
\begin{aligned}
L(\boldsymbol{\theta}|\{y_T\}) &= L(\boldsymbol{\theta}|y_1, y_2, ..., y_T) \\
&= f(y_1, y_2, ..., y_T|\boldsymbol{\theta}) \\
&= f(y_T|y_{T-1}, ..., y_1, \boldsymbol{\theta})f(y_{T-1}|y_{T-2}, ..., y_1|\boldsymbol{\theta})...f(y_1|\boldsymbol{\theta}) \\
&= \Pi_{i=1}^{T} f(y_i|\{y_{i-1}\}, \boldsymbol{\theta}). \qquad (23)
\end{aligned}
$$

Since $P(AB|Z) = P(A|B, Z)P(B|Z)$ and so on.

Maximum Likelihood Estimation

- However, the derivation of eq. (23) is not always straightforward.
- Consider the following example with unobserved components:

$$\begin{aligned} y_t &= x_t + e_t, \quad (24) \\ x_t &= \delta + \phi x_{t-1} + v_t, \quad (25) \end{aligned}$$

  where $e_t \sim N(0, \sigma_e^2)$ and $v_t \sim N(0, \sigma_v^2)$.
- The parameter vector $\boldsymbol{\theta} = (\delta, \phi, \sigma_e, \sigma_v)$.
- The conditional density $f(y_t | y_{t-1}, \boldsymbol{\theta})$ is not directly obtained.

Maximum Likelihood Estimation

- However, due to the normal assumption, we can express $\{y_T\}$ in a multivariate normal distribution

$$y_T \sim N(\mu, \Omega),$$

with the likelihood function

$$L(\boldsymbol{\theta}|y_T) = (2\pi)^{-\frac{T}{2}}|\Omega|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y_T-\mu)'\Omega^{-1}(y_T-\mu)), \quad (26)$$

with all elements in μ and Ω are complicated functions of $\boldsymbol{\theta}$.

- Even when μ and Ω can be specified explicitly, maximizing the log likelihood function could be troublesome because the inversion of the $T \times T$ matrix Ω.
- Harvey (1980) provides a solution based on the prediction error decomposition.

Maximum Likelihood Estimation

- Note that Ω is symmetric, thus we can decompose it as follows

$$\Omega = AfA', \tag{27}$$

where

$$f = \begin{bmatrix} f_1 & 0 & 0 & \cdots & 0 \\ 0 & f_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & f_T \end{bmatrix}, \text{ and } A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{T1} & a_{T2} & a_{T3} & \cdots & 1 \end{bmatrix}.$$

Maximum Likelihood Estimation

- Substitute (27) into (26), we get

$$
\begin{aligned}
L(\boldsymbol{\theta}|y_T) &= (2\pi)^{-\frac{T}{2}}|AfA'|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y_T-\mu)'(AfA')^{-1}(y_T-\mu)) \\
&= (2\pi)^{-\frac{T}{2}}|f|^{-\frac{1}{2}}\exp(-\frac{1}{2}\eta' f^{-1}\eta) \\
&= (2\pi)^{-\frac{T}{2}}\Pi_{t=1}^{T} f_t^{-\frac{1}{2}}\exp(-\frac{1}{2}\sum_{t=1}^{T}\eta_t' f_t^{-1}\eta_t) \\
&= \Pi_{t=1}^{T}[\frac{1}{\sqrt{2\pi f_t}}\exp(-\frac{1}{2}\frac{\eta_t^2}{f_t})],
\end{aligned}
$$

where $\eta = A^{-1}(y_T-\mu)$ and η_t is the $t-$th element of η.

Maximum Likelihood Estimation

- Since A is a lower trangular matrix, it is easily shown that

$$\eta_t = y_t - y_{t|t-1},$$

where $y_{t|t-1}$ is the prediction of y_t given the information up to y_{t-1}, $I_t = \{y_1, y_2, ..., y_t\}$, since

$$y_{t|t-1} = \sum_{i=1}^{t-1} a^*_{t,i} y_i,$$

where $a^*_{t,i}$ is the $(t, i)-$th element of A^{-1}.

- In summary, if we know the prediction error η_t and their variances, the log likelihood function can be easily calculated.

$$y_t | I_{t-1} \sim N(y_{t|t-1}, f_t),$$

where f_t is the variance of the prediction error η_t.

Parameter Constraints

- We know that the MLE estimator can be obtained by setting

$$\frac{\partial \ell(\widehat{\boldsymbol{\theta}}_{ML}|\mathbf{y})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

- In most cases, the closed-form solution for $\widehat{\boldsymbol{\theta}}_{ML}$ is not available.
- We rely on numerical optimization procedure to estimate $\widehat{\boldsymbol{\theta}}_{ML}$.
- When numerical optimization is employed, the computer searches the parameter space ranges between $(-\infty, +\infty)$.
- But some parameters may lie in an interval $[m, n]$.
- We can use the following transformation

$$\theta = g(\psi),$$

where $g(\cdot)$ is a continuous function.

Parameter Constraints

- For example, if $\theta_j > 0$, then we can let

$$\theta_j = \psi_j^2, \text{ or } \theta_j = \exp(\psi_j).$$

- If θ_j represents a probability term, then $0 < \theta_j < 1$, we can let

$$\theta_j = \frac{1}{1 + \exp(\psi_j^{-1})}.$$

- If θ_j represents an autoregressive parameter in an $AR(1)$ model, i.e. $-1 < \theta_j < 1$, we can apply

$$\theta_j = \frac{\psi_j}{1 + |\psi_j|}.$$

Parameter Constraints

- If $\theta = (\phi_1, \phi_2)$ are the autoregressive coefficents in an $AR(2)$ model, then we know they should satisfy that roots of

$$(1 - \phi_1 L - \phi_2 L^2) = 0$$

lie outside the unit circle.

- In this case, we may define

$$z_1 = \frac{\psi_1}{1 + |\psi_1|}, \; z_2 = \frac{\psi_2}{1 + |\psi_2|},$$

and let

$$\phi_1 = z_1 + z_2, \; \phi_2 = -1 * z_1 * z_2.$$

Parameter Constraints

- Finally, consider the following GARCH model

$$h_t = a_0 + a_1 e_{t-1}^2 + a_2 h_{t-1},$$

  where h_t is the conditional variance. We generally want $a_1 > 0$, $a_2 > 0$, and $a_1 + a_2 < 1$.
- The following transformation achieves this goal:

$$\begin{aligned} a_1 &= \frac{\exp(\psi_1)}{1 + \exp(\psi_1) + \exp(\psi_2)}, \\ a_2 &= \frac{\exp(\psi_2)}{1 + \exp(\psi_2) + \exp(\psi_2)}. \end{aligned}$$

Covariance Matrix

- Consider the transformation for constrained parameters

$$\theta = g(\psi).$$

- We need to calcuate $Cov(\theta)$, but we may only know $Cov(\psi)$ from the estimation.
- How to solve the problem?
- From the Delta Method, we can easily show that

$$Cov(\theta) = (\frac{\partial g(\psi)}{\partial \psi}) Cov(\psi) (\frac{\partial g(\psi)}{\partial \psi})'.$$

The Delta Method

- Suppose that $\boldsymbol{\theta}$ is a vector of random variables, which follows a normal distribution

$$\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, V_{\boldsymbol{\theta}}). \tag{28}$$

- Now assume $f(\boldsymbol{\theta})$ is a nonlinear function of $\boldsymbol{\theta}$, then we have

$$\sqrt{T}(f(\widehat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta})) \sim N(0, V_f), \tag{29}$$

  where $V_f = \frac{\partial f}{\partial \boldsymbol{\theta}'} V_{\boldsymbol{\theta}} \frac{\partial f}{\partial \boldsymbol{\theta}}$.
- This can be derived from the first order approximation of Taylor's expansion.

Taylor Expansion

- A real function $f(x)$, which is infinitely differentiable in a neighborhood of x_0, can be written as

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2!}f''(x_0)(x - x_0)^2 \\ &+ \frac{1}{3!}f'''(x_0)(x - x_0)^3 + \cdots \end{aligned} \quad (30)$$

- Example
  $f(x) = e^x$, we Taylor expand $f(x)$ at $x_0 = 0$, thus

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots$$

- Example: Derive the inference for maximum likelihood estimator. We know that maximum likelihood estimator is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{y}), \tag{31}$$

  where $\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}; \mathbf{y}_i)$.
- It is useful to work with the average log-likelihood directly. Define $\overline{\ell}_n(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}; \mathbf{y}_i)$, so that $\overline{\ell}_n(\boldsymbol{\theta}; \mathbf{y})$ will converge to $E[\ell(\boldsymbol{\theta}; \mathbf{y})]$ if $n \rightarrow \infty$.
- The scores of the average log-likelihood are

$$\frac{\partial \overline{\ell}_n(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta}}.$$

- Since y_i are i.i.d., the scores will be i.i.d., thus the average scores will follow the law of large numbers for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$, that is, $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta}} \longrightarrow E\left[\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta}}\right]$.

- Applying CLT, we know that

$$\sqrt{n}\nabla\bar{\ell}_n(\boldsymbol{\theta}_0) \sim N(0, J),$$

where we define the operator $\nabla\bar{\ell}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta};\mathbf{y}_i)}{\partial \boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ as the score.

- Take a mean value expansion around $\boldsymbol{\theta}_0$

$$\sqrt{n}\nabla\bar{\ell}_n(\widehat{\boldsymbol{\theta}}) = \sqrt{n}\nabla\bar{\ell}_n(\boldsymbol{\theta}_0) + \sqrt{n}\nabla^2\bar{\ell}_n(\bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),$$

we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = [-\nabla^2\bar{\ell}_n(\bar{\boldsymbol{\theta}})]^{-1}\sqrt{n}\nabla\bar{\ell}_n(\boldsymbol{\theta}_0), \tag{32}$$

where $\nabla^2\bar{\ell}_n(\bar{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^{n} \frac{\partial \ell_i^2(\boldsymbol{\theta};\mathbf{y}_i)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}$.

Distribution Theory

- Eq. (32) implies that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \sim N(0, I^{-1}JI^{-1}), \tag{33}$$

where $I = -E[\frac{\partial \ell^2(\boldsymbol{\theta};\mathbf{y}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}]$, $J = E[\frac{\partial \ell(\boldsymbol{\theta};\mathbf{y}_i)}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y}_i)}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}]$.

- Under the assumption of i.i.d., we have $I \longrightarrow J$ in probability, thus we may simplify eq. (33) as

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \sim N(0, I^{-1}). \tag{34}$$

- Quasi Maximum Likelihood (QML).

- Example: Derive the inference in the Normal MLE.
- Recall that the MLE estimators of the mean and variance are

$$\begin{aligned}\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2,\end{aligned}$$

and the log-likelihood is (eq. (20))

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \mu)^2}{2\sigma^2}.$$

Distribution Theory

- Taking derivative with respect to the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)'$ (eqs. (21) and (22)),

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu} &= \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma^2}, \\
\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2} &= -\frac{2n}{\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^4}.
\end{aligned}
$$

- The second derivatives are

$$
\begin{aligned}
\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu \partial \mu} &= -\sum_{i=1}^{n} \frac{1}{\sigma^2}, \\
\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu \partial \sigma^2} &= -\sum_{i=1}^{n} \frac{(y_i - \mu)}{\sigma^4}, \\
\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2 \partial \sigma^2} &= \frac{n}{2\sigma^4} - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^6}.
\end{aligned}
$$

Distribution Theory

- Taking expectations on these second derivatives:

$$
\begin{aligned}
E[\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu \partial \sigma^2}] &= 0, \\
E[\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2 \partial \sigma^2}] &= -\frac{n}{2\sigma^4}.
\end{aligned}
$$

- Putting these together, the Hessian can be formed:

$$
E[\frac{\partial \ell^2(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}] = \begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}.
$$

- So the asymptotic covariance is

$$-E\left[\frac{\partial \ell^2(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1} = -\left[\begin{array}{cc} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{array}\right]^{-1} = \left[\begin{array}{cc} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{array}\right].$$

- Thus the asymptotic distribution is

$$\sqrt{n}\left(\left[\begin{array}{c} \widehat{\mu} \\ \widehat{\sigma}^2 \end{array}\right] - \left[\begin{array}{c} \mu \\ \sigma^2 \end{array}\right]\right) \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{array}\right]\right).$$

- Exercise: Let $\{x_t\}$, $t = 1, 2, \cdots, T$, be a random sample from the following probability density function (pdf):

$$f(x|\theta) = \theta x^{\theta-1},\ 0 \leqslant x \leqslant 1,\ 0 < \theta < \infty.$$

1. Derive the maximum likelihood estimator (MLE) for θ.
2. Derive the variance of the MLE for θ.
3. Show that the MLE for θ is a consistent estimator (you should use the definition for a consistent estimator).

Hypothesis Testing

- Econometrics models are estimated in order to test hypothesis.
- Null hypothesis (H_0): statement about the population values of some parameters to be tested.
  For example, the model is specified as follows

$$y_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \varepsilon_i,$$

  the null can be simple: $H_0 : \theta_0 = 0$, or composite, $H_0 : \theta_1 = \theta_2 = 0$.
- The null cannot be accepted: the data can either lead to rejection of the null or a failure to reject the null.

Hypothesis Testing

- Alternative hypothesis (H_1): specify the population values of parameters for which the null should be rejected. In most cases, the alternative is the natural complement to the null.
  For example, when testing whether a random variable has mean of 0, the null is $H_0 : \mu = 0$, the alternative is $H_1 : \mu \neq 0$.
- In some situations, we may also consider a one-sided test, i.e.
  $H_0 : \mu = 0$ v.s. $H_1 : \mu > 0$.
- Hypothesis test: a rule that specifies which values to reject H_0 in favor of H_1.
- Critical value:
  1. denote a $\alpha-$sized critical value as C_α, is the value where a test statistic, T, indicates rejection of a null when it is true, that is, if $|T| > |C_\alpha|$, then the null is rejected.

Hypothesis Testing

- Type I error: the null is rejected when it is true.
- Size of test: the probability of rejecting the null when it is true, i.e. the probability of type I error. The typical size is 1%, 5%, and 10%.
- Type II error: the null is not rejected when the alternative is true.
- Power of test: the probability of rejecting the null when the alternative is true.
- The power is equivalently defined as $1-$ Pr(Type II error). There exists conflict between controlling for Type I error and Type II error. A perfect test should have correct size and unit power against any alternatives.

Hypothesis Testing

- Confidence interval: the range of values, $\theta_0 \in [\underline{C}_\alpha, \overline{C}_a]$ where the null $H_0 : \theta = \theta_0$ cannot be rejected for a size of α.
- Example.
  Suppose n i.i.d. normal random variables have unknown mean μ but known variance σ^2 and so the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, is then distributed $N(\mu, \sigma^2/n)$.
  We want to test the null $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$.
  For a given size α, $\Pr(\widehat{\mu} \in [\underline{C}_\alpha, \overline{C}_a] | \mu = \mu_0) = 1 - \alpha$, a natural choice for choosing the lower and upper limit is that $\underline{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2})$ and $\overline{C}_a = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \frac{\alpha}{2})$, where $\Phi(\cdot)$ is the CDF for a standard normal.

- Now suppose the population mean is μ_1 instead of μ_0. The power of the test will depend on μ_1, the sample size, the test size, and the specified mean by the null, μ_0.
- When testing using an $\alpha-$sized test, the rejection will occur when $\widehat{\mu} < \underline{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2})$ or $\widehat{\mu} > \overline{C}_a = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$.
- Under the alternative $\widehat{\mu} \sim N(\mu_1, \sigma^2/n)$, these probabilities will be

$$\Phi(\frac{\underline{C}_\alpha - \mu_1}{\sigma/\sqrt{n}}) \text{ and } 1-\Phi(\frac{\overline{C}_a - \mu_1}{\sigma/\sqrt{n}}),$$

  the power function is then

$$Power(\mu_0, \mu_1, \sigma, \alpha, n) = \Phi(\frac{\underline{C}_\alpha - \mu_1}{\sigma/\sqrt{n}}) + 1 - \Phi(\frac{\overline{C}_a - \mu_1}{\sigma/\sqrt{n}}).$$

Hypothesis Testing

- Consider the following null hypothesis for testing parameters in $\boldsymbol{\theta}$:

$$H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}, \tag{35}$$

  where $\mathbf{R}(\cdot)$ is a function from $\mathbb{R}^k$ to $\mathbb{R}^m$, k is the number of parameters and m represents the number of hypotheses in a composite null.
- Eq. (35) is very flexible. A subset of eq. (35) is the linear class, which can be specified as

$$H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0}. \tag{36}$$

Wald Test

- Recall that a χ^2_v random variable is defined to be the sum of v independent standard normals squared

$$\sum_{i=1}^{v} z_i^2 \sim \chi^2_v \text{ if } z_i \sim N(0,1).$$

Also note that if z is a $m-$dimension normal vector with mean μ and covariance Σ, then

$$\Sigma^{-\frac{1}{2}}(z-\mu) \sim N(0, I).$$

- Now we can see that if

$$\sqrt{n}(\widehat{\theta}-\theta_0) \sim N(0, V_\theta),$$

then by applying the Delta method, we have

$$\sqrt{n}(R\widehat{\theta}-r) \sim N(0, RV_\theta R').$$

Wald Test

- We form the Wald-statistic as

$$w = n(R\widehat{\theta} - r)'(RV_\theta R')^{-1}(R\widehat{\theta} - r), \tag{37}$$

by construction, eq. (37) is the sum of squares of m random variables, each follows a standard normal. Thus $w \sim \chi^2_m$.
- T-test is a special case of Wald test. Suppose the null is $H_0 : R\theta - r = 0$ against $H_1 : R\theta - r \neq 0$.

$$\sqrt{n}(R\widehat{\theta} - r) \sim N(0, RV_\theta R'),$$

the t-test is defined as

$$t = \frac{\sqrt{n}(R\widehat{\theta} - r)}{\sqrt{RV_\theta R'}} \sim N(0, 1).$$

Likelihood Ratio Test

- Likelihood ratio test examines how "likely" the data are under the null and alternative. If the null is valid, then the data should be (approximately) equally likely under each.
- The LR test statistic is defined as

$$LR = -2(\ell(\widetilde{\theta}; y) - \ell(\widehat{\theta}; y)), \tag{38}$$

where $\widetilde{\theta}$ is the maximum likelihood estimator from the constrained model with $R\theta - r = 0$, and $\widehat{\theta}$ is the maximum likelihood estimator from the unconstrained model.
- We can prove that under the null, $H_0 : R\theta - r = 0$,

$$LR \sim \chi^2_m,$$

where m is the number of constraints.

Lagrange Multiplier (LM), Score and Rao Tests

- LM, Score and Rao Tests are all the same statistic. Under the unconstrained estimator $\widehat{\theta}$, the score must be zero

$$\frac{\partial \ell(\theta; y)}{\partial \theta}\Big|_{\theta=\widehat{\theta}} = 0.$$

- The Score test examines whether the score are "close" to zero when the parameters $\widetilde{\theta}$, are evaluated under the null. Define

$$s_i(\widetilde{\theta}) = \frac{\partial \ell_i(\theta; y)}{\partial \theta}\Big|_{\theta=\widetilde{\theta}},$$

as the i^{th} score evaluated with the restricted parameters $\widetilde{\theta}$. If the null is true, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} s_i(\widetilde{\theta})\right) \sim N(0, \Sigma).$$

Lagrange Multiplier (LM), Score and Rao Tests

- The LM or score test is constructed as

$$LM = n\bar{s}(\widetilde{\theta})'\Sigma^{-1}\bar{s}(\widetilde{\theta}),$$

  where $\bar{s}(\widetilde{\theta}) = n^{-1}\sum_{i=1}^{n} s_i(\widetilde{\theta})$.

- We can replace Σ with a consistent estimator and to compute the feasible score statistic

$$LM = n\bar{s}(\widetilde{\theta})'\widehat{\Sigma}^{-1}\bar{s}(\widetilde{\theta}),$$

  where

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n} s_i(\widetilde{\theta})s_i(\widetilde{\theta})'.$$

- Under the null, LM follows a χ^2_m distribution.
- Comparing and choosing the tests.