Lecture Notes on

ADVANCED ECONOMETRICS

Yongmiao Hong

Department of Economics and
Department of Statistical Sciences
Cornell University

Emails: yh20@cornell.edu & yhong.cornell@gmail.com

SPRING 2016

Tables of Contents

Preface

Modern economies are full of uncertainties and risk. Economics studies resource allocations in an uncertain market environment. As a generally applicable quantitative analytic tool for uncertain events, probability and statistics have been playing an important role in economic research. Econometrics is statistical analysis of economic and financial data. It has become an integral part of training in modern economics and business. This book develops a coherent set of econometric theory and methods for economic models. It is written for an advanced econometrics course for doctoral students in economics, business, management, statistics, applied mathematics, and related fields. It can also be used as a reference book on econometric theory by scholars who may be interested in both theoretical and applied econometrics.

The book is organized in a coherent manner. Chapter 1 is a general introduction to econometrics. It first describes the two most important features of modern economics, namely mathematical modeling and empirical validation, and then discusses the role of econometrics as a methodology in empirical studies. A few motivating economic examples are given to illustrate how econometrics can be used in empirical studies. Finally, it points out the limitations of econometrics and economics due to the fact that an economy is not a repeatedly controlled experiment. Assumptions and careful interpretations are needed when conducting empirical studies in economics and finance.

Chapter 2 introduces a general regression analysis. Regression analysis is modeling, estimation, inference, and specification analysis of the conditional mean of economic variables of interest given a set of explanatory variables. It is most widely applied in economics. Among other things, this chapter interprets the mean squared error and its optimizer, which lays down the probability-theoretic foundation for least squares estimation. In particular, it provides an interpretation for the least squares estimator and its relationship with the true parameter value of a correctly specified regression model.

Chapter 3 introduces the classical linear regression analysis. A set of classical assumptions are given and discussed, and conventional statistical procedures for estimation, inference, and hypothesis testing are introduced. The roles of conditional homoskedasticity, serial uncorrelatedness, and normality of the disturbance of a linear regression model are analyzed in a finite sample econometric theory. We also discuss the generalized least squares estimation as an efficient estimation method of a linear regression model when the variance-covariance matrix is known up to a constant. In particular, the generalized least squares estimation is embedded as an ordinary least squares estimation of a suitably transformed regression model via conditional variance scaling and autocorrelation filtering.

The subsequent chapters 4–7 are the generalizations of classical linear regression analysis when various classical assumptions fail. Chapter 4 first relaxes the normality and conditional homoskedasticity assumptions, two key conditions assumed in the classical linear regression modeling. A large sample theoretic approach is taken. For simplicity, it is assumed that the observed data are generated from an independent and identically distributed random sample. It is shown that while the finite distributional theory is no longer valid, the classical statistical procedures are still approximately applicable when the sample size is large, provided conditional homoskedasticity holds. In contrast, if the data display conditional heteroskedasticity, classical statistical procedures are not applicable even for large samples, and heteroskedasticity-robust procedures will be called for. Tests for existence of conditional heteroskedasticity in a linear regression framework are introduced.

Chapter 5 extends the linear regression theory to time series data. First, it introduces a variety of basic concepts in time series analysis. Then it shows that the large sample theory for i.i.d. random samples carries over to stationary ergodic time series data if the regression error follows a martingale difference sequence. We introduce tests for serial correlation, and tests for conditional heteroskedasticity and autoregressive conditional heteroskedasticity in a time series regression framework. We also discuss the impact of autoregressive conditional heteroskedasticity on inferences of static time series regressions and dynamic time series regressions.

Chapter 6 extends the large sample theory to a very general case where there exist conditional heteroskedasticity and autocorrelation. In this case, the classical regression theory cannot be used, and a long-run variance-covariance matrix estimator is called for to validate statistical inferences in a time series regression framework.

Chapter 7 is the instrumental variable estimation for linear regression models, where the regression error is correlated with the regressors. This can arise due to measurement errors, simultaneous equation biases, and other various reasons. Two- stage least squares estimation method and related statistical inference procedures are fully exploited. We describe tests for endogeneity.

Chapter 8 introduces the generalized method of moments, which is a popular estimation method for possibly nonlinear econometric models characterized as a set of moment conditions. Indeed, most economic theories, such as rational expectations, can be formulated by a moment condition. The generalized method of moments is particularly suitable to estimate model parameters contained in the moment conditions for which the conditional distribution is usually not available.

Chapter 9 introduces the maximum likelihood estimation and the quasi-maximum likelihood estimation methods for conditional probability models and other nonlinear econometric mod-

els. We exploit the important implications of correctly specification of a conditional distribution model, especially the analogy between the martingale difference sequence property of the score function and serial uncorrelatedness, and the analogy between the conditional information equality and conditional homoskedasticity. These links can provide a great help in understanding the large sample properties of the maximum likelihood estimator and the quasi-maximum likelihood estimator.

Chapter 10 concludes the book by summarizing the main econometric theory and methods covered in this book, and pointing out directions for further build-up in econometrics.

This book has several important features. It covers, in a progressive manner, various econometrics models and related methods from conditional means to possibly nonlinear conditional moments to the entire conditional distributions, and this is achieved in a unified and coherent framework. We also provide a brief review of asymptotic analytic tools and show how they are used to develop the econometric theory in each chapter. By going through this book progressively, readers will learn how to do asymptotic analysis for econometric models. Such skills are useful not only for those students who intend to work on theoretical econometrics, but also for those who intend to work on applied subjects in economics because with such analytic skills, readers will be able to understand more specialized or more advanced econometrics textbooks.

This book is based on my lecture notes taught at Cornell University, Renmin University of China, Shandong University, Shanghai Jiao Tong University, Tsinghua University, and Xiamen University, where the graduate students provide detailed comments on my lecture notes.

CHAPTER 1 INTRODUCTION TO ECONOMETRICS

Abstract: Econometrics has become an integral part of training in modern economics and business. Together with microeconomics and macroeconomics, econometrics has been taught as one of the three core courses in most undergraduate and graduate economic programs in North America. This chapter discusses the philosophy and methodology of econometrics in economic research, the roles and limitations of econometrics, and the differences between econometrics and mathematical economics as well as mathematical statistics. A variety of illustrative econometric examples are given, which cover various fields of economics and finance.

Key Words: Data generating process, Econometrics, Probability law, Quantitative analysis, Statistics.

1.1 Introduction

Econometrics has become an integrated part of teaching and research in modern economics and business. The importance of econometrics has been increasingly recognized over the past several decades. In this chapter, we will discuss the philosophy and methodology of econometrics in economic research. First, we will discuss the quatitative feature of modern economics, and the differences between econometrics and mathematical economics as well as mathematical statistics. Then we will focus on the important roles of econometrics as a fundamental methodology in economic research via a variety of illustrative economic examples including the consumption function, marginal propensity to consume and multipliers, rational expectations models and dynamic asset pricing, the constant return to scale and regulations, evaluation of effects of economic reforms in a transitional economy, the efficient market hypothesis, modeling uncertainty and volatility, and duration analysis in labor economics and finance. These examples range from econometric analysis of the conditional mean to the conditional variance to the conditional distribution of economic variables of interest. we will also discuss the limitations of econometrics, due to the nonexperimental nature of economic data and the time-varying nature of econometric structures.

1.2 Quantitative Features of Modern Economics

Modern market economies are full of uncertainties and risk. When economic agents make a decision, the outcome is usually unknown in advance and economic agents will take this uncertainty into account in their decision-making. Modern economics is a study on scarce resource allocations in an uncertain market environment. Generally speaking, modern economics can be

roughly classified into four categories: macroeconomics, microeconomics, financial economics, and econometrics. Of them, macroeconomics, microeconomics and econometrics now constitute the core courses for most economic doctoral programs in North America, while financial economics is now mainly being taught in business and management schools.

Most doctoral programs in economics in the U.S. emphasize quantitative analysis. Quantitative analysis consists of mathematical modeling and empirical studies. To understand the roles of quantitative analysis, it may be useful to first describe the general process of modern economic research. Like most natural science, the general methodology of modern economic research can be roughly summarized as follows:

- Step 1: Data collection and summary of empirical stylized facts. The so-called stylized facts are often summarized from observed economic data. For example, in microeconomics, a well-known stylized fact is the Engel's curve, which characterizes that the share of a consumer's expenditure on a commodity out of her or his total income will vary as his/her income changes; in macroeconomics, a well-known stylized fact is the Phillips Curve, which characterizes a negative correlation between the inflation rate and the unemployment rate in an aggregate economy; and in finance, a well-known stylized fact about financial markets is volatility clustering, that is, a high volatility today tends to be followed by another high volatility tomorrow, a low volatility today tends to be followed by another low volatility tomorrow, and both alternate over time. The empirical stylized facts often serve as a starting point for economic research. For example, the development of unit root and cointegration econometrics was mainly motivated by the empirical study of Nelson and Plosser (1982) who found that most macroeconomic time series are unit root processes.

- Step 2: Development of economic theories/models. With the empirical stylized facts in mind, economists then develop an economic theory or model in order to explain them. This usually calls for specifying a mathematical model of economic theory. In fact, the objective of economic modeling is not merely to explain the stylized facts, but to understand the mechanism governing the economy and to forecast the future evolution of the economy.

- Step 3: Empirical verification of economic models. Economic theory only suggests a qualitative economic relationship. It does not offer any concrete functional form. In the process of transforming a mathematical model into a testable empirical econometric model, one often has to assume some functional form, up to some unknown model parameters. One needs to estimate unknown model parameters based on the observed data, and check whether the econometric model is adequate. An adequate model should be at least consistent with the empirical stylized facts.

- Step 4: Applications. After an econometric model passes the empirical evaluation, it can

then be used to test economic theory or hypotheses, to forecast future evolution of the economy, and to make policy recommendations.

For an excellent example highlighting these four steps, see Gujarati (2006, Section 1.3) on labor force participation. We note that not every economist or every research paper has to complete these four steps. In fact, it is not uncommon that each economist may only work on research belonging to a certain stage in his/her entire academic lifetime.

From the general methodology of economic research, we see that modern economics has two important features: one is mathematical modeling for economic theory, and the other is empirical analysis for economic phenomena. These two features arise from the effort of several generations of economists to make economics a "science". To be a science, any theory must fulfill two criteria: one is logical consistency and coherency in theory itself, and the other is consistency between theory and stylized facts. Mathematics and econometrics serve to help fulfill these two criteria respectively. This has been the main objective of the econometric society. The setup of the Nobel Memorial Prize in economics in 1969 may be viewed as the recognition of economics as a science in the academic profession.

1.3 Mathematical Modeling

We first discuss the role of mathematical modeling in economics. Why do we need mathematics and mathematical models in economics? It should be pointed out that there are many ways or tools (e.g., graphical methods, verbal discussions, mathematical models) to describe economic theory. Mathematics is just one of them. To ensure logical consistency of the theory, it is not necessary to use mathematics. Chinese medicine is an excellent example of science without using mathematical modeling. However, mathematics is well-known as the most rigorous logical language. Any theory, when it can be represented by the mathematical language, will ensure its logical consistency and coherency, thus indicating that it has achieved a rather sophisticated level. Indeed, as Karl Marx pointed out, the use of mathematics is an indication of the mature development of a science.

It has been a long history to use mathematics in economics. In his *Mathematical Principles of the Wealth Theory*, Cournot (1838) was among the earliest to use mathematics in economic analysis. Although the *marginal revolution*, which provides a cornerstone for modern economics, was not proposed using mathematics, it was quickly found in the economic profession that the marginal concepts, such as marginal utility, marginal productivity, and marginal cost, correspond to the derivative concepts in calculus. Walras (1874), a mathematical economist, heavily used mathematics to develop his general equilibrium theory. The game theory, which was proposed by Von Neumann and Morgenstern (1944) and now becomes a core in modern microeconomics, originated from a branch in mathematics.

Why does economics need mathematics? Briefly speaking, mathematics plays a number of important roles in economics. First, the mathematical language can summarize the essence of a theory in a very concise manner. For example, macroeconomics studies relationships between aggregate economic variables (e.g., GDP, consumption, unemployment, inflation, interest rate, exchange rate, etc.) A very important macroeconomic theory was proposed by Keynes (1936). The classical Keynesian theory can be summarized by two simple mathematical equations:

$$\begin{aligned} \text{National Income identity:} \quad & Y = C + I + G, \\ \text{Consumption function:} \quad & C = \alpha + \beta Y, \end{aligned}$$

where Y is income, C is consumption, I is private investment, G is government spending, α is the "survival level" consumption, and β is the marginal propensity to consume. Substituting the consumption function into the income identity, arranging terms, and taking a partial derivative, we can obtain the multiplier effect of (e.g.) government spending

$$\frac{\partial Y}{\partial G} = \frac{1}{1-\beta}.$$

Thus, the Keynesian theory can be effectively summarized by two mathematical equations.

Second, complicated logical analysis in economics can be greatly simplified by using mathematics. In introductory economics, economic analysis can be done by verbal descriptions or graphical methods. These methods are very intuitive and easy to grasp. One example is the partial equilibrium analysis where a market equilibrium can be characterized by the intersection of the demand curve and the supply curve. However, in many cases, economic analysis cannot be done easily by verbal languages or graphical methods. One example is the general equilibrium theory first proposed by Walras (1874). This theory addresses a fundamental problem in economics, namely whether the market force can achieve an equilibrium for a competitive market economy where there exist many markets and when there exist mutual interactions between different markets. Suppose there are n goods, with demand $D_i(P)$, supply $S_i(P)$ for good i, where $P = (P_1, P_2, ..., P_n)'$ is a price vector for n goods. Then the general equilibrium analysis addresses whether there exists an equilibrium price vector P^* such that all markets are clear simultaneously:

$$D_i(P^*) = S_i(P^*) \text{ for all } i \in \{1, ..., n\}.$$

Conceptually simple, it is rather challenging to give a definite answer because both the demand and supply functions could be highly nonlinear. Indeed, Walras was unable to establish this theory formally. It was satisfactorily solved by Arrow and Debreu many years later, when they used the fixed point theorem in mathematics to prove the existence of an equilibrium price vector. The power and magic of mathematics was clearly demonstrated in the development of the general

equilibrium theory.

Third, mathematical modeling is a necessary path to empirical verification of an economic theory. Most economic and financial phenomena are in form of data (indeed we are in a digital era!). We need "digitalize" economic theory so as to link the economic theory to data. In particular, one needs to formulate economic theory into a testable mathematical model whose functional form or important structural model parameters will be estimated from observed data.

1.4 Empirical Validation

We now turn to discuss the second feature of modern economics: empirical analysis of an economic theory. Why is empirical analysis of an economic theory important? The use of mathematics, although it can ensure logical consistency of a theory itself, cannot ensure that economics is a science. An economic theory would be useless from a practical point of view if the underlying assumptions are incorrect or unrealistic. This is the case even if the mathematical treatment is free of errors and elegant. As pointed out earlier, to be a science, an economic theory must be consistent with reality. That is, it must be able to explain historical stylized facts and predict future economic phenomena.

How to check a theory or model empirically? Or how to validate an economic theory? In practice, it is rather difficult or even impossible to check whether the underlying assumptions of an economic theory or model are correct. Nevertheless, one can confront the implications of an economic theory with the observed data to check if they are consistent. In the early stage of economics, empirical validation was often conducted by case studies or indirect verifications. For example, in his well-known *Wealth of Nations*, Adam Smith (1776) explained the advantage of specialization using a case study example. Such a method is still useful nowadays, but is no longer sufficient for modern economic analysis, because economic phenomena are much more complicated while data may be limited. For rigorous empirical analysis, we need to use econometrics. Econometrics is the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory. It was founded as a scientific discipline around 1930 as marked by the founding of the econometric society and the creation of the most influential economic journal—*Econometrica* in 1933.

Econometrics has witnessed a rather rapid development in the past several decades, for a number of reasons. First, there is a need for empirical verification of economic theory, and for forecasting using economic models. Second, there are more and more high-quality economic data available. Third, advance in computing technology has made the cost of computation cheaper and cheaper over time. The speed of computing grows faster than the speed of data accumulation.

Although not explicitly stated in most of the econometric literature, modern econometrics is essentially built upon the following fundamental axioms:

- Any economy can be viewed as a stochastic process governed by some probability law.
- Economic phenomenon, as often summarized in form of data, can be reviewed as a realization of this stochastic data generating process.

There is no way to verify these axioms. They are the philosophic views of econometricians toward an economy. Not every economist or even econometrician agrees with this view. For example, some economists view an economy as a deterministic chaotic process which can generate seemingly random numbers. However, most economists and econometricians (e.g., Granger and Teräsvirta 1993, Lucas 1977) view that there are a lot of uncertainty in an economy, and they are best described by stochastic factors rather than deterministic systems. For instance, the multiplier-accelerator model of Samuelson (1939) is characterized by a deterministic second-order difference equation for aggregate output. Over a certain range of parameters, this equation produces deterministic cycles with a constant period of business cycles. Without doubt this model sheds deep insight into macroeconomic fluctuations. Nevertheless, a stochastic framework will provide a more realistic basis for analysis of periodicity in economics, because the observed periods of business cycles never occur evenly in any economy. Frisch (1933) demonstrates that a structural propagation mechanism can convert uncorrelated stochastic impulses into cyclical outputs with uneven, stochastic periodicity. Indeed, although not all uncertainties can be well characterized by probability theory, probability is the best quantitative analytic tool to describe uncertainties. The probability law of this stochastic economic system, which characterizes the evolution of the economy, can be viewed as the "law of economic motions." Accordingly, the tools and methods of mathematical statistics will provide the operating principles.

One important implication of the fundamental axioms is that one should not hope to determine precise, deterministic economic relationships, as do the models of demand, production, and aggregate consumption in standard micro- and macro-economic textbooks. No model could encompass the myriad essentially random aspects of economic life (i.e., no precise point forecast is possible, using a statistical terminology). Instead, one can only postulate some stochastic economic relationships. The purpose of econometrics is to infer the probability law of the economic system using observed data. Economic theory usually takes a form of imposing certain restrictions on the probability law. Thus, one can test economic theory or economic hypotheses by checking the validity of these restrictions.

It should be emphasized that the role of mathematics is different from the role of econometrics. The main task of mathematical economics is to express economic theory in the mathematical form of equations (or models) without regard to measurability or empirical verification of economic theory. Mathematics can check whether the reasoning process of an economic theory is correct and sometime can give surprising results and conclusions. However, it cannot check whether an economic theory can explain reality. To check whether a theory is consistent with reality, one needs econometrics. Econometrics is a fundamental methodology in the process of economic

analysis. Like the development of a natural science, the development of economic theory is a process of refuting the existing theories which cannot explain newly arising empirical stylized facts and developing new theories which can explain them. Econometrics rather than mathematics plays a crucial role in this process. There is no absolutely correct and universally applicable economic theory. Any economic theory can only explain the reality at certain stage, and therefore, is a "relative truth" in the sense that it is consistent with historical data available at that time. An economic theory may not be rejected due to limited data information. It is possible that more than one economic theory or model coexist simultaneously, because data does not contain sufficient information to distinguish the true one (if any) from false ones. When new data become available, a theory that can explain the historical data well may not explain the new data well and thus will be refuted. In many cases, new econometric methods can lead to new discovery and call for new development of economic theory.

Econometrics is not simply an application of a general theory of mathematical statistics. Although mathematical statistics provides many of the operating tools used in econometrics, econometrics often needs special methods because of the unique nature of economic data, and the unique nature of economic problems at hand. One example is the generalized method of moment estimation (Hansen 1982), which was proposed by econometricians aiming to estimate rational expectations models which only impose certain conditional moment restrictions characterized by the Euler equation and the conditional distribution of economic processes is unknown (thus, the classical maximum likelihood estimation cannot be used). The development of unit root and cointegration (e.g., Engle and Granger 1987, Phillips 1987), which is a core in modern time series econometrics, has been mainly motivated from Nelson and Plosser's (1982) empirical documentation that most macroeconomic time series display unit root behaviors. Thus, it is necessary to provide an econometric theory for unit root and cointegrated systems because the standard statistical inference theory is no longer applicable. The emergence of financial econometrics is also due to the fact that financial time series display some unique features such as persistent volatility clustering, heavy tails, infrequent but large jumps, and serially uncorrelated but not independent asset returns. Financial applications, such as financial risk management, hedging and derivatives pricing, often call for modeling for volatilities and the entire conditional probability distributions of asset returns. The features of financial data and the objectives of financial applications make the use of standard time series analysis quite limited, and therefore, call for the development of financial econometrics. Labor economics is another example which shows how labor economics and econometrics have benefited from each other. Labor economics has advanced quickly over the last few decades because of availability of high-quality labor data and rigorous empirical verification of hypotheses and theories on labor economics. On the other hand, microeconometrics, particularly panel data econometrics, has also advanced quickly due to the increasing availability of microeconomic data and the need to develop econometric theory

to accommondate the features of microeconomic data (e.g., censoring and endogeneity).

In the first issue of *Econometrica*, the founder of the econometric society, Fisher (1933), nicely summarizes the objective of the econometric society and main features of econometrics: "Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences.

But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous [sic] with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics."

1.5 Illustrative Examples

Specifically, econometrics can play the following roles in economics:

- Examine how well an economic theory can explain historical economic data (particularly the important stylized facts);
- Test validity of economic theories and economic hypotheses;
- Predict the future evolution of the economy.

To appreciate the roles of modern econometrics in economic analysis, we now discuss a number of illustrative econometric examples in various fields of economics and finance.

Example 1: The Keynes Model, the Multiplier and Policy Recommendation
The simplest Keynes model can be described by the system of equations

$$\begin{cases} Y_t = C_t + I_t + G_t, \\ C_t = \alpha + \beta Y_t + \varepsilon_t, \end{cases}$$

where Y_t is aggregate income, C_t is private consumption, I_t is private investment, G_t is government spending, and ε_t is consumption shock. The parameters α and β can have appealing economic interpretations: α is survival level consumption, and β is the marginal propensity to consume. The multiplier of the income with respect to government spending is

$$\frac{\partial Y_t}{\partial G_t} = \frac{1}{1-\beta},$$

which depends on the marginal propensity to consume β.

To assess the effect of fiscal policies on the economy, it is important to know the magnitude of β. For example, suppose the Chinese government wants to maintain a steady growth rate (e.g., an annual 8%) for its economy by active fiscal policy. It has to figure out how many government bonds to be issued each year. Insufficient government spending will jeopardize the goal of achieving the desired growth rate, but excessive government spending will cause budget deficit in the long run. The Chinese government has to balance these conflicting effects and this crucially depends on the knowledge of the value of β. Economic theory can only suggest a positive qualitative relationship between income and consumption. It never tells exactly what β should be for a given economy. It is conceivable that β differs from country to country, because cultural factors may have impact on the consumption behavior of an economy. It is also conceivable that β will depend on the stage of economic development in an economy. Fortunately, econometrics offers a feasible way to estimate β from observed data. In fact, economic theory even does not suggest a specific functional form for the consumption function. The linear functional form for the consumption is assumed for convenience, not implied by economic theory. Econometrics can provide a consistent estimation procedure for the unknown consumption function. This is called the nonparametric method (see, e.g., Hardle 1990, Pagan and Ullah 1999).

Example 2: Rational Expectations and Dynamic Asset Pricing Models

Suppose a representative agent has a constant relative risk aversion utility

$$U = \sum_{t=0}^{n} \beta^t u(C_t) = \sum_{t=0}^{n} \beta^t \frac{C_t^\gamma - 1}{\gamma},$$

where $\beta > 0$ is the agent's time discount factor, $\gamma \geq 0$ is the risk aversion parameter, $u(\cdot)$ is the agent's utility function in each time period, and C_t is consumption during period t. Let the information available to the agent at time t be represented by the σ-algebra I_t —in the sense that any variable whose value is known at time t is presumed to be I_t-measurable, and let $R_t = P_t/P_{t-1}$ be the gross return to an asset acquired at time $t-1$ at a price of P_{t-1}. The agent's optimization problem is to choose a sequence of consumptions $\{C_t\}$ over time to

$$\max_{\{C_t\}} E(U)$$

subject to the intertemporal budget constraint

$$C_t + P_t q_t \leq W_t + P_t q_{t-1},$$

where q_t is the quantity of the asset purchased at time t and W_t is the agent's period t income.

Define the marginal rate of intertemporal substitution

$$\text{MRS}_{t+1}(\theta) = \frac{\frac{\partial u(C_{t+1})}{\partial C_{t+1}}}{\frac{\partial u(C_t)}{\partial C_t}} = \left(\frac{C_{t+1}}{C_t}\right)^{\gamma-1},$$

where model parameter vector $\theta = (\beta, \gamma)'$. Then the first order condition of the agent's optimization problem can be characterized by

$$E\left[\beta \text{MRS}_{t+1}(\theta) R_{t+1} | I_t\right] = 1.$$

That is, the marginal rate of intertemporal substitution discounts gross returns to unity. This FOC is usually called the Euler equation of the economic system (see Hansen and Singelton 1982 for more discussion).

How to estimate this model? How to test validity of a rational expectations model? Here, the traditional popular maximum likelihood estimation method cannot be used, because one does not know the conditional distribution of economic variables of interest. Nevertheless, econometricians have developed a consistent estimation method based on the conditional moment condition or the Euler equation, which does not require knowledge of the conditional distribution of the data generating process. This method is called the generalized method of moments (see Hansen 1982).

In the empirical literature, it was documented that the empirical estimates of risk aversion parameter γ are often too small to justify the substantial difference between the observed returns on stock markets and bond markets (e.g., Mehra and Prescott 1985). This is the well-known equity premium puzzle. To resolve this puzzle, effort has been devoted to the development of new economic models with time-varying, large risk aversion. An example is Campbell and Cochrance's (1999) consumption-based capital asset pricing model. This story confirms our earlier statement that econometric analysis calls for new economic theory after documenting the inadequacy of the existing model.

Example 3: The Production Function and the Hypothesis on Constant Return to Scale

Suppose that for some industry, there are two inputs—labor L_i and capital stock K_i, and one output Y_i, where i is the index for firm i. The production function of firm i is a mapping from inputs (L_i, K_i) to output Y_i :

$$Y_i = \exp(\varepsilon_i) F(L_i, K_i),$$

where ε_i is a stochastic factor (e.g., the uncertain weather condition if Y_i is an agricultural product). An important economic hypothesis is that the production technology displays a constant return to scale (CRS), which is defined as follows:

$$\lambda F(L_i, K_i) = F(\lambda L_i, \lambda K_i) \text{ for all } \lambda > 0.$$

CRS is a necessary condition for the existence of a long-run equilibrium of a competitive market economy. If CRS does not hold for some industry, and the technology displays the increasing return to scale (IRS), the industry will lead to natural monopoly. Government regulation is then necessary to protect consumers' welfare. Therefore, testing CRS versus IRS has important policy implication, namely whether regulation is necessary.

A conventional approach to testing CRS is to assume that the production function is a Cobb-Douglas function:

$$F(L_i, K_i) = A \exp(\varepsilon_i) L_i^{\alpha} K_i^{\beta}.$$

Then CRS becomes a mathematical restriction on parameters (α, β) :

$$\mathbf{H}_0 : \alpha + \beta = 1.$$

If $a + \beta > 1$, the production technology displays IRS.

In statistics, a popular procedure to test one-dimensional parameter restriction is Student's t-test. Unfortunately, this procedure is not suitable for many cross-sectional economic data, which usually display conditional heteroskedasticity (e.g., a larger firm has a larger output variation). One needs to use a robust, heteroskedasticity-consistent test procedure, originally proposed in White (1980).

It should be emphasized that CRS is equivalent to the statistical hypothesis $\mathbf{H}_0 : \alpha + \beta = 1$ under the assumption that the production technology is a Cobb-Douglas function. This additional condition is not part of the CRS hypothesis and is called an auxiliary assumption. If the auxiliary assumption is incorrect, the statistical hypothesis $\mathbf{H}_0 : \alpha + \beta = 1$ will not be equivalent to CRS. Correct model specification is essential here for a valid conclusion and interpretation for the econometric inference.

Example 4: Effect of Economic Reforms on a Transitional Economy

We now consider an extended Cobb-Dauglas production function (after taking a logarithmic operation)

$$\ln Y_{it} = \ln A_{it} + \alpha \ln L_{it} + \beta \ln K_{it} + \gamma \text{Bonus}_{it} + \delta \text{Contract}_{it} + \varepsilon_{it},$$

where i is the index for firm $i \in \{1, ..., N\}$, and t is the index for year $t \in \{1, ..., T\}$, Bonus_{it} is the proportion of bonus out of total wage bill, and Contract_{it} is the proportion of workers who have signed a fixed-term contract. This is an example of the so-called panel data model (see, e.g., Hsiao 2003).

Paying bonus and signing fixed-term contracts were two innovative incentive reforms in the Chinese state-owned enterprises in the 1980s, compared to the fixed wage and life-time employment systems in the pre-reform era. Economic theory predicts that the introduction of the bonus and contract systems provides stronger incentives for workers to work harder, thus increasing

the productivity of a firm (see Groves, Hong, McMillan and Naughton 1994).

To examine the effects of these incentive reforms, we consider the null statistical hypothesis

$$\mathbf{H}_0 : \ \gamma = \delta = 0.$$

It appears that conventional t-tests or F-tests would serve our purpose here, if we can assume conditional homoskedasticity. Unfortunately, this cannot be used because there may well exist the other way of causation from Y_{it} to Bonus_{it} : a productive firm may pay its workers higher bonuses regardless of their efforts. This will cause correlation between the bonuses and the error term u_{it}, rendering the OLS estimator inconsistent and invalidating the conventional t-tests or F-tests. Fortunately, econometricians have developed an important estimation procedure called Instrumental Variables estimation, which can effectively filter out the impact of the causation from output to bonus and obtain a consistent estimator for the bonus parameter. Related hypothesis test procedures can be used to check whether bonus and contract reforms can increase firm productivity.

In evaluating the effect of economic reforms, we have turned an economic hypothesis—that introducing bonuses and contract systems has no effect on productivity—into a statistical hypothesis $\mathbf{H}_0 : \delta = \gamma = 0$. When the hypothesis $\mathbf{H}_0 : \delta = \gamma = 0$ is not rejected, we should not conclude that the reforms have no effect. This is because the extended production function model, where the reforms are specified additively, is only one of many ways to check the effect of the reforms. For example, one could also specify the model such that the reforms affect the marginal produtivities of labor and capital (i.e., the coefficients of labor and capital). Thus, when the hypothesis $\mathbf{H}_0 : \delta = \gamma = 0$ is not rejected, we can only say that we do not find evidence against the economic hypothesis that the reforms have no effect. We should not conclude that the reforms have no effect.

Example 5: The Efficient Market Hypothesis and Predictability of Financial Returns

Let Y_t be the stock return in period t, and let $I_{t-1} = \{Y_{t-1}, Y_{t-2}, ...\}$ be the information set containing the history of past stock returns. The weak form of efficient market hypothesis (EMH) states that it is impossible to predict future stock returns using the history of past stock returns:

$$E(Y_t|I_{t-1}) = E(Y_t).$$

The LHS, the so-called conditional mean of Y_t given I_{t-1}, is the expected return that can be obtained when one is fully using the information available at time $t-1$. The RHS, the unconditional mean of Y_t, is the expected market average return in the long-run; it is the expected return of a buy-and-hold trading strategy. When EMH holds, the past information of stock returns has no predictive power for future stock returns. An important implication of EMH is that mutual

fund managers will have no informational advantage over layman investors.

One simple way to test EMH is to consider the following autoregression

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

where p is a pre-selected number of lags, and ε_t is a random disturbance. EMH implies

$$\mathbf{H}_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0.$$

Any nonzero coefficient $\alpha_j, 1 \leq j \leq p$, is evidence against EMH. Thus, to test EMH, one can test whether the α_j are jointly zero. The classical F-test in a linear regression model can be used to test the hypothesis $\mathbf{H}_0$ when $\text{var}(\varepsilon_t|I_{t-1}) = \sigma^2$, i.e., when there exists conditional homoskedasticity. However, EMH may coexist with volatility clustering (i.e., $\text{var}(\varepsilon_t|I_{t-1})$ may be time-varying), which is one of the most important empirical stylized facts of financial markets (see Chen and Hong (2003) for more discussion). This implies that the standard F-test statistic cannot be used here, even asymptotically. Similarly, the popular Box and Pierce's (1970) portmanteau Q test, which is based on the sum of the first p squared sample autocorrelations, also cannot be used, because its asymptotic χ^2 distribution is invalid in presence of autoregressive conditional heteroskedasticity. One has to use procedures that are robust to conditional heteroskedasticity.

Like the discussion in Subsection 5.4, when one rejects the null hypothesis $\mathbf{H}_0$ that the α_j are jointly zero, we have evidence against EMH. Furthermore, the linear AR(p) model has predictive ability for asset returns. However, when one fails to reject the hypothesis $\mathbf{H}_0$ that the α_j are jointly zero, one can only conclude that we do not find evidence against EMH. One cannot conclude that EMH holds. The reason is, again, that the linear AR(p) model is one of many possibilities to check EMH (see, e.g., Hong and Lee 2005, for more discussion).

Example 6: Volatility Clustering and ARCH Models

Since the 1970s, oil crisis, the floating foreign exchanges system, and the high interest rate policy in the U.S. have stimulated a lot of uncertainty in the world economy. Economic agents have to incorporate these uncertainty in their decision-making. How to measure uncertainty has become an important issue.

In economics, volatility is a key instrument for measuring uncertainty and risk in finance. This concept is important to investigate information flows and volatility spillover, financial contagions between financial markets, options pricing, and calculation of Value at Risk.

Volatility can be measured by the conditional variance of asset return Y_t given the information available at time $t-1$:

$$\sigma_t^2 \equiv var(Y_t|I_{t-1}) = E\left[(Y_t - E(Y_t|I_{t-1}))^2|I_{t-1}\right].$$

An example of the conditional variance is the AutoRegressive Conditional Heteroskedasticity (ARCH) model, originally proposed by Engle (1982). An ARCH(q) model assumes that

$$\begin{cases} Y_t = \mu_t + \varepsilon_t, \\ \varepsilon_t = \sigma_t z_t, \\ \mu_t = E(Y_t|I_{t-1}), \\ \sigma_t^2 = \alpha + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}^2, \qquad \alpha > 0, \beta > 0, \\ \{z_t\} \sim i.i.d.(0,1). \end{cases}$$

This model can explain a well-known stylized fact in financial markets—volatility clustering: a high volatility tends to be followed by another high volatility, and a small volatility tends to be followed by another small volatility. It can also explain the non-Gaussian heavy tail of asset returns. More sophisticated volatility models, such as Bollerslev's (1986) Generalized ARCH or GARCH model, have been developed in time series econometrics.

In practice, an important issue is how to estimate a volatility model. Here, the models for the conditional mean μ_t and the conditional variance σ_t^2 are assumed to be correctly specified, but the conditional distribution of Y_t is unknown, because the distribution of the standardized innovation $\{z_t\}$ is unknown. Thus, the popular maximum likelihood estimation (MLE) method cannot be used. Nevertheless, one can assume that $\{z_t\}$ is i.i.d.$N(0,1)$ or follows other plausible distribution. Under this assumption, we can obtain a conditional distribution of Y_t given I_{t-1} and estimate model parameters using the MLE procedure. Although $\{z_t\}$ is not necessarily i.i.d.$N(0,1)$ and we know this, the estimator obtained this way is still consistent for the true model parameters. However, the asymptotic variance of this estimator is larger than that of the MLE (i.e., when the true distribution of $\{z_t\}$ is known), due to the effect of not knowing the true distribution of $\{z_t\}$. This method is called the quasi-MLE, or QMLE (see, e.g., White 1994). Inference procedures based on the QMLE are different from those based on the MLE. For example, the popular likelihood ratio test cannot be used. The difference comes from the fact that the asymptotic variance of the QMLE is different from that of the MLE, just like the fact that the asymptotic variance of the OLS estimator under conditional heteroskedasticity is different from that of the OLS under conditional homoskedasticity. Incorrect calculation of the asymptotic variance estimator for the QMLE will lead to misleading inference and conclusion (see White 1982, 1994 for more discussion).

Example 7: Modeling Economic Durations

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.

In practice, the main interest often lies in the question of how long a duration will continue,

given that it has not finished yet. The so-called 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 T_i is the duration from a population with the probability density function $f(t)$ and probability distribution function $F(t)$. Then the survival function is

$$S(t) = P(T_i > t) = 1 - F(t),$$

and the hazard rate

$$\lambda(t) = \lim_{\delta \to 0^+} \frac{P(t < T_i \leq t + \delta | T_i > t)}{\delta} = \frac{f(t)}{S(t)}.$$

Intuitively, the hazard rate $\lambda(t)$ is the instantaneous probability that an event of interest will end at time t given that it has lasted for period t. Note that the specification of $\lambda(t)$ is equivalent to a specification of the probability density $f(t)$. But $\lambda(t)$ is more interpretable from an economic point of view.

The hazard rate may not be the same for all individuals. To control heterogeneity across individuals, we assume that the individual-specific hazard rate depends on some individual characteristics X_i via the form

$$\lambda_i(t) = \exp(X_i'\beta)\lambda(t).$$

This is called the proportional hazard model, originally proposed by Cox (1972). The parameter

$$\beta = \frac{\partial}{\partial X_i} \ln \lambda_i(t) = \frac{1}{\lambda_i(t)} \frac{\partial}{\partial X_i} \lambda_i(t)$$

can be interpreted as the marginal relative effect of X_i on the hazard rate of individual i. Inference of β will allow one to examine how individual characteristics affect the duration of interest. For example, suppose T_i is the unemployment duration for individual i, then the inference of β will allow us to examine how individual characteristics, such as age, education, gender, and etc, can affect the unemployment duration. This will provide important policy implication on labor markets.

Because one can obtain the conditional probability density function of Y_i given X_i

$$f_i(t) = \lambda_i(t) S_i(t),$$

where the survival function $S_i(t) = \exp[-\int_0^t \lambda_i(s)ds]$, we can estimate β by the maximum likelihood estimation method.

For an excellent survey on duration analysis in labor economics, see Kiefer (1988), and for a complete and detailed account, see Lancaster (1990). Duration analysis has been also widely used in credit risk modeling in the recent financial literature.

The above examples, although not exhaustive, illustrate how econometric models and tools can be used in economic analysis. As noted earlier, an economy can be completely characterized by the probability law governing the economy. In practice, which attributes (e.g., conditional moments) of the probability law should be used depends on the nature of the economic problem at hand. In other words, different economic problems will require modeling different attributes of the probability law and thus require different econometric models and methods. In particular, it is not necessary to specify a model for the entire conditional distribution function for all economic applications. This can be seen clearly from the above examples.

1.6 Limitatons of Econometric Analysis

Although the general methodology of economic research is very similar to that of natural science, in general, economics and finance have not reached the mature stage that natural science (e.g., physics) has achieved. In particular, the prediction in economics and finance is not as precise as natural science (see, e.g., Granger 2001, for an assessment of macroeconomic forecasting practice).

Why?

Like any other statistical analysis, econometrics is the analysis of the "average behavior" of a large number of realizations, or the outcomes of a large number of random experiments with the same or similar features. However, economic data are not produced by a large number of repeated random experiments, due to the fact that an economy is not a controlled experiment. Most economic data are nonexperimental in their nature. This imposes some limitations on econometric analysis.

First, as a simplification of reality, economic theory or model can only capture the main or most important factors, but the observed data is the joint outcome of many factors together, and some of them are unknown and unaccounted for. These unknown factors are well present but their influences are ignored in economic modeling. This is unlike natural science, where one can remove secondary factors via controlled experiments. In the realm of economics, we are only passive observers; most data collected in economics are nonexperimental in that the data collecting agency may not have direct control over the data. The recently emerging field of experimental economics can help somehow, because it studies the behavior of economic agents under controlled experiments (see, e.g., Samuelson 2005). In other words, experimental economics controls the data generating process so that data is produced by the factors under study. Nevertheless, the scope of experimental economics is limited. One can hardly imagine how an economy with 1.3 billions of people can be experimented. For example, can we repeat the economic reforms in China and former Eastern European Socialist countries?

Second, an economy is an irreversible or non-repeatable system. A consequence of this is that data observed are a single realization of economic variables. For example, we consider the

annual Chinese GDP growth rate $\{Y_t\}$ over the past several years:

Y_{1997}	Y_{1998}	Y_{1999}	Y_{2000}	Y_{2001}	Y_{2002}	Y_{2003}	Y_{2004}	Y_{2005}
9.3%	7.8%	7.6%	8.4%	8.3%	9.1%	10.0%	10.1%	9.9%

.

GDP growths in different years should be viewed as different random variables, and each variable Y_t only has one realization! There is no way to conduct statistical analysis if one random variable only has a single realization. As noted earlier, statistical analysis studies the "average" behavior of a large number of realizations from the same data generating process. To conduct statistical analysis of economic data, economists and econometricians often assume some time-invarying "common features" of an economic system so as to use time series data or cross-sectional data of different economic variables. These common features are usually termed as "stationarity" or "homogeneity" of the economic system. With these assumptions, one can consider that the observed data are generated from the same population or populations with similar characters. Economists and econometricians assume that the conditions needed to employ the tools of statistical inference hold, but this is rather difficult, if not impossible, to check in practice.

Third, economic relationships are often changing over time for an economy. Regime shifts and structural changes are rather a rule than an exception, due to technology shocks and changes in preferences, population structure and institution arrangements. An unstable economic relationship makes it difficult for out-of-sample forecasts and policy-making. With a structural break, an economic model that was performing well in the past may not forecast well in the future. Over the past several decades, econometricians have made some progress to copy with the time-varying feature of an economic system. Chow's (1960) test, for example, can be used to check whether there exist structural breaks. Engle's (1982) volatility model can be used to forecast time-varying volatility using historical asset returns. Nevertheless, the time-varying feature of an economic system always imposes a challenge for economic forecasts. This is quite different from natural sciences, where the structure and relationships are more or less stable over time.

Fourth, data quality. The success of any econometric study hinges on the quantity as well as the quality of data. However, economic data may be subject to various defects. The data may be badly measured or may correspond only vaguely to the economic variables defined in the model. Some of the economic variables may be inherently unmeasurable, and some relevant variables may be missing from the model. Moreover, sample selection bias will also cause a problem. In China, there may have been a tendency to over-report or estimate the GDP growth rates given the existing institutional promotion mechanism for local government officials. Of course, the advances in computer technology, the development of statistical sampling theory and practice can help improve the quality of economic data. For example, the use of scanning machines makes every transaction data available.

The above features of economic data and economic systems together unavoidably impose some limitations for econometrics to achieve the same mature stage as the natural science.

1.7 Conclusion

In this chapter, we have discussed the philosophy and methodology of econometrics in economic research, and the differences between econometrics and mathematical economics and mathematical statistics. I first discussed two most important features of modern economics, namely mathematical modeling and empirical analysis. This is due to the effort of several generations of economists to make economics a science. As the methodology for empirical analysis in economics, econometrics is an interdisciplinary field. It uses the insights from economic theory, uses statistics to develop methods, and uses computers to estimate models. We then discussed the roles of econometrics and its differences from mathematics, via a variety of illustrative examples in economics and finance. Finally, we pointed out some limitations of econometric analysis, due to the fact that any economy is not a controlled experiment. It should be emphasized that these limitations are not only the limitations of econometrics, but of economics as a whole.

EXERCISES

1.1. Discuss the differences of the roles of mathematics and econometrics in economic research.

1.2. What are the fundamental axioms of econometrics? Discuss their roles and implications.

1.3. What are the limitations of econometric analysis? Discuss possible ways to alleviate the impact of these limits.

1.4. How do you perceive the roles of econometrics in decision-making in economics and business?

CHAPTER 2 GENERAL REGRESSION ANALYSIS

Abstract: This chapter introduces *regression analysis*, the most popular statistical tool to explore the dependence of one variable (say Y) on others (say X). The variable Y is called the dependent variable, and X is called the independent variable or explanatory variable. The regression relationship between X and Y can be used to study the effect of X on Y or to predict Y using X. We motivate the importance of the regression function from both the economic and statististical perspectives, and characterize the condition for correct specification of a linear model for the regression function, which is shown to be crucial for a valid economic interpretation of model parameters.

Key words: Conditional distribution, Conditional mean, Consumption function, Linear regression model, Marginal propensity to consume, Model specification, Regression

2.1 Conditional Probability Distribution

Notational Convention: Throughout this book, capital letters ($e.g., Y$) denote random variables or random vectors, lower case letters ($e.g., y$) denote realizations of random variables.

We assume that $Z = (Y, X')'$ is a random vector with $E(Y^2) < \infty$, where Y is a scalar, X is a $(k+1) \times 1$ vector of economic variables with its first component being a constant, and X' denotes the transpose of X. Given this assumption, the conditional mean $E(Y|X)$ exists and is well-defined.

Statistically speaking, the relationship between two random variables or vectors X (e.g., oil price change) and Y (e.g., economic growth) can be characterized by their joint distribution function. Suppose $(X', Y)'$ are continuous random vectors, and the joint probability density function (*pdf*) of $(X', Y)'$ is $f(x, y)$. Then the marginal *pdf* of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy,$$

and the conditional *pdf* of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)},$$

provided $f_X(x) > 0$. The conditional pdf $f_{Y|X}(y|x)$ completely describes how Y depends on X. In other words, it characterizes a predictive relationship of Y using X. With this conditional *pdf* $f_{Y|X}(y|x)$, we can compute the following quantities:

- The conditional mean

$$\begin{aligned} E(Y|x) &\equiv E(Y|X=x) \\ &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy; \end{aligned}$$

- the conditional variance

$$\begin{aligned} \text{var}(Y|x) &\equiv \text{var}(Y|X=x) \\ &= \int_{-\infty}^{\infty} [y - E(Y|x)]^2 f_{Y|X}(y|x)dy \\ &= E(Y^2|x) - [E(Y|x)]^2; \end{aligned}$$

- the conditional skewness

$$S(Y|x) = \frac{E[(Y - E(Y|x))^3|x]}{[\text{var}(Y|x)]^{3/2}};$$

- the conditional kurtosis

$$K(Y|x) = \frac{E[(Y - E(Y|x))^4|x]}{[\text{var}(Y|x)]^2};$$

- the α-conditional quantile $Q(x, \alpha)$:

$$P\left[Y \leq Q(X, \alpha)|X = x\right] = \alpha \in (0, 1).$$

  Note that when $\alpha = 0.5$, $Q(x, 0.5)$ is the conditional median, which is the cutoff point or threshold that divides the population into two equal halves, conditional on $X = x$.

The class of conditional moments is a summary characterization of the conditional distribution $f_{Y|X}(y|x)$. A mathematical model (i.e., an assumed functional form with a finite number of unknown parameters) for a conditional moment is called an econometric model for that conditional moment.

Question: Which moment to model and use in practice?

It depends on economic applications. For some applications, we only need to model the first conditional moment, namely the conditional mean. For example, asset pricing aims at explaining excess asset returns by systematic risk factors. An asset pricing model is essentially a model for the conditional mean of asset returns on risk factors. For others, we may have to model higher order conditional moments and even the entire conditional distribution. In econometric practice, the most popular models are the first two conditional moments, namely the conditional mean

and conditional variance. There is no need to model the entire conditional distribution of Y given X when only certain conditional moments are needed. For example, when the conditional mean is of concern, there is no need to model the conditional variance or impose restrictive conditions on it.

The conditional moments, and more generally the conditional probability distribution of Y given X, are not the causal relationship from X to Y. They are a predictive relationship. That is, one can use the information on X to predict the distribution of Y or its attributes. These probability concepts cannot tell whether the change in Y is caused by the change in X. Such causal interpretation has to reply on economic theory. Economic theory usually hypothesizes that a change in Y is caused by a change in X, i.e., there exists a causal relationship from X to Y. If such an economic causal relationship exists, we will find a predictive relationship from X to Y. On the other hand, a documented predictive relationship from X to Y may not be caused by an economic causal relationship from X to Y. For example, it is possible that both X and Y are positively correlated due to their dependence on a common factor. As a result, we will find a predictive relationship from X to Y, although they do not have any causal relationship. In fact, it is well-known in econometrics that some economic variables that trend consistently upwards over time are highly correlated even in the absence of any causal relationship between them. Such strong correlations are called spurious relationships.

2.2 Regression Analysis

We now focus on the first conditional moment, $E(Y|X)$, which is called the regression function of Y on X, where Y is called the regressand, and X is called the regressor vector. The term "regression" is used to signify a predictive relationship between Y and X.

Definition 2.1 [Regression Function]: The conditional mean $E(Y|X)$ is called a regression function of Y on X.

Many economic theories can be characeterized by the conditional mean $E(Y|X)$ of Y given X, provided X and Y are suitably defined. Most, though not all, of dynamic economic theories and/or dynamic optimization models, such as rational expectations, efficient markets hypothesis, expectations hypothesis, and optimal dynamic asset pricing, have important implications on (and only on) the conditional mean of underlying economic variables given the information available to economic agents (e.g., Cochrane 2001, Sargent and Ljungqvist 2002). For example, the classical efficient market hypothesis states that the expected asset return given the information available, is zero, or at most, is constant over time; the optimal dynamic asset pricing theory implies that the expectation of the pricing error given the information available is zero for each asset (Cochrane 2001). Although economic theory may suggest a nonlinear relationship, it does not

give a completely specified functional form for the conditional mean of economic variables. It is therefore important to model the conditional mean properly.

Before modeling $E(Y|X)$, we first discuss some probabilistic properties of $E(Y|X)$.

Lemma 2.1: $E[E(Y|X)] = E(Y)$.

Proof: The result follows immediately from applying the law of iterated expectations below.

Lemma 2.2 [Law of Iterated Expectations (LIE)]: *For any measurable function* $G(X, Y)$,

$$E\left[G(X,Y)\right] = E\left\{E\left[G(X,Y)|X\right]\right\},$$

provided the expectation $E[G(X,Y)]$ *exists.*

Proof: We consider the case of the continuous distribution of $(Y, X')'$ only. By the multiplication rule that the joint *pdf* $f(x,y) = f_{Y|X}(y|x)f_X(x)$, we have

$$\begin{aligned}
E[G(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)f_{Y|X}(y|x)f_X(x)dxdy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} G(x,y)f_{Y|X}(y|x)dy\right] f_X(x)dx \\
&= \int_{-\infty}^{\infty} E[G(X,Y)|X=x]f_X(x)dx \\
&= E\{E[G(X,Y)|X]\},
\end{aligned}$$

where the operator $E(\cdot|X)$ is the expectation with respect to $f_{Y|X}(\cdot|X)$, and the operator $E(\cdot)$ is the expectation with respect to $f_X(\cdot)$. This completes the proof.

Interpretation of $E(Y|X)$ and LIE:

Example 1: Suppose Y is wage, and X is a gender dummy variable, taking value 1 if an employee is female and value 0 if an employee is male. Then

$$\begin{aligned}
E(Y|X &= 1) = \text{ average wage of a female worker,} \\
E(Y|X &= 0) = \text{ average wage of a male worker,}
\end{aligned}$$

and the overall average wage

$$\begin{aligned} E(Y) &= E[E(Y|X)] \\ &= P(X=1)E(Y|X=1) + P(X=0)E(Y|X=0), \end{aligned}$$

where $P(X=1)$ is the proportion of female employees in the labor force, and $P(X=0)$ is the proportion of the male employees in the labor force. The use of LIE here thus provides some insight into the income distribution between genders.

Example 2: Suppose Y is an asset return and we have two information sets: X and $\tilde{X}$, where $X \subset \tilde{X}$ so that all information in X is also in $\tilde{X}$ but $\tilde{X}$ contains some extra information. Then we have a conditional version of the law of iterated expectations says that

$$E(Y|X) = E[E(Y|\tilde{X})|X]$$

or equivalently

$$E\left\{\left[Y - E(Y|\tilde{X})|X\right]\right\} = 0.$$

where $Y - E(Y|\tilde{X})$ is the prediction error using the superior information set $\tilde{X}$. The conditional LIE says that one cannot use limited information X to predict the prediction error one would make if one had superior information $\tilde{X}$. See Campbell, Lo and MacKinlay (1997, p.23) for more discussion.

Question: Why is $E(Y|X)$ important from a statistical perspective?

Suppose we are interested in predicting Y using some function $g(X)$ of X, and we use a so-called Mean Squared Error (MSE) criterion to evaluate how well $g(X)$ approximates Y. Then the optimal predictor under the MSE criterion is the conditional mean, as will be shown below.

We first define the MSE criterion. Intuitively, MSE is the average of the squared deviations between the predictor $g(X)$ and the actual outcome Y.

Definition 2.2 [MSE]: Suppose function $g(X)$ is used to predict Y. Then the mean squared error of function $g(X)$ is defined as

$$MSE(g) = E\left[Y - g(X)\right]^2,$$

provided the expectation exists.

The theorem below states that $E(Y|X)$ minimizes the MSE.

Theorem 2.3 [Optimality of $E(Y|X)$]: The regression function $E(Y|X)$ *is the solution to the optimization problem*

$$\begin{aligned} E(Y|X) &= \arg\min_{g\in\mathbb{F}} MSE(g) \\ &= \arg\min_{g\in\mathbb{F}} E[Y-g(X)]^2, \end{aligned}$$

where $\mathbb{F}$ *is the space of all measurable and square-integrable functions*

$$\mathbb{F} = \{g(\cdot): \int_{-\infty}^{\infty} g^2(x) f_X(x) dx < \infty\}.$$

Proof: We will use the variance and squared-bias decomposition technique. Put

$$g_o(X) \equiv E(Y|X).$$

Then

$$\begin{aligned} MSE(g) &= E[Y-g(X)]^2 \\ &= E\left[Y-g_o(X)+g_o(X)-g(X)\right]^2 \\ &= E\left[Y-g_o(X)\right]^2 + E\left[g_o(X)-g(X)\right]^2 \\ &\quad +2E\{\left[Y-g_o(X)\right]\left[g_o(X)-g(X)\right]\} \\ &= E\left[Y-g_o(X)\right]^2 + E\left[g_o(X)-g(X)\right]^2, \end{aligned}$$

where the cross-product term

$$E\{\left[Y-g_o(X)\right]\left[g_o(X)-g(X)\right]\} = 0$$

by LIE and the fact that $E\{[Y-g_o(X)]|X\} = 0$ a.s.

In the above MSE decomposition, the first term $E[Y-g_o(X)]^2$ is the quadratic variation of the prediction error of the regression function $g_o(X)$. This does not depend on the choice of function $g(X)$. The second term $E[g_o(X)-g(X)]^2$ is the is the quadratic variation of the approximation error of $g(X)$ for $g_o(X)$. This term achieves its minimum of zero if and only if one chooses $g(X) = g_o(X)$ a.s. Because the first term $E[Y-g_o(X)]^2$ does not depend on $g(X)$, minimizing MSE(g) is equivalent to minimizing the second term $E[g_o(X)-g(X)]^2$. Therefore, the optimal solution for minimizing MSE(g) is given by $g^*(X) = g_o(X)$. This completes the proof.

Remarks:

MSE is a popular criterion for measuring precision of a predictior $g(X)$ for Y. It has at least two advantages: first, it can be analyzed conveniently, and second, it has a nice decomposition of a variance component and a squared-bias component.

However, MSE is one of many possible criteria for measuring goodness of the predictor $g(X)$ for Y. In general, any increasing function of the absolute value $|Y - g(X)|$ can be used to measure the goodness of fit for the predictor $g(X)$. For example, the Mean Absolute Error

$$MAE(g) = E|Y - g(X)|$$

is also a reasonable criterion.

It should be emphasized that different criteria have different optimizers. For example, the optimizer for $MAE(g)$ is the conditional median, rather than the conditional mean. The conditional median, say $m(x)$, is defined as the solution to

$$\int_{-\infty}^{m} f_{Y|X}(y|x)dy = 0.5.$$

In other words, $m(x)$ divides the conditional population into two equal halves.

Example 3: Let the joint *pdf* $f_{XY}(x, y) = e^{-y}$ for $0 < x < y < \infty$. Find $E(Y|X)$ and $\text{var}(Y|X)$.

Solution: We first find the conditional pdf $f_{Y|X}(y|x)$. The marginal *pdf* of X

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy \\ &= \int_{x}^{\infty} e^{-y}dy \\ &= e^{-x} \text{ for } 0 < x < \infty. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x, y)}{f_X(x)} \\ &= e^{-(y-x)} \text{ for } 0 < x < y < \infty. \end{aligned}$$

Then

$$\begin{aligned} E(Y|x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \\ &= \int_{x}^{\infty} y e^{-(y-x)}dy \\ &= e^{x}\int_{x}^{\infty} y e^{-y}dy \\ &= -e^{x}\int_{x}^{\infty} y de^{-y} \\ &= 1+x. \end{aligned}$$

Thus, the regression function $E(Y|X)$ is linear in X.

To compute $\text{var}(Y|X)$, we will use the formula

$$\text{var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2 .$$

Because

$$\begin{aligned} E(Y^2|x) &= \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)dy \\ &= \int_{x}^{\infty} y^2 e^{-(y-x)}dy \\ &= e^{x}\int_{x}^{\infty} y^2 e^{-y}dy \\ &= -e^{x}\int_{x}^{\infty} y^2 de^{-y} \text{ where } de^{-y} = -e^{-y}dy. \\ &= (-e^{x})\left[y^2 e^{-y}|_{x}^{\infty} - \int_{x}^{\infty} e^{-y}dy^2\right] \\ &= [-e^{x}]\left[0 - x^2 e^{-x} - 2\int_{x}^{\infty} y e^{-y}dy\right] \\ &= x^2 + 2e^{x}\int_{x}^{\infty} y e^{-y}dy \\ &= x^2 + 2\int_{x}^{\infty} y e^{-(y-x)}dy \\ &= x^2 + 2(1+x), \end{aligned}$$

we have

$$\begin{aligned} \text{var}(Y|x) &= E(Y^2|x) - [E(Y|x)]^2 \\ &= x^2 + 2(1+x) - (1+x)^2 \\ &= 1. \end{aligned}$$

The conditional variance of Y given X does not depend on X. That is, X has no effect on the conditional variance of Y.

The above example shows that while the conditional mean of Y given X is a linear function of X, the conditional variance of Y may not depend on X. This is essentially the assumption made in the classical linear regression model (see Chapter 3). Another example for which we have a linear regression function with constant conditional variance is when X and Y are jointly normally distributed (see Exercise 2 at the end of this chapter).

Theorem 2.4 [Regression Identity]: *Suppose $E(Y|X)$ exists. Then we can always write*

$$Y = E(Y|X) + \varepsilon,$$

where ε is called the regression disturbance and has the property that

$$E(\varepsilon|X) = 0.$$

Proof: Put $\varepsilon = Y - E(Y|X)$. Then

$$Y = E(Y|X) + \varepsilon,$$

where

$$\begin{aligned} E(\varepsilon|X) &= E\{[Y - E(Y|X)]|X\} \\ &= E(Y|X) - E[g_o(X)|X] \\ &= E(Y|X) - g_o(X) \\ &= 0. \end{aligned}$$

Remarks:

The regression function $E(Y|X)$ can be used to predict the expected value of Y using the information of X. In regression analysis, an important issue is the direction of causation between Y and X.In practice, one often hope to check whether Y "depends" on or can be "explained" by X, with help of economic theory. For this reason, Y is called the dependent variable, and X is

called the explanatory variable or vector. However, it should be emphasized that the regression function $E(Y|X)$ itself does not tell any causal relationship between Y and X.

The random variable ε represents the part of Y that is not captured by $E(Y|X)$. It is usually called a *noise* or a *disburbance*, because it "disturbs" an otherwise stable relationship between Y and X. On the other hand, the regression function $E(Y|X)$ is called a *signal*.

The property that $E(\varepsilon|X) = 0$ implies that the regression disturbance ε contains no systematic information of X that can be used to predict the expected value of Y. In other words, all information of X that can be used to predict the expectation of Y has been completely summarized by $E(Y|X)$. The condition $E(\varepsilon|X) = 0$ is crucial for the validity of economic interpretation of model parameters, as will be seen shortly.

$E(\varepsilon|X) = 0$ implies that the unconditional mean of ε is zero:

$$E(\varepsilon) = E[E(\varepsilon|X)] = 0$$

and that ε is orthogonal to X :

$$\begin{aligned} E(X\varepsilon) &= E\left[E(X\varepsilon|X)\right] \\ &= E\left[XE(\varepsilon|X)\right] \\ &= E(X \cdot 0) \\ &= 0. \end{aligned}$$

Since $E(\varepsilon) = 0$, we have $E(X\varepsilon) = \text{cov}(X, \varepsilon)$. Thus, orthogonality ($E(X\varepsilon) = 0$) means that X and ε are uncorrelated.

In fact, ε is orthogonal to any measurable function of X, i.e., $E[\varepsilon h(X)] = 0$ for any measurable function $h(\cdot)$. This implies that we cannot predict the mean of ε by using any possible model $h(X)$, no matter it is linear or nonlinear.

Question: Is $E(\varepsilon|X) = 0$ equivalent to $E[\varepsilon h(X)] = 0$ for all measurable $h(\cdot)$?
Answer: Yes. How to show it? See Exercise 11 at the end of this chapter for more discussion.

It is possible that $E(\varepsilon|X) = 0$ but $\text{var}(\varepsilon|X)$ is a function of X. If $\text{var}(\varepsilon|X) = \sigma^2 > 0$, we say that there exists **conditional homoskedasticity** for ε. In this case, X cannot be used to predict the (quadratic) variation of Y. On the other hand, if $\text{var}(\varepsilon|X) \neq \sigma^2$ for any constant $\sigma^2 > 0$, we say that there exists **conditional heteroskedasticity**. Econometric procedures of regression analysis are usually different, depending on whether there exists conditional heteroskedasticity. For example, the so-called conventional t-test and F-test are invalid under conditional heteroskedasticity (see Chapter 3 for the introduction of the t-test and F-test). This will be discussed in detail in subsequent chapters.

Example 4: Suppose

$$\varepsilon = \eta\sqrt{\beta_0 + \beta_1 X^2},$$

where random variables X and η are independent, and $E(\eta) = 0$,$\text{var}(\eta) = 1$. Find $E(\varepsilon|X)$ and $\text{var}(\varepsilon|X)$.

Solution:

$$\begin{aligned} E(\varepsilon|X) &= E\left[\eta\sqrt{\beta_0 + \beta_1 X^2}\,|X\right] \\ &= \sqrt{\beta_0 + \beta_1 X^2}E(\eta|X) \\ &= \sqrt{\beta_0 + \beta_1 X^2}E(\eta) \\ &= \sqrt{\beta_0 + \beta_1 X^2}\cdot 0 \\ &= 0. \end{aligned}$$

Next,

$$\begin{aligned} \text{var}(\varepsilon|X) &= E\left\{[\varepsilon - E(\varepsilon|X)]^2|X\right\} \\ &= E(\varepsilon^2|X) \\ &= E[\eta^2(\beta_0 + \beta_1 X^2)|X] \\ &= (\beta_0 + \beta_1 X^2)E(\eta^2|X) \\ &= (\beta_0 + \beta_1 X^2)\cdot 1 \\ &= \beta_0 + \beta_1 X^2. \end{aligned}$$

Although the conditional mean ε given X is identically zero, the conditional variance of ε given X depends on X.

The regression analysis (conditional mean analysis) is the most popular statistical method in econometrics. It has been applied widely to economics. For example, it can be used to

- estimate the relationship between economic variables.
- test economic hypotheses.
- forecast future values of Y.

Example 5: Let Y=consumption, X=disposable income. Then the regression function $E(Y|X) = C(X)$ is the so-called consumption function, and the marginal propensity to consume (MPC) is the derivative

$$MPC = C'(X) = \frac{d}{dX}E(Y|X).$$

MPC is an important concept in the "multiplier effect" analysis. The magnitude of MPC is important in macroeconomic policy analysis and forecasting. On the other hand, when Y is consumption on food only, then Engle's law implies that MPC must be a decreasing function of X. Therefore, we can test Engle's law by testing whether $C'(X) = \frac{d}{dX}E(Y|X)$ is a decreasing function of X.

Example 6: Y=output, X=(labor, capital, raw material)$'$, then the regression $E(Y|X) = F(X)$ is the so-called production function. This can be used to test the hypothesis of constant return to scale (CRS), which is defined as

$$\lambda F(X) = F(\lambda X) \text{ for all } \lambda > 0.$$

Example 7: Let Y be the cost of producing certain output X. Then the regression function $E(Y|X) = C(X)$ is the cost function. For a monopoly firm or industry, the marginal cost must be declining in output X. That is,

$$\begin{aligned} \frac{d}{dX}E(Y|X) &= C'(X) > 0, \\ \frac{d^2}{dX^2}E(Y|X) &= C''(X) < 0. \end{aligned}$$

These imply that the cost function of a monopoly is a nonlinear function of X.

Question: Why may there exist conditional heteroskedasticity?

Generally speaking, given that $E(Y|X)$ depends on X, it is conceivable that $\text{var}(Y|X)$ and other higher order conditional moments may also depend on X. In fact, conditional heteroskedasticity may arise from different sources. For example, a larger firm may have a larger output variation. Granger and Machina (2006) explain why economic variables may display volatility clustering from an econometric structral perspective.

The following example shows that conditional heteroskedasticity may arise due to random coefficients in a data generating process.

Example 8 [Random Coefficient Process]: Suppose

$$Y = \beta_0 + (\beta_1 + \beta_2\eta)X + \eta,$$

where X and η are independent, and $E(\eta) = 0$,$\text{var}(\eta) = \sigma_\eta^2$. Find the conditional mean $E(Y|X)$ and conditional variance $\text{var}(Y|X)$.

Solution: (i)

$$\begin{aligned} E(Y|X) &= \beta_0 + E[(\beta_1 + \beta_2\eta)X|X] + E(\eta|X) \\ &= \beta_0 + \beta_1 X + \beta_2 X E(\eta|X) + E(\eta|X) \\ &= \beta_0 + \beta_1 X + \beta_2 X E(\eta) + E(\eta) \\ &= \beta_0 + \beta_1 X + \beta_2 X \cdot 0 + 0 \\ &= \beta_0 + \beta_1 X. \end{aligned}$$

(ii)

$$\begin{aligned} \text{var}(Y|X) &= E\left[(Y - E(Y|X))^2|X\right] \\ &= E\left[(\beta_0 + (\beta_1 + \beta_2\eta)X + \eta - \beta_0 - \beta_1 X)^2|X\right] \\ &= E\left[(\beta_2 X\eta + \eta)^2|X\right] \\ &= E\left[(\beta_2 X + 1)^2\eta^2|X\right] \\ &= (1 + \beta_2 X)^2 E(\eta^2|X) \\ &= (1 + \beta_2 X)^2 E(\eta^2) \\ &= (1 + \beta_2 X)^2 \sigma_\eta^2. \end{aligned}$$

The random coefficient process has been used to explain why the conditional variance may depend on the regressor X. We can write this process as

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where

$$\varepsilon = (1 + \beta_2 X)\eta.$$

Note that $E(\varepsilon|X) = 0$ but $\text{var}(\varepsilon|X) = (1 + \eta_2 X)^2\sigma_\eta^2$.

2.3 Linear Regression Modeling

As we have known above, the conditional mean $g_o(X) \equiv E(Y|X)$ is the solution to the MSE optimization problem

$$\min_{g \in \mathbb{F}} E[Y - g(X)]^2,$$

where $\mathbb{F}$ is a class of functions that includes all measurable and square-integrable functions, i.e.,

$$\mathbb{F} = \left\{ g(\cdot) : \mathbb{R}^{k+1} \to \mathbb{R} \mid \int g^2(x) f_X(x) dx < \infty \right\}.$$

In general, the regression function $E(Y|X)$ is an unknown functional form of X. Economic theory usually suggests a qualitative relationship between X and Y (e.g., the cost of production is an increasing function of output X), but it never suggests a concrete functional form. One needs to use some mathematical model to approximate $g_o(X)$.

Question: How to model $g_o(X)$?

In econometrics, a most popular modeling strategy is the parametric approach, which assumes a known functional form for $g_o(X)$, up to some unknown parameters. In particular, one usually uses a class of linear functions to approximate $g_o(x)$, which is simple and easy to interpret. This is the approach we will take in most of this book.

We first introduce a class of affine functions.

Definition 2.3 [Affine Functions]: Denote

$$X = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_k \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Then the class of affine functions is defined as

$$\begin{aligned} \mathbb{A} &= \{g : \mathbb{R}^{k+1} \to \mathbb{R} : g(X) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j, \beta_j \in \mathbb{R}\} \\ &= \{g : \mathbb{R}^{k+1} \to \mathbb{R} \mid g(X) = \beta' X\}. \end{aligned}$$

Here, there is no restriction on the values of parameter vector β. For this class of functions, the functional form is known to be linear in both explanatory varibles X and parameters β; the unknown is the $(k+1) \times 1$ vector β.

Remarks:

From an econometric point of view, the key feature of $\mathbb{A}$ is that $g(X) = X'\beta$ is linear in β, not in X. Later, we will generalize $\mathbb{A}$ so that $g(X) = X'\beta$ is linear in β but is possibly nonlinear in X. For example, when $k = 1$, we can generalize $\mathbb{A}$ to include

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2,$$

or

$$g(X) = \beta_0 + \beta_1 \ln X_1.$$

These possibilities are included in $\mathbb{A}$ if we properly redefine X as $X = (1, X_1, X_1^2)'$ or $X = (1, \ln X_1)'$. Therefore, our econometric theory to be developed in subsequent chapters are actually applicable to all regression models that are linear in β but not necessarily linear in X. Such models are called linear regression models. Conversely, a nonlinear regression model for $g_o(X)$ means a known parametric functional form $g(X, \beta)$ which is nonlinear in β. An example is the so-called logistic regression model

$$g(X, \beta) = \frac{1}{1 + \exp(-X'\beta)}.$$

Nonlinear regression models can be handled using the analytic tools developed in Chapter 8. See more discussions there.

We now solve the constrained minimization problem

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2.$$

The solution $g^*(X) = X'\beta^*$ is called the **Best Linear Least Squares** Predictor for Y, and β^* is called the best LS approximation coefficient vector.

Theorem 2.5 [Best Linear LS Prediction]: *Suppose $E(Y^2) < \infty$ and the $(k+1) \times (k+1)$ matrix $E(XX')$ is nonsingular. Then the best linear LS predictor that solves*

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2$$

is the linear function

$$g^*(X) = X'\beta^*,$$

where the optimizing coefficient vector

$$\beta^* = [E(XX')]^{-1} E(XY).$$

Proof: First, noting that

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2,$$

we first find the FOC:

$$\frac{d}{d\beta} E(Y - X'\beta)^2|_{\beta = \beta^*} = 0.$$

The left hand side

$$\begin{aligned}\frac{d}{d\beta}E(Y-X'\beta)^2 &= E\left[\frac{\partial}{\partial\beta}(Y-X'\beta)^2\right] \\ &= E\left[2(Y-X'\beta)\frac{\partial}{\partial\beta}(-X'\beta)\right] \\ &= -2E\left[(Y-X'\beta)\frac{\partial}{\partial\beta}(X'\beta)\right] \\ &= -2E[X(Y-X'\beta)].\end{aligned}$$

Therefore, FOC implies that

$$\begin{aligned}E[X(Y-X'\beta^*)] &= 0 \text{ or} \\ E(XY) &= E(XX')\beta^*.\end{aligned}$$

Multiplying the inverse of $E(XX')$, we obtain

$$\beta^* = [E(XX')]^{-1}E(XY).$$

It remains to check SOC: The Hessian matrix

$$\frac{d^2}{d\beta d\beta'}E(Y-X'\beta)^2 = 2E(XX')$$

is positive definite provided $E(XX')$ is nonsingular (why?). Therefore, β^* is a global minimizer. This completes the proof.

Remarks:

The moment condition $E(Y^2) < \infty$ ensures that $E(Y|X)$ exists and is well-defined. When the $(k+1)\times(k+1)$ matrix

$$E(XX') = \begin{bmatrix} 1 & E(X_1) & E(X_2) & \cdots & E(X_k) \\ E(X_1) & E(X_1^2) & E(X_1X_2) & \cdots & E(X_1X_k) \\ E(X_2) & E(X_2X_1) & E(X_2^2) & \cdots & \\ \vdots & \vdots & & & \\ E(X_k) & E(X_kX_1) & & & E(X_k^2) \end{bmatrix}$$

is nonsingular and $E(XY)$ exists, the best linear LS approximation coefficient β^* is always well-defined, no matter whether $E(Y|X)$ is linear or nonlinear in X.

To gain insight into the nature of β^*, we consider a simple case where $\beta = (\beta_0, \beta_1)'$ and

$X = (1, X_1)'$. Then the slope coefficient and the intercept coefficient are, respectively,

$$\begin{aligned} \beta_1^* &= \frac{\text{cov}(Y, X_1)}{\text{var}(X_1)}, \\ \beta_0^* &= E(Y) - \beta_1^* E(X_1). \end{aligned}$$

Thus, the best linear LS approximation coefficient β_1^* is proportional to $\text{cov}(Y, X_1)$. In other words, β_1^* captures the dependence between Y and X_1 that is measurable by $\text{cov}(Y, X_1)$. It will miss the dependence between Y and X_1 that cannot be measured by $\text{cov}(Y, X_1)$. Therefore, linear regression analysis is essentially *correlation analysis*.

In general, the best linear LS predictor $g^*(X) \equiv X'\beta^* \neq E(Y|X)$. An important question is what happens if $g^*(X) = X'\beta^* \neq E(Y|X)$? In particular, what is the interpretation of β^*?

We now discuss the relationship between the best linear LS prediction and a linear regression model.

Definition 2.4 [Linear Regression Model]: The specification

$$Y = X'\beta + u, \ \beta \in \mathbb{R}^{k+1},$$

is called a linear regression model, where u is the regression model disturbance or regression model error. If $k = 1$, it is called a bivariate linear regression model or a straight line regression model. If $k > 1$, it is called a multiple linear regression model.

The linear regression model is an artificial specification. Nothing ensures that the regression function is linear, namely $E(Y|X) = X'\beta^o$ for some β^o. In other words, the linear model may not contain the true regression function $g_o(X) \equiv E(Y|X)$. However, even if $g_o(X)$ is not a linear function of X, the linear regresson model $Y = X'\beta + u$ may still have some predictive ability although it is a misspecified model.

We first characterize the relationship between the best linear LS approximation and the linear regression model.

Theorem 2.6: *Suppose the conditions of the previous theorem hold. Let*

$$Y = X'\beta + u,$$

and let β^* be the best linear least squares approximation coefficient. Then

$$\beta = \beta^*$$

if and only if the following orthogonality condition holds:

$$E(Xu) = 0.$$

Proof: From the linear regression model $Y = X'\beta + u$, we have $u = Y - X'\beta$, and so

$$E(Xu) = E(XY) - E(XX')\beta.$$

(a) Necessarity: If $\beta = \beta^*$, then

$$\begin{aligned} E(Xu) &= E(XY) - E(XX')\beta^* \\ &= E(XY) - E(XX')[E(XX')]^{-1}E(XY) \\ &= 0. \end{aligned}$$

(b) Sufficiency: If $E(Xu) = 0$, then

$$\begin{aligned} E(Xu) &= E(XY) - E(XX')\beta \\ &= 0. \end{aligned}$$

From this and the fact that $E(XX')$ is nonsingular, we have

$$\beta = [E(XX')]^{-1}E(XY) \equiv \beta^*.$$

This completes the proof.

Remarks:

This theorem implies that no matter whether $E(Y|X)$ is linear or nonlinear in X, we can always write

$$Y = X'\beta + u$$

for some $\beta = \beta^*$ such that the orthogonality condition $E(Xu) = 0$ holds, where $u = Y - X'\beta^*$.

The orthogonality condition $E(Xu) = 0$ is fundamentally linked with the best least squares optimizer. If β is the best linear LS coefficient β^*, then the disturbance u must be orthogonal to X. On the other hand, if X is orthogonal to u, then β must be the least squares minimizer β^*. Essentially the orthogonality between X and ε is the FOC of the best linear LS problem! In other words, the orthoganality condition $E(Xu) = 0$ will always hold as long as the MSE ceriterion is used to obtain the best linear prediction. Note that when X contains an intercept, the orthogonality condition $E(Xu) = 0$ implies that $E(u) = 0$. In this case, we have $E(Xu) = \text{cov}(X, u)$. In other words, the orthogonality condition is equivalent to uncorrelatedness between X and u. This implies that u does not contain any component that can be predicted by a linear function

of X.

The condition $E(Xu) = 0$ is fundamentally different from $E(u|X) = 0$. The latter implies the former but not vice versa. In other words, $E(u|X) = 0$ implies $E(Xu) = 0$ but it is possible that $E(Xu) = 0$ and $E(u|X) \neq 0$. This can be illustrated by the following example.

Example 1: Suppose $u = (X^2 - 1) + \varepsilon$, where X and ε are independent N(0,1) random variables. Then

$$\begin{aligned} E(u|X) &= X^2 - 1 \neq 0, \text{ but} \\ E(Xu) &= E[X(X^2 - 1)] + E(X\varepsilon) \\ &= E(X^3) - E(X) + E(X)E(\varepsilon) \\ &= 0. \end{aligned}$$

2.4 Correct Model Specification for Condtional Mean

Question: What is the characterization for correct model specification in conditional mean?

Definition 2.5 [Correct Model Specification in Conditional Mean]: *The linear regression model*

$$Y = X'\beta + u, \ \beta \in \mathbb{R}^{k+1},$$

is said to be correctly specified for $E(Y|X)$ *if*

$$E(Y|X) = X'\beta^o \text{ for some } \beta^o \in \mathbb{R}^{k+1}.$$

On the other hand, if

$$E(Y|X) \neq X'\beta \text{ for all } \beta \in \mathbb{R}^{k+1},$$

then the linear model is said to be misspecified for $E(Y|X)$.

Remarks:

The class of linear regression models contains an infinite number of linear functions, each corresponding to a particular value of β. When the linear model is correctly specified, a linear function corresponding to some β^o will coincide with $g_o(X)$. The coefficient β^o is called the "true parameter", because now it has a meaningful economic interpretation as the expected marginal effect of X on Y :

$$\beta^o = \frac{d}{dX} E(Y|X).$$

For example, when Y is comsumption and X is income, β^o is the marginal propensity to consume (MPC).

When β^o is a vector, the component

$$\beta_j^o = \frac{\partial E(Y|X)}{\partial X_j}, \qquad 1 \leq j \leq k,$$

is the partial marginal effect of X_j on Y when holding all other explanatory variables in X fixed.

Question: What is the interpretation of the intercept coefficient β_0^o when a linear regression model is correctly specified for $g_o(X)$?

Answer: The intercept β_0^o corresponds to the variable $X_0 = 1$, which is always uncorrelated with any other random variables. It captures the "average effect" on Y from all possible factors rather than the explanatory variables in X_t. For example, consider the standard Capital Asset Pricing Model (CAPM)

$$E(Y|X) = \beta_0^o + \beta_1^o X_1,$$

where Y is the excess portfolio return (i.e., the difference between a portfolio return and a risk-free rate) and X_1 is the excess market portfolio return (i.e., the difference between the market portfolio return and a risk-free rate). Here, β_0^o represents the average pricing error. When CAPM holds, $\beta_0^o = 0$. Thus, if the data generating process has $\beta_0^o > 0$, CAPM underprices the portfolio. If $\beta_0^o < 0$, CAPM overprices the portfolio.

No economic theory ensures that the functional form of $E(Y|X)$ must be linear in X. Non-linear functional form in X is a generic possibility. Therefore, we must be very cautious about the economic interpretation of linear coefficients.

Theorem 2.7: *If the linear model*

$$Y = X'\beta + u$$

is correctly specified for $E(Y|X)$, *then*

(a) $Y = X'\beta^o + \varepsilon$ *for some* β^o *and* ε, *where* $E(\varepsilon|X) = 0$;

(b) $\beta^* = \beta^o$.

Proof: (a) If the linear model is correctly specified for $E(Y|X)$, then $E(Y|X) = X'\beta^o$ for some β^o.

On the other hand, we always have the regression identity $Y = E(Y|X)+\varepsilon$, where $E(\varepsilon|X) = 0$. Combining these two equations gives result (a) immediately.

(b) From part (a) we have

$$\begin{aligned} E(X\varepsilon) &= E[XE(\varepsilon|X)] \\ &= E(X \cdot 0) \\ &= 0. \end{aligned}$$

It follows that the orthogonality condition holds for $Y = X'\beta^o + \varepsilon$. Therefore, we have $\beta^* = \beta^o$ by the previous theorem (which one?).

Remarks:

Theorem (a) implies $E(Y|X) = X'\beta^o$ under correct model specification for $E(Y|X)$. This, together with Theorem (b), implies that when a linear regression model is correctly specified, the conditional mean $E(Y|X)$ will coincide with the best linear least squares predictor $g^*(X) = X'\beta^*$.

Under correct model specification, the best linear LS approximation coefficient β^* is equal to the true marginal effect parameter β^o. In other words, β^* can be interpreted as the true parameter β^o when (and only when) the linear regression model is correctly specified.

Question: What happens if the linear regression model

$$Y = X'\beta + u,$$

where $E(Xu) = 0$, is misspecified for $E(Y|X)$? In other words, what happens if $E(Xu) = 0$ but $E(u|X) \neq 0$?

Answer: The regression function

$$\begin{aligned} E(Y|X) &= X'\beta + E(u|X) \\ &\neq X'\beta. \end{aligned}$$

There exists some neglected structure in u that can be exploited to improve the prediction of Y using X. A misspecified model always yields suboptimal predictions. A correctly specified model yields optimal predictions in terms of MSE.

Example 1: Consider the following data generating process (DGP)

$$Y = 1 + \frac{1}{2}X_1 + \frac{1}{4}(X_1^2 - 1) + \varepsilon,$$

where X_1 and ε are mutually independent $N(0, 1)$.

(a) Find the conditional mean $E(Y|X_1)$ and $\frac{d}{dX_1}E(Y|X_1)$, the marginal effect of X_1 on Y.

Suppose now a linear regression model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + u \\ &= X'\beta + u, \end{aligned}$$

where $X = (X_0, X_1)' = (1, X_1)'$, is specified to approximate this DGP.

(b) Find the best LS approximation coefficient β^* and the best linear LS predictor $g^*_{\mathbb{A}}(X) = X'\beta^*$.

(c) Let $u = Y - X'\beta^*$. Show $E(Xu) = 0$.

(d) Check if the true marginal effect $\frac{d}{dX_1}E(Y|X_1)$ is equal to β^*_1, the model-implied marginal effect.

Solution: (a) Given that X_1 and u are independent, we obtain

$$\begin{aligned} E(Y|X_1) &= 1 + \frac{1}{2}X_1 + \frac{1}{4}(X_1^2 - 1), \\ \frac{d}{dX_1}E(Y|X_1) &= \frac{1}{2} + \frac{1}{2}X_1. \end{aligned}$$

(b) Using the best LS approximation formula, we have

$$\begin{aligned} \beta^* &= [E(XX')]^{-1} E(XY) \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}. \end{aligned}$$

Hence, we have

$$g^*(X) = X'\beta^* = 1 + \frac{1}{2}X_1.$$

(c) By definition and part (b), we have

$$\begin{aligned} u &= Y - X'\beta^* \\ &= Y - (\beta_0^* + \beta_1^* X_1) \\ &= \frac{1}{4}(X_1^2 - 1) + \varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned} E(Xu) &= E\begin{bmatrix} 1\cdot(\frac{1}{4}(X_1^2-1)+\varepsilon) \\ X_1\cdot(\frac{1}{4}(X_1^2-1)+\varepsilon) \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \end{aligned}$$

although

$$E(u|X_1) = \frac{1}{4}(X_1^2-1) \neq 0.$$

(d) No, because

$$\frac{d}{dX_1}E(Y|X_1) = \frac{1}{2}+\frac{1}{2}X_1 \neq \beta_1^* = \frac{1}{2}.$$

The marginal effect depends on the level of X_1, rather than only on a constant. Therefore, the condition $E(Xu) = 0$ is not sufficient for the validity of the economic interpretation for β_1^* as the marginal effect.

Any parametric regression model is subject to potential model misspecification. This can occur due to the use of a misspecified functional form, as well as the existence of omitted variables which are correlated with the existing regressors, among other things. In econometrics, there exists a modeling strategy which is free of model misspecification when a data set is sufficiently large. This modeling strategy is called a nonparametric approach, which does not assume any functional form for $E(Y|X)$ but let data speak for the true relationship. We now introduce the basic idea of a nonparametric approach.

Nonparametric modeling is a statistical method that can model the unknown function arbitrarily well without having to know the functional form of $E(Y|X)$. To illustrate the basic idea of nonparametric modeling, suppose $g_o(x)$ is a smooth function of x. Then we can expand $g_o(x)$ using a set of orthonormal "basis" functions $\{\psi_j(x)\}_{j=0}^{\infty}$:

$$g_o(x) = \sum_{j=0}^{\infty}\beta_j\psi_j(x) \text{ for } x \in \text{support}(X),$$

where the Fourier coefficient

$$\beta_j = \int_{-\infty}^{\infty} g_o(x)\psi_j(x)dx$$

and

$$\int_{-\infty}^{\infty}\psi_i(x)\psi_j(x)dx = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i=j, \\ 0 & \text{if } i\neq j. \end{cases}$$

The function δ_{ij} is called the Kronecker delta.

Example 2: Suppose $g_o(x) = x^2$ where $x \in [-\pi, \pi]$. Then

$$\begin{aligned} g_o(x) &= \frac{\pi^2}{3} - 4\left[\cos(x) - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \cdots\right] \\ &= \frac{\pi^2}{3} - 4\sum_{j=1}^{\infty}(-1)^{j-1}\frac{\cos(jx)}{j^2}. \end{aligned}$$

Example 3: Suppose

$$g_o(x) = \begin{cases} -1 & \text{if } -\pi < x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } 0 < x < \pi. \end{cases}$$

Then

$$\begin{aligned} g_o(x) &= \frac{4}{\pi}\left[\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \cdots\right] \\ &= \frac{4}{\pi}\sum_{j=0}^{\infty}\frac{\sin[(2j+1)x]}{(2j+1)}. \end{aligned}$$

Generally, suppose $g_o(x)$ is square-integrable. We have

$$\begin{aligned} \int_{-\pi}^{\pi} g_o^2(x)dx &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\beta_j\beta_k\int_{-\pi}^{\pi}\psi_j(x)\psi_k(x)dx \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\beta_j\beta_k\delta_{jk} \text{ by orthonormality of } \{\psi_j(\cdot)\} \\ &= \sum_{j=0}^{\infty}\beta_j^2 < \infty, \end{aligned}$$

Therefore, $\beta_j \to 0$ as $j \to \infty$. That is, the Fourier coefficient β_j will eventually vanish to zero as the order j goes to infinity. This motivates us to use the following truncated approximation:

$$g_p(x) = \sum_{j=0}^{p}\beta_j\psi_j(x),$$

where p is the order of bases. The approximation bias of $g_p(x)$ for $g_o(x)$ is

$$\begin{aligned} B_p(x) &= g_o(x) - g_p(x) \\ &= \sum_{j=p+1}^{\infty} \beta_j \psi_j(x) \\ &= \text{Bias}. \end{aligned}$$

The coefficients $\{\beta_j\}$ are unknown in practice, so we have to estimate them from an observed data $\{Y_t, X_t\}_{t=1}^n$, where n is the sample size. We consider a linear regression

$$Y_t = \sum_{j=0}^{p} \beta_j \psi_j(X_t) + u_t, \qquad t = 1, ..., n.$$

Obviously, we need to let $p = p(n) \to \infty$ as $n \to \infty$ to ensure that the bias $B_p(x)$ vanishes to zero as $n \to \infty$. However, we should not let p grow to infinity too fast, because otherwise there will be too much sampling variation in parameter estimators (due to too many unknown parameters). This requires $p/n \to 0$ as $n \to \infty$.

The nonparametric approach just described is called **nonparametric series regression** (see, e.g., Andrews 1991, Hong and White 1995). There are many nonparametric methods available in the literature. Another popular nonparametric method is called **kernel method**, which is based on the idea of the Taylor series expansion in a local region. See Härdle (1990), *Applied Nonparametric Regression*, for more discussion on kernel smoothing. The key feature of nonparametric modeling is that it does not specify a concrete functional form or model but rather estimates the unknown true function from data. As can be seen above, nonparametric series regression is easy to use and understand, because it is a natural extension of linear regression with the number of regressors increasing with the sample size n.

The nonparametric approach is flexible and powerful, but it generally requires a large data set for precise estimation because there is a large number of unknown parameters. Moreover, there is little economic interpretation for it (for example, it is difficult to give economic interpretation for the coefficients $\{\beta_j\}$). Nonparametric analysis is usually treated in a separate, more advanced econometric course (see more discussion in Chapter 10).

2.5 Conclusion

Most economic theories (e.g., rational expectations theory) have implications on and only on the conditional mean of the underlying economic variable given some suitable information set. The conditional mean $E(Y|X)$ is called the regression function of Y on X. In this chapter, we have shown that the regression function $E(Y|X)$ is the optimal solution to the MSE minimization

problem

$$\min_{g\in\mathbb{F}} E[Y - g(X)]^2,$$

where $\mathbb{F}$ is the space of measurable and square-integrable functions.

The regression function $E(Y|X)$ is generally unknown, because economic theory usually does not tell a concrete functional form. In practice, one usually uses a parametric model for $E(Y|X)$ that has a known functional form but with a finite number of unknown parameters. When we restrict $g(X)$ to $\mathbb{A} = \{g : \mathbb{R}^K \to \mathbb{R} \mid g(x) = x'\beta\}$, a class of affine functions, the optimal predictor that solves

$$\min_{g\in\mathbb{A}} E[Y - g(X)]^2 = \min_{\beta\in R^K} E(Y - X'\beta)^2$$

is $g^*(X) = X'\beta^*$, where

$$\beta^* = [E(XX')]^{-1}E(XY)$$

is called the best linear least squares approximation coefficient. The best linear least squares predictor $g_A^*(X) = X'\beta^*$ is always well-defined, no matter whether $E(Y|X)$ is linear in X.

Suppose we write

$$Y = X'\beta + u.$$

Then $\beta = \beta^*$ if and only if

$$E(Xu) = 0.$$

This orthogonality condition is actually the first order condition for the best linear least squares minimization problem. It does not guarantee correct specification of a linear regression model. A linear regression model is correctly specified for $E(Y|X)$ if $E(Y|X) = X'\beta^o$ for some β^o, which is equivalent to the condition that

$$E(u|X) = 0,$$

where $u = Y - X'\beta^o$. That is, correct model specification for $E(Y|X)$ holds if and only if the conditional mean of the linear regression model error is zero when evaluated at some parameter β^o. Note that $E(u|X) = 0$ is equivalent to the condition that $E[uh(X)] = 0$ for all measurable functions $h(\cdot)$. When $E(Y|X) = X'\beta^o$ for some β^o, we have $\beta^* = \beta^o$. That is, the best linear least squares approximation coefficient β^* will coincide with the true model parameter β^o and can be interpreted as the marginal effect of X on Y. The condition $E(u|X) = 0$ fundamentally differs from $E(Xu) = 0$. The former is crucial for validity of economic interpretation of the coefficient β^* as the true coefficient β^o. The orthogonality condition $E(Xu) = 0$ does not guarantee this interpretation. Correct model specification is important for economic interpretation of model coefficient and for optimal predictions.

An econometric model aims to provide a concise and reasonably accurate reflection of the data generating process. By disregarding less relevant aspects of the data, the model helps to

obtain a better understanding of the main aspects of the DGP. This implies that an econometric model will never provide a completely accurate description of the DGP. Therefore, the concept of a "true model" does not make much practical sense. It reflects an idealized situation that allows us to obtain mathematically exact results. The idea is that similar results hold approximately true if the model is a reasonably accurate approximation of the DGP.

The main purpose of this chapter is to provide a general idea of regression analysis and to shed some light on the nature and limitation of linear regression models, which have been popularly used in econometrics and will be the subject of study in Chapters 3 to 7.

EXERCISES

2.1. Put $\varepsilon = Y - E(Y|X)$. Show $\text{var}(Y|X) = \text{var}(\varepsilon|X)$.

2.2. Show $\text{var}(Y) = \text{var}[E(Y|X)] + \text{var}[Y - E(Y|X)]$.

2.3. Suppose (X, Y) follows a bivariate normal distribution with joint pdf

$$\begin{aligned} & f_{XY}(x,y) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}, \end{aligned}$$

where $-1 < \rho < 1, -\infty < \mu_1, \mu_2 < \infty, 0 < \sigma_1, \sigma_2 < \infty$. Find

(a) $E(Y|X)$.

(b) $\text{var}(Y|X)$. (Hint: Use the change of variable method for integration and the fact that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx = 1$.)

2.4. Suppose $Z \equiv (Y, X')'$ is a stochastic process such that the conditional mean $g_o(X) \equiv E(Y|X)$ exists, where X is a $(k+1) \times 1$ random vector. Suppose one uses a model (or a function) $g(X)$ to predict Y. A popular evaluation criterion for model $g(X)$ is the mean squared error $MSE(g) \equiv E[Y - g(X)]^2$.

(a) Show that the optimal predictor $g^*(X)$ for Y that minimizes $MSE(g)$ is the conditional mean $g_o(X)$; namely, $g^*(X) = g_o(X)$.

(b) Put $\varepsilon \equiv Y - g_o(X)$, which is called the true regression disturbance. Show that $E(\varepsilon|X) = 0$ and interpret this result.

2.5. The choices of model $g(X)$ in Exercise 2.4 are very general. Suppose that we now restrict our choice of $g(X)$ to a linear (or affine) models $\{g_{\mathbb{A}}(X) = X'\beta\}$, where β is a $(k+1) \times 1$ parameter. One can choose a linear function $g_{\mathbb{A}}(X)$ by choosing a value for parameter β. Different values of β give different linear functions $g_{\mathbb{A}}(X)$. The best linear predictor g_L^* that minimizes the mean squared error criterion is defined as $g_{\mathbb{A}}^*(X) \equiv X'\beta^*$, where

$$\beta^* \equiv \arg \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2$$

is called the optimal linear coefficient.

(a) Show that

$$\beta^* = [E(XX')]^{-1} E(XY).$$

(b) Define $u^* \equiv Y - X'\beta^*$. Show that $E(Xu^*) = 0$, where 0 is a $(k+1) \times 1$ zero vector.

(c) Suppose the conditional mean $g_o(X) = X'\beta^o$ for some given β^o. Then we say that the linear model $g_{\mathbb{A}}(X)$ is correctly specified for conditional mean $g_o(X)$, and β^o is the true parameter of the data generating process. Show that $\beta^* = \beta^o$ and $E(u^*|X) = 0$.

(d) Suppose the conditional mean $g_o(X) \neq X'\beta$ for any value of β. Then we say that the linear model $g_{\mathbb{A}}(X)$ is misspecified for conditional mean $g_o(X)$. Check if $E(u^*|X) = 0$ and discuss its implication.

2.6. Suppose $Y = \beta_0^* + \beta_1^* X_1 + u$, where Y and X_1 are scalars, and $\beta^* = (\beta_0^*, \beta_1^*)'$ is the best linear least squares approximation coefficient.

(a) Show that $\beta_1^* = \text{cov}(Y, X_1)/\sigma_{X_1}^2$ and $\beta_0^* = E(Y) - \beta_1^* E(X)$, and the mean squared error

$$E[Y - (\beta_0^* + \beta_1^* X_1)]^2 = \sigma_Y^2(1 - \rho_{X_1 Y}^2),$$

where $\sigma_Y^2 = \text{var}(Y)$ and $\rho_{X_1 Y}$ is the correlation coefficient between Y and X_1.

(b) Suppose in addition Y and X_1 follow a bivariate normal distribution. Show $E(Y|X_1) = \beta_0^* + \beta_1^* X_1$ and $\text{var}(Y|X_1) = \sigma_Y^2(1 - \rho_{X_1 Y}^2)$. That is, the conditional mean of Y given X_1 coincides with the best linear least squares predictor and the conditional variance of Y given X_1 is equal to the mean squared error of the best linear least squares predictor.

2.7. Suppose

$$Y = \beta_0 + \beta_1 X_1 + |X_1|\varepsilon,$$

where $E(X_1) = 0$, $\text{var}(X_1) = \sigma_{X_1}^2 > 0$, $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma_\varepsilon^2 > 0$, and ε and X_1 are independent. Both β_0 and β_1 are scalar constants.

(a) Find $E(Y|X_1)$.

(b) Find $\text{var}(Y|X_1)$.

(c) Show that $\beta_1 = 0$ if and only if $\text{cov}(X_1, Y) = 0$.

2.8. Suppose an aggregate consumption function is given by

$$Y = 1 + 0.5X_1 + \frac{1}{4}(X_1^2 - 1) + \varepsilon,$$

where $X_1 \sim N(0, 1), \varepsilon \sim N(0, 1)$, and X_1 is independent of ε.

(a) Find the conditional mean $g_o(X) \equiv E(Y|X)$, where $X \equiv (1, X_1)'$.

(b) Find the marginal propensity to consume (MPC) $\frac{d}{dX_1} g_o(X)$.

(c) Suppose we use a linear model

$$Y = X'\beta + u = \beta_0 + \beta_1 X_1 + u$$

where $\beta \equiv (\beta_0, \beta_1)'$ to predict Y. Find the optimal linear coefficient β^* and the optimal linear predictor $g_{\mathbb{A}}^*(X) \equiv X'\beta^*$.

(d) Compute the partial derivative of the linear model $\frac{d}{dX_1}g^*_{\mathbb{A}}(X)$, and compare it with the MPC in part (b). Discuss the results you obtain.

2.9. Put $g_o(X) = E(Y|X)$,where $X = (1, X_1)'$. Then we have

$$Y = g_o(X) + \varepsilon,$$

where $E(\varepsilon|X) = 0$.

Consider a first order Taylor series expansion of $g_o(X)$ around $\mu_1 = E(X_1)$:

$$\begin{aligned} g_o(X) &\approx g_o(\mu_1) + g_o'(\mu_1)(X_1 - \mu_1) \\ &= [g_o(\mu_1) - \mu g_o'(\mu_1)] + g_o'(\mu_1)X_1. \end{aligned}$$

Suppose $\beta^* = (\beta_0^*, \beta_1^*)'$ is the best linear least squares approximation coefficient. Is it true that $\beta_1^* = g_o'(\mu_1)$? Provide your reasoning.

2.10. Suppose a data generating process is given by

$$Y = 0.8X_1X_2 + \varepsilon,$$

where $X_1 \sim N(0,1), X_2 \sim N(0,1), \varepsilon \sim N(0,1)$, and X_1, X_2 and ε are mutually independent. Put $X = (1, X_1, X_2)'$.

(a) Is Y predictable in mean using information X?

(b) Suppose we use a linear model

$$\begin{aligned} g_{\mathbb{A}}(X) &= X'\beta + u \\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + u \end{aligned}$$

to predict Y. Does this linear model has any predicting power? Explain.

2.11. Show that $E(u|X) = 0$ if and only if $E[h(X)u] = 0$ for any measurable functions $h(\cdot)$.

2.13. Suppose $E(u|X)$ exists, X is a bounded random variable, and $h(X)$ is an arbitrary measurable function. Put $g(X) = E(\varepsilon|X)$ and assume that $E[g^2(X)] < \infty$.

(a) Show that if $g(X) = 0$, then $E[\varepsilon h(X)] = 0$.

(b) Show that if $E[\varepsilon h(X)] = 0$, then $E(\varepsilon|X) = 0$. [Hint: Consider $h(X) = e^{tX}$ for t in a small neighborhood containing 0. Given that X is bounded, we can expand

$$g(X) = \sum_{j=0}^{\infty} \beta_j X^j$$

where $\beta_j = \int_{-\infty}^{\infty} g(x)x^j f_X(x)dx$ is the Fourier coefficient. Then

$$\begin{aligned} E(\varepsilon e^{tX}) &= E\left[E(\varepsilon|X)e^{tX}\right] \\ &= E\left[g(X)e^{tX}\right] \\ &= \sum_{j=0}^{\infty} \frac{t^j}{j!} E\left[g(X)X^j\right] \\ &= \sum_{j=0}^{\infty} \frac{t^j}{j!}\beta_j \end{aligned}$$

for all t in a small neighborhood containing 0.]

2.14. Consider the following nonlinear least squares problem

$$\min_{\beta \in \mathbf{R}^{k+1}} E\left[Y - g(X,\beta)\right]^2,$$

where $g(X,\beta)$ is possibly a nonlinear function of β. [An example is a logistic regression model where $g(X,\beta) = \frac{1}{1+\exp(-X'\beta)}$.] Suppose $E\left[\frac{\partial}{\partial\beta}g(X,\beta)\frac{\partial}{\partial\beta'}g(X,\beta)\right]$ is a $(k+1)\times(k+1)$ bounded and nonsingular matrix for all $\beta \in \mathbf{R}^{k+1}$, where $\frac{\partial}{\partial\beta'}g(X,\beta)$ is the transpose of the $(k+1)\times 1$ column vector $\frac{\partial}{\partial\beta}g(X,\beta)$.

(a) Derive the first order condition for the best nonlinear least squares approximation coefficient β^* (say).

(b) Put $Y = g(X,\beta) + u$. Show that $\beta = \beta^*$ if and only if $E[u\frac{\partial}{\partial\beta}g(X,\beta^*)] = 0$. Do we have $E(Xu) = 0$ when $g(X,\beta)$ is nonlinear in β?

(c) The nonlinear regression model $g(X,\beta)$ is said to be correctly specified for $E(Y|X)$ if there exists some unknown β^o such that $E(Y|X) = g(X,\beta^o)$ almost surely. Here, β^o can be interpreted as a true model parameter. Show that $\beta^* = \beta^o$ if and only if the model $g(X,\beta)$ is correctly specified for $E(Y|X)$.

(d) Do we have $E(u|X) = 0$, where $u = Y - g(X,\beta^o)$, for some β^o, when the model $g(X,\beta)$ is correctly specified?

(e) If $E(u|X) = 0$, where $u = Y - g(X,\beta^o)$ for some β^o, is $g(X,\beta)$ correctly specified for $E(Y|X)$?

2.15. Comment on the following statement: "All econometric models are approximations of the economic system of interest and are therefore misspecified. Therefore, there is no need to check correct model specification in practice."

CHAPTER 3 CLASSICAL LINEAR REGRESSION MODELS

Abstract: In this chapter, we will introduce the classical linear regression theory, including the classical model assumptions, the statistical properties of the OLS estimator, the t-test and the F-test, as well as the GLS estimator and related statistical procedures. This chapter will serve as a starting point from which we will develop the modern econometric theory.

Key words: Classical linear regression, Conditional heteroskedasticity, Conditional homoskedasticity, F-test, GLS, Hypothesis testing, Model selection criterion, OLS, R^2, t-test

3.1 Framework and Assumptions

Suppose we have an observed random sample $\{Z_t\}_{t=1}^n$ of size n, where $Z_t = (Y_t, X_t')'$, Y_t is a scalar, $X_t = (1, X_{1t}, X_{2t}, ..., X_{kt})'$ is a $(k+1) \times 1$ vector, t is an index (either cross-sectional unit or time period) for observations, and n is the sample size. We are interested in the conditional mean $E(Y_t|X_t)$ using an observed realization (i.e., a data set) of the random sample $\{Y_t, X_t'\}', t = 1, ..., n$.

Notations:

Throughtout this book, we set $K \equiv k + 1$, the number of regressors which contains k economic variables and an intercept. The index t may denote an individual unit (e.g., a firm, a household, a country) for cross-sectional data, or denote a time period (e.g., day, weak, month, year) in a time series context.

We first list and discuss the assumptions of the classical linear regression theory.

Assumption 3.1 [Linearity]:

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

where β^o is a $K \times 1$ unknown parameter vector, and ε_t is an unobservable disturbance.

Remarks:

In Assumption 3.1, Y_t is the dependent variable (or regressand), X_t is the vector of regressors (or independent variables, or explanatory variables), and β^o is the regression coefficient vector. When the linear model is correctly specified for the conditional mean $E(Y_t|X_t)$, i.e., when $E(\varepsilon_t|X_t) = 0$, the parameter $\beta^o = \frac{\partial}{\partial X_t}E(Y_t|X_t)$ can be interpreted as the marginal effect of X_t on Y_t.

The key notion of *linearity* in the classical linear regression model is that the regression model is linear in β^o rather than in X_t.In other words, linear regression models cover some models for Y_t which have a nonlinear relationship with X_t.

Question: Does Assumption 3.1 imply a causal relationship from X_t to Y_t?

Not necessarily. As Kendall and Stuart (1961, Vol.2, Ch. 26, p.279) point out, "a statistical relationship, however strong and however suggestive, can never establish causal connection. Our ideas of causation must come from outside statistics ultimately, from some theory or other." Assumption 3.1 only implies a predictive relationship: Given X_t, can we predict Y_t linearly?

Denote

$$\begin{aligned} Y &= (Y_1, ..., Y_n)', \qquad n \times 1, \\ \varepsilon &= (\varepsilon_1, ..., \varepsilon_n)', \qquad n \times 1, \\ \mathbf{X} &= (X_1, ..., X_n)', \qquad n \times K. \end{aligned}$$

where the t-th row of $\mathbf{X}$ is $X_t' = (1, X_{1t}, ..., X_{kt})$. With these matrix notations, we have a compact expression for Assumption 3.1:

$$\begin{aligned} Y &= \mathbf{X}\beta^o + \varepsilon, \\ n \times 1 &= (n \times K)(K \times 1) + n \times 1. \end{aligned}$$

The second assumption is a strict exogeneity condition.

Assumption 3.2 [Strict Exogeneity]:

$$E(\varepsilon_t|\mathbf{X}) = E(\varepsilon_t|X_1, ..., X_t, ..., X_n) = 0, \qquad t = 1, ..., n.$$

Remarks:

Among other things, Assumption 3.2 implies correct model specification for $E(Y_t|X_t)$. This is because Assumption 3.2 implies $E(\varepsilon_t|X_t) = 0$ by conditional expectation. It also implies $E(\varepsilon_t) = 0$ by the law of iterated expectations.

Under Assumption 3.2, we have $E(X_s\varepsilon_t) = 0$ for any (t, s), where $t, s \in \{1, ..., n\}$. This follows because

$$\begin{aligned} E(X_s\varepsilon_t) &= E[E(X_s\varepsilon_t|\mathbf{X})] \\ &= E[X_s E(\varepsilon_t|\mathbf{X})] \\ &= E(X_s \cdot 0) \\ &= 0. \end{aligned}$$

Note that (i) and (ii) imply $\text{cov}(X_s, \varepsilon_t) = 0$ for all $t, s \in \{1, ..., n\}$.

Because $\mathbf{X}$ contains regressors $\{X_s\}$ for both $s \leq t$ and $s > t$, Assumption 3.2 essentially requires that the error ε_t do not depend on the past and future values of regressors if t is a time index. This rules out dynamic time series models for which ε_t may be correlated with the future values of regressors (because the future values of regressors depend on the current shocks), as is illustrated in the following example.

Example 1: Consider a so-called AutoRegressive AR(1) model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, \qquad t = 1, ..., n, \\ &= X_t'\beta + \varepsilon_t, \\ \{\varepsilon_t\} &\sim \text{i.i.d.}(0, \sigma^2), \end{aligned}$$

where $X_t = (1, Y_{t-1})'$. This is a dynamic regression model because the term $\beta_1 Y_{t-1}$ represents the "memory" or "feedback" of the past into the present value of the process, which induces a correlation between Y_t and the past. The term autoregression refers to the regression of Y_t on its own past values. The parameter β_1 determines the amount of feedback, with a large absolute value of β_1 resulting in more feedback. The disturbance ε_t can be viewed as representing the effect of "new information" that is revealed at time t. Information that is truly new cannot be anticipated so that the effects of today's new information should be unrelated to the effects of yesterday's news in the sense that $E(\varepsilon_t|X_t) = 0$. Here, we make a stronger assumption that we can model the effect of new information as an i.i.d.$(0, \sigma^2)$ sequence.

Obviously, $E(X_t\varepsilon_t) = E(X_t)E(\varepsilon_t) = 0$ but $E(X_{t+1}\varepsilon_t) \neq 0$. Thus, we have $E(\varepsilon_t|\mathbf{X}) \neq 0$, and so Assumption 3.2 does not hold. Here, the lagged dependent variable Y_{t-1} in the regressor vector X_t is called a predetermined variable, since it is orthogonal to ε_t but depends on the past history of $\{\varepsilon_t\}$.

In Chapter 5 later, we will consider linear regression models with dependent observations, which will include this example as a special case. In fact, the main reason of imposing Assumption 3.2 is to obtain a finite sample distribution theory. For a large sample theory (i.e., an asymptotic theory), the strict exogeneity condition will not be needed.

In econometrics, there are some alternative definitions of strict exogeneity. For example, one definition assumes that ε_t and $\mathbf{X}$ are independent. Another example is that $\mathbf{X}$ is nonstochastic. This rules out conditional heteroskedasticity (i.e., $\text{var}(\varepsilon_t|\mathbf{X})$ depends on $\mathbf{X}$). In Assumption 3.2, we still allow for conditional heteroskedasticity, because we

do not assume that ε_t and $\mathbf{X}$ are independent. We only assume that the conditional mean $E(\varepsilon_t|\mathbf{X})$ does not depend on $\mathbf{X}$.

Question: What happens to Assumption 3.2 if $\mathbf{X}$ is nonstochastic?

If $\mathbf{X}$ is nonstochastic, Assumption 3.2 becomes

$$E(\varepsilon_t|\mathbf{X}) = E(\varepsilon_t) = 0.$$

An example of nonstochastic $\mathbf{X}$ is $X_t = (1, t, ..., t^k)'$. This corresponds to a time-trend regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \sum_{j=0}^{k} \beta_j^o t^j + \varepsilon_t. \end{aligned}$$

Question: What happens to Assumption 3.2 if $Z_t = (Y_t, X_t')'$ is an independent random sample (i.e., Z_t and Z_s are independent whenever $t \neq s$, although Y_t and X_t may not be independent)?

When $\{Z_t\}$ is i.i.d., Assumption 3.2 becomes

$$\begin{aligned} E(\varepsilon_t|\mathbf{X}) &= E(\varepsilon_t|X_1, X_2, ...X_t, ..., X_n) \\ &= E(\varepsilon_t|X_t) \\ &= 0. \end{aligned}$$

In other words, when $\{Z_t\}$ is i.i.d., $E(\varepsilon_t|\mathbf{X}) = 0$ is equivalent to $E(\varepsilon_t|X_t) = 0$.

Assumption 3.3 [Nonsingularity]: *(a) The minimum eigenvalue of the $K \times K$ square matrix $X'X = \sum_{t=1}^n X_t X_t'$ is nonsingular, and (b)*

$$\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty \text{ as } n \to \infty$$

with probability one.

Remarks:

Assumption 3.3(a) rules out multicollinearity among the $(k+1)$ regressors in X_t. We say that there exists multicollinearity (sometimes called the exact or perfect multicolinearity in the literature) among the X_t if for all $t \in \{1, ..., n\}$, the variable X_{jt} for some $j \in \{0, 1, ..., k\}$ is a linear combination of the other $K-1$ column variables $\{X_{it}, i \neq j\}$.

In this case, the matrix $\mathbf{X}'\mathbf{X}$ is singular, and as a consequence, the true model parameter β^o in Assumption 3.1 is not identifiable.

The nonsingularity of $\mathbf{X}'\mathbf{X}$ implies that $\mathbf{X}$ must be of full rank of $K = k + 1$. Thus, we need $K \leq n$. That is, the number of regressors cannot be larger than the sample size. This is a necessary condition for identification of parameter β^o.

The eigenvalue λ of a square matrix A is characterized by the system of linear equations:

$$\det(A - \lambda I) = 0,$$

where $\det(\cdot)$ denotes the determinant of a square matrix, and I is an identity matrix with the same dimension as A.

It is well-known that the eigenvalue λ can be used to summarize information contained in a matrix (recall the popular principal component analysis). Assumption 3.3 implies that new information must be available as the sample size $n \to \infty$ (i.e., X_t should not only have same repeated values as t increases).

Intuitively, if there are no variations in the values of the X_t, it will be difficult to determine the relationship between Y_t and X_t (indeed, the purpose of classical linear regression is to investigate how a change in $\mathbf{X}$ causes a change in Y). In certain sense, one may call $\mathbf{X}'\mathbf{X}$ the "information matrix" of the random sample $\mathbf{X}$ because it is a measure of the information contained in $\mathbf{X}$. The magnitude of $\mathbf{X}'\mathbf{X}$ will affect the preciseness of parameter estimation for β^o. Indeed, as will be shown below, the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ ensures that variance of the OLS estimator will vanish to zero as $n \to \infty$. This rule out a possibility called near-multicolinearity that there exists an approximate linear relationship among the sample values of explanatory variables in X_t such that although $\mathbf{X}'\mathbf{X}$ is nonsingular, its minimum eigenvalue $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n. When $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with n, the OLS estimator is well-defined and has a well-behaved finite sample distribution, but its variance never vanishes to zero as $n \to \infty$. In other words, in the near multicolinearity case where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with n, the OLS estimator will never converge to the true parameter value β^o, although it will still have a well-defined finite sample distribution.
Question: Why can the eigenvalue λ be used as a measure of the information contained in $\mathbf{X}'\mathbf{X}$?

Assumption 3.4 [Spherical error variance]:

(a) [conditional homoskedasticity]:

$$E(\varepsilon_t^2|\mathbf{X}) = \sigma^2 > 0, \qquad t = 1, ..., n;$$

(b) [conditional non-autocorrelation]:

$$E(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0, \qquad t \neq s, t, s \in \{1, ..., n\}.$$

Remarks:

We can write Assumption 3.4 as

$$E(\varepsilon_t\varepsilon_s|\mathbf{X}) = \sigma^2\delta_{ts},$$

where $\delta_{ts} = 1$ if $t = s$ and $\delta_{ts} = 0$ otherwise. In mathematics, δ_{ts} is called the Kronecker delta function. Under this assumption, we have

$$\begin{aligned} var(\varepsilon_t|\mathbf{X}) &= E(\varepsilon_t^2|\mathbf{X}) - [E(\varepsilon_t|\mathbf{X})]^2 \\ &= E(\varepsilon_t^2|\mathbf{X}) \\ &= \sigma^2 \end{aligned}$$

and

$$\begin{aligned} cov(\varepsilon_t, \varepsilon_s|\mathbf{X}) &= E(\varepsilon_t\varepsilon_s|\mathbf{X}) \\ &= 0 \text{ for all } t \neq s. \end{aligned}$$

By the law of iterated expectations, Assumption 3.4(b) implies that $\text{var}(\varepsilon_t) = \sigma^2$ for all $t = 1, ..., n$, the so-called unconditional homoskedasticity. Similarly, Assumption 3.4(a) implies $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$. Thus, there exists no serial correlation between ε_t and its lagged values when t is an index for time, or there exists no spatial correlation between the disburbances associated with different cross-sectional units when t is an index for the cross-sectional unit (e.g., consumer, firm, household, etc).

Assumption 3.4 does not imply that ε_t and $\mathbf{X}$ are independent. It allows the possibility that the conditional higher order moments (e.g., skewness and kurtosis) of ε_t depend on $\mathbf{X}$.

We can write Assumptions 3.2 and 3.4 compactly as follows:

$$E(\varepsilon|\mathbf{X}) = 0 \text{ and } E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I,$$

where $I \equiv I_n$ is a $n \times n$ identity matrix.

3.2 OLS Estimation

Question: How to estimate β^o using an observed data set generated from the random sample $\{Z_t\}_{t=1}^n$, where $Z_t = (Y_t, X_t')'$?

Definition 3.1 [OLS estimator]: *Suppose Assumptions 3.1 and 3.3(a) hold. Define the sum of squared residuals (SSR) of the linear regression model* $Y_t = X_t'\beta + u_t$ *as*

$$\begin{aligned} SSR(\beta) &\equiv (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) \\ &= \sum_{t=1}^{n}(Y_t - X_t'\beta)^2. \end{aligned}$$

Then the Ordinary Least Squares (OLS) estimator $\hat{\beta}$ *is the solution to*

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^K} SSR(\beta).$$

Note that $SSR(\beta)$ is the sum of squred model errors $\{u_t = Y_t - X_t'\beta\}$, with equal weighting for each t.

Theorem 3.1 [Existence of OLS]: *Under Assumptions 3.1 and 3.3, the OLS estimator* $\hat{\beta}$ *exists and*

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= \left(\frac{1}{n}\sum_{t=1}^{n} X_t X_t'\right)^{-1} \frac{1}{n}\sum_{t=1}^{n} X_t Y_t. \end{aligned}$$

The last expression will be useful for our asymptotic analysis in subsequent chapters.

Proof: Using the formula that for an $K \times 1$ vector A and and $K \times 1$ vector β, the derivative

$$\frac{\partial(A'\beta)}{\partial\beta} = A,$$

we have

$$\begin{aligned} \frac{dSSR(\beta)}{d\beta} &= \frac{d}{d\beta}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2 \\ &= \sum_{t=1}^{n}\frac{\partial}{\partial\beta}(Y_t - X_t'\beta)^2 \\ &= \sum_{t=1}^{n} 2(Y_t - X_t'\beta)\frac{\partial}{\partial\beta}(Y_t - X_t'\beta) \\ &= -2\sum_{t=1}^{n} X_t(Y_t - X_t'\beta) \\ &= -2\mathbf{X}'(Y - \mathbf{X}\beta). \end{aligned}$$

The OLS must satisfy the FOC:

$$\begin{aligned} -2\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) &= 0, \\ \mathbf{X}'(Y - \mathbf{X}\hat{\beta}) &= 0, \\ \mathbf{X}'Y - (\mathbf{X}'\mathbf{X})\hat{\beta} &= 0. \end{aligned}$$

It follows that

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'Y.$$

By Assumption 3.3, $\mathbf{X}'\mathbf{X}$ is nonsingular. Thus,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y.$$

Checking the SOC, we have the $K \times K$ Hessian matrix

$$\begin{aligned} \frac{\partial^2 SSR(\beta)}{\partial\beta\partial\beta'} &= -2\sum_{t=1}^{n} \frac{\partial}{\partial\beta'}\left[(Y_t - X_t'\beta)X_t\right] \\ &= 2\mathbf{X}'\mathbf{X} \\ &\sim \text{positive definite} \end{aligned}$$

given $\lambda_{\min}(\mathbf{X}'\mathbf{X}) > 0$. Thus, $\hat{\beta}$ is a global minimizer. Note that for the existence of $\hat{\beta}$, we only need that $\mathbf{X}'\mathbf{X}$ is nonsingular, which is implied by the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ but it does not require that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$. This completes the proof.

Remarks:

Suppose $Z_t = \{Y_t, X_t'\}', t = 1, ..., n$, is an independent and identically distributed (i.i.d.) random sample of size n. Consider the sum of squared residual scaled by n^{-1} :

$$\frac{SSR(\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - X_t'\beta)^2$$

and its minimizer

$$\hat{\beta} = \left(\frac{1}{n}\sum_{t=1}^{n} X_t X_t'\right)^{-1} \frac{1}{n}\sum_{t=1}^{n} X_t Y_t.$$

These are the sample analogs of the population MSE criterion

$$MSE(\beta) = E(Y_t - X_t'\beta)^2$$

and its minimizer

$$\beta^* \equiv \left[E(X_t X_t')\right]^{-1} E(X_t Y_t).$$

That is, $SSR(\beta)$, after scaled by n^{-1}, is the sample analogue of $MSE(\beta)$, and the OLS $\hat{\beta}$ is the sample analogue of the best LS approximation coefficient β^*.

Put $\hat{Y}_t \equiv X_t'\hat{\beta}$. This is called the fitted value (or predicted value) for observation Y_t, and $e_t \equiv Y_t - \hat{Y}_t$ is the estimated residual (or prediction error) for observation Y_t. Note that

$$\begin{aligned} e_t &= Y_t - \hat{Y}_t \\ &= (X_t'\beta^o + \varepsilon_t) - X_t'\hat{\beta} \\ &= \varepsilon_t - X_t'(\hat{\beta} - \beta^o), \end{aligned}$$

where ε_t is the unavoidable true disturbance ε_t, and $X_t'(\hat{\beta} - \beta^o)$ is an estimation error, which is smaller when a larger data set is available (so $\hat{\beta}$ becomes closer to β^o).

The FOC implies that the estimated residual $e = Y - \mathbf{X}\hat{\beta}$ is orthogonal to regressors $\mathbf{X}$ in the sense that

$$\mathbf{X}'e = \sum_{t=1}^{n} X_t e_t = 0.$$

This is the consequence of the very nature of OLS, as implied by the FOC of $\min_{\beta \in R^K} SSR(\beta)$. It always holds no matter whether $E(\varepsilon_t|\mathbf{X}) = 0$ (recall that we do not impose Assumption 3.2 in the Theorem above). Note that if X_t contains the intercept, then $\mathbf{X}'e = 0$ implies $\Sigma_{t=1}^n e_t = 0$.

Some useful identities

To investigate the statistical properties of $\hat{\beta}$, we first state some useful lemmas.

Lemma 3.2: *Under Assumptions 3.1 and 3.3(a), we have:*

(i)

$$\mathbf{X}'e = 0;$$

(ii)

$$\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon;$$

(iii) Define a $n \times n$ projection matrix

$$P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and

$$M = I_n - P.$$

Then both P and M are symmetric (i.e., $P = P'$ and $M = M'$) and idempotent (i.e., $P^2 = P, M^2 = M$), with

$$\begin{aligned} P\mathbf{X} &= \mathbf{X}, \\ M\mathbf{X} &= 0. \end{aligned}$$

(iv)

$$SSR(\hat{\beta}) = e'e = Y'MY = \varepsilon' M \varepsilon.$$

Proof: (i) The result follows immediately from the FOC of the OLS estimator.

(ii) Because $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ and $Y = \mathbf{X}\beta^o + \varepsilon$, we have

$$\begin{aligned} \hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta^o + \varepsilon) - \beta^o \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon. \end{aligned}$$

(iii) P is idempotent because

$$\begin{aligned} P^2 &= PP \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= P. \end{aligned}$$

Similarly we can show $M^2 = M$.

(iv) By the definition of M, we have

$$\begin{aligned} e &= Y - \mathbf{X}\hat{\beta} \\ &= Y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= [I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']Y \\ &= MY \\ &= M(\mathbf{X}\beta^o + \varepsilon) \\ &= M\mathbf{X}\beta^o + M\varepsilon \\ &= M\varepsilon \end{aligned}$$

given $M\mathbf{X} = 0$. It follows that

$$\begin{aligned} SSR(\hat{\beta}) &= e'e \\ &= (M\varepsilon)'(M\varepsilon) \\ &= \varepsilon' M^2 \varepsilon \\ &= \varepsilon' M \varepsilon, \end{aligned}$$

where the last equality follows from $M^2 = M$.

3.3 Goodness of Fit and Model Selection Criteria

Question: How well does the linear regression model fit the data? That is, how well does the linear regression model explain the variation of the observed data of $\{Y_t\}_{t=1}^n$?

We need some criteria or some measures to characterize goodness of fit.

We first introduce two measures for goodness of fit. The first measure is called the uncentered squared multi-correlation coefficient R^2

Definition 3.2 [Uncentered R^2] : *The uncentered squared multi-correlation coefficient is defined as*

$$R_{uc}^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{e'e}{Y'Y},$$

where the second equality follows from the first order condition of the OLS estimation.

Remarks:

The measure R_{uc}^2 has a nice interpretation: The proportion of the uncentered sample quadratic variation in the dependent variables $\{Y_t\}$ that can be attributed to the uncentered sample quadratic variation of the predicted values $\{\hat{Y}_t\}$. Note that we always have $0 \leq R_{uc}^2 \leq 1$.

Next, we define a closely related measure called Centered R^2.

Definition 3.3 [Centered R^2 : Coefficient of Determination]: *The coefficient of determination*

$$R^2 \equiv 1 - \frac{\sum_{t=1}^n e_t^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2},$$

where $\bar{Y} = n^{-1} \sum_{t=1}^n Y_t$ is the sample mean.

Remarks:

When X_t contains the intercept, we have the following orthogonal decomposition:

$$\begin{aligned}\sum_{t=1}^{n}(Y_t-\bar{Y})^2 &= \sum_{t=1}^{n}(\hat{Y}_t-\bar{Y}+Y_t-\hat{Y}_t)^2 \\ &= \sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2 \\ &\quad +2\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})e_t \\ &= \sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2,\end{aligned}$$

where the cross-product term

$$\begin{aligned}\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})e_t &= \sum_{t=1}^{n}\hat{Y}_te_t-\bar{Y}\sum_{t=1}^{n}e_t \\ &= \hat{\beta}'\sum_{t=1}^{n}X_te_t-\bar{Y}\sum_{t=1}^{n}e_t \\ &= \hat{\beta}'(\mathbf{X}'e)-\bar{Y}\sum_{t=1}^{n}e_t \\ &= \hat{\beta}'\cdot 0-\bar{Y}\cdot 0 \\ &= 0,\end{aligned}$$

where we have made use of the facts that $\mathbf{X}'e=0$ and $\sum_{t=1}^{n}e_t=0$ from the FOC of the OLS estimation and the fact that X_t contains the intercept (i.e., $X_{0t}=1$). It follows that

$$\begin{aligned}R^2 &\equiv 1-\frac{e'e}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \\ &= \frac{\sum_{t=1}^{n}(Y_t-\bar{Y})^2-\sum_{t=1}^{n}e_t^2}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \\ &= \frac{\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}.\end{aligned}$$

and consequently we have

$$0\le R^2\le 1.$$

Question: Can R^2 be negative?

Yes, it is possible! If X_t does not contain the intercept, then the orthogonal decomposition identity

$$\sum_{t=1}^{n}(Y_t-\bar{Y})^2=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2$$

no longer holds. As a consequence, R^2 may be negative when there is no intercept! This is because the cross-product term

$$2\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})e_t$$

may be negative.

When X_t contains an intercept, the centered R^2 has a similar interpretation to the uncentered R^2_{uc}. That is, R^2 measures the proportion of the sample variance of $\{Y_t\}_{t=1}^n$ that can be explained by the linear predictor of X_t.

Example 1 [Capital Asset Pricing Model (CAPM)]: The classical CAPM is characterized by the equation

$$r_{pt} - r_{ft} = \alpha_p + \beta_p(r_{mt} - r_{ft}) + \varepsilon_{pt}, \qquad t = 1, ..., n,$$

where r_{pt} is the return on portfolio (or asset) p, r_{ft} is the return on a risk-free asset, and r_{mt} is the return on the market portfolio. Here, $r_{pt} - r_{ft}$ is the risk premium of portfolio p, $r_{mt} - r_{ft}$ is the risk premium of the market portfolio, which is the only systematic market risk factor, and ε_{pt} is the individual-specific risk which can be eliminated by diversification if the ε_{pt} are uncorrelated across different assets. In this model, R^2 has an interesting economic interpretation: it is the proportion of the risk of portfolio p (as measured by the sample variance of its risk premium $r_{pt} - r_{ft}$) that is attributed to the market risk factor $(r_{mt} - r_{ft})$. In contrast, $1 - R^2$ is the proportion of the risk of portfolio p that is contributed by individual-specific risk factor ε_{pt}.

For any given random sample $\{Y_t, X_t'\}', t = 1, ..., n$, R^2 is nondecreasing in the number of explanatory variables X_t. In other words, the more explanatory variables are added in the linear regression, the higher R^2 is. This is always true no matter whether X_t has any true explanatory power for Y_t.

Theorem 3.3: *Suppose* $\{Y_t, X_{1t}, ..., X_{(k+q)t}\}', t = 1, ..., n$, *is a random sample, and Assumptions 3.1 and 3.3(a) hold. Let* R_1^2 *be the centered* R^2 *from the linear regression*

$$Y_t = X_t'\beta + u_t,$$

where $X_t = (1, X_{1t}, ..., X_{kt})'$, *and* β *is a* $K \times 1$ *parameter vector; also,* R_2^2 *is the centered* R^2 *from the extended linear regression*

$$Y_t = \tilde{X}_t'\gamma + v_t,$$

where $\tilde{X}_t = (1, X_{1t}, ..., X_{kt}, X_{(k+1)t}, ..., X_{(k+q)t})'$, *and* γ *is a* $(K+q)\times 1$ *parameter vector. Then* $R_2^2 \geq R_1^2$.

Proof: By definition, we have

$$\begin{aligned} R_1^2 &= 1 - \frac{e'e}{\sum_{t=1}^n (Y_t - \bar{Y})^2}, \\ R_2^2 &= 1 - \frac{\tilde{e}'\tilde{e}}{\sum_{t=1}^n (Y_t - \bar{Y})^2}, \end{aligned}$$

where e is the estimated residual vector from the regression of Y on $\mathbf{X}$, and $\tilde{e}$ is the estimated residual vector from the regression of Y on $\tilde{\mathbf{X}}$. It suffices to show $\tilde{e}'\tilde{e} \leq e'e$. Because the OLS estimator $\hat{\gamma} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Y$ minimizes $SSR(\gamma)$ for the extended model, we have

$$\tilde{e}'\tilde{e} = \sum_{t=1}^n (Y_t - \tilde{X}_t'\hat{\gamma})^2 \leq \sum_{t=1}^n (Y_t - \tilde{X}_t'\gamma)^2 \text{ for all } \gamma \in \mathbb{R}^{K+q}.$$

Now we choose

$$\gamma = (\hat{\beta}', 0')',$$

where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ is the OLS from the first regression. It follows that

$$\begin{aligned} \tilde{e}'\tilde{e} &\leq \sum_{t=1}^n \left(Y_t - \sum_{j=0}^k \hat{\beta}_j X_{jt} - \sum_{j=k+1}^{k+q} 0 \cdot X_{jt} \right)^2 \\ &= \sum_{t=1}^n (Y_t - X_t'\hat{\beta})^2 \\ &= e'e. \end{aligned}$$

Hence, we have $R_1^2 \leq R_2^2$. This completes the proof.

Question: What is the implication of this theorem?

The measure R^2 can be used to compare models with the same number of predictors, but it is not a useful criterion for comparing models of different sizes because it is biased in favor of large models.

The measure R^2 is not a suitable criterion for correct model specification. It is a measure for sampling variation rather than a measure of population. A high value of R^2 does not necessarily imply correct model specification, and correct model specification also does not necessarily imply a high value of R^2.

Strictly speaking, R^2 is a measure merely of association with nothing to say about causality. High values of R^2 are often very easy to achieve when dealing with economic

time series data, even when the causal link between two variables is extremely tenuous or perhaps nonexistent. For example, in the spurious regressions where the dependent variable Y_t and the regressors X_t have no causal relationship but they diaplay similar trending behaviors over time, it is often found that R^2 is close to unity.

Finally, R^2 is a measure of the strength of linear association between the dependent variable Y_t and the regressors X_t (see Exercise 3.2). It is not a suitable measure for goodness of fit of a nonlinear regression model where $E(Y_t|X_t)$ is a nonlinear function of X_t.

Question: How to interpret R^2 for the linear regression model

$$\ln Y_t = \beta_0 + \beta_1 \ln L_t + \beta_2 \ln K_t + \varepsilon_t,$$

where Y_t is output, L_t is labor and K_t is capital?

Answer: R^2 is the proportion of the total sample variations in $\ln Y_t$ that can be attributed to the sample variations in $\ln L_t$ and $\ln K_t$. It is not the proportion of the sample quadratic variations in Y_t that can be attributed to the sample variations of L_t and K_t.

Question: Does a high R^2 value imply a precise estimation for β^o?

Two popular model selection criteria

Often, a large number of potential predictors are available, but we do not necessarily want to include all of them. There are two conflicting factors to consider: on one hand, a larger model has less systematic bias and it would give the best predictions if all parameters could be estimated without error. On the other hand, when unknown parameters are replaced by estimates, the prediction becomes less accurate, and this effect is worse when there are more parameters to estimate. An important idea in statistics is to use a simple model to capture essential information contained in data as much as possible. This is often called the KISS principle, namely "Keep It Sophisticatedly Simple"!

Below, we introduce two popular model selection criteria that reflect such an idea.

Akaike Information Criterion [AIC]:

A linear regression model can be selected by minimizing the following AIC criterion with a suitable choice of K :

$$\begin{aligned} AIC &= \ln(s^2) + \frac{2K}{n} \\ &\sim \text{goodness of fit + model complexity} \end{aligned}$$

where

$$s^2 = e'e/(n-K),$$

is called the residual variance estimator for $E(\varepsilon_t^2) = \sigma^2$and $K = k+1$ is the number of regressors. AIC is proposed by Akaike (1973).

Bayesian Information Criterion [BIC, Schwarz (1978)]:

A linear regression model can be selected by minimizing the following criterion with a suitable choice of K :

$$BIC = \ln(s^2) + \frac{K\ln(n)}{n}.$$

This is called the Baysian information criterion (BIC), proposed by Schwarz (1978).

Both AIC and BIC try to trade off the goodness of fit to data measured by $\ln(s^2)$ with the desire to use as few paramerers as possible. When $\ln n \geq 2$, which is the case when $n > 7$, BIC gives a heavier penalty for model complexity than AIC, which is measured by the number of estimated parameters (relative to the sample size n). As a consequence, BIC will choose a more parsimonious linear regression model than AIC.

The difference between AIC and BIC is due to the way they are constructed. AIC is designed to select a model that will predict best and is less concerned than BIC with having a few too many parameters. BIC is designed to select the true value of K exactly. Under certain regularity conditions, BIC is strongly consistent in the sense that it determines the true model asymptotically (i.e., as $n \to \infty$), whereas for AIC an overparameterized model will emerge no matter how large the sample is. Of course, such properties are not necessarily guaranteed in finite samples. In practice, te best AIC model is usually close to the best BIC model and often they deliver the same model.

In addition to AIC and BIC, there are other criteria such as $\bar{R}^2$, the so-called adjusted R^2 that can also be used to select a linear regression model. The adjusted $\bar{R}^2$ is defined as

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{(Y-\bar{Y})'(Y-\bar{Y})/(n-1)}.$$

This differs from

$$R^2 = 1 - \frac{e'e}{(Y-\bar{Y})'(Y-\bar{Y})}.$$

In $\bar{R}^2$, the adjustmet is made according to the degrees of freedom, or the number of explanatory variables in X_t. It may be shown that

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-K}(1-R^2)\right].$$

we note that $\bar{R}^2$ may take a negative value although there is an intercept in X_t.

All model criteria are structured in terms of the estimated residual variance $\hat{\sigma}^2$ plus a penalty adjustment involving the number of estimated parameters, and it is in the extent of this penalty that the criteria differ from. For more discussion about these and other selection criteria, see Judge *et al.* (1985, Section 7.5).

Question: Why is it not a good practice to use a complicated model?

A complicated model contains many unknown parameters. Given a fixed amount of data information, parameter estimation will become less precise if more parameters have to be estimated. As a consequence, the out-of-sample forecast for Y_t may become less precise than the forecast of a simpler model. The latter may have a larger bias but more precise parameter estimates. Intuitively, a complicated model is too flexible in the sense that it may capture not only systematic components but also some features in the data which will not show up again. Thus, it cannot forecast futures well.

3.4 Consistency and Efficiency of OLS

We now investigate the statistical properties of $\hat{\beta}$. We are interested in addressing the following basic questions:

- Is $\hat{\beta}$ a good estimator for β^o(consistency)?
- Is $\hat{\beta}$ the best estimator (efficiency)?
- What is the sampling distribution of $\hat{\beta}$ (normality)?

Question: What is the sampling distribution of $\hat{\beta}$?

The distribution of $\hat{\beta}$ is called the sampling distribution of $\hat{\beta}$, because $\hat{\beta}$ is a function of the random sample $\{Z_t\}_{t=1}^n$, where $Z_t = (Y_t, X_t')'$.

The sampling distribution of $\hat{\beta}$ is useful for any statistical inference involving $\hat{\beta}$, such as confidence interval estimation and hypothesis testing.

We first investigate the statistical properties of $\hat{\beta}$.

Theorem 3.4: *Suppose Assumptions 3.1-3.3(a) and 3.4 hold. Then*
(i) [Unbiasedness] $E(\hat{\beta}|\mathbf{X}) = \beta^o$ *and* $E(\hat{\beta}) = \beta^o$.
(ii) [Vanishing Variance]

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= E\left[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})'|\mathbf{X}\right] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

If in addition Assumption 3.3(b) holds, then for any $K \times 1$ vector τ such that $\tau'\tau = 1$, we have

$$\tau' \text{var}(\hat{\beta}|\mathbf{X})\tau \to 0 \text{ as } n \to \infty.$$

(iii) [Orthogonality between e and $\hat{\beta}$]

$$\text{cov}(\hat{\beta}, e|\mathbf{X}) = E\{[\hat{\beta} - E(\hat{\beta}|\mathbf{X})]e'|\mathbf{X}\} = 0.$$

(iv) [Gauss-Markov]

$$\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) \textit{ is positive semi-definite (p.s.d.)}$$

for any unbiased estimator $\hat{b}$ that is linear in Y with $E(\hat{b}|\mathbf{X}) = \beta^o$.

(v) [Residual variance estimator]

$$s^2 = e'e/(n-K) = \frac{1}{n-K}\sum_{t=1}^{n} e_t^2$$

is unbiased for $\sigma^2 = E(\varepsilon_t^2)$. That is, $E(s^2|\mathbf{X}) = \sigma^2$.

Proof: (i) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, we have

$$\begin{aligned} E[(\hat{\beta} - \beta^o)|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 \\ &= 0. \end{aligned}$$

(ii) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$ and $E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I$, we have

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &\equiv E\left[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})'|\mathbf{X}\right] \\ &= E\left[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}\right] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Note that Assumption 3.4 is crucial here to obtain the expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ for $\text{var}(\hat{\beta}|\mathbf{X})$. Moreover, for any $\tau \in \mathbb{R}^K$ such that $\tau'\tau = 1$, we have

$$\begin{aligned}
\tau'\text{var}(\hat{\beta}|\mathbf{X})\tau &= \sigma^2\tau'(\mathbf{X}'\mathbf{X})^{-1}\tau \\
&\leq \sigma^2\lambda_{\max}[(\mathbf{X}'\mathbf{X})^{-1}] \\
&= \sigma^2\lambda_{\min}^{-1}(\mathbf{X}'\mathbf{X}) \\
&\to 0
\end{aligned}$$

given $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ with probability one. Note that the condition that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ ensures that $\text{var}(\hat{\beta}|\mathbf{X})$ vanishes to zero as $n \to \infty$.

(iii) Given $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, $e = Y - \mathbf{X}\hat{\beta} = MY = M\varepsilon$ (since $M\mathbf{X} = 0$), and $E(e) = 0$, we have

$$\begin{aligned}
\text{cov}(\hat{\beta}, e|\mathbf{X}) &= E\left[(\hat{\beta} - E\hat{\beta})(e - Ee)'|\mathbf{X}\right] \\
&= E\left[(\hat{\beta} - \beta^o)e'|\mathbf{X}\right] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'M|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})M \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 IM \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'M \\
&= 0.
\end{aligned}$$

Again, Assumption 3.4 plays a crucial role in ensuring zero correlation between $\hat{\beta}$ and e.

(iv) Consider a linear estimator

$$\hat{b} = C'Y,$$

where $C = C(\mathbf{X})$ is a $n \times K$ matrix depending on $\mathbf{X}$. It is unbiased for β^o regardless of the value of β^o if and only if

$$\begin{aligned}
E(\hat{b}|\mathbf{X}) &= C'\mathbf{X}\beta^o + C'E(\varepsilon|\mathbf{X}) \\
&= C'\mathbf{X}\beta^o \\
&= \beta^o.
\end{aligned}$$

This follows if and only if

$$C'\mathbf{X} = I.$$

Because

$$\begin{aligned}\hat{b} &= C'Y \\ &= C'(\mathbf{X}\beta^o + \varepsilon) \\ &= C'\mathbf{X}\beta^o + C'\varepsilon \\ &= \beta^o + C'\varepsilon,\end{aligned}$$

the variance of $\hat{b}$

$$\begin{aligned}\text{var}(\hat{b}) &= E\left[(\hat{b}-\beta^o)(\hat{b}-\beta^o)'|\mathbf{X}\right] \\ &= E\left[C'\varepsilon\varepsilon'C|\mathbf{X}\right] \\ &= C'E(\varepsilon\varepsilon'|\mathbf{X})C \\ &= C'\sigma^2 IC \\ &= \sigma^2 C'C.\end{aligned}$$

Using $C'\mathbf{X} = I$, we now have

$$\begin{aligned}\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2 C'C - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2[C'C - C'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'C] \\ &= \sigma^2 C'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']C \\ &= \sigma^2 C'MC \\ &= \sigma^2 C'MMC \\ &= \sigma^2 C'M'MC \\ &= \sigma^2 (MC)'(MC) \\ &= \sigma^2 D'D \\ &= \sigma^2 \sum_{t=1}^{n} D_t D_t' \\ &\sim \text{p.s.d.}\end{aligned}$$

where we have used the fact that for any real-valued matrix D, the squared matrix $D'D$ is always p.s.d. [**Question:** How to show this?]

(v) Now we show $E[e'e/(n-K)] = \sigma^2$. Because $e'e = \varepsilon' M\varepsilon$ and tr(AB) =tr(BA),

we have

$$\begin{aligned} E(e'e|\mathbf{X}) &= E(\varepsilon' M\varepsilon|\mathbf{X}) \\ &= E[\text{tr}(\varepsilon' M\varepsilon)|\mathbf{X}] \\ [\text{putting } A &= \varepsilon' M, B = \varepsilon] \\ &= E[\text{tr}(\varepsilon\varepsilon' M)|\mathbf{X}] \\ &= \text{tr}[E(\varepsilon\varepsilon'|\mathbf{X})M] \\ &= \text{tr}(\sigma^2 IM) \\ &= \sigma^2 \text{tr}(M) \\ &= \sigma^2(n-K) \end{aligned}$$

where

$$\begin{aligned} \text{tr}(M) &= \text{tr}(I_n) - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \text{tr}(I_n) - \text{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= n-K, \end{aligned}$$

using $\text{tr}(AB) = \text{tr}(BA)$ again. It follows that

$$\begin{aligned} E(s^2|\mathbf{X}) &= \frac{E(e'e|\mathbf{X})}{n-K} \\ &= \frac{\sigma^2(n-K)}{(n-K)} \\ &= \sigma^2. \end{aligned}$$

This completes the proof.

Remarks:

Both Theorem 3.4 (i) and (ii) imply that the conditional MSE

$$\begin{aligned} MSE(\hat{\beta}|\mathbf{X}) &= E[(\hat{\beta}-\beta^o)(\hat{\beta}-\beta^o)'|\mathbf{X}] \\ &= \text{var}(\hat{\beta}|\mathbf{X}) + \text{Bias}(\hat{\beta}|\mathbf{X})\text{Bias}(\hat{\beta}|\mathbf{X})' \\ &= \text{var}(\hat{\beta}|\mathbf{X}) \\ &\to 0 \text{ as } n \to \infty, \end{aligned}$$

where we have used the fact that

$$Bias(\hat{\beta}|\mathbf{X}) \equiv E(\hat{\beta}|\mathbf{X}) - \beta^o = 0.$$

Recall that MSE measures how close an estimator $\hat{\beta}$ is to the target parameter β^o.

Theorem (iv) implies that $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β^o because $\text{var}(\hat{\beta}|\mathbf{X})$ is the smallest among all unbiased linear estimators for β^o.

Formally, we can define a related concept for comparing two unbiased estimators:

Definition 3.4 [Efficiency]: *An unbiased estimator $\hat{\beta}$ of parameter β^o is more efficient than another unbiased estimator $\hat{b}$ of parameter β^o if*

$$\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X}) \text{ is p.s.d.}$$

When $\hat{\beta}$ is more efficient than $\hat{b}$, we have that for any $\tau \in \mathbb{R}^K$ such that $\tau'\tau = 1$,

$$\tau' \left[\text{var}(\hat{b}|\mathbf{X}) - \text{var}(\hat{\beta}|\mathbf{X})\right] \tau \geq 0.$$

Choosing $\tau = (1, 0, ..., 0)'$, for example, we have

$$\text{var}(\hat{b}_1) - \text{var}(\hat{\beta}_1) \geq 0.$$

We note that the OLS estimator $\hat{\beta}$ is still BLUE even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n, and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$. Near-multicolinearity is essentially a sample or data problem which we cannot remedy or improve upon when the objective is to estimate the unknown parameter β^o.

3.5 Sampling Distribution of OLS

To obtain the finite sample sampling distribution of $\hat{\beta}$, we impose the normality assumption on ε.

Assumption 3.5: $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$.

Remarks:

Assumption 3.5 implies both Assumptions 3.2 ($E(\varepsilon|\mathbf{X}) = 0$) and 3.4 ($E(\varepsilon\varepsilon|\mathbf{X}) = \sigma^2 I$). Moreover, under Assumption 3.5, the conditional *pdf* of ε given $\mathbf{X}$ is

$$f(\varepsilon|\mathbf{X}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right) = f(\varepsilon),$$

which does not depend on $\mathbf{X}$, so the disturbance ε is independent of $\mathbf{X}$. Thus, every conditional moment of ε given $\mathbf{X}$ does not depend on $\mathbf{X}$.

The normal distribution is also called the Gaussian distribution named after the German mathematician and astronomer Carl F. Gauss. It is assumed here so that we can derive the finite sample distributions of $\hat{\beta}$ and related statistics, i.e., the distributions of $\hat{\beta}$ and related statistics when the sample size n is a finite integer. This assumption may be reasonable for observations that are computed as the averages of the outcomes of many repeated experiments, due to the effect of the so-called central limit theorem (CLT). This may occur in physics, for example. In economics, the normality assumption may not always be reasonable. For example, many high-frequency financial time series usually display heavy tails (with kurtosis larger than 3).

Question: What is the sampling distribution of $\hat{\beta}$?

We write

$$\begin{aligned}\hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \\ &= \sum_{t=1}^{n} C_t\varepsilon_t,\end{aligned}$$

where the weighting vector

$$C_t = (\mathbf{X}'\mathbf{X})^{-1}X_t$$

is called the leverage of observation X_t.

Theorem 3.5 [Normality of $\hat{\beta}$]: *Under Assumptions 3.1, 3.3(a) and 3.5,*

$$(\hat{\beta} - \beta^o)|\mathbf{X} \sim N[0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

Proof: Conditional on $\mathbf{X}$, $\hat{\beta} - \beta^o$ is a weighted sum of independent normal random variables $\{\varepsilon_t\}$, and so it is also normally distributed.

We note that the OLS estimator $\hat{\beta}$ still has the conditional finite sample normal distribution $N(\beta^o, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$.

The corollary below follows immediately.

Corollary 3.6 [Normality of $R(\hat{\beta} - \beta^o)$]: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then for any nonstochastic $J \times K$ matrix R, we have*

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'].$$

Proof: Conditional on $\mathbf{X}$, $\hat{\beta} - \beta^o$ is normally distributed. Therefore, conditional on $\mathbf{X}$, the linear combination $R(\hat{\beta} - \beta^o)$ is also normally distributed, with

$$E[R(\hat{\beta} - \beta^o)|\mathbf{X}] = RE[(\hat{\beta} - \beta^o)|\mathbf{X}] = R \cdot 0 = 0$$

and

$$\begin{aligned}
\text{var}[R(\hat{\beta} - \beta^o)|\mathbf{X}] &= E\left[R(\hat{\beta} - \beta^o)(R(\hat{\beta} - \beta^o))'|\mathbf{X}\right] \\
&= E\left[R(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'R'|\mathbf{X}\right] \\
&= RE\left[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}\right]R' \\
&= R\text{var}(\hat{\beta}|\mathbf{X})R' \\
&= \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'.
\end{aligned}$$

It follows that

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R').$$

Question: What is the role of the $J \times K$ nonstochastic matrix R?
Answer: The $J \times K$ matrix R is a selection matrix. For example, when $R = (1, 0, ..., 0)$, we then have $R(\hat{\beta} - \beta^o) = \hat{\beta}_0 - \beta_0^o$.

Question: Why would we like to know the sampling distribution of $R(\hat{\beta} - \beta^o)$?

This is mainly for confidence interval estimation and hypothesis testing.

3.6 Variance Matrix Estimator for OLS

Since $\text{var}(\varepsilon_t) = \sigma^2$ is unknown, $\text{var}[R(\hat{\beta} - \beta^o)|\mathbf{X}] = \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'$ is unknown. We need to estimate σ^2. We can use the residual variance estimator

$$s^2 = e'e/(n - K).$$

Theorem 3.7 [Residual Variance Estimator]: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then we have for all $n > K$, (i)*

$$\frac{(n-K)s^2}{\sigma^2}|\mathbf{X} = \frac{e'e}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K},$$

where χ^2_{n-K} denotes the Chi-square distribution with $n - K$ degrees of freedom;

(ii) conditional on $\mathbf{X}$, s^2 and $\hat{\beta}$ are independent.

Proof: (i) Because $e = M\varepsilon$, we have

$$\frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right).$$

In addition, because $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I_n)$, and M is an idempotent matrix with rank $= n - K$ (as has been shown before), we have the quadratic form

$$\frac{e'e}{\sigma^2} = \frac{\varepsilon' M \varepsilon}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K}$$

by the following lemma.

Lemma 3.8 [Quadratic form of normal random variables]: *If $v \sim N(0, I_n)$ and Q is an $n \times n$ nonstochastic symmetric idempotent matrix with rank $q \leq n$, then the quadratic form*

$$v'Qv \sim \chi^2_q.$$

In our application, we have $v = \varepsilon/\sigma \sim N(0, I)$, and $Q = M$. Since $\text{rank}(M) = n - K$, we have

$$\frac{e'e}{\sigma^2}|\mathbf{X} \sim \chi^2_{n-K}.$$

(ii) Next, we show that s^2 and $\hat{\beta}$ are independent. Because $s^2 = e'e/(n-K)$ is a function of e, it suffices to show that e and $\hat{\beta}$ are independent. This follows immediately because both e and $\hat{\beta}$ are jointly normally distributed and they are uncorrelated. It is well-known that for a joint normal distribution, zero correlation is equivalent to independence.

It remains to show that e and $\hat{\beta}$ jointly normally distributed? For this purpose, we write

$$\begin{bmatrix} e \\ \hat{\beta} - \beta^o \end{bmatrix} = \begin{bmatrix} M\varepsilon \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \end{bmatrix} = \begin{bmatrix} M \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \varepsilon.$$

Because $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$, the linear combination of

$$\begin{bmatrix} M \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{bmatrix} \varepsilon$$

is also normally distributed conditional on $\mathbf{X}$. It follows that e and $\hat{\beta}$ are independent given $\text{cov}(\hat{\beta}, e|\mathbf{X}) = 0$. This completes the proof.

Question: What is a χ^2_q distribution?

Definition 3.5 [Chi-square Distribution, χ^2_q] *Suppose $\{Z_i\}_{i=1}^q$ are i.i.d.N(0,1) random variables. Then the random variable*

$$\chi^2 = \sum_{i=1}^{q} Z_i^2$$

will follow a χ^2_q distribution.

The χ^2_q distribution is nonsymmetric and has long right tails. For a χ^2_q random variable, we have $E(\chi^2_q) = q$ and $\text{var}(\chi^2_q) = 2q$.

Based on these properties of a χ^2 distribution, Theorem 3.7(i) implies

$$\begin{aligned} E\left[\frac{(n-K)s^2}{\sigma^2}|\mathbf{X}\right] &= n-K. \\ \frac{(n-K)}{\sigma^2}E(s^2|\mathbf{X}) &= n-K. \end{aligned}$$

It follows that $E(s^2|\mathbf{X}) = \sigma^2$. Note that we have shown this result with a different method but under a more general condition.

Theorem 3.7(i) also implies

$$\begin{aligned} \text{var}\left[\frac{(n-K)s^2}{\sigma^2}|\mathbf{X}\right] &= 2(n-K), \\ \text{var}(s^2|\mathbf{X}) &= \frac{2\sigma^4}{n-K} \\ &\to 0 \end{aligned}$$

as $n \to \infty$.

Both Theorems 3.7(i) and (ii) imply that the conditional MSE of s^2

$$\begin{aligned} MSE(s^2|\mathbf{X}) &= E\left[(s^2-\sigma^2)^2|\mathbf{X}\right] \\ &= \text{var}(s^2|\mathbf{X}) + [E(s^2|\mathbf{X}) - \sigma^2]^2 \\ &\to 0. \end{aligned}$$

Thus, s^2 is a good estimator for σ^2.

The independence between s^2 and $\hat{\beta}$ is crucial for us to obtain the sampling distribution of the popular t-test and F-test statistics, which will be introduced shortly.

The sample residual variance $s^2 = e'e/(n-K)$ is a generalization of the sample variance $S_n^2 = (n-1)^{-1}\Sigma_{t=1}^n (Y_t - \bar{Y})^2$ for the random sample $\{Y_t\}_{t=1}^n$. The factor $n-K$ is called the degrees of freedom of the estimated residual sample $\{e_t\}_{t=1}^n$. To gain intuition why the degrees of freedom is equal to $n-K$, note that the orginal sample $\{Z_t\}_{t=1}^n = \{Y_t, X_t'\}_{t=1}^{'n}$ has n observations, which can be viewed to have n degrees of freedom. Now when estimating σ^2, we have to use the estimated residual sample $\{e_t\}_{t=1}^n$. These n estimated residuals are not linearly independent because they have to satisfy the FOC of the OLS estimation, namely,

$$\begin{aligned} \mathbf{X}'e &= 0. \\ (K \times n) \times (n \times 1) &= K \times 1. \end{aligned}$$

The FOC imposes K restrictions on $\{e_t\}_{t=1}^n$, conditional on $\mathbf{X}$. These K restrictions are needed in order to estimate K unknown parameters β^o. They can be used to obtain the remaining K estmated residuals $\{e_{T-K+1}, ..., e_T\}$ from the first $n-K$ estimated residuals $\{e_1, ..., e_{n-K}\}$ if the latter have been available. Thus, the remaining degrees of freedom of e is $n-K$. Note that the sample variance S_n^2 is the residual variance estimator with $Y_t = \beta_0^o + \varepsilon_t$.

Question: Why are these sampling distributions of $\hat{\beta}$ and s^2 useful in practice?

They are useful in confidence interval estimation and hypothesis testing on model parameters. In this book, we will focus on hypothesis testing on model parameters. Statistically speaking, confidence interval estimation and hypothesis testing on model parameters are just two sides of the same coin.

3.7 Hypothesis Testing

We now use the sampling distributions of $\hat{\beta}$ and s^2 to develop test procedures for hypotheses of interest. We consider testing the following linear hypothesis in form of

$$\begin{aligned} \mathbf{H}_0 &: R\beta^o = r, \\ (J \times K)(K \times 1) &= J \times 1, \end{aligned}$$

where R is called the selection matrix, and J is the number of restrictions. We assume $J \leq K$.

It is important to emphasize that we will test $\mathbf{H}_0$ under correct model specification for $E(Y_t|X_t)$.

Motivation

We first provide a few motivating examples for hypothesis testing.

Example 1 [Reforms have no effect]: Consider the extended production function

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \varepsilon_t,$$

where AU_t is a dummy variable indicating whether firm t is granted autonomy, and PS_t is the profit share of firm t with the state.

Suppose we are interested in testing whether autonomy AU_t has an effect on productivity. Then we can write the null hypothesis

$$\mathbf{H}_0^a : \beta_3^o = 0$$

This is equivalent to the choices of:

$$\begin{aligned} \beta^o &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'. \\ R &= (0, 0, 0, 1, 0), \\ r &= 0. \end{aligned}$$

If we are interested in testing whether profit sharing has an effect on productivity, we can consider the null hypothesis

$$\mathbf{H}_0^b : \beta_4^o = 0.$$

Alternatively, to test whether the production technology exhibits the constant return to scale (CRS), we can write the null hypothesis as follows:

$$\mathbf{H}_0^c : \beta_1^o + \beta_2^o = 1.$$

This is equivalent to the choice of $R = (0, 1, 1, 0, 0)$ and $r = 1$.

Finally, if we are interested in examining the joint effect of both autonomy and profit sharing, we can test the hypothesis that neither autonomy nor profit sharing has impact:

$$\mathbf{H}_0^d : \beta_3^o = \beta_4^o = 0.$$

This is equivalent to the choice of

$$\begin{aligned} R &= \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\ r &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

Example 2 [Optimal Predictor for Future Spot Exchange Rate]: Consider

$$S_{t+\tau} = \beta_0 + \beta_1 F_t(\tau) + \varepsilon_{t+\tau}, \qquad t = 1, ..., n,$$

where $S_{t+\tau}$ is the spot exchange rate at period $t+\tau$, and $F_t(\tau)$ is the forward exchange rate, namely the period t's price for the foreign currency to be delivered at period $t+\tau$. The null hypothesis of interest is that the forward exchange rate $F_t(\tau)$ is an optimal predictor for the future spot rate $S_{t+\tau}$ in the sense that $E(S_{t+\tau}|I_t) = F_t(\tau)$, where I_t is the information set available at time t. This is actually called the *expectations hypothesis* in economics and finance. Given the above specification, this hypothesis can be written as

$$\mathbf{H}_0^e : \beta_0^o = 0, \beta_1^o = 1,$$

and $E(\varepsilon_{t+\tau}|I_t) = 0$. This is equivalent to the choice of

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

All examples considered above can be formulated with a suitable specification of R, where R is a $J \times K$ matrix in the null hypothesis

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

where r is a $J \times 1$ vector.

Basic Ideas of Hypothesis Testing

To test the null hypothesis

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

we can consider the statistic:

$$R\hat{\beta} - r$$

and check if this difference is significantly different from zero.

Under $\mathbf{H}_0 : R\beta^o = r$, we have

$$\begin{aligned} R\hat{\beta} - r &= R\hat{\beta} - R\beta^o \\ &= R(\hat{\beta} - \beta^o) \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

because $\hat{\beta} - \beta^o \to 0$ as $n \to \infty$ in terms of MSE.

Under the alternative to $\mathbf{H}_0$, $R\beta^o \neq r$, but we still have $\hat{\beta} - \beta^o \to 0$ in terms of MSE. It follows that

$$\begin{aligned} R\hat{\beta} - r &= R(\hat{\beta} - \beta^o) + R\beta^o - r \\ &\to R\beta^o - r \neq 0 \end{aligned}$$

as $n \to \infty$, where the convergence is in terms of MSE. In other words, $R\hat{\beta} - r$ will converge to a nonzero limit, $R\beta^o - r$.

The fact that the behavior of $R\hat{\beta} - r$ is different under $\mathbf{H}_0$ and under the alternative hypothesis to $\mathbf{H}_0$ provides a basis to construct hypothesis tests. In particular, we can test $\mathbf{H}_0$ by examining whether $R\hat{\beta} - r$ is significantly different from zero.

Question: How large should the magnitude of the absolute value of the difference $R\hat{\beta} - r$ be in order to claim that $R\hat{\beta} - r$ is significantly different from zero?

For this purpose, we need a decision rule which specifies a threshold value with which we can compare the (absolute) value of $R\hat{\beta} - r$. Because $R\hat{\beta} - r$ is a random variable and so it can take many (possibly an infinite number of) values. Given a data set, we only obtain one realization of $R\hat{\beta} - r$. Whether a realization of $R\hat{\beta} - r$ is close to zero should be judged using the critical value of its sampling distribution, which depends on the sample size n and the significance level $\alpha \in (0, 1)$ one preselects.

Question: What is the sampling distribution of $R\hat{\beta} - r$ under $\mathbf{H}_0$?

Because

$$R(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1} R'),$$

we have that conditional on $\mathbf{X}$,

$$\begin{aligned} R\hat{\beta} - r &= R(\hat{\beta} - \beta^o) + R\beta^o - r \\ &\sim N(R\beta^o - r, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1} R') \end{aligned}$$

Corollary 3.9: *Under Assumptions 3.1, 3.3 and 3.5, and* $\mathbf{H}_0 : R\beta^o = r$, *we have for each* $n > K$,

$$(R\hat{\beta} - r)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1} R').$$

The difference $R\hat{\beta} - r$ cannot be used as a test statistic for $\mathbf{H}_0$, because σ^2 is unknown and there is no way to calculate the critical values of the sampling distribution of $R\hat{\beta} - r$.

Question: How to construct a feasible (i.e., computable) test statistic?

The forms of test statistics will differ depending on whether we have $J = 1$ or $J > 1$. We first consider the case of $J = 1$.

Case I: t-Test ($J = 1$):

Recall that we have

$$(R\hat{\beta} - r)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'),$$

When $J = 1$, the conditional variance

$$\text{var}[(R\hat{\beta} - r)|\mathbf{X}] = \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'$$

is a scalar (1×1). It follows that conditional on $\mathbf{X}$, we have

$$\begin{aligned} \frac{R\hat{\beta} - r}{\sqrt{\text{var}[(R\hat{\beta} - r)|\mathbf{X}]}} &= \frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &\sim N(0, 1). \end{aligned}$$

Question: What is the unconditional distribution of

$$\frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}?$$

The unconditional distribution is also N(0,1).

However, σ^2 is unknown, so we cannot use the ratio

$$\frac{R\hat{\beta} - r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}$$

as a test statistic. We have to replace σ^2 by s^2, which is a good estimator for σ^2. This gives a feasible (i.e., computable) test statistic

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}.$$

However, the test statistic T will be no longer normally distributed. Instead,

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\ &= \frac{\frac{R\hat{\beta}-r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2}/(n-K)}} \\ &\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-K}/(n-K)}} \\ &\sim t_{n-K}, \end{aligned}$$

where t_{n-K} denotes a Student's t-distribution with $n-K$ degrees of freedom. Note that the numerator and denominator are mutually independent conditional on $\mathbf{X}$, because $\hat{\beta}$ and s^2 are mutually independent conditional on $\mathbf{X}$. The feasible statistic T is called a t-test statistic because it follows a t_{n-K} distribution.

Question: What is the Student t_q distribution?

Definition 3.6 [Student's t-distribution]: *Suppose $Z \sim N(0,1)$ and $V \sim \chi^2_q$, and both Z and V are independent. Then the ratio*

$$\frac{Z}{\sqrt{V/q}} \sim t_q.$$

The t_q-distribution is symmetric about 0 with heavier tails than the $N(0,1)$ distribution. The smaller number of the degrees of freedom, the heavier tails it has. When $q \to \infty$, $t_q \xrightarrow{d} N(0,1)$, where $\xrightarrow{d}$ denotes convergence in distribution. This implies that we have

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

This result has a very important implication in practice: for a large sample size n, it makes no difference to use either the critical values from t_{n-K} or from $N(0,1)$.

Question: What is convergence in distribution?

Definition 3.7 [Convergence in distribution]: *Suppose $\{Z_n, n = 1, 2, \cdots\}$ is a sequence of random variables/vectors with distribution functions $F_n(z) = P(Z_n \leq z)$, and Z is a random variable/vector with distribution $F(z) = P(Z \leq z)$. We say that Z_n*

converges to Z in distribution if the distribution of Z_n converges to the distribution of Z at all continuity points; namely,

$$\begin{aligned} \lim_{n\to\infty} F_n(z) &= F(z) \text{ or} \\ F_n(z) &\to F(z) \text{ as } n \to \infty \end{aligned}$$

for any continuity point z (i.e., for any point at which $F(z)$ is continuous). We use the notation $Z_n \xrightarrow{d} Z$. The distribution of Z is called the asymptotic or limiting distribution of Z_n.

In practice, Z_n is a test statistic or a parameter estimator, and often its sampling distribution $F_n(z)$ is either unknown or very complicated, but $F(z)$ is known or very simple. As long as $Z_n \xrightarrow{d} Z$, then we can use $F(z)$ as an approximation to $F_n(z)$. This gives a convenient procedure for statistical inference. The potential cost is that the approximation of $F_n(z)$ to $F(z)$ may not be good enough in finite samples (i.e., when n is finite). How good the approximation is will depend on the data generating process and the sample size n.

Example 3: Suppose $\{\varepsilon_n, n = 1, 2, \cdots\}$ is an *i.i.d.* sequence with distribution function $F(z)$. Let ε be a random variable with the same distribution function $F(z)$. Then $\varepsilon_n \xrightarrow{d} \varepsilon$.

With the obtained sampling distribution for the test statistic T, we can now describe a decision rule for testing $\mathbf{H}_0$ when $J = 1$.

Decision Rule of the T-test Based on Critical Values

(i) Reject $\mathbf{H}_0 : R\beta^o = r$ at a prespecified significance level $\alpha \in (0, 1)$ if

$$|T| > C_{t_{n-K,\frac{\alpha}{2}}},$$

where $C_{t_{n-K,\frac{\alpha}{2}}}$ is the so-called upper-tailed critical value of the t_{n-K} distribution at level $\frac{\alpha}{2}$, which is determined by

$$P\left[t_{n-K} > C_{t_{n-K,\frac{\alpha}{2}}}\right] = \frac{\alpha}{2}$$

or equivalently

$$P\left[|t_{n-K}| > C_{t_{n-K,\frac{\alpha}{2}}}\right] = \alpha.$$

(ii) Do not reject $\mathbf{H}_0$ at the significance level α if

$$|T| \leq C_{t_{n-K,\frac{\alpha}{2}}}.$$

Remarks:

In testing $\mathbf{H}_0$, there exist two types of errors, due to the limited information about the population in a given random sample $\{Z_t\}_{t=1}^n$. One possibility is that $\mathbf{H}_0$ is true but we reject it. This is called the "Type I error". The significance level α is the probability of making the Type I error. If

$$P\left[|T| > C_{t_{n-K,\frac{\alpha}{2}}}|\mathbf{H}_0\right] = \alpha,$$

we say that the decision rule is a test with size α.

On the other hand, the probability $P[|T| > C_{t_{n-K,\frac{\alpha}{2}}}|\mathbf{H}_0 \text{ is false}]$ is called the power function of a size α test. When

$$P\left[|T| > C_{t_{n-K,\frac{\alpha}{2}}}|\mathbf{H}_0 \text{ is false}\right] < 1,$$

there exists a possibility that one may fail to reject $\mathbf{H}_0$ when it is false. This is called the "Type II error".

Ideally one would like to minimize both the Type I error and Type II error, but this is impossible for any given finite sample. In practice, one usually presets the level for Type I error, the so-called significance level, and then minimizes the Type II error. Conventional choices for significance level α are 10%, 5% and 1% respectively.

Next, we describe an alternative decision rule for testing $\mathbf{H}_0$ when $J = 1$, using the so-called p-value of test statistic T.

An Equivalent Decision Rule Based on p-values

Given a data set $\mathbf{z}^n = \{y_t, x_t'\}_{t=1}^n$, which is a realization of the random sample $\mathbf{Z}^n = \{Y_t, X_t'\}_{t=1}^n$, we can compute a realization (i.e., a number) for the t-test statistic T, namely

$$T(\mathbf{z}^n) = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{x}'\mathbf{x})^{-1}R'}}.$$

Then the probability

$$p(\mathbf{z}^n) = P\left[|t_{n-K}| > |T(\mathbf{z}^n)|\right],$$

is called the p-value (i.e., probability value) of the test statistic T given that $\{Y_t, X_t'\}_{t=1}^n = \{y_t, x_t'\}_{t=1}^n$ is observed, where t_{n-K} is a Student's t random variable with $n-K$ degrees of freedom, and $T(\mathbf{z}^n)$ is a realization for test statistic $T = T(\mathbf{Z}^n)$ given the observed data $\mathbf{z}^n$. Intuitively, the p-value is the smallest value of significance level α for which the null hypothesis is rejected. Here, it is the tail probability that the absolute value of a Student's t_{n-K} random variable takes values larger than the absolute value of the test statistic $T(\mathbf{z}^n)$. If this probability is very small relative to the significance level, then it

is unlikely that the test statistic $T(\mathbf{Z}^n)$ will follow a Student's t_{n-K} distribution. As a consequence, the null hypothesis is likely to be false.

The above decision rule can be described equivalently as follows:

Decision Rule Based on the p-value

(i) Reject $\mathbf{H}_0$ at the significance level α if $p(\mathbf{z}^n) < \alpha$.

(ii) Do not reject $\mathbf{H}_0$ at the significance level α if $p(\mathbf{z}^n) \geq \alpha$.

Remarks:

A small p-value is evidence against the null hypothesis. A large p-value shows that the data are consistent with the null hypothesis.

Question: What are the advantages/disadvantages of using p-values versus using critical values?

p-values are more informative than only rejecting/accepting the null hypothesis at some significance level α. A p-value is the smallest significance level at which a null hypothesis can be rejected. It not only tells us whether the null hypothesis should be accepted or rejected, but it also tells us whether the decision to accept or reject the null hypothesis is a close call.

Most statistical software reports p-values of parameter estimates. This is much more convenient than asking the user to specify significance level α and then reporting whether the null hypothesis is accepted or rejected for that α.

When we reject a null hypothesis, we often say there is a statistically significant effect. This does not mean that there is an effect of practical importance (i.e., an effect of economic importance). This is because when large samples are used, small and practically unimportant effects are likely to be statistically significant.

The t-test and associated procedures just introduced are valid even when there exists near-multicolinearity, where $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ does not grow with the sample size n and $\text{var}(\hat{\beta}|\mathbf{X})$ does not vanish to zero as $n \to \infty$. However, the degree of near-multicolineary, as measured by sample correlations between explanatory variables, will affect the the precision of the OLS estimator $\hat{\beta}$. Other things being equal, the higher degree of near-multicolinearity, the larger the variance of $\hat{\beta}$. As a result, the t-statistic is often insignificant even when the null hypothesis $\mathbf{H}_0$ is false.

Examples of t-tests

Example 4 [Reforms have no effects (continued.)]

We first consider testing the null hypothesis

$$\mathbf{H}_0^a : \beta_3 = 0,$$

where β_3 is the coefficient of the autonomy AU_t in the extended production function regression model. This is equivalent to the selection of $R = (0, 0, 0, 1, 0)$. In this case, we have

$$\begin{aligned} s^2 R(\mathbf{X}'\mathbf{X})^{-1} R' &= \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(4,4)} \\ &= S^2_{\hat{\beta}_3} \end{aligned}$$

which is the estimator of $\text{var}(\hat{\beta}_3|\mathbf{X})$. The squared root of $\text{var}(\hat{\beta}_3|X)$ is called the standard error of estimator $\hat{\beta}_3$, and $S_{\hat{\beta}_3}$ is called the estimated standard error of $\hat{\beta}_3$. The t-test statistic

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1} R'}} \\ &= \frac{\hat{\beta}_3}{\sqrt{S^2_{\hat{\beta}_3}}} \\ &\sim t_{n-K}. \end{aligned}$$

Next, we consider testing the CRS hypothesis

$$\mathbf{H}_0^c : \beta_1 + \beta_2 = 1,$$

which corresponds to $R = (0, 1, 1, 0, 0)$ and $r = 1$. In this case,

$$\begin{aligned} s^2 R(\mathbf{X}'\mathbf{X})^{-1} R' &= S^2_{\hat{\beta}_1} + S^2_{\hat{\beta}_2} + 2\hat{\text{c}}\text{ov}(\hat{\beta}_1, \hat{\beta}_2) \\ &= \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(2,2)} \\ &\quad + \left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(3,3)} \\ &\quad + 2\left[s^2(\mathbf{X}'\mathbf{X})^{-1}\right]_{(2,3)} \\ &= S^2_{\hat{\beta}+\hat{\beta}_2}, \end{aligned}$$

which is the estimator of $\text{var}(\hat{\beta}_1 + \hat{\beta}_2|\mathbf{X})$. Here, $\hat{\text{c}}\text{ov}(\hat{\beta}_1, \hat{\beta}_2)$ is the estimator for $\text{cov}(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X})$, the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ conditional on $\mathbf{X}$.

The t-test statistic is

$$\begin{aligned} T &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1} R'}} \\ &= \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{S_{\hat{\beta}_1+\hat{\beta}_2}} \\ &\sim t_{n-K}. \end{aligned}$$

Case II: F-testing ($J > 1$)

Question: How to construct a test statistic for H_0 if $J > 1$?

We first state a useful lemma.

Lemma 3.10: *If $Z \sim N(0, V)$, where $V = var(Z)$ is a nonsingular $J \times J$ variance-covariance matrix, then*

$$Z'V^{-1}Z \sim \chi^2_J.$$

Proof: Because V is symmetric and positive definite, we can find a symmetric and invertible matrix $V^{1/2}$ such that

$$\begin{aligned} V^{1/2}V^{1/2} &= V, \\ V^{-1/2}V^{-1/2} &= V^{-1}. \end{aligned}$$

(Question: What is this decomposition called?) Now, define

$$Y = V^{-1/2}Z.$$

Then we have $E(Y) = 0$, and

$$\begin{aligned} \text{var}(Y) &= E\left\{[Y - E(Y)][Y - E(Y)]'\right\} \\ &= E(YY') \\ &= E(V^{-1/2}ZZ'V^{-1/2}) \\ &= V^{-1/2}E(ZZ')V^{-1/2} \\ &= V^{-1/2}VV^{-1/2} \\ &= V^{-1/2}V^{1/2}V^{1/2}V^{-1/2} \\ &= I. \end{aligned}$$

It follows that $Y \sim N(0, I)$. Therefore, we have

$$Y'Y \sim \chi^2_J.$$

Applying this lemma, and using the result that

$$(R\hat{\beta} - r)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']$$

under $\mathbf{H}_0$, we have the quadratic form

$$(R\hat{\beta} - r)'[\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r) \sim \chi^2_J$$

conditional on $\mathbf{X}$, or

$$\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2} \sim \chi^2_J$$

conditional on $\mathbf{X}$.

Because χ^2_J does not depend on $\mathbf{X}$, therefore, we also have

$$\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2} \sim \chi^2_J$$

unconditionally.

Like in constructing a t-test statistic, we should replace σ^2 by s^2 in the left hand side:

$$\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2}.$$

The replacement of σ^2 by s^2 renders the distribution of the quadratic form no longer Chi-squared. Instead, after proper scaling, the quadratic form will follow a so-called F-distribution with degrees of freedom equal to $(J, n-K)$.

Why?

To explain this, we observe

$$\begin{aligned} & \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2} \\ = & J \cdot \frac{\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2}/J}{\frac{(n-K)s^2}{\sigma^2}/(n-K)} \\ \sim & J \cdot F_{J,n-K}, \end{aligned}$$

where $F_{J,n-K}$ denotes the F distribution with degrees of J and $n-K$ distributions.

Question: What is a $F_{J,n-K}$ distribution?

Definition 3.8: *Suppose $U \sim \chi^2_p$ and $V \sim \chi^2_q$, and both U and V are independent. Then the ratio*

$$\frac{U/p}{V/q} \sim F_{p,q}$$

is called to follow a $F_{p,q}$ distribution with degrees of freedom (p, q).

This distribution is called F-distribution because it is named after Professor Fisher, a well-known statistician in the 20th century. It is similar to the shape of the χ^2 distribution with a long right tail. An $F_{p,q}$ random variable has the following properties:

(i) If $F \sim F_{p,q}$, then $F^{-1} \sim F_{q,p}$.

(ii) $t_q^2 \sim F_{1,q}$.

$$t_q^2 = \frac{\chi_1^2/1}{\chi^2/q} \sim F_{1,q}$$

(iii) Given any fixed integer p, $p \cdot F_{p,q} \to \chi_p^2$ as $q \to \infty$.

Property (ii) implies that when $J = 1$, using either the t-test or the F-test will deliver the same conclusion. Property (iii) implies that the conclusions based on $F_{p,q}$ and on $p \cdot F_{p,q}$ using the χ_p^2 approximation will be approximately the same when q is sufficiently large.

We now define the following F-test statistic to test H_0 :

$$\begin{aligned} F &\equiv \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} \\ &\sim F_{J,n-K}. \end{aligned}$$

Theorem 3.11: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then under* $\mathbf{H}_0$: $R\beta^o = r$, *we have*

$$\begin{aligned} F &= \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} \\ &\sim F_{J,n-K} \end{aligned}$$

for all $n > K$.

Alternative Expression for the F-Test Statistic

A practical issue now is how to compute the F-statistic. One can of course compute the F-test statistic using the above definition of the F test statistic. However, there is a very convenient way to compute the F-test statistic. We now introduce this method.

Theorem 3.12: *Suppose Assumptions 3.1 and 3.3(a) hold. Let* $SSR_u = e'e$ *be the sum of squared residuals from the unrestricted model*

$$Y = \mathbf{X}\beta^o + \varepsilon.$$

Let $SSR_r = \tilde{e}'\tilde{e}$ *be the sum of squared residuals from the restricted model*

$$Y = \mathbf{X}\beta^o + \varepsilon$$

subject to

$$R\beta^o = r,$$

where $\tilde{\beta}$ is the restricted OLS estimator. Then under $\mathbf{H}_0$, the F-test statistic can be written as

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n-K)} \sim F_{J,n-K}.$$

Proof: Let $\tilde{\beta}$ be the OLS under $\mathbf{H}_0$; that is,

$$\tilde{\beta} = \arg\min_{\beta \in R^K} (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta)$$

subject to the constraint that $R\beta = r$. We first form the Lagrangian function

$$L(\beta, \lambda) = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) + 2\lambda'(r - R\beta),$$

where λ is a $J \times 1$ vector called the Lagrange multiplier vector.

We have the following FOC:

$$\begin{aligned} \frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \beta} &= -2\mathbf{X}'(Y - \mathbf{X}\tilde{\beta}) - 2R'\tilde{\lambda} = 0, \\ \frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \lambda} &= 2(r - R\tilde{\beta}) = 0. \end{aligned}$$

With the unconstrained OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$, and from the first equation of FOC, we can obtain

$$\begin{aligned} -(\hat{\beta} - \tilde{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}R'\tilde{\lambda}, \\ R(\mathbf{X}'\mathbf{X})^{-1}R'\tilde{\lambda} &= -R(\hat{\beta} - \tilde{\beta}). \end{aligned}$$

Hence, the Lagrange multiplier

$$\begin{aligned} \tilde{\lambda} &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}R(\hat{\beta} - \tilde{\beta}). \\ &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r), \end{aligned}$$

where we have made use of the constraint that $R\tilde{\beta} = r$. It follows that

$$\hat{\beta} - \tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}R'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r).$$

Now,

$$\begin{aligned} \tilde{e} &= Y - \mathbf{X}\tilde{\beta} \\ &= Y - \mathbf{X}\hat{\beta} + \mathbf{X}(\hat{\beta} - \tilde{\beta}) \\ &= e + \mathbf{X}(\hat{\beta} - \tilde{\beta}). \end{aligned}$$

It follows that

$$\begin{aligned}\tilde{e}'\tilde{e} &= e'e + (\hat{\beta}-\tilde{\beta})'\mathbf{X}'\mathbf{X}(\hat{\beta}-\tilde{\beta}) \\ &= e'e + (R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r).\end{aligned}$$

We have

$$(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r) = \tilde{e}'\tilde{e} - e'e$$

and

$$\begin{aligned}F &= \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)/J}{s^2} \\ &= \frac{(\tilde{e}'\tilde{e}-e'e)/J}{e'e/(n-K)}.\end{aligned}$$

This completes the proof.

Remarks:

The F-statistic is a convenient test statistic! One only needs to compute SSR in order to compute the F-test statistic. Intuitively, the sum of squared residuals SSR_u of the unrestricted regression model is always larger than or at least equal to that of the restricted regression model. When the null hypothesis $\mathbf{H}_0$ is true (i.e., when the parameter restriction is valid), the sum of squared residuals SSR_r of the restricted model is more or less similar to that of the unrestricted model, subject to the difference due to sampling variations. If SSR_r is sufficiently larger than SSR_u, then there exists evidence against $\mathbf{H}_0$.How large a difference between SSR_r and SSR_u is considered as sufficiently large to reject $\mathbf{H}_0$ is determined by the critical value of the associated F distribution.

Question: What is the interpretation for the Lagrange multiplier $\tilde{\lambda}$?

Recall that we have obtained the relation that

$$\begin{aligned}\tilde{\lambda} &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}R(\hat{\beta}-\tilde{\beta}) \\ &= -[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r).\end{aligned}$$

Thus, $\tilde{\lambda}$ is an indicator of the departure of $R\hat{\beta}$ from r. That is, the value of $\tilde{\lambda}$ will indicate whether $R\hat{\beta}-r$ is significantly different from zero.

Question: What happens to the distribution of F when $n \to \infty$?

Recall the important property of the $F_{p,q}$ distribution that $p \cdot F_{p,q} \xrightarrow{d} \chi^2_p$ when $q \to \infty$. Since our F-statistic for $\mathbf{H}_0$ follows a $F_{J,n-K}$ distribution, it follows that under $\mathbf{H}_0$, the quadratic form

$$J \cdot F = (R\hat{\beta} - r)' \left[s^2 R(\mathbf{X}'\mathbf{X})^{-1} R'\right]^{-1} (R\hat{\beta} - r) \xrightarrow{d} \chi^2_J$$

as $n \to \infty$. We formally state this result below.

Theorem 3.13: *Suppose Assumptions 3.1, 3.3(a) and 3.5 hold. Then under* $\mathbf{H}_0$, *we have the Wald test statistic*

$$W = \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{s^2} \xrightarrow{d} \chi^2_J$$

as $n \to \infty$.

This result implies that when n is sufficiently large, using the F-statistic and the exact $F_{J,n-K}$ distribution and using the quadratic form W and the simpler χ^2_J approximation will make no essential difference in statistical inference.

3.8 Applications

We now consider some special but important cases often encountered in economics and finance.

Case I: Testing for the Joint Significance of Explanatory Variables

Consider a linear regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0^o + \sum_{j=1}^{k} \beta_j^o X_{jt} + \varepsilon_t. \end{aligned}$$

We are interested in testing the combined effect of all the regressors except the intercept. The null hypothesis is

$$\mathbf{H}_0 : \beta_j^o = 0 \text{ for } 1 \le j \le k,$$

which implies that none of the explanatory variables influences Y_t.
The alternative hypothesis is

$$\mathbb{H}_A : \ \beta_j^o \neq 0 \text{ at least for some } \beta_j^o, \qquad j = 1, \cdots, k.$$

One can use the F-test and

$$F \sim F_{k,n-(k+1)}.$$

In fact, the restricted model under $\mathbf{H}_0$ is very simple:

$$Y_t = \beta_0^o + \varepsilon_t.$$

The restricted OLS estimator $\tilde{\beta} = (\bar{Y}, 0, \cdots, 0)'$. It follows that

$$\tilde{e} = Y - \mathbf{X}\tilde{\beta} = Y - \bar{Y}.$$

Hence, we have

$$\tilde{e}'\tilde{e} = (Y - \bar{Y})'(Y - \bar{Y}).$$

Recall the definition of R^2 :

$$\begin{aligned} R^2 &= 1 - \frac{e'e}{(Y - \bar{Y})'(Y - \bar{Y})} \\ &= 1 - \frac{e'e}{\tilde{e}'\tilde{e}}. \end{aligned}$$

It follows that

$$\begin{aligned} F &= \frac{(\tilde{e}'\tilde{e} - e'e)/k}{e'e/(n-k-1)} \\ &= \frac{(1 - \frac{e'e}{\tilde{e}'\tilde{e}})/k}{\frac{e'e}{\tilde{e}'\tilde{e}}/(n-k-1)} \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)}. \end{aligned}$$

Thus, it suffices to run one regression, namely the unrestricted model in this case. We emphasize that this formula is valid only when one is testing for $\mathbf{H}_0 : \beta_j^o = 0$ for all $1 \leq j \leq k$.

Example 1 [Efficient Market Hypothesis]: Suppose Y_t is the exchange rate return in period t, and I_{t-1} is the information available at time $t-1$. Then a classical version of the efficient market hypothesis (EMH) can be stated as follows:

$$E(Y_t|I_{t-1}) = E(Y_t)$$

To check whether exchange rate changes are unpredictable using the past history of exchange rate changes, we specify a linear regression model:

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

where

$$X_t = (1, Y_{t-1}, ..., Y_{t-k})'.$$

Under EMH, we have

$$\mathbf{H}_0 : \beta_j^o = 0 \text{ for all } j = 1, ..., k.$$

If the alternative

$$\mathbb{H}_A : \beta_j^o \neq 0 \text{ at least for some } j \in \{1, ..., k\}$$

holds, then exchange rate changes are predictable using the past information.

Question: What is the appropriate interpretation if $\mathbf{H}_0$ is not rejected?

Note that there exists a gap between the efficiency hypothesis and $\mathbf{H}_0$, because the linear regression model is just one of many ways to check EMH. Thus, $\mathbf{H}_0$ is not rejected, at most we can only say that no evidence against the efficiency hypothesis is found. We should not conclude that EMH holds.

Strictly speaking, the current theory (Assumption 3.2: $E(\varepsilon_t|\mathbf{X}) = 0$) rules out this application, which is a dynamic time series regression model. However, we will justify in Chapter 5 that

$$\begin{aligned} k \cdot F &= \frac{R^2}{(1-R^2)/(n-k-1)} \\ &\xrightarrow{d} \chi_k^2 \end{aligned}$$

under conditional homoskedasticity even for a linear dynamic regression model.

In fact, we can use a simpler version when n is large:

$$(n-k-1)R^2 \xrightarrow{d} \chi_k^2.$$

This follows from the Slutsky theorem because $R^2 \xrightarrow{p} 0$ under $\mathbf{H}_0$. Although Assumption 3.5 is not needed for this result, conditional homoskedasticity is still needed, which rules out autoregressive conditional heteroskedasticity (ARCH) in the time series context.

Below is a concrete numerical example.

Example 2 [Consumption Function and Wealth Effect]: Let Y_t = consumption, X_{1t} = labor income, X_{2t} = liquidity asset wealth. A regression estimation gives

$$\begin{aligned} Y_t &= 33.88 - 26.00X_{1t} + 6.71X_{2t} + e_t, \qquad R^2 = 0.742, n = 25. \\ &\quad [1.77] \qquad [-0.74] \qquad [0.77] \end{aligned}$$

where the numbers inside $[\cdot]$ are t-statistics.

Suppose we are interested in whether labor income or liquidity asset wealth has impact on consumption. We can use the F-test statistic,

$$\begin{aligned} F &= \frac{R^2/2}{(1-R^2)/(n-3)} \\ &= (0.742/2)/[(1-0.742)/(25-3)] \\ &= 31.636 \\ &\sim F_{2,22} \end{aligned}$$

Comparing it with the critical value of $F_{2,22}$ at the 5% significance level, we reject the null hypothesis that neither income nor liquidity asset has impact on consumption at the 5% significance level.

Case II: Testing for Omitted Variables (or Testing for No Effect)

Suppose $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, where $\mathbf{X}^{(1)}$ is a $n \times (k_1 + 1)$ matrix and $\mathbf{X}^{(2)}$ is a $n \times k_2$ matrix.

A random vector $X_t^{(2)}$ has no explanatory power for the conditional expectation of Y_t if

$$E(Y_t|X_t) = E(Y_t|X_t^{(1)}).$$

Alternatively, it has explanatory power for the conditional expectation of Y_t if

$$E(Y_t|X_t) \neq E(Y_t|X_t^{(1)}).$$

When $X_t^{(2)}$ has explaining power for Y_t but is not included in the regression, we say that $X_t^{(2)}$ is an omitted random variable or vector.

Question: How to test whether $X_t^{(2)}$ is an omitted variable in the linear regression context?

Consider the restricted model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_{k_1} X_{k_1 t} + \varepsilon_t.$$

Suppose we have additional k_2 variables $(X_{(k_1+1)t}, \cdots, X_{(k_1+k_2)t})$, and so we consider the unrestricted regression model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{1t} + ... + \beta_{k_1} X_{k_1 t} \\ &\quad + \beta_{k_1+1} X_{(k_1+1)t} + \cdots + \beta_{(k_1+k_2)} X_{(k_1+k_2)t} + \varepsilon_t. \end{aligned}$$

The null hypothesis is that the additional variables have no effect on Y_t. If this is the case, then

$$\mathbf{H}_0 : \beta_{k_1+1} = \beta_{k_1+2} = \cdots = \beta_{k_1+k_2} = 0.$$

The alternative is that at least one of the additional variables has effect on Y_t.

The F-Test statistic is

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/k_2}{e'e/(n - k_1 - k_2 - 1)} \sim F_{k_2, n-(k_1+k_2+1)}.$$

Question: Suppose we reject the null hypothesis. Then some important explanatory variables are omitted, and they should be included in the regression. On the other hand, if the F-test statistic does not reject the null hypothesis $\mathbf{H}_0$, can we say that there is no omitted variable?

No. There may exist a nonlinear relationship for additional variables which a linear regression specification cannot capture.

Example 3 [Testing for the Effect of Reforms]:

Consider the extended production function

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) \\ &\quad + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t, \end{aligned}$$

where AU_t is the autonomy dummy, PS_t is the profit sharing ratio, and CM_t is the dummy for change of manager. The null hypothesis of interest here is that none of the three reforms has impact:

$$\mathbf{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0.$$

We can use the F-test, and $F \sim F_{3,n-6}$ under $\mathbf{H}_0$.

Suppose rejection occurs. Then there exists evidence against $\mathbf{H}_0$. However, if no rejection occurs, then we can only say that we find no evidence against $\mathbf{H}_0$ (which is not the same as the statement that reforms have no effect). It is possible that the effect of $X_t^{(2)}$ is of nonlinear form. In this case, we may obtain a zero coefficient for $X_t^{(2)}$, because the linear specification may not be able to capture it.

Example 4 [Testing for Granger Causality]:

Consider two time series $\{Y_t, Z_t\}$, where t is the time index, $I_{t-1}^Y = \{Y_{t-1}, ..., Y_1\}$ and $I_{t-1}^Z = \{Z_{t-1}, ..., Z_1\}$. For example, Y_t is the GDP growth, and Z_t is the money supply growth. We say that Z_t does not Granger-cause Y_t in conditional mean with respect to $I_{t-1} = \{I_{t-1}^{(Y)}, I_{t-1}^{(Z)}\}$ if

$$E(Y_t | I_{t-1}^{(Y)}, I_{t-1}^{(Z)}) = E(Y_t | I_{t-1}^{(Y)}).$$

In other words, the lagged variables of Z_t have no impact on the level of Y_t.

In time series analysis, Granger causality is defined in terms of incremental predictability rather than the real cause-effect relationship. From an econometric point of view, it is a test of omitted variables in a time series context. It is first introduced by Granger (1969).

Question: How to test Granger causality?

Consider now a linear regression model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} \\ &\quad + \beta_{p+1} Z_{t-1} + \cdots + \beta_{p+q} Z_{t-q} + \varepsilon_t. \end{aligned}$$

Under non-Granger causality, we have

$$\mathbf{H}_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0.$$

The F-test statistic

$$F \sim F_{q,n-(p+q+1)}.$$

The current econometric theory (Assumption 3.2: $E(\varepsilon_t|\mathbf{X}) = 0$) actually rules out this application, because it is a dynamic regression model. However, we will justify in Chapter 5 that under H_0,

$$q \cdot F \xrightarrow{d} \chi^2_q$$

as $n \to \infty$ under conditional homoskedasticity even for a linear dynamic regression model.

Example 5 [Testing for Structural Change (or testing for regime shift)]
Consider a bivariate regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t,$$

where t is a time index, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent. Suppose there exist changes after $t = t_0$, i.e., there exist structural changes. We can consider the extended regression model:

$$\begin{aligned} Y_t &= (\beta_0 + \alpha_0 D_t) + (\beta_1 + \alpha_1 D_t) X_{1t} + \varepsilon_t \\ &= \beta_0 + \beta_1 X_{1t} + \alpha_0 D_t + \alpha_1 (D_t X_{1t}) + \varepsilon_t, \end{aligned}$$

where $D_t = 1$ if $t > t_0$ and $D_t = 0$ otherwise. The variable D_t is called a dummy variable, indicating whether it is a pre- or post-structral break period.

The null hypothesis of no structral change is

$$\mathbf{H}_0 : \alpha_0 = \alpha_1 = 0.$$

The alternative hypothesis that there exists a structral change is

$$\mathbb{H}_A : \alpha_0 \neq 0 \text{ or } \alpha_1 \neq 0.$$

The F-test statistic

$$F \sim F_{2,n-4}.$$

The idea of such a test is first proposed by Chow (1960).

Case III: Testing for linear restrictions

Example 6 [Testing for CRS]:

Consider the extended production function

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t.$$

We will test the null hypothesis of CRS:

$$\mathbf{H}_0 : \beta_1 + \beta_2 = 1.$$

The alternative hypothesis is

$$\mathbf{H}_0 : \beta_1 + \beta_2 \neq 1.$$

What is the restricted model under $\mathbf{H}_0$? It is given by

$$\ln(Y_t) = \beta_0 + \beta_1 \ln(L_t) + (1 - \beta_1) \ln(K_t) + \beta_3 AU_t + \beta_4 PS_t + \beta_5 CM_t + \varepsilon_t$$

or equivanetly

$$\ln(Y_t/K_t) = \beta_0 + \beta_1 \ln(L_t/K_t) + \beta_3 AU_t + \beta_4 CON_t + \beta_5 CM_t + \varepsilon_t.$$

The F-test statistic

$$F \sim F_{1,n-6}.$$

Because there is only one restriction, both t- and F- tests are applicable to test CRS.

Example 7 [Wage Determination]: Consider the wage function

$$\begin{aligned} W_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 U_t \\ &\quad + \beta_4 V_t + \beta_5 W_{t-1} + \varepsilon_t, \end{aligned}$$

where W_t = wage, P_t = price, U_t = unemployment, and V_t = unfilled vacancies.

We will test the null hypothesis

$$\mathbf{H}_0 : \beta_1 + \beta_2 = 0, \beta_3 + \beta_4 = 0, \text{ and } \beta_5 = 1.$$

Question: What is the economic interpretation of the null hypothesis $\mathbf{H}_0$?

Under $\mathbf{H}_0$, we have the restricted wage equation:

$$\Delta W_t = \beta_0 + \beta_1 \Delta P_t + \beta_4 D_t + \varepsilon_t,$$

where $\Delta W_t = W_t - W_{t-1}$ is the wage growth rate, $\Delta P_t = P_t - P_{t-1}$ is the inflation rate, and $D_t = V_t - U_t$ is an index for job market situation (excess job supply). This implies that the wage increase depends on the inflation rate and the excess labor supply.

The F-test statistic for $\mathbf{H}_0$ is

$$F \sim F_{3,n-6}.$$

Case IV: Testing for Near-Multicollinearity

Example 8 [Consumption Function (Cont.)]:

Consider the following estimation results for three separate regressions based on the same data set with $n = 25$. The first is a regression of consumption on income:

$$\begin{aligned} Y_t &= 36.74 + 0.832 X_{1t} + e_{1t}, \qquad R^2 = 0.735 \\ &\quad [1.98][7.98] \end{aligned}$$

The second is a regression of consumption on wealth:

$$\begin{aligned} Y_t &= 36.61 + 0.208 X_{2t} + e_{2t}, \qquad R^2 = 0.735 \\ &\quad [1.97][7.99] \end{aligned}$$

The third is a regression of consumption on both income and wealth:

$$\begin{aligned} Y &= 33.88 - 26.00 X_{1t} + 6.71 X_{2t} + e_t, \qquad R^2 = 0.742, \\ &\quad [1.77][-0.74][0.77] \end{aligned}$$

Note that in the first two separate regressions, we can find significant t-test statistics for income and wealth, but in the third joint regression, both income and wealth are

insignificant. This may be due to the fact that income and wealth are highly multicolinear! To test neither income nor wealth has impact on consumption, we can use the F-test:

$$\begin{aligned} F &= \frac{R^2/2}{(1-R^2)/(n-3)} \\ &= \frac{0.742/2}{(1-0.742)/(25-3)} \\ &= 31.636 \\ &\sim F_{2,22}\,. \end{aligned}$$

This F-test shows that the null hypothesis is firmly rejected at the 5% significance level, because the critical value of $F_{2,22}$ at the 5% level is 3.44.

3.9 Generalized Least Squares (GLS) Estimation

Question: The classical linear regression theory crucially depends on the assumption that $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$, or equivalently $\{\varepsilon_t\} \sim i.i.d.N(0, \sigma^2)$, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent. What may happen if some classical assumptions do not hold?

Question: Under what conditions, the existing procedures and results are still approximately true?

Assumption 3.5 is unrealistic for many economic and financial data. Suppose Assumption 3.5 is replaced by the following condition:

Assumption 3.6: $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 V)$, *where* $0 < \sigma^2 < \infty$ *is unknown and* $V = V(\mathbf{X})$ *is a known* $n \times n$ *symmetric, finite and positive definite matrix.*

Remarks:

Assumption 3.6 implies that

$$\begin{aligned} \text{var}(\varepsilon|\mathbf{X}) &= E(\varepsilon\varepsilon'|\mathbf{X}) \\ &= \sigma^2 V = \sigma^2 V(\mathbf{X}) \end{aligned}$$

is known up to a constant σ^2. It allows for conditional heteroskedasticity of known form.

In Assumption 3.6, it is possible that V is not a diagonal matrix. Thus, $\text{cov}(\varepsilon_t, \varepsilon_s|\mathbf{X})$ may not be zero. In other words, Assumption 3.6 allows conditional autocorrelation of known form. If t is a time index, this implies that there exists serial correlation of

known form. If t is an index for cross-sectional units, this implies that there exists spatial correlation of known form.

However, the assumption that V is known is still very restrictive from a practical point of view. In practice, V usually has an unknown form.

Question: What is the statistical property of OLS $\hat{\beta}$ under Assumption 3.6?

Theorem 3.14: *Suppose Assumptions 3.1, 3.3(a) and 3.6 hold. Then*
(i) unbiasedness: $E(\hat{\beta}|\mathbf{X}) = \beta^o$.
(ii) variance:

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

(iii)

$$(\hat{\beta} - \beta^o)|\mathbf{X} \sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

(iv) $cov(\hat{\beta}, e|\mathbf{X}) \neq 0$ *in general.*

Proof: (i) Using $\hat{\beta} - \beta^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, we have

$$\begin{aligned} E[(\hat{\beta} - \beta^o)|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 \\ &= 0. \end{aligned}$$

(ii)

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= E[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Note that we cannot further simplify the expression here because $V \neq I$.
(iii) Because

$$\begin{aligned} \hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= \sum_{t=1}^{n} C_t\varepsilon_t, \end{aligned}$$

where the weighting vector

$$C_t = (\mathbf{X}'\mathbf{X})^{-1}X_t,$$

$\hat{\beta} - \beta^o$ follows a normal distribution given $\mathbf{X}$, because it is a sum of a normal random variables. As a result,

$$\hat{\beta} - \beta^o \sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

(iv)

$$\begin{aligned}
\text{cov}(\hat{\beta}, e|\mathbf{X}) &= E[(\hat{\beta} - \beta^o)e'|\mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'M|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon'|\mathbf{X})M \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'VM \\
&\neq 0
\end{aligned}$$

because $\mathbf{X}'VM \neq 0$. We can see that it is conditional heteroskedasticity and/or autocorrelation in $\{\varepsilon_t\}$ that cause $\hat{\beta}$ to be correlated with e.

Remarks:

OLS $\hat{\beta}$ is still unbiased and one can show that its variance goes to zero as $n \to \infty$ (see Question 6, Problem Set 03). Thus, it converges to β^o in the sense of MSE.

However, the variance of the OLS estimator $\hat{\beta}$ does no longer have the simple expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ under Assumption 3.6. As a consequence, the classical t- and F-test statistics are invalid because they are based on an incorrect variance-covariance matrix of $\hat{\beta}$. That is, they use an incorrect expression of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ rather than the correct variance formula of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem (iv) implies that even if we can obtain a consistent estimator for $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and use it to construct tests, we can no longer obtain the Student t-distribution and F-distribution, because the numerator and the denominator in defining the t- and F-test statistics are no longer independent.

Generalized Least Squares (GLS) Estimation

To introduce GLS, we first state a useful lemma.

Lemma 3.15: *For any symmetric positive definite matrix V, we can always write*

$$\begin{aligned}
V^{-1} &= C'C, \\
V &= C^{-1}(C')^{-1}
\end{aligned}$$

where C is a $n \times n$ nonsingular matrix.

Question: What is this decomposition called? Note that C may not be symmetric.

Consider the original linear regression model:

$$Y = \mathbf{X}\beta^o + \varepsilon.$$

If we mutitiply the equation by C, we obtain the transformed regression model

$$\begin{aligned} CY &= (C\mathbf{X})\beta^o + C\varepsilon, \text{ or} \\ Y^* &= \mathbf{X}^*\beta^o + \varepsilon^*, \end{aligned}$$

where $Y^* = CY, \mathbf{X}^* = C\mathbf{X}$ and $\varepsilon^* = C\varepsilon$. Then the OLS of this transformed model

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y^* \\ &= (\mathbf{X}'C'C\mathbf{X})^{-1}(\mathbf{X}'C'CY) \\ &= (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y \end{aligned}$$

is called the Generalized Least Squares (GLS) estimator.

Question: What is the nature of GLS?

Observe that

$$\begin{aligned} E(\varepsilon^*|\mathbf{X}) &= E(C\varepsilon|\mathbf{X}) \\ &= CE(\varepsilon|\mathbf{X}) \\ &= C \cdot 0 \\ &= 0. \end{aligned}$$

Also, note that

$$\begin{aligned} \text{var}(\varepsilon^*|\mathbf{X}) &= E[\varepsilon^*\varepsilon^{*\prime}|\mathbf{X}] \\ &= E[C\varepsilon\varepsilon'C'|\mathbf{X}] \\ &= CE(\varepsilon\varepsilon'|\mathbf{X})C' \\ &= \sigma^2 CVC' \\ &= \sigma^2 C[C^{-1}(C')^{-1}]C' \\ &= \sigma^2 I. \end{aligned}$$

It follows from Assumption 3.6 that

$$\varepsilon^*|\mathbf{X} \sim N(0, \sigma^2 I).$$

The transformation makes the new error ε^* conditionally homoskedastic and serially uncorrelated, while maintaining the normality distribution. Suppose that for t, ε_t has a large variance σ_t^2. The transformation $\varepsilon_t^* = C\varepsilon_t$ will discount ε_t by dividing it by its conditional standard deviation so that ε_t^* becomes conditionally homoskedastic. In addition, the transformation also removes possible correlation between ε_t and $\varepsilon_s, t \neq s$. As a consequence, GLS becomes the best linear LS estimator for β^o in term of the Gauss-Markov theorem.

To appreciate how the transformation by matrix C removes conditional heteroskedasticity and eliminates serial correlation, we now consider two examples.

Example 1 [Removing Heteroskedasticity]: Suppose

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \sigma_n^2 \end{bmatrix},$$

Then

$$C = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \sigma_n^{-1} \end{bmatrix}$$

where $\sigma_i^2 = \sigma_i^2(\mathbf{X}), i = 1, ..., n$, and

$$\varepsilon^* = C\varepsilon = \begin{bmatrix} \frac{\varepsilon_1}{\sigma_1} \\ \frac{\varepsilon_2}{\sigma_2} \\ \cdots \\ \frac{\varepsilon_n}{\sigma_n} \end{bmatrix}.$$

The transformed regression model is

$$Y_t^* = X_t^{*\prime}\beta^o + \varepsilon_t^*, \qquad t = 1, ..., n,$$

where

$$\begin{aligned} Y_t^* &= Y_t/\sigma_t, \\ X_t^* &= X_t/\sigma_t, \\ \varepsilon_t^* &= \varepsilon_t/\sigma_t. \end{aligned}$$

Example 2 [Eliminating Serial Correlation] Suppose

$$V = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-3} & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-4} & \rho^{n-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & \rho & 1 \end{bmatrix}.$$

Then we have

$$V^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & \rho^{n-3} & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & \rho^{n-4} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

and

$$C = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

It follows that

$$\varepsilon^* = C\varepsilon = \begin{bmatrix} \sqrt{1-\rho^2}\varepsilon_1 \\ \varepsilon_2 - \rho\varepsilon_1 \\ \cdots \\ \varepsilon_n - \rho\varepsilon_{n-1} \end{bmatrix}.$$

The transformed regression model is

$$Y_t^* = X_t^{*\prime}\beta^o + \varepsilon_t^*, t = 1, ..., n,$$

where

$$\begin{aligned} Y_1^* &= \sqrt{1-\rho^2}Y_1, \qquad Y_t^* = Y_t - \rho Y_{t-1}, t = 2, ..., n, \\ X_1^* &= \sqrt{1-\rho^2}X_1, \qquad X_t^* = X_t - \rho X_{t-1}, t = 2, ..., n, \\ \varepsilon_1^* &= \sqrt{1-\rho^2}\varepsilon_1, \qquad \varepsilon_t^* = \varepsilon_t - \rho\varepsilon_{t-1}, t = 2, ..., n. \end{aligned}$$

The $\sqrt{1-\rho^2}$transformation for $t = 1$ is called the Prais-Winsten transformation.

Theorem 3.16: *Under Assumptions* $3.1, 3.3(a)$ *and 3.6,*

(i) $E(\hat{\beta}^*|\mathbf{X}) = \beta^o$;
(ii) $var(\hat{\beta}^*|\mathbf{X}) = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'V^{-1}\mathbf{X})^{-1}$;
(iii) $cov(\hat{\beta}^*, e^*|\mathbf{X}) = 0$, *where* $e^* = Y^* - \mathbf{X}^*\hat{\beta}^*$;
(iv) $\hat{\beta}^*$ *is BLUE.*
(v) $E(s^{*2}|\mathbf{X}) = \sigma^2$, *where* $s^{*2} = e^{*\prime}e^*/(n-K)$.

Proof: Results in (i)–(iii) follow because the GLS is the OLS of the transformed model.

(iv) The transformed model satisfies 3.1, 3.3 and 3.5 of the classical regression assumptions with $\varepsilon^*|\mathbf{X}^* \sim N(0, \sigma^2 I_n)$. It follows that GLS is BLUE by the Gauss-Markov theorem. Result (v) also follows immediately. This completes the proof.

Remarks:

Because $\hat{\beta}^*$ is the OLS of the transformed regression model with i.i.d. $N(0, \sigma^2 I)$ errors, the t-test and F-test are applicable, and these test statistics are defined as follows:

$$\begin{aligned} T^* &= \frac{R\hat{\beta}^* - r}{\sqrt{s^{*2}R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R'}} \sim t_{n-K}, \\ F^* &= \frac{(R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r)/J}{s^{*2}} \\ &\sim F_{J,n-K}. \end{aligned}$$

It is very important to note that we still have to estimate the proportionality σ^2 in spite of the fact that $V = V(X)$ is known.

When testing whether all coefficients except the intercept are jointly zero, we have $(n-K)R^{*2} \xrightarrow{d} \chi^2_k$.

Because GLS $\hat{\beta}^*$ is BLUE and OLS $\hat{\beta}$ differs from $\hat{\beta}^*$, OLS $\hat{\beta}$ cannot be BLUE.

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y^*, \\ &= (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y. \\ \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \end{aligned}$$

In fact, the most important message of GLS is the insight it provides into the impact of conditional heteroskedasticity and serial correlation on the estimation and inference of the linear regression model. In practice, GLS is generally not feasible, because the $n \times n$ matrix V is of unknown form, where $\text{var}(\varepsilon|\mathbf{X}) = \sigma^2 V$.

Question: What are feasible solutions?

Two Approaches

(i) First Approach: Adaptive feasible GLS

In some cases with additional assumptions, we can use a nonparametric estimator $\hat{V}$ to replace the unknown V, we obtain the adaptive feasible GLS

$$\hat{\beta}_a^* = (\mathbf{X}'\hat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{V}^{-1}Y,$$

where $\hat{V}$ is an estimator for V. Because V is an $n \times n$ unknown matrix and we only have n data points, it is impossible to estimate V consistently using a sample of size n if we do not impose any restriction on the form of V. In other words, we have to impose some restrictions on V in order to estimate it consistently. For example, suppose we assume

$$\begin{aligned} \sigma^2 V &= \text{diag}\{\sigma_1^2(\mathbf{X}), ..., \sigma_n^2(\mathbf{X})\} \\ &= \text{diag}\{\sigma^2(X_1), ..., \sigma^2(X_n)\}, \end{aligned}$$

where $\text{diag}\{\cdot\}$ is a $n \times n$ diagonal matrix and $\sigma^2(X_t) = E(\varepsilon_t^2|X_t)$ is unknown. The fact that $\sigma^2 V$ is a diagonal matrix can arise when $\text{cov}(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$, i.e., when there is no serial correlation. Then we can use the nonparametric kernel estimator

$$\begin{aligned} \hat{\sigma}^2(x) &= \frac{\frac{1}{n}\sum_{t=1}^n e_t^2 \frac{1}{b}K\left(\frac{x-X_t}{b}\right)}{\frac{1}{n}\sum_{t=1}^n \frac{1}{b}K\left(\frac{x-X_t}{b}\right)} \\ &\xrightarrow{p} \sigma^2(x), \end{aligned}$$

where e_t is the estimated OLS residual, and $K(\cdot)$ is a kernel function which is a specified symmetric density function (e.g., $K(u) = (2\pi)^{-1/2}\exp(-\frac{1}{2}u^2)$ if x is a scalar), and $b = b(n)$ is a bandwidth such that $b \to 0, nb \to \infty$ as $n \to \infty$. The finite sample distribution of $\hat{\beta}_a^*$ will be different from the finite sample distribution of $\hat{\beta}^*$, which assumes that V were known. This is because the sampling errors of the estimator $\hat{V}$ have some impact on the estimator $\hat{\beta}_a^*$. However, under some suitable conditions on $\hat{V}$, $\hat{\beta}_a^*$ will share the same asymptotic property as the infeasible GLS $\hat{\beta}^*$ (i.e., the MSE of $\hat{\beta}_a^*$ is approximately equal to the MSE of $\hat{\beta}^*$). In other words, the first stage estimation of $\sigma^2(\cdot)$ has no impact on the asymptotic distribution of $\hat{\beta}_a^*$. For more discussion, see Robinson (1988) and Stinchcombe and White (1991).

(ii) Second Approach

Continue to use OLS $\hat{\beta}$, obtaining the correct formula for

$$\text{var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

as well as a consistent estimator for $\text{var}(\hat{\beta}|\mathbf{X})$. The classical definitions of t and F-tests cannot be used, because they are based on an incorrect formula for $\text{var}(\hat{\beta}|\mathbf{X})$. However,

some modified tests can be obtained by using a consistent estimator for the correct formula for $\text{var}(\hat{\beta}|\mathbf{X})$. The trick is to estimate $\sigma^2\mathbf{X}'V\mathbf{X}$, which is a $K \times K$ unknown matrix, rather than to estimate V, which is a $n \times n$ unknown matrix. However, only asymptotic distributions can be used in this case.

Question: Suppose we assume

$$\begin{aligned} E(\varepsilon\varepsilon'|\mathbf{X}) &= \sigma^2 V \\ &= \text{diag}\{\sigma_1^2(\mathbf{X}), ..., \sigma_n^2(\mathbf{X})\}. \end{aligned}$$

As pointed out earlier, this essentially assumes $E(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$. That is, there is no serial correlation in $\{\varepsilon_t\}$ conditional on $\mathbf{X}$. Instead of estimating $\sigma_t^2(\mathbf{X})$, one can estimate the $K \times K$ matrix $\sigma^2\mathbf{X}'V\mathbf{X}$ directly.

Then, how to estimate

$$\sigma^2\mathbf{X}'V\mathbf{X} = \sum_{t=1}^{n} X_t X_t' \sigma_t^2(\mathbf{X})?$$

We can use the folllowing estimator

$$\mathbf{X}'\text{D}(e)\text{D}(e)'\mathbf{X} = \sum_{t=1}^{n} X_t X_t' e_t^2,$$

where $\text{D}(e) = \text{diag}(e_1, ..., e_n)$ is a $n \times n$ diagonal matrix with all off-diagonal elements being zero. This is called White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator. See more discussion in Chapter 4.

Question: For $J = 1$, do we have

$$\frac{R\hat{\beta} - r}{\sqrt{R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{D}(e)\text{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'}} \sim t_{n-K}?$$

For $J > 1$, do we have

$$\begin{aligned} & (R\hat{\beta} - r)' \left[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{D}(e)\text{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} (R\hat{\beta} - r)/J \\ \sim \;& F_{J,n-K}? \end{aligned}$$

No. Although we have standardized both test statistics by the correct variance estimators, we still have $\text{cov}(\hat{\beta}, e|\mathbf{X}) \neq 0$ under Assumption 3.6. This implies that $\hat{\beta}$ and e are not independent, and therefore, we no longer have a t-distribution or an F-distribution in finite samples.

However, when $n \to \infty$, we have
(i) Case I ($J = 1$) :

$$\frac{R\hat{\beta} - r}{\sqrt{R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0,1).$$

This can be called a robust t-test.
(ii) Case II ($J > 1$) :

$$(R\hat{\beta} - r)' \left[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} (R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

This is a robust Wald test statistic.

The above two feasible solutions are based on the assumption that $E(\varepsilon_t\varepsilon_s|\mathbf{X}) = 0$ for all $t \neq s$.

In fact, we can also consistently estimate the limit of $X'VX$ when there exists conditional heteroskedasticity and autocorrelation. This is called heteroskedasticity and autocorrelation consistent variance-covariance estimation. When there exists serial correlation of unknown form, an alternative solution should be provided. This is discussed in Chapter 6. See also Andrews (1991) and Newey and West (1987, 1994).

3.10 Conclusion

In this chapter, we have presented the econometric theory for the classical linear regression models. We first provide and discuss a set of assumptions on which the classical linear regression model is built. This set of regularity conditions will serve as the starting points from which we will develop modern econometric theory for linear regression models.

We derive the statistical properties of the OLS estimator. In particular, we point out that R^2 is not a suitable model selection criterion, because it is always nondecreasing with the dimension of regressors. Suitable model selection criteria, such as AIC and BIC, are discussed. We show that conditional on the regressor matrix $\mathbf{X}$, the OLS estimator $\hat{\beta}$ is unbiased, has a vanishing variance, and is BLUE. Under the additional conditional normality assumption, we derive the finite sample normal distribution for $\hat{\beta}$, the Chi-squared distribution for $(n-K)s^2/\sigma^2$, as well as the independence between $\hat{\beta}$ and s^2.

Many hypotheses encountered in economics can be formulated as linear restrictions on model parameters. Depending on the number of parameter restrictions, we derive the t-test and the F-test. In the special case of testing the hypothesis that all slope

coefficients are jointly zero, we also derive an asymptotically Chi-squared test based on R^2.

When there exist conditional heteroskedasticity and/or autocorrelation, the OLS estimator is still unbiased and has a vanishing variance, but it is no longer BLUE, and $\hat{\beta}$ and s^2 are no longer mutually independent. Under the assumption of a known variance-covariance matrix up to some scale parameter, one can transform the linear regression model by correcting conditional heteroskedasticity and eliminating autocorrelation, so that the transformed regression model has conditionally homoskedastic and uncorrelated errors. The OLS estimator of this transformed linear regression model is called the GLS estimator, which is BLUE. The t-test and F-test are applicable. When the variance-covariance structure is unknown, the GLS estimator becomes infeasible. However, if the error in the original linear regression model is serially uncorrelated (as is the case with independent observations across t), there are two feasible solutions. The first is to use a nonparametric method to obtain a consistent estimator for the conditional variance $\text{var}(\varepsilon_t|X_t)$, and then obtain a feasible plug-in GLS. The second is to use White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator for the OLS estimator $\hat{\beta}$. Both of these two methods are built on the asymptotic theory. When the error of the original linear regression model is serially correlated, a feasible solution to estimate the variance-covariance matrix is provided in Chapter 6.

EXERCISES

3.1. Consider a bivariate linear regression model

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

where $X_t = (X_{0t}, X_{1t})' = (1, X_{1t})'$, and ε_t is a regression error.

(a) Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ be the OLS estimator. Show that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$, and

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1)(Y_t - \bar{Y})}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2} \\ &= \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1) Y_t}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2} \\ &= \sum_{t=1}^n C_t Y_t, \end{aligned}$$

where $C_t = (X_{1t} - \bar{X}_1)/\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2$.

(b) Suppose $\mathbf{X} = (X_{11}, ..., X_{1n})'$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ are independent. Show that $\text{var}(\hat{\beta}_1 | \mathbf{X}) = \sigma_\varepsilon^2/[(n-1)S_{X_1}^2]$, where $S_{X_1}^2$ is the sample variance of $\{X_{1t}\}_{t=1}^n$ and σ_ε^2 is the variance of ε_t.Thus, the more variations in $\{X_{1t}\}$, the more accurate estimation for β_1^o.

(c) Let $\hat{\rho}$ denote the sample correlation between Y_t and X_{1t}; namely,

$$\hat{\rho} = \frac{\sum_{t=1}^n (X_{1t} - \bar{X}_1)(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2 \sum_{t=1}^n (Y_t - \bar{Y})^2}}.$$

Show that $R^2 = \hat{\rho}^2$.Thus, the squared sample correlation between Y and X_1 is the fraction of the sample variation in Y that can be predicted using the linear predictor of X_1.This resuly also implies that R^2 is a measure of the strength of sample linear association between Y_t and X_{1t}.

3.2. For the OLS estimation of the linear regression model $Y_t = X_t'\beta^o + \varepsilon$, where X_t is a $K \times 1$ vector, show $R^2 = \hat{\rho}_{Y\hat{Y}}^2$, the squared sample correlation between Y_t and $\hat{Y}_t$.

3.2. Suppose $X_t = Q$ for all $t \geq m$, where m is a fixed integer, and Q is a $K \times 1$ constant vector. Do we have $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$? Explain.

3.3. The adjusted R^2, denoted as $\bar{R}^2$, is defined as follows:

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{(Y - \bar{Y})'(Y - \bar{Y})/(n-1)}.$$

Show

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-K}(1 - R^2)\right].$$

3.4. [Effect of Multicolinearity] Consider a regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t.$$

Suppose Assumptions 3.1–3.4 hold. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$ be the OLS estimator. Show

$$\begin{aligned} \text{var}(\hat{\beta}_1|X) &= \frac{\sigma^2}{(1-\hat{r}^2)\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2}, \\ \text{var}(\hat{\beta}_2|X) &= \frac{\sigma^2}{(1-\hat{r}^2)\sum_{t=1}^n (X_{2t} - \bar{X}_2)^2}, \end{aligned}$$

where $\bar{X}_1 = n^{-1}\sum_{t=1}^n X_{1t}, \bar{X}_2 = n^{-1}\sum_{t=1}^n X_{2t}$, and

$$\hat{r}^2 = \frac{\left[\sum_{t=1}^n (X_{1t} - \bar{X}_1)(X_{2t} - \bar{X}_2)\right]^2}{\sum_{t=1}^n (X_{1t} - \bar{X}_1)^2 \sum_{t=1}^n (X_{2t} - \bar{X}_2)^2}.$$

3.5. Consider the linear regression model

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

where $X_t = (1, X_{1t}, ..., X_{kt})'$. Suppose Assumptions 31.–3.3 hold. Let R_j^2 is the coefficient of determination of regressing variable X_{jt} on all the other explanatory variables $\{X_{it}, 0 \le i \le k, i \ne j\}$. Show

$$\text{var}(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{(1-R_j^2)\sum_{t=1}^n (X_{jt} - \bar{X}_j)^2},$$

where $\bar{X}_j = n^{-1}\sum_{t=1}^n X_{jt}$. The factor $1/(1-R_j^2)$ is called the variance inflation factor (VIF); it is used to measure the degree of multicolinearity among explanatory variables in X_t.

3.6. Consider the following linear regression model

$$Y_t = X_t'\beta^o + u_t, \qquad t = 1, ..., n, \tag{4.1}$$

where

$$u_t = \sigma(X_t)\varepsilon_t,$$

where $\{X_t\}$ is a nonstochastic process, and $\sigma(X_t)$ is a positive function of X_t such that

$$\Omega = \begin{bmatrix} \sigma^2(X_1) & 0 & 0 & ... & 0 \\ 0 & \sigma^2(X_2) & 0 & ... & 0 \\ 0 & 0 & \sigma^2(X_3) & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \sigma^2(X_n) \end{bmatrix} = \Omega^{\frac{1}{2}}\Omega^{\frac{1}{2}},$$

with

$$\Omega^{\frac{1}{2}} = \begin{bmatrix} \sigma(X_1) & 0 & 0 & ... & 0 \\ 0 & \sigma(X_2) & 0 & ... & 0 \\ 0 & 0 & \sigma(X_3) & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \sigma(X_n) \end{bmatrix}.$$

Assume that $\{\varepsilon_t\}$ is i.i.d. $N(0,1)$. Then $\{u_t\}$ is i.i.d. $N(0, \sigma^2(X_t))$. This differs from Assumption 3.5 of the classical linear regression analysis, because now $\{u_t\}$ exhibits conditional heteroskedasticity.

Let $\hat{\beta}$ denote the OLS estimator for β^o.

(a) Is $\hat{\beta}$ unbiased for β^o?

(b) Show that $\text{var}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

Consider an alternative estimator

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}Y \\ &= \left[\sum_{t=1}^{n} \sigma^{-2}(X_t)X_tX_t'\right]^{-1} \sum_{t=1}^{n} \sigma^{-2}(X_t)X_tY_t. \end{aligned}$$

(c) Is $\tilde{\beta}$ unbiased for β^o?

(d) Show that $\text{var}(\tilde{\beta}) = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$.

(e) Is $\text{var}(\hat{\beta}) - \text{var}(\tilde{\beta})$ positive semi-definite (p.s.d.)? Which estimator, $\hat{\beta}$ or $\tilde{\beta}$, is more efficient?

(f) Is $\tilde{\beta}$ the Best Linear Unbiased Estimator (BLUE) for β^o? [Hint: There are several approaches to this question. A simple one is to consider the transformed model

$$Y_t^* = X_t^{*\prime}\beta^o + \varepsilon_t, \qquad t = 1, ..., n, \tag{4.2}$$

where $Y_t^* = Y_t/\sigma(X_t), X_t^* = X_t/\sigma(X_t)$. This model is obtained from model (4.1) after dividing by $\sigma(X_t)$. In matrix notation, model (4.2) can be written as

$$Y^* = \mathbf{X}^*\beta^o + \varepsilon,$$

where the $n \times 1$ vector $Y^* = \Omega^{-\frac{1}{2}}Y$ and the $n \times k$ matrix $\mathbf{X}^* = \Omega^{-\frac{1}{2}}\mathbf{X}$.]

(g) Construct two test statistics for the null hypothesis of interest $\mathbf{H}_0 : \beta_2^o = 0$. One test is based on $\hat{\beta}$, and the other test is based on $\tilde{\beta}$. What are the finite sample distributions of your test statistics under $\mathbf{H}_0$? Can you tell which test is better?

(h) Construct two test statistics for the null hypothesis of interest $\mathbf{H}_0 : R\beta^o = r$, where R is a $J \times k$ matrix with $J > 0$. One test is based on $\hat{\beta}$, and the other test is based on $\tilde{\beta}$. What are the finite sample distributions of your test statistics under $\mathbf{H}_0$?

3.7. Consider the following classical regression model

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

Suppose that we are interested in testing the null hypothesis

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

where R is a $J \times K$ matrix, and r is a $J \times 1$ vector. The F-test statistic is defined as

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/J}{e'e/(n-k-1)}.$$

where $e'e$ is the sum of squared residuals from the unrestricted model, and $\tilde{e}'\tilde{e}$ is the sum of squared residuals from the restricted regression model subject to the restriction $R\beta = r$.

3.8. The F-test statistic is defined as follows:

$$F = \frac{(R\hat{\beta} - r)'[R(\mathbf{X'X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that

$$\begin{aligned} F &= \frac{\sum_{t=1}^n (\hat{Y}_t - \tilde{Y}_t)^2/J}{s^2} \\ &= \frac{(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/J}{s^2}, \end{aligned}$$

where $\hat{Y}_t = X_t'\hat{\beta}, \tilde{Y}_t = X_t'\tilde{\beta}$, and $\hat{\beta}, \tilde{\beta}$ are the unrestricted and restricted OLS estimators respectively.

3.9. Consider the following classical regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0^o + \sum_{j=1}^k \beta_j^o X_{jt} + \varepsilon_t, \qquad t = 1, ..., n. \end{aligned} \tag{7.1}$$

Suppose that we are interested in testing the null hypothesis

$$\mathbf{H}_0 : \beta_1^o = \beta_2^o = \cdots = \beta_k^o = 0.$$

Then the F-statistic can be written as

$$F = \frac{(\tilde{e}'\tilde{e} - e'e)/k}{e'e/(n-k-1)}.$$

where $e'e$ is the sum of squared residuals from the unrestricted model (7.1), and $\tilde{e}'\tilde{e}$ is the sum of squared residuals from the restricted model (7.2)

$$Y_t = \beta_0^o + \varepsilon_t. \tag{7.2}$$

(a) Show that under Assumptions 3.1 and 3.3,

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)},$$

where R^2 is the coefficient of determination of the unrestricted model (7.1).

(b) Suppose in addition Assumption 3.5 holds. Show that under $\mathbf{H}_0$,

$$(n-k-1)R^2 \xrightarrow{d} \chi_k^2.$$

3.10. The F-test statistic is defined as folllows:

$$F = \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2}.$$

Show that F

$$F = \frac{(1/J)\sum_{t=1}^{n}(\hat{Y}_t - \tilde{Y}_t)^2}{s^2} = \frac{(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/J}{s^2},$$

where $\hat{Y}_t = X_t'\hat{\beta}, \tilde{Y}_t = X_t'\tilde{\beta}$, and $\hat{\beta}, \tilde{\beta}$ are the unrestricted and restricted OLS estimators respectively.

3.11. **[Structral Change]** Suppose Assumptions 3.1 and 3.3 hold. Consider the following model on the whole sample:

$$Y_t = X_t'\beta^o + (D_t X_t)'\alpha^o + \varepsilon_t, t = 1, ..., n,$$

where the time dummy variable $D_t = 0$ if $t \leq n_1$ and $D_t = 1$ if $t > n_1$. This model can be written as two separate models:

$$Y_t = X_t'\beta^o + \varepsilon_t, t = 1, ..., n_1$$

and

$$Y_t = X_t'(\beta^o + \alpha^o) + \varepsilon_t, t = n_1 + 1, ..., n.$$

Let SSR_u, SSR_1, SSR_2 denotes the sums of squared residuals of the above three regression models via OLS. Show

$$SSR_u = SSR_1 + SSR_2.$$

This identity implies that estimating the first regression mdoel with time dummy variable D_t via OLS is equivalent to estimating two separate regression models over two subsample periods respectively.

3.12. A quadratic polynomial regression moel

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t$$

is fit to data. Suppose the p-value for the OLS estimator of β_1 was 0.67 and for β_2 was 0.84. Can we accept the hypothesis that β_1 and β_2 are both 0? Explain.

3.13. Suppose $\mathbf{X}'\mathbf{X}$ is a $K \times K$ matrix, and V is a $n \times n$ matrix, and both $\mathbf{X}'\mathbf{X}$ and V are symmetric and nonsingular, with the minimum eigenvalue $\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$ and $0 < c \leq \lambda_{\max}(V) \leq C < \infty$. Show that for any $\tau \in R^K$ such that $\tau'\tau = 1$,

$$\tau' \text{var}(\hat{\beta}|\mathbf{X})\tau = \sigma^2 \tau'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\tau \to 0$$

as $n \to \infty$. Thus, $\text{var}(\hat{\beta}|\mathbf{X})$ vanishes to zero as $n \to \infty$ under conditional heteroskedasticity.

3.14. Suppose the conditions in 3.9 hold. It can be shown that the variances of the OLS $\hat{\beta}$ and GLS $\hat{\beta}^*$ are respectively:

$$\begin{aligned} \text{var}(\hat{\beta}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \\ \text{var}(\hat{\beta}^*|\mathbf{X}) &= \sigma^2(\mathbf{X}'V^{-1}\mathbf{X})^{-1}. \end{aligned}$$

Show that $\text{var}(\hat{\beta}|\mathbf{X}) - \text{var}(\hat{\beta}^*|\mathbf{X})$ is positive semi-definite.

3.15. Suppose a data generating process is given by

$$Y_t = \beta_1^o X_{1t} + \beta_2^o X_{2t} + \varepsilon_t = X_t'\beta^o + \varepsilon_t,$$

where $X_t = (X_{1t}, X_{2t})'$, $E(X_t X_t')$ is nonsingular, and $E(\varepsilon_t|X_t) = 0$. For simplicity, we further assume $E(X_{2t}) = 0$ and $E(X_{1t}X_{2t}) \neq 0$.

Now consider the following bivariate linear regression model

$$Y_t = \beta_1^o X_{1t} + u_t.$$

(a) Show that if $\beta_2^o \neq 0$, then $E(Y_1|X_t) = X_t'\beta^o \neq E(Y_{1t}|X_{1t})$. That is, there exists an omitted variable (X_{2t}) in the bivariate regression model.

(b) Show that $E(Y_t|X_{1t}) \neq \beta_1 X_{1t}$ for all β_1. That is, the bivariate linear regression model is misspecified for $E(Y_t|X_{1t})$.

(c) Is the best linear least squares approximation cofficient β_1^* in the bivariate linear regression model equal to β_1^o?

3.16. Suppose a data generating process is given by

$$Y_t = \beta_1^o X_{1t} + \beta_2^o X_{2t} + \varepsilon_t = X_t'\beta^o + \varepsilon_t,$$

where $X_t = (X_{1t}, X_{2t})'$, and Assumptions 3.1–3.4 hold. (For simplicity, we have assumed no intercept.) Denote the OLS estimator by $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$.

If $\beta_2^o = 0$ and we know it. Then we can consider a simpler regression

$$Y_t = \beta_1^o X_{1t} + \varepsilon_t.$$

Denote the OLS of this simpler regression as $\tilde{\beta}_1$.

Please compare the relative efficiency between $\hat{\beta}_1$ and $\tilde{\beta}_1$. That is, which estimator is better for β_1^o? Give your reasoning.

3.17. Suppose Assumption 3.6 is replaced by the following assumption:
Assumption 3.6′ : $\varepsilon|\mathbf{X} \sim N(0, V)$, *where* $V = V(\mathbf{X})$ *is a known* $n \times n$ *finite and positive definite matrix.*

Compared to Assumption 3.6, Assumption 3.6′ assumes that $\text{var}(\varepsilon|\mathbf{X}) = V$ is completely known and there is no unknown proportionality σ^2. Define GLS $\hat{\beta}^* = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}Y$.

(a) Is $\hat{\beta}^*$ BLUE?

(b) Put $X^* = CX$ and $s^{*2} = e^{*\prime}e^*/(n-K)$, where $e^* = Y - X^*\hat{\beta}^*$, $C'C = V$. Do the the usual t-test and F-test defined as

$$\begin{aligned} T^* &= \frac{R\hat{\beta}^* - r}{\sqrt{s^{*2}R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R'}}, \text{ for } J = 1, \\ F^* &= \frac{(R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r)/J}{s^{*2}} \end{aligned}$$

follow the t_{n-K} and $F_{J,n-K}$ distributions respectively under the null hypothesis that $R\beta = r$? Explain.

(c) Construct two new test statistics:

$$\begin{aligned} \tilde{T}^* &= \frac{R\hat{\beta}^* - r}{\sqrt{R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R'}}, \text{ for } J = 1, \\ \tilde{Q}^* &= (R\hat{\beta}^* - r)'[R(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}R']^{-1}(R\hat{\beta}^* - r). \end{aligned}$$

What distributions will these test statistics follow under the null hypothesis that $R\beta = r$? Explain.

(d) Which set of tests, (T^*, F^*) or $(\tilde{T}^*, \tilde{Q}^*)$, are more powerful at the same significance level? Explain. [*Hint: The t-distribution has a heavier tail than* $N(0,1)$ *and so has a larger critical value at a given significance level.*]

3.18. Consider a linear regression model

$$Y_t = X_t'\beta^0 + \varepsilon_t, \qquad t = 1, 2, ..., n,$$

where $\varepsilon_t = \sigma(X_t)v_t$, X_t is a $K \times 1$ nonstochastic vector, and $\sigma(X_t)$ is a positive function of X_t, and $\{v_t\}$ is i.i.d. $N(0,1)$.

Let $\hat{\beta} = (X'X)^{-1}X'Y$ denote the OLS estimator for β^0, where X is a $n \times K$ matrix whose t-th row is X_t, and Y is a $n \times 1$ vector whose t-th component is Y_t.

(a) Is $\hat{\beta}$ unbiased for β^0?

(b) Find $\text{var}(\hat{\beta}) = E[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})']$. You may find the following notation useful: $\Omega = \text{diag}\{\sigma^2(X_1), \sigma^2(X_2), ..., \sigma^2(X_n)\}$, i.e., Ω is a $n \times n$ diagonal matrix with the t-th diagonal component equal to $\sigma^2(X_t)$ and all off-diagonal components equal to zero.

Consider the transformed regression model

$$\frac{1}{\sigma(X_t)}Y_t = \frac{1}{\sigma(X_t)}X_t'\beta^0 + v_t$$

or

$$Y_t^* = X_t^{*\prime}\beta^0 + v_t,$$

where $Y_t^* = \sigma^{-1}(X_t)Y_t$ and $X_t^* = \sigma^{-1}(X_t)X_t$.

Denote the OLS estimator of this transformed model as $\tilde{\beta}$.

(c) Show

$$\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

(d) Is $\tilde{\beta}$ unbiased for β^0?

(e) Find $\text{var}(\tilde{\beta})$.

(f) Which estimator, $\hat{\beta}$ or $\tilde{\beta}$, is more efficient in terms of the mean squared error criterion? Give your reasoning.

(g) Use the difference $R\tilde{\beta} - r$ to construct a test statistic for the null hypothesis of interest $\mathbf{H}_0 : R\beta^0 = r$, where R is a $J \times K$ matrix, r is $K \times 1$, and $J > 1$. What is the finite sample distribution of your test statistic under $\mathbf{H}_0$?

CHAPTER 4 LINEAR REGRESSION MODELS WITH I.I.D. OBSERVATIONS

Abstract: When the conditional normality assumption on the regression error does not hold, the OLS estimator no longer has the finite sample normal distribution, and the t-test statistics and F-test statistics no longer follow the Student t-distribution and a F-distribution in finite samples respectively. In this chapter, we show that under the assumption of i.i.d. observations with conditional homoskedasticity, the classical t-test and F-test are approximately applicable in large samples. However, under conditional heteroskedasticity, the t-test statistics and F-test statistics are not applicable even when the sample size goes to infinity. Instead, White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator should be used, which yields asymptotically valid hypothesis test procedures. A direct test for conditional heteroskedasticity due to White (1980) is presented. To facilitate asymptotic analysis in this and subsequent chapters, we first introduce some basic tools in asymptotic analysis.

Key words: Asymptotic analysis, Almost sure convergence, Central limit theorem, Convergence in distribution, Convergence in quadratic mean, Convergence in probability, I.I.D., Law of large numbers, the Slutsky theorem, White's heteroskedasticity-consistent variance-covariance matrix estimator.

Motivation

The assumptions of classical linear regression models are rather strong and one may have a hard time finding practical applications where all these assumptions hold exactly. For example, it has been documented that most economic and financial data have heavy tails, and so they are not normally distributed. An interesting question now is whether the estimators and tests which are based on the same principles as before still make sense in this more general setting. In particular, what happens to the OLS estimator, the t- and F-tests if any of the following assumptions fails:

- strict exogeneity $E(\varepsilon_t|\mathbf{X}) = 0$?
- normality $(\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I))$?
- conditional homoskedasticity $(\text{var}(\varepsilon_t|\mathbf{X}) = \sigma^2)$?
- serial uncorrelatedness $(\text{cov}(\varepsilon_t, \varepsilon_s|\mathbf{X}) = 0 \text{ for } t \neq s)$?

When classical assumptions are violated, we do not know the finite sample statistical properties of the estimators and test statistics anymore. A useful tool to obtain the

understanding of the properties of estimators and tests in this more general setting is to pretend that we can obtain a limitless number of observations. We can then pose the question how estimators and test statistics would behave when the number of observations increases without limit. This is called *asymptotic analysis*. In practice, the sample size is always finite. However, the asymptotic properties translate into results that hold true approximately in finite samples, provided that the sample size is large enough. We now need to introduce some basic analytic tools for asymptotic theory. For more systematic introduction of asymptotic theory, see, for example, White (1994, 1999).

4.1 Introduction to Asymptotic Theory

In this section, we introduce some important convergence concepts and limit theorems. First, we introduce the concept of convergence in mean squares, which is a distance measure of a sequence of random variables from a random variable.

Definition 4.1 [Convergence in mean squares (or in quadratic mean)] *A sequence of random variables/vectors/matrices $Z_n, n = 1, 2, ...,$ is said to converge to Z in mean squares as $n \to \infty$ if*

$$E||Z_n - Z||^2 \to 0 \text{ as } n \to \infty,$$

where $||\cdot||$ is the sum of the absolute value of each component in $Z_n - Z$.

When Z_n is a vector or matrix, convergence can be understood as convergence in each element of Z_n.When $Z_n - Z$ is a $l \times m$ matrix, where l and m are fixed positive integers, then we can also define the squared norm as

$$||Z_n - Z||^2 = \sum_{t=1}^{l}\sum_{s=1}^{m}[Z_n - Z]^2_{(t,s)}.$$

Note that Z_n converges to Z in mean squares if and only if each component of Z_n converges to the corresponding component of Z in mean squares.

Example 1: Suppose $\{Z_t\}$ is i.i.d.(μ, σ^2), and $\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t$. Then

$$\bar{Z}_n \overset{q.m.}{\to} \mu.$$

Solution: Because $E(\bar{Z}_n) = \mu$, we have

$$\begin{aligned}
E(\bar{Z}_n - \mu)^2 &= \text{var}(\bar{Z}_n) \\
&= \text{var}\left(n^{-1}\sum_{t=1}^{n} Z_t\right) \\
&= \frac{1}{n^2}\text{var}\left(\sum_{t=1}^{n} Z_t\right) \\
&= \frac{1}{n^2}\sum_{t=1}^{n}\text{var}(Z_t) \\
&= \frac{\sigma^2}{n} \\
&\to 0 \text{ as } n \to \infty.
\end{aligned}$$

It follows that

$$E(\bar{Z}_n - \mu)^2 = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty.$$

Next, we introduce the concept of convergence in probability that is another popular distance measure between a sequence of random variables and a random variable.

Definition 4.2 [Convergence in probability] *Z_n converges to Z in probability if for any given constant $\epsilon > 0$,*

$$\begin{aligned} \Pr[||Z_n - Z|| &> \epsilon] \to 0 \text{ as } n \to \infty \text{ or} \\ \Pr[||Z_n - Z|| &\leq \epsilon] \to 1 \text{ as } n \to \infty. \end{aligned}$$

For convergence in probability, we can also write

$$Z_n - Z \xrightarrow{p} 0 \text{ or } Z_n - Z = o_P(1),$$

The notation $o_P(1)$ means that $Z_n - Z$ vanishes to zero in probability. When $Z = b$ is a constant, we can write $Z_n \xrightarrow{p} b$ and $b = p\lim Z_n$ is called the probability limit of Z_n.

Convergence in probability is also called weak convergence or convergence with probability approaching one. When $Z_n \xrightarrow{p} Z$, the probability that the difference $||Z_n - Z||$ exceeds any given small constant ϵ is rather small for all n sufficiently large. In other words, Z_n will be very close to Z with very high probability, when the sample size n is sufficiently large.

To gain more intuition of the convergence in probability, we define the event

$$A_n(\epsilon) = \{\omega \in \Omega : |Z_n(\omega) - Z(\omega)| > \epsilon\},$$

where ω is a basic outcome in sample space Ω. Then convergence in probability says that the probability of event $A_n(\epsilon)$ may be nonzero for any finite n, but such a probability will eventually vanish to zero as $n \to \infty$. In other words, it becomes less and less likely that the difference $|Z_n - Z|$ is larger than a prespecified constant $\epsilon > 0$. Or, we have more and more confidence that the difference $|Z_n - Z|$ will be smaller than ϵ as $n \to \infty$. The constant ϵ can be viewed as a prespecified tolerance level.

Lemma 4.1 [Weak Law of Large Numbers (WLLN) for I.I.D. Sample] *Suppose $\{Z_t\}$ is i.i.d.(μ, σ^2), and define $\bar{Z}_n = n^{-1}\sum_{t=1}^n Z_t, n = 1, 2, \ldots$. Then*

$$\bar{Z}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

Proof: For any given constant $\epsilon > 0$, we have by Chebyshev's inequality

$$\begin{aligned} \Pr(|\bar{Z}_n - \mu| &> \epsilon) \leq \frac{E(\bar{Z}_n - \mu)^2}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty. \end{aligned}$$

Hence,

$$\bar{Z}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

This is the so-called weak law of large numbers (WLLN). In fact, we can weaken the moment condition.

We now provide an economic interpretation of the WLLN using an example. In finance, there is a popular trading strategy called buy-and-hold trading startegy. An investor buys a stock at some day and then hold it for a long time period before he sells it out. This is called a buy-and-hold trading strategy. How is the average return of this trading strategy?

Suppose Z_t is the return of the stock on period t, and the returns over different time periods are i.i.d.(μ, σ^2). Also assume the investor holds the stock for a total of n period. Then the average return over each time period is the sample mean

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t.$$

When the number n of holding periods is large, we have

$$\bar{Z} \xrightarrow{p} \mu = E(Z_t)$$

as $n \to \infty$. Thus, the average return of the buy-and-hold trading startegy is approximately equal to μ when n is sufficiently large.

Lemma 4.2 [WLLN for I.I.D. Random Sample] *Suppose $\{Z_t\}$ is i.i.d. with $E(Z_t) = \mu$ and $E|Z_t| < \infty$. Define $\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t$. Then*

$$\bar{Z}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

Question: Why do we need the moment condition $E|Z_t| < \infty$?

We can consider a counter example: Suppose $\{Z_t\}$ is a sequence of i.i.d. Cauchy$(0, 1)$ random variables whose moments do not exist. Then $\bar{Z}_n \sim$ Cauchy$(0, 1)$ for all $n \geq 1$, and so it does not converge in probability to some constant as $n \to \infty$.

We now introduce a useful related concept:

Definition 4.3 [Boundedness in Probability] *A sequence of random variables/vectors/matrices $\{Z_n\}$ is bounded in probability if for any small constant $\delta > 0$, there exists a constant $C < \infty$ such that*

$$P(||Z_n|| > C) \leq \delta$$

as $n \to \infty$. *We denote*

$$Z_n = O_P(1).$$

Intuitively, when $Z_n = O_P(1)$, the probability that $||Z_n||$ exceeds a very large constant is small as $n \to \infty$. Or, equivalently, $||Z_n||$ is smaller than C with a very high probability as $n \to \infty$.

Example 2: Suppose $Z_n \sim N(\mu, \sigma^2)$ for all $n \geq 1$. Then

$$Z_n = O_P(1).$$

Solution: For any $\delta > 0$, we always have a sufficiently large constant $C = C(\delta) > 0$ such that

$$\begin{aligned} P(|Z_n| &> C) = 1 - P(-C \leq Z_n \leq C) \\ &= 1 - P\left[\frac{-C-\mu}{\sigma} \leq \frac{Z_n - \mu}{\sigma} \leq \frac{C-\mu}{\sigma}\right] \\ &= 1 - \Phi\left(\frac{C-\mu}{\sigma}\right) + \Phi\left(-\frac{C+\mu}{\sigma}\right) \\ &\leq \delta, \end{aligned}$$

where $\Phi(z) = P(Z \leq z)$ is the CDF of $N(0,1)$. [We can choose C such that $\Phi[(C-\mu)/\sigma] \geq 1 - \frac{1}{2}\delta$ and $\Phi[-(C+\mu)/\sigma] \leq \frac{1}{2}\delta$.]

A Special Case: What happens to C if $Z_n \sim N(0,1)$?

In this case,

$$\begin{aligned} P(|Z_n| &> C) = 1 - \Phi(C) + \Phi(-C) \\ &= 2[1 - \Phi(C)]. \end{aligned}$$

Suppose we set

$$2[1 - \Phi(C)] = \delta,$$

that is, we set

$$C = \Phi^{-1}\left(1 - \frac{\delta}{2}\right),$$

where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$. Then we have

$$P(|Z_n| > C) = \delta.$$

The following lemma provides a convenient way to verify convergence in probability.

Lemma 4.3: If $Z_n - Z \xrightarrow{q.m.} 0$, *then* $Z_n - Z \xrightarrow{p} 0$.

Proof: By Chebyshev's inequality, we have

$$P(|Z_n - Z| > \epsilon) \leq \frac{E[Z_n - Z]^2}{\epsilon^2} \to 0$$

for any given $\epsilon > 0$ as $n \to \infty$. This completes the proof.

Example 3: Suppose Assumptions 3.1–3.4 hold. Does the OLS estimator $\hat{\beta}$ converges in probability to β^o?

Solution: From Theorem 3.4, we have

$$\begin{aligned} \tau' E[(\hat{\beta} - \beta^o)(\hat{\beta} - \beta^o)'|\mathbf{X}]\tau &= \sigma^2 \tau'(X'X)^{-1}\tau \\ &\to 0 \end{aligned}$$

for any $\tau \in R^K, \tau'\tau = 1$ as $n \to \infty$ with probability one. It follows that $E||\hat{\beta} - \beta^o||^2 = E\{E[||\hat{\beta} - \beta^o||^2|X]\} \to 0$ as $n \to \infty$. Therefore, by Lemma 4.3, we have $\hat{\beta} \xrightarrow{p} \beta^o$.

Example 4: Suppose Assumptions 3.1, 3.3 and 3.5 hold. Does s^2 converge in probability to σ^2?

Solution: Under the given assumptions,

$$(n - K)\frac{s^2}{\sigma^2} \sim \chi^2_{n-K},$$

and therefore we have $E(s^2) = \sigma^2$ and $\text{var}(s^2) = \frac{2\sigma^4}{n-K}$. It follows that $E(s^2 - \sigma^2)^2 = 2\sigma^4/(n - K) \to 0$, $s^2 \xrightarrow{q.m.} \sigma^2$ and so $s^2 \xrightarrow{p} \sigma^2$ because convergence in quadratic mean implies convergence in probability.

While convergence in mean squares implies convergence in probability, the converse is not true. We now give an example.

Example 5: Suppose

$$Z_n = \begin{cases} 0 & \text{with prob } 1 - \frac{1}{n} \\ n & \text{with prob } \frac{1}{n}. \end{cases}$$

Then $Z_n \xrightarrow{p} 0$ as $n \to \infty$ but $E(Z_n - 0)^2 = n \to \infty$. Please verify it.

Solution:

(i) For any given $\varepsilon > 0$, we have

$$P(|Z_n - 0| > \varepsilon) = P(Z_n = n) = \frac{1}{n} \to 0.$$

(ii)

$$\begin{aligned} E(Z_n - 0)^2 &= \sum_{z_n \in \{0,n\}} (z_n - 0)^2 f(z_n) \\ &= (0-0)^2 \cdot \left(1 - n^{-1}\right) + (n-0)^2 \cdot n^{-1} \\ &= n \to \infty. \end{aligned}$$

Next, we provide another convergence concept called almost sure convergence.

Definition 4.4 [Almost Sure Convergence] $\{Z_n\}$ *converges to* Z *almost surely if*

$$\Pr\left[\lim_{n\to\infty} ||Z_n - Z|| = 0\right] = 1.$$

We denote $Z_n - Z \overset{a.s.}{\to} 0$.

To gain intuition for the concept of almost sure convergence, recall the definition of a random variable: any random variable is a mapping from the sample space Ω to the real line, namely $Z : \Omega \to \mathbb{R}$. Let ω be a basic outcome in the sample space Ω. Define a subset in Ω :

$$A^c = \{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}.$$

That is, A^c is the set of basic outcomes on which the sequence of $\{Z_n(\cdot)\}$ converges to $Z(\cdot)$ as $n \to \infty$. Then almost sure convergence can be stated as

$$P(A^c) = 1.$$

In other words, the convergent set A^c has probability one to occur.

Example 6: Let ω be uniformly distributed on $[0, 1]$, and define

$$Z(\omega) = \omega \text{ for all } \omega \in [0, 1].$$

and

$$Z_n(\omega) = \omega + \omega^n \text{ for } \omega \in [0, 1].$$

Is $Z_n - Z \overset{a.s.}{\to} 0$?

Solution: Consider

$$A^c = \{\omega \in \Omega : \lim_{n\to\infty} |Z_n(\omega) - Z(\omega)| = 0\}.$$

Because for any given $\omega \in [0, 1)$, we always have

$$\begin{aligned} \lim_{n\to\infty} |Z_n(\omega) - Z(\omega)| &= \lim_{n\to\infty} |(\omega + \omega^n) - \omega| \\ &= \lim_{n\to\infty} \omega^n = 0. \end{aligned}$$

In contrast, for $\omega = 1$, we have

$$\lim_{n\to\infty} |Z_n(1) - Z(1)| = 1^n = 1 \neq 0.$$

Thus, $A^c = [0, 1)$ and $P(A^c) = 1$. We also have $P(A) = P(\omega = 1) = 0$.

In probability theory, almost sure convergence is closely related to pointwise convergence (almost everywhere). It is also called strong convergence.

Lemma 4.4 [Strong Law of Large Numbers (SLLN) for I.I.D. Random Samples] *Suppose* $\{Z_t\}$ *be i.i.d. with* $E(Z_t) = \mu$ *and* $E|Z_t| < \infty$. *Then*

$$\bar{Z}_n \xrightarrow{a.s.} \mu \text{ as } n \to \infty.$$

Almost sure convergence implies convergence in probability but not vice versa.

Question: If $s^2 \xrightarrow{p} \sigma^2$, do we have $s \xrightarrow{p} \sigma$?

Answer: Yes. It follows from the following continuity lemma with the choice of $g(s^2) = \sqrt{s^2} = s$.

Lemma 4.5 [Continuity]: *(i) Suppose* $a_n \xrightarrow{p} a$ *and* $b_n \xrightarrow{p} b$, *and* $g(\cdot)$ *and* $h(\cdot)$ *are continuous functions. Then*

$$\begin{aligned} &g(a_n) + h(b_n) \xrightarrow{p} g(a) + h(b), \text{ and} \\ &g(a_n)h(b_n) \xrightarrow{p} g(a)h(b). \end{aligned}$$

(ii) Similar results hold for almost sure convergence.

The last convergence concept we will introduce is called convergenve in distribution.

It should be emphasized that convergence in mean squres, convergence in probability and almost sure convergence all measure the closeness between the random variable Z_n

and the random variable Z. This differs from the concept of convergence in distribution introduced in Chapter 3. There, convergence in distribution is defined in terms of the closeness of the CDF $F_n(z)$ of Z_t to the CDF $F(z)$ of Z, not between the closeness of the random variable Z_n to the random variable Z. As a result, for convergence in mean squares, convergence in probability and almost sure convergence, Z_n converges to Z if and only if convergence of Z_n to Z occurs element by element (that is, each element of Z_n converges to the corresponding element of Z). For the convergence in distribution of Z_n to Z, however, element by element convergence does not imply convergence in distribution of Z_n to Z,because element-wise convergence in distribution ignores the relationships among the components of Z_n. Nevertheless, $Z_n \xrightarrow{d} Z$ does imply element by element convergence in distribution. That is, convergence in joint distribution implies convergence in marginal distribution.

The main purpose of asymptotic analysis is to derive the large sample distribution of the estimator or statistic of interest and use it as an approximation in statistical inference. For this purpose, we need to make use of an important limit theorem, namely Central Limit Theorem (CLT). We now state and prove the CLT for i.i.d. random samples, a fundamental limit theorem in probability theory.

Lemma 4.6 [Central Limit Theorem (CLT) for I.I.D. Random Samples]: *Suppose* $\{Z_t\}$ *is i.i.d.*(μ, σ^2), *and* $\bar{Z}_n = n^{-1}\sum_{t=1}^n Z_t$. *Then as* $n \to \infty$,

$$\begin{aligned}\frac{\bar{Z}_n - E(\bar{Z}_n)}{\sqrt{\text{var}(\bar{Z}_n)}} &= \frac{\bar{Z}_n - \mu}{\sqrt{\sigma^2/n}} \\ &= \frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma} \\ &\xrightarrow{d} N(0,1).\end{aligned}$$

Proof: Put

$$Y_t = \frac{Z_t - \mu}{\sigma},$$

and $\bar{Y}_n = n^{-1}\sum_{t=1}^n Y_t$. Then

$$\frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma} = \sqrt{n}\bar{Y}_n.$$

The characteristic function of $\sqrt{n}\bar{Y}_n$

$$\begin{aligned}
\phi_n(u) &= E[\exp(iu\sqrt{n}\bar{Y}_n)], \qquad i = \sqrt{-1} \\
&= E\left[\exp\left(\frac{iu}{\sqrt{n}}\sum_{t=1}^{n} Y_t\right)\right] \\
&= \prod_{t=1}^{n} E\left[\exp\left(\frac{iu}{\sqrt{n}}Y_t\right)\right] \text{ by independence} \\
&= \left[\phi_Y\left(\frac{u}{\sqrt{n}}\right)\right]^n \text{ by identical distribution.} \\
&= \left[\phi_Y(0) + \phi'(0)\frac{u}{\sqrt{n}} + \frac{1}{2}\phi''(0)\frac{u^2}{n} + \cdots\right]^n \\
&= \left(1 - \frac{u^2}{2n}\right)^n + o(1) \\
&\to \exp\left(-\frac{u^2}{2}\right) \text{ as } n \to \infty,
\end{aligned}$$

where the third equality follows from independence, the fourth equality follows from identical distribution, the fifth equality follows from the Taylor series expansion, and $\phi(0) = 1, \phi'(0) = 0, \phi''(0) = -1$. Note that $o(1)$ means a reminder term that vanishes to zero as $n \to \infty$, and we have also made use of the fact that $\left(1 + \frac{a}{n}\right)^n \to e^a$.

More rigorously, we can show

$$\begin{aligned}
\ln \phi_n(u) &= n \ln \phi_Y\left(\frac{u}{\sqrt{n}}\right) \\
&= \frac{\ln \phi_Y\left(\frac{u}{\sqrt{n}}\right)}{n^{-1}} \\
&\to \frac{u}{2}\lim_{n\to\infty} \frac{\frac{\phi_Y'(u/\sqrt{n})}{\phi_Y(u/\sqrt{n})}}{n^{-1/2}} \\
&= \frac{u^2}{2}\lim_{n\to\infty} \frac{\phi_Y''(u/\sqrt{n})\phi_Y(u/\sqrt{n}) - [\phi_Y'(u/\sqrt{n})]^2}{\phi_Y^2(u/\sqrt{n})} \\
&= -\frac{u^2}{2}.
\end{aligned}$$

It follows that

$$\lim_{n\to\infty} \phi_n(u) = e^{-\frac{1}{2}u^2}.$$

This is the characteristic function of $N(0,1)$. By the uniqueness of the characteristic function, the asymptotic distribution of

$$\frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma}$$

is $N(0,1)$. This completes the proof.

Lemma 4.7 [Cramer-Wold Device] *A $p \times 1$ random vector $Z_n \xrightarrow{d} Z$ if and only if for any nonzero $\lambda \in R^p$ such that $\lambda'\lambda = \Sigma_{j=1}^p \lambda_j^2 = 1$, we have*

$$\lambda' Z_n \xrightarrow{d} \lambda' Z.$$

This lemma is useful for obtaining asymptotic multivariate distributions.

Lemma 4.8 [Slutsky Theorem] *Let $Z_n \xrightarrow{d} Z$, $a_n \xrightarrow{p} a$ and $b_n \xrightarrow{p} b$, where a and b are constants. Then*

$$a_n + b_n Z_n \xrightarrow{d} a + bZ \text{ as } n \to \infty.$$

Question: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. Is $X_n + Y_n \xrightarrow{d} X + Y$?

Answer: No. We consider two examples:

Example 7: X_n and Y_n are independent N(0,1). Then

$$X_n + Y_n \xrightarrow{d} N(0,2).$$

Example 8: $X_n = Y_n \sim N(0,1)$ for all $n \geq 1$. Then

$$X_n + Y_n = 2X_n \sim N(0,4).$$

Example 9: Suppose Assumptions 3.1, 3.3(a) and 3.5, and the hypothesis $\mathbf{H}_0 : R\beta^o = r$ hold, where R is a $J \times K$ nonstochastic matrix with rank J, r is a $J \times 1$ nonstochastic vector, and $J \leq K$. Then the quadratic form

$$\frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi_J^2.$$

Suppose now we replace σ^2 by s^2. What is the asymptotic distribution of the quadratic form

$$\frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{s^2}?$$

Finally, we introduce a lemma which is very useful in deriving the asymptotic distributions of nonlinear statistics (i.e., nonlinear functions of the random sample).

Lemma 4.9 [Delta Method] *Suppose $\sqrt{n}(\bar{Z}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$, and $g(\cdot)$ is continuously differentiable with $g'(\mu) \neq 0$. Then as $n \to \infty$,*

$$\sqrt{n}[g(\bar{Z}_n) - g(\mu)] \xrightarrow{d} N(0, [g'(\mu)]^2\sigma^2).$$

Proof: First, because $\sqrt{n}(\bar{Z}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$ implies $\sqrt{n}(\bar{Z}_n - \mu)/\sigma = O_P(1)$, we have $\bar{Z}_n - \mu = O_P(n^{-1/2}) = o_P(1)$.

Next, by a first order Taylor series expansion, we have

$$Y_n = g(\bar{Z}_n) = g(\mu) + g'(\bar{\mu}_n)(\bar{Z}_n - \mu),$$

where $\bar{\mu}_n = \lambda\mu + (1-\lambda)\bar{Z}_n$ for $\lambda \in [0,1]$. It follows by the Slutsky theorem that

$$\begin{aligned} \sqrt{n}\frac{g(\bar{Z}_n) - g(\mu)}{\sigma} &= g'(\bar{\mu}_n)\sqrt{n}\frac{\bar{Z}_n - \mu}{\sigma} \\ &\xrightarrow{d} N(0, [g'(\mu)]^2), \end{aligned}$$

where $g'(\bar{\mu}_n) \xrightarrow{p} g'(\mu)$ given $\bar{\mu}_n \xrightarrow{p} \mu$.

By the Slutsky theorem again, we have

$$\sqrt{n}[Y_n - g(\mu)] \xrightarrow{d} N(0, \sigma^2[g'(\mu)]^2).$$

This completes the proof.

The Delta method is a Taylor series approximation in a statistical context. It linearizes a smooth (i.e., differentiable) nonlinear statistic so that the CLT can be applied to the linearized statistic. Therefore, it can be viewed as a generalization of the CLT from a sample average to a nonlinear statistic. This method is very useful when more than one parameter makes up the function to be estimated and more than one random variable is used in the estimator.

Example 10: Suppose $\sqrt{n}(\bar{Z}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$ and $\mu \neq 0$ and $0 < \sigma < \infty$. Find the limiting distribution of $\sqrt{n}(\bar{Z}_n^{-1} - \mu^{-1})$.

Solution: Put $g(\bar{Z}_n) = \bar{Z}_n^{-1}$. Because $\mu \neq 0$, $g(\cdot)$ is continuous at μ. By a first order Taylor series expansion, we have

$$\begin{aligned} g(\bar{Z}_n) &= g(\mu) + g'(\bar{\mu}_n)(\bar{Z}_n - \mu), \text{ or} \\ \bar{Z}_n^{-1} - \mu^{-1} &= (-\bar{\mu}_n^{-2})(\bar{Z}_n - \mu) \end{aligned}$$

where $\bar{\mu}_n = \lambda\mu + (1-\lambda)\bar{Z}_n \xrightarrow{p} \mu$ given $\bar{Z}_n \xrightarrow{p} \mu$ and $\lambda \in [0,1]$. It follows that

$$\begin{aligned} \sqrt{n}(\bar{Z}_n^{-1} - \mu^{-1}) &= -\frac{\sigma}{\bar{\mu}_n^2}\frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma} \\ &\xrightarrow{d} N(0, \sigma^2/\mu^4). \end{aligned}$$

Taylor series expansions, various convergence concepts, laws of large numbers, central limit theorems, and slutsky theorem constitute a tool kit of asymptotic analysis. We now use these asymptotic tools to investigate the large sample behavior of the OLS estimator and related statistics in subsequent chapters.

4.2 Framework and Assumptions

We first state the assumptions under which we will establish the asymptotic theory for linear regression models.

Assumption 4.1 [I.I.D.]: $\{Y_t, X_t'\}_{t=1}^{'n}$ is an i.i.d. random sample.

Assumption 4.2 [Linearity]:

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

for some unknown $K \times 1$ parameter β^o and some unobservable random variable ε_t.

Assumption 4.3 [Correct Model Specification]: $E(\varepsilon_t|X_t) = 0$ a.s. with $E(\varepsilon_t^2) = \sigma^2 < \infty$.

Assumption 4.4 [Nonsingularity]: The $K \times K$ matrix

$$Q = E(X_t X_t')$$

is nonsingular and finite.

Assumption 4.5: The $K \times K$ matrix $V \equiv \text{var}(X_t\varepsilon_t) = E(X_t X_t'\varepsilon_t^2)$ is finite and positive definite (p.d.).

Remarks:

The i.i.d. observations assumption in Assumption 4.1 implies that the asymptotic theory developed in this chapter will be applicable to cross-sectional data, but not time series data. The observations of the later are usually correlated and will be considered in Chapter 5. Put $Z_t = (Y_t, X_t')'$. Then I.I.D. implies that Z_t and Z_s are independent when $t \neq s$, and the Z_t have the same distribution for all t. The identical distribution means that the observations are generated from the same data generating process, and independence means that different observations contain new information about the data generating process.

Assumptions 4.1 and 4.3 imply the strict exogeneity condition (Assumption 3.2) holds, because we have

$$\begin{aligned} E(\varepsilon_t|\mathbf{X}) &= E(\varepsilon_t|X_1, X_2, ...X_t, ...X_n) \\ &= E(\varepsilon_t|X_t) \\ &= 0 \ a.s. \end{aligned}$$

As a most important feature of Assumptions 4.1–4.5 together, we allow for conditional heteroskedasticity (i.e., $\text{var}(\varepsilon_t|X_t) \neq \sigma^2$ a.s.), and do not assume normality for the conditional distribution of $\varepsilon_t|X_t$. It is possible that $\text{var}(\varepsilon_t|X_t)$ may be correlated with X_t. For example, the variation of the output of a firm may depend on the size of the firm, and the variation of a household may depend on its income level. In economics and finance, conditional heteroskedasticity is more likely to occur in cross-sectional observations than in time series observations, and for time series observations, conditional heteroskedasticity is more likely to occur for high-frequency data than low-frequency data. In this chapter, we will consider the effect of conditional heteroskedasticity in cross-section observations. The effect of conditional heteroskedasticity in time series observations will be considered in Chapter 5.

On the other hand, relaxation of the normality assumption is more realistic for economic and financial data. For example, it has been well documented (Mandelbrot 1963, Fama 1965, Kon 1984) that returns on financial assets are not normally distributed. However, the I.I.D. assumption implies that $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$. That is, there exists no serial correlation in the regression disturbance.

Among other things, Assumption 4.4 implies $E(X_{jt}^2) < \infty$ for $0 \leq j \leq k$. By the SLLN for i.i.d. random samples, we have

$$\frac{\mathbf{X}'\mathbf{X}}{n} = \frac{1}{n}\sum_{t=1}^{n} X_t X_t' \xrightarrow{a.s.} E(X_t X_t') = Q$$

as $n \to \infty$. Hence, when n is large, the matrix $\mathbf{X}'\mathbf{X}$ behaves approximately like nQ, whose minimum eigenvalue $\lambda_{\min}(nQ) = n\lambda_{\min}(Q) \to \infty$ at the rate of n. Thus, Assumption 4.4 implies Assumption 3.3.

When $X_{0t} = 1$, Assumption 4.5 implies $E(\varepsilon_t^2) < \infty$. If $E(\varepsilon_t^2|X_t) = \sigma^2 < \infty$ a.s., i.e., there exists conditional homoskedasticity, then Assumption 4.5 can be ensured by Assumption 4.4. More generally, there exists conditional heteroskedasticity, the moment condition in Assumption 4.5 can be ensured by the moment conditions that $E(\varepsilon_t^4) < \infty$ and $E(X_{jt}^4) < \infty$ for $0 \leq j \leq k$, because by repeatedly using the Cauchy-Schwarz inequality twice, we have

$$\begin{aligned} |E(\varepsilon_t^2 X_{jt} X_{lt})| &\leq [E(\varepsilon_t^4)]^{1/2}[E(X_{jt}^2 X_{lt}^2)]^{1/2} \\ &\leq [E(\varepsilon_t^4)]^{1/2}[E(X_{jt}^4)E(X_{lt}^4)]^{1/4} \end{aligned}$$

where $0 \leq j, l \leq k$ and $1 \leq t \leq n$.

We now address the following questions:

- Consistency of OLS?
- Asymptotic normality?
- Asymptotic efficiency?
- Confidence interval estimation?
- Hypothesis testing?

In particular, we are interested in knowing whether the statistical properties of OLS $\hat{\beta}$ and related test statistics derived under the classical linear regression setup are still valid under the current setup, at least when n is large.

4.3 Consistency of OLS

Suppose we have a random sample $\{Y_t, X_t'\}_{t=1}^{'n}$. Recall that the OLS estimator:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'Y}{n} \\ &= \hat{Q}^{-1}n^{-1}\sum_{t=1}^{n}X_tY_t,\end{aligned}$$

where

$$\hat{Q} = n^{-1}\sum_{t=1}^{n}X_tX_t'.$$

Substituting $Y_t = X_t'\beta^o + \varepsilon_t$, we obtain

$$\hat{\beta} = \beta^o + \hat{Q}^{-1}n^{-1}\sum_{t=1}^{n}X_t\varepsilon_t.$$

We will consider the consistency of $\hat{\beta}$ directly.

Theorem 4.10 [Consistency of OLS] *Under Assumptions 4.1-4.4, as* $n \to \infty$,

$$\hat{\beta} \xrightarrow{p} \beta^o \text{ or } \hat{\beta} - \beta^o = o_P(1).$$

Proof: Let $C > 0$ be some bounded constant. Also, recall $X_t = (X_{0t}, X_{1t}, ..., X_{kt})'$. First, the moment condition holds: for all $0 \leq j \leq k$,

$$\begin{aligned}E|X_{jt}\varepsilon_t| &\leq (EX_{jt}^2)^{\frac{1}{2}}(E\varepsilon_t^2)^{\frac{1}{2}} \text{ by the Cauchy-Schwarz inequality} \\ &\leq C^{\frac{1}{2}}C^{\frac{1}{2}} \\ &\leq C\end{aligned}$$

where $E(X_{jt}^2) \leq C$ by Assumption 4.4, and $E(\varepsilon_t^2) \leq C$ by Assumption 4.3. It follows from WLLN (with $Z_t = X_t\varepsilon_t$) that

$$n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \xrightarrow{p} E(X_t\varepsilon_t) = 0,$$

where

$$\begin{aligned} E(X_t\varepsilon_t) &= E[E(X_t\varepsilon_t|X_t)] \text{ by the law of iterated expectations} \\ &= E[X_t E(\varepsilon_t|X_t)] \\ &= E(X_t \cdot 0) \\ &= 0. \end{aligned}$$

Applying WLLN again (with $Z_t = X_t X_t'$) and noting that

$$E|X_{jt}X_{lt}| \leq [E(X_{jt}^2)E(X_{lt}^2)]^{\frac{1}{2}} \leq C$$

by the Cauchy-Schwarz inequality for all pairs (j, l), where $0 \leq j, l \leq k$, we have

$$\hat{Q} \xrightarrow{p} E(X_t X_t') = Q.$$

Hence, we have $\hat{Q}^{-1} \xrightarrow{p} Q^{-1}$ by continuity. It follows that

$$\begin{aligned} \hat{\beta} - \beta^o &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= \hat{Q}^{-1}n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \\ \xrightarrow{p} Q^{-1} \cdot 0 &= 0. \end{aligned}$$

This completes the proof. ■

4.4 Asymptotic Normality of OLS

Next, we derive the asymptotic distribution of $\hat{\beta}$. We first provide a multivariate CLT for I.I.D. random samples.

Lemma 4.11 [Multivariate Central Limit Theorem (CLT) for I.I.D. Random Samples]: *Suppose $\{Z_t\}$ is a sequence of i.i.d. random vectors with $E(Z_t) = 0$ and $var(Z_t) = E(Z_t Z_t') = V$ is finite and positive definite. Define*

$$\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t.$$

Then as $n \to \infty$,

$$\sqrt{n}\bar{Z}_n \xrightarrow{d} N(0, V)$$

or

$$V^{-\frac{1}{2}}\sqrt{n}\bar{Z}_n \xrightarrow{d} N(0, I).$$

Question: What is the variance-covariance matrix of $\sqrt{n}\bar{Z}_n$?
Answer: Noting that $E(Z_t) = 0$, we have

$$\begin{aligned}
\text{var}(\sqrt{n}\bar{Z}_n) &= \text{var}\left(n^{-\frac{1}{2}}\sum_{t=1}^{n} Z_t\right) \\
&= E\left[\left(n^{-\frac{1}{2}}\sum_{t=1}^{n} Z_t\right)\left(n^{-\frac{1}{2}}\sum_{s=1}^{n} Z_s\right)'\right] \\
&= n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} E(Z_t Z_s') \\
&= n^{-1}\sum_{t=1}^{n} E(Z_t Z_t') \quad \text{(because } Z_t \text{ and } Z_s \text{ are independent for } t \neq s\text{)} \\
&= E(Z_t Z_t') \\
&= V.
\end{aligned}$$

In other words, the variance of $\sqrt{n}\bar{Z}_n$ is identical to the variance of each individual random vector Z_t.

Theorem 4.12 [Asymptotic Normality of OLS] *Under Assumptions 4.1-4.5, we have*

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

as $n \to \infty$, *where* $V \equiv \text{var}(X_t\varepsilon_t) = E(X_t X_t'\varepsilon_t^2)$.

Proof: Recall that

$$\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}^{-1}n^{-\frac{1}{2}}\sum_{t=1}^{n} X_t\varepsilon_t.$$

First, we consider the second term

$$n^{-\frac{1}{2}}\sum_{t=1}^{n} X_t\varepsilon_t.$$

Noting that $E(X_t\varepsilon_t) = 0$ by Assumption 4.3, and $\text{var}(X_t\varepsilon_t) = E(X_t X_t'\varepsilon_t^2) = V$, which is finite and p.d. by Assumption 4.5. Then, by the CLT for i.i.d. random sequences

$\{Z_t = X_t\varepsilon_t\}$, we have

$$\begin{aligned} n^{-\frac{1}{2}}\sum_{t=1}^{n} X_t\varepsilon_t &= \sqrt{n}\left(n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t\right) \\ &= \sqrt{n}\bar{Z}_n \\ &\xrightarrow{d} Z \sim N(0, V). \end{aligned}$$

On the other hand, as shown earlier, we have

$$\hat{Q} \xrightarrow{p} Q,$$

and so

$$\hat{Q}^{-1} \xrightarrow{p} Q^{-1}$$

given that Q is nonsingular so that the inverse function is continuous and well defined. It follows by the Slutsky Theorem that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta^o) &= \hat{Q}^{-1}n^{-\frac{1}{2}}\sum_{t=1}^{n} X_t\varepsilon_t \\ \xrightarrow{d} Q^{-1}Z &\sim N(0, Q^{-1}VQ^{-1}). \end{aligned}$$

This completes the proof. ■

Remarks:

The theorem implies that the asymptotic mean of $\sqrt{n}(\hat{\beta} - \beta^o)$ is equal to 0. That is, the mean of $\sqrt{n}(\hat{\beta} - \beta^o)$ is approximately 0 when n is large.

It also implies that the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is $Q^{-1}VQ^{-1}$. That is, the variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is approximately $Q^{-1}VQ^{-1}$. Because the asymptotic variance is a different concept from the variance of $\sqrt{n}(\hat{\beta} - \beta^o)$, we denote the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ as follows: $\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1}$.

We now consider a special case under which we can simplfy the expression of $\text{avar}(\sqrt{n}\hat{\beta})$.

Special Case: Conditional Homoskedasticity

Assumption 4.6: $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.

Theorem 4.13: *Suppose Assumptions 4.1–4.6 hold. Then as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

Proof: Under Assumption 4.6, we can simplify

$$\begin{aligned} V &= E(X_t X_t' \varepsilon_t^2) \\ &= E[E(X_t X_t' \varepsilon_t^2 | X_t)] \text{ by the law of iterated expectations} \\ &= E[X_t X_t' E(\varepsilon_t^2 | X_t)] \\ &= \sigma^2 E(X_t X_t') \\ &= \sigma^2 Q. \end{aligned}$$

The results follow immediately because

$$Q^{-1} V Q^{-1} = Q^{-1} \sigma^2 Q Q^{-1} = \sigma^2 Q^{-1}.$$

Remarks:
Under conditional homoskedasticity, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is

$$\text{avar}(\sqrt{n}\hat{\beta}) = \sigma^2 Q^{-1}.$$

Question: Is the OLS estimator $\hat{\beta}$ the BLUE estimator asymptotically (i.e., when $n \to \infty$)?

4.5 Asymptotic Variance Estimator

To construct confidence interval estimators or hypothesis tests, we need to estimate the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$, $\text{avar}(\sqrt{n}\hat{\beta})$. Because the expression of $\text{avar}(\sqrt{n}\hat{\beta})$ differs under conditional homoskedasticity and conditional heteroskedasticity respectively, we consider the estimator for $\text{avar}(\sqrt{n}\hat{\beta})$ under these two cases separately.

Case I: Conditional Homoskedasticity
In this case, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is

$$\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1} V Q^{-1} = \sigma^2 Q^{-1}.$$

Question: How to estimate Q?

Lemma 4.14: *Suppose Assumptions 4.1, 4.2 and 4.4 hold. Then*

$$\hat{Q} = n^{-1} \sum_{t=1}^{n} X_t X_t' \xrightarrow{p} Q.$$

Question: How to estimate σ^2?

Recalling that $\sigma^2 = E(\varepsilon_t^2)$, we use the sample residual variance estimator

$$\begin{aligned} s^2 &= e'e/(n-K) \\ &= \frac{1}{n-K}\sum_{t=1}^{n} e_t^2 \\ &= \frac{1}{n-K}\sum_{t=1}^{n}(Y_t - X_t'\hat{\beta})^2. \end{aligned}$$

Theorem 4.15 [Consistent Estimator for σ^2]: *Under Assumptions 4.1-4.4,*

$$s^2 \xrightarrow{p} \sigma^2.$$

Proof: Given that $s^2 = e'e/(n-K)$ and

$$\begin{aligned} e_t &= Y_t - X_t'\hat{\beta} \\ &= \varepsilon_t + X_t'\beta^o - X_t'\hat{\beta} \\ &= \varepsilon_t - X_t'(\hat{\beta} - \beta^o), \end{aligned}$$

we have

$$\begin{aligned} s^2 &= \frac{1}{n-K}\sum_{t=1}^{n}[\varepsilon_t - X_t'(\hat{\beta} - \beta^o)]^2 \\ &= \frac{n}{n-K}\left(n^{-1}\sum_{t=1}^{n}\varepsilon_t^2\right) \\ &\quad +(\hat{\beta} - \beta^o)'\left[(n-K)^{-1}\sum_{t=1}^{n} X_t X_t'\right](\hat{\beta} - \beta^o) \\ &\quad -2(\hat{\beta} - \beta^o)'(n-K)^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \\ &\quad \xrightarrow{p} 1\cdot\sigma^2 + 0\cdot Q\cdot 0 - 2\cdot 0\cdot 0 \\ &= \sigma^2 \end{aligned}$$

given that K is a fixed number (i.e., K does not grow with the sample size n), where we have made use of the WLLN in three places respectively.

We can then consistently estimate $\sigma^2 Q^{-1}$ by $s^2\hat{Q}^{-1}$.

Theorem 4.16 [Asymptotic Variance Estimator of $\sqrt{n}(\hat{\beta} - \beta^o)$] *Under Assumptions 4.1-4.4, we have*

$$s^2\hat{Q}^{-1} \xrightarrow{p} \sigma^2 Q^{-1}.$$

Remarks:

The asymptotic variance estimator of $\sqrt{n}(\hat{\beta} - \beta^o)$ is

$$s^2\hat{Q}^{-1} = s^2(\mathbf{X}'\mathbf{X}/n)^{-1}.$$

This is equivalent to saying that the variance estimator of $\hat{\beta} - \beta^o$ is approximately equal to

$$s^2\hat{Q}^{-1}/n = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

when for a large n. Thus, when $n \to \infty$ and there exists conditional homoskedasticity, the variance estimator of $\hat{\beta} - \beta^o$ coincides with the form of the variance estimator for $\hat{\beta} - \beta^o$ in the classical regression case. Because of this, as will be seen below, the conventional t-test and F-test are still valid for large samples under conditional homoskedasticity.

Case II: Conditional Heteroskedasticity

In this case,

$$\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1},$$

which cannot be simplified.

Question: We can still use $\hat{Q}$ to estimate Q. How to estimate $V = E(X_tX_t'\varepsilon_t^2)$?

We can use its sample analog

$$\hat{V} = n^{-1}\sum_{t=1}^{n} X_tX_t'e_t^2 = \frac{\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}}{n},$$

where

$$\mathrm{D}(e) = diag(e_1, e_2, ..., e_n)$$

is an $n \times n$ diagonal matrix with diagonal elements equal to e_t for $t = 1, ..., n$. To ensure consistency of $\hat{V}$ to V, we impose the following additional moment conditions.

Assumption 4.7: (i) $E(X_{jt}^4) < \infty$ for all $0 \le j \le k$; and (ii) $E(\varepsilon_t^4) < \infty$.
Lemma 4.17: *Suppose Assumptions 4.1–4.5 and 4.7 hold. Then*

$$\hat{V} \xrightarrow{p} V.$$

Proof: Because $e_t = \varepsilon_t - (\hat{\beta} - \beta^o)'X_t$, we have

$$\begin{aligned}\hat{V} &= n^{-1}\sum_{t=1}^{n} X_t X_t'\varepsilon_t^2 \\ &\quad +n^{-1}\sum_{t=1}^{n} X_t X_t'[(\hat{\beta}-\beta^o)'X_t X_t'(\hat{\beta}-\beta^o)] \\ &\quad -2n^{-1}\sum_{t=1}^{n} X_t X_t'[\varepsilon_t X_t'(\hat{\beta}-\beta^o)] \\ &\xrightarrow{p} \quad V+0-2\cdot 0,\end{aligned}$$

where for the first term, we have

$$n^{-1}\sum_{t=1}^{n} X_t X_t'\varepsilon_t^2 \xrightarrow{p} E(X_t X_t'\varepsilon_t^2) = V$$

by the WLLN and Assumption 4.7, which implies

$$E|X_{it}X_{jt}\varepsilon_t^2| \leq [E(X_{it}^2 X_{jt}^2)E(\varepsilon_t^4)]^{\frac{1}{2}}.$$

For the second term, we have

$$\begin{aligned}& n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}(\hat{\beta}-\beta^o)'X_t X_t'(\hat{\beta}-\beta^o) \\ = \;& \sum_{l=0}^{k}\sum_{m=0}^{k}(\hat{\beta}_l-\beta_l^o)(\hat{\beta}_m-\beta_m^o)\left(n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}X_{mt}\right) \\ & \xrightarrow{p} 0\end{aligned}$$

given $\hat{\beta}-\beta^o \xrightarrow{p} 0$, and

$$n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}X_{mt} \xrightarrow{p} E\left(X_{it}X_{jt}X_{lt}X_{mt}\right) = O(1)$$

by the WLLN and Assumption 4.7.

Similarly, for the last term, we have

$$\begin{aligned}& n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}\varepsilon_t X_t'(\hat{\beta}-\beta^o) \\ = \;& \sum_{l=0}^{k}(\hat{\beta}_l-\beta_l^o)\left(n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}\varepsilon_t\right) \\ & \xrightarrow{p} \quad 0\end{aligned}$$

given $\hat{\beta} - \beta^o \xrightarrow{p} 0$, and

$$n^{-1} \sum_{t=1}^{n} X_{it} X_{jt} X_{lt} \varepsilon_t \xrightarrow{p} E\left(X_{it} X_{jt} X_{lt} \varepsilon_t\right) = 0$$

by the WLLN and Assumption 4.7. This completes the proof.

We now construct a consistent estimator for avar$(\sqrt{n}\hat{\beta})$ under conditional heteroskedasticity.

Theorem 4.18 [Asymptotic variance estimator for $\sqrt{n}(\hat{\beta} - \beta^o)$]: *Under Assumptions 4.1–4.5 and 4.7, we have*

$$\hat{Q}^{-1} \hat{V} \hat{Q}^{-1} \xrightarrow{p} Q^{-1} V Q^{-1}.$$

Remarks:

This is the so-called White's (1980) heteroskedasticity-consistent variance-covariance matrix of the estimator $\sqrt{n}(\hat{\beta} - \beta^o)$. It follows that when there exists conditional heteroskedasticity, the estimator for the variance of $\hat{\beta} - \beta^o$ is

$$\begin{aligned} & (\mathbf{X}'\mathbf{X}/n)^{-1} \hat{V} (\mathbf{X}'\mathbf{X}/n)^{-1}/n \\ = \; & (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathrm{D}(e) \mathrm{D}(e)' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

which differs from the estimator $s^2(\mathbf{X}'\mathbf{X})^{-1}$ in the case of conditional homoskedasticity.

Question: What happens if we use $s^2 \hat{Q}^{-1}$ as an estimator for the avar$[\sqrt{n}(\hat{\beta} - \beta^o)]$ while there exists conditional heteroskedasticity?

Observe that

$$\begin{aligned} V &\equiv E(X_t X_t' \varepsilon_t^2) \\ &= \sigma^2 Q + \mathrm{cov}(X_t X_t', \varepsilon_t^2) \\ &= \sigma^2 Q + \mathrm{cov}[X_t X_t', \sigma^2(X_t)], \end{aligned}$$

where $\sigma^2 = E(\varepsilon_t^2)$, $\sigma^2(X_t) = E(\varepsilon_t^2 | X_t)$, and the last equality follows from the LIE. Thus, if $\sigma^2(X_t)$ is positively correlated with $X_t X_t'$, $\sigma^2 Q$ will underestimate the true variance-covariance $E(X_t X_t' \varepsilon_t^2)$ in the sense that $V - \sigma^2 Q$ is a positive definite matrix. Consequently, the standard t-test and F-test will overreject the correct null hypothesis at any given significance level. There will exist substantial Type I errors.

Question: What happens if one use the asymptotic variance estimator $\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}$ but there exists conditional homoskedasticity?

The asymptotic variance estimator is asymptotically valid, but it will not perform as well as the estimator $s^2\hat{Q}^{-1}$ in finite samples, because the latter exploits the information of conditonal homoskedasticity.

4.6 Hypothesis Testing

Question: How to construct a test statistic for the null hypothesis

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

where R is a $J \times K$ constant matrix, and r is a $J \times 1$ constant vector?

We first consider

$$R\hat{\beta} - r = R(\hat{\beta} - \beta^o) + R\beta^o - r.$$

It follows that under $\mathbf{H}_0 : R\beta^o = r$, we have

$$\sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} N(0, RQ^{-1}VQ^{-1}R').$$

The test procedures will differ depending on whether there exists conditional heteroskedasticity. We first consider the case of conditional homoskedasticity.

Case I: Conditional Homoskedasticity

Under conditional homoskedasticity, we have $V = \sigma^2 Q$ and so

$$\sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} N(0, \sigma^2 RQ^{-1}R')$$

when $\mathbf{H}_0$ holds.

When $J = 1$, we can use the conventional t-test statistic for large sample inference.

Theorem 4.19 [t-test]: *Suppose Assumptions 4.1-4.4 and 4.6 hold. Then under* $\mathbf{H}_0$ with $J = 1$,

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0, 1)$$

as $n \to \infty$.

Proof: Give $R\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2 RQ^{-1}R')$, $R\beta^o = r$ under $\mathbf{H}_0$, and $J = 1$, we have

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{\sigma^2 RQ^{-1}R'}} = \frac{R\sqrt{n}(\hat{\beta} - \beta^o)}{\sqrt{\sigma^2 RQ^{-1}R'}} \xrightarrow{d} N(0, 1).$$

By the Slutsky theorem and $\hat{Q} = \mathbf{X}'\mathbf{X}/n$, we obtain

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{s^2 R\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1).$$

This ratio is the conventional t-test statistic we examined in Chapter 3, namely:

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{s^2 R\hat{Q}^{-1}R'}} = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} = T.$$

For $J > 1$, we use a quadratic form test statistic.

Theorem 4.20 [Asymptotic χ^2 Test] *Suppose Assumptions 4.1–4.4 and 4.6 hold. Then under* $\mathbf{H}_0$,

$$\begin{aligned} J \cdot F &\equiv (R\hat{\beta} - r)' \left[s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} (R\hat{\beta} - r) \\ &\xrightarrow{d} \chi^2_J \end{aligned}$$

as $n \to \infty$.

Proof: Under $\mathbf{H}_0$, the quadratic form

$$\sqrt{n}(R\hat{\beta} - r)' \left(\sigma^2 RQ^{-1}R'\right)^{-1} \sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

Also, $s^2\hat{Q}^{-1} \xrightarrow{p} \sigma^2 Q^{-1}$, so we have by the Slutsky theorem

$$\sqrt{n}(R\hat{\beta} - r)' \left(s^2 R\hat{Q}^{-1}R'\right)^{-1} \sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

or equivalently

$$J \cdot \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} = J \cdot F \xrightarrow{d} \chi^2_J,$$

namely

$$J \cdot F \xrightarrow{d} \chi^2_J.$$

Remarks:

When $\{\varepsilon_t\}$ is not i.i.d.$N(0, \sigma^2)$ conditional on X_t, we cannot use the F distribution, but we can still compute the F-statistic and the appropriate test statistic is J times the F-statistic, which is asymptotically χ^2_J. That is,

$$J \cdot F = \frac{(\tilde{e}'\tilde{e} - e'e)}{e'e/(n-K)} \xrightarrow{d} \chi^2_J.$$

Because $J \cdot F_{J,n-K}$ approaches χ^2_J as $n \to \infty$, we may interpret the above theorem in the following way: the classical results for the F-test are still approximately valid under conditional homoskedasticity when n is large.

When the null hypothesis is that all slope coefficients except the intercept are jointly zero, we can use a test statistic based on R^2.

A Special Case: Testing for Joint Significance of All Economic Variables

Theorem 4.21 [$(n-K)R^2$ **Test**]: *Suppose Assumption 4.1-4.6 hold, and we are interested in testing the null hypothesis that*

$$\mathbf{H}_0 : \beta_1^o = \beta_2^o = \cdots = \beta_k^o = 0,$$

where the β are the regression coefficients from

$$Y_t = \beta_0^o + \beta_1^o X_{1t} + \cdots + \beta_k^o X_{kt} + \varepsilon_t.$$

Let R^2 be the coefficient of determination from the unrestricted regression model

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

Then under $\mathbf{H}_0$,

$$(n-K)R^2 \xrightarrow{d} \chi^2_k,$$

where $K = k + 1$.

Proof: First, recall that in this special case we have

$$\begin{aligned} F &= \frac{R^2/k}{(1-R^2)/(n-k-1)} \\ &= \frac{R^2/k}{(1-R^2)/(n-K)}. \end{aligned}$$

By the above theorem and noting $J = k$, we have

$$k \cdot F = \frac{(n-K)R^2}{1-R^2} \xrightarrow{d} \chi^2_k$$

under $\mathbf{H}_0$. This implies that $k \cdot F$ is bounded in probability; that is,

$$\frac{(n-K)R^2}{1-R^2} = O_P(1).$$

Consequently, given that k is a fixed integer,

$$\frac{R^2}{1-R^2} = O_P(n^{-1}) = o_P(1)$$

or

$$R^2 \xrightarrow{p} 0.$$

Therefore, $1 - R^2 \xrightarrow{p} 1$. By the Slutsky theorem, we have

$$\begin{aligned}(n-K)R^2 = &\quad k \cdot \frac{(n-K)R^2/k}{1-R^2}(1-R^2) \\ = &\quad (k \cdot F)(1-R^2) \\ &\xrightarrow{d} \chi^2_k,\end{aligned}$$

or asymptotically equivalently,

$$(n-K)R^2 \xrightarrow{d} \chi^2_k.$$

This completes the proof. ■

Question: Do we have $nR^2 \xrightarrow{d} \chi^2_k$?

Yes, we have

$$nR^2 = \frac{n}{n-K}(n-K)R^2 \text{ and } \frac{n}{n-K} \to 1.$$

Case II: Conditional Heteroskedasticity

Recall that under $\mathbf{H}_0$,

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

where

$$V = E(X_t X_t' \varepsilon_t^2).$$

Therefore, when $J = 1$,we have

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{RQ^{-1}VQ^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Given $\hat{Q} \xrightarrow{p} Q$ and $\hat{V} \xrightarrow{p} V$, where $\hat{V} = \mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}/n$, and the Slutsky theorem, we can define a robust t-test statistic

$$T_r = \frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

when $\mathbf{H}_0$ holds. By robustness, we mean that T_r is valid no matter whether there exists conditional heteroskedasticity.

Theorem 4.22 [Robust t-Test Under Conditional Heteroskedasticity] *Suppose Assumptions 4.1–4.5 and 4.7 hold. Then under* $\mathbf{H}_0$ with $J = 1$, *as* $n \to \infty$, the robust t-test statistic

$$T_r = \frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1).$$

When $J > 1$, we have the quadratic form

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

under $\mathbf{H}_0$. Given $\hat{Q} \xrightarrow{p} Q$ and $\hat{V} \xrightarrow{p} V$, the robust Wald test statistic

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

by the Slutsky theorem.

We can write W equivalently as follows:

$$W = (R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r),$$

where we have used the fact that

$$\begin{aligned} \hat{V} &= \frac{1}{n}\sum_{t=1}^{n} X_t e_t e_t X_t' \\ &= \frac{\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}}{n}, \end{aligned}$$

where $\mathrm{D}(e) = diag(e_1, e_2, ..., e_n)$.

Theorem 4.23 [Robust Wald Test Under Conditional Heteroskedasticity] *Suppose Assumptions 4.1–4.5 and 4.7 hold. Then under* $\mathbf{H}_0$, *as* $n \to \infty$,

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

Remarks:

Under conditional heteroskedasticity, the test statistics $J \cdot F$ and $(n-K)R^2$ cannot be used.

Question: What happens if there exists conditional heteroskedasticity but $J \cdot F$ or $(n-K)R^2$ is used.

There will exist Type I errors because $J \cdot F$ or $(n-K)R^2$ will be no longer asymptotically χ^2-distributed under $\mathbf{H}_0$.

Although the general form of the Wald test statistic developed here can be used no matter whether there exists conditional homoskedasticity, this general form of test statistic may perform poorly in small samples. Thus, if one has information that the error term is conditionally homoskedastic, one should use the test statistics derived under conditional homoskedasticity, which will perform better in small sample sizes. Because of this reason, it is important to test whether conditional homoskedasticity holds.

4.7 Testing for Conditional Homoskedasticity

We now introduce a method to test conditional heteroskedasticity.

Question: How to test conditional homoskedasticity for $\{\varepsilon_t\}$ in a linear regression model?

There have been many tests for conditional homoskedasticity. Here, we introduce a popular one due to White (1980).

White's (1980) test

The null hypothesis

$$\mathbf{H}_0 : E(\varepsilon_t^2 | X_t) = \sigma^2,$$

where ε_t is the regression error in the linear regression model

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

First, suppose ε_t were observed, and we consider the auxiliary regression

$$\begin{aligned} \varepsilon_t^2 &= \gamma_0 + \sum_{j=1}^{k} \gamma_j X_{jt} + \sum_{1 \le j \le l \le k} \gamma_{jl} X_{jt} X_{lt} + v_t \\ &= \gamma' \text{vech}(X_t X_t') + v_t \\ &= \gamma' U_t + v_t, \end{aligned}$$

where $\text{vech}(X_tX_t')$ is an operator stacks all lower triangular elements of matrix X_tX_t' into a $\frac{K(K+1)}{2} \times 1$ column vector. For example, when $X_t = (1, X_{1t}, X_{2t})'$, we have

$$\text{vech}(X_tX_t') = (1, X_{1t}, X_{2t}, X_{1t}^2, X_{1t}X_{2t}, X_{2t}^2)'.$$

For the auxiliary regression, there is a total of $\frac{K(K+1)}{2}$ regressors in U_t. This is essentially regressing ε_t^2 on the intercept, X_t, and the quadratic terms and cross-product terms of X_t. Under $\mathbf{H}_0$, all coefficients except the intercept are jointly zero. Any nonzero coefficients will indicate the existence of conditional heteroskedasticity. Thus, we can test $\mathbf{H}_0$ by checking whether all coefficients except the intercept are jointly zero. Assuming that $E(\varepsilon_t^4|X_t) = \mu_4$ (which implies $E(v_t^2|X_t) = \sigma_v^2$ under $\mathbf{H}_0$), then we can run an OLS regression and construct a R^2-based test statistic. Under $\mathbf{H}_0$, we can obtain

$$(n - J - 1)\tilde{R}^2 \xrightarrow{d} \chi_J^2,$$

where $J = \frac{K(K+1)}{2} - 1$ is the number of the regressors except the intercept.
Unfortunately, ε_t is not observable. However, we can replace ε_t with $e_t = Y_t - X_t'\hat{\beta}$, and run the following feasible auxiliary regression

$$\begin{aligned} e_t^2 &= \gamma_0 + \sum_{j=1}^{k} \gamma_j X_{jt} + \sum_{1 \le j \le l \le k} \gamma_{jl} X_{jt} X_{lt} + \tilde{v}_t \\ &= \gamma' \text{vech}(X_tX_t') + \tilde{v}_t, \end{aligned}$$

the resulting test statistic

$$(n - J - 1)R^2 \xrightarrow{d} \chi_J^2.$$

It can be shown that the replacement of ε_t^2 by e_t^2 has no impact on the asymptotic χ_J^2 distribution of $(n - J - 1)R^2$. The proof, however, is rather tedious. For the details of the proof, see White (1980). Below, we provide some intuition.

Question: Why does the use of e_t^2 in place of ε_t^2 have no impact on the asymptotic distribution of $(n - J - 1)R^2$?

To explain this, we put $U_t = \text{vech}(X_tX_t')$. Then the infeasible auxiliary regression is

$$\varepsilon_t^2 = U_t'\gamma^0 + v_t.$$

We have $\sqrt{n}(\tilde{\gamma} - \gamma^0) \xrightarrow{d} N(0, \sigma_v^2 Q_{\text{uu}}^{-1})$, where $Q_{\text{uu}} = E(U_tU_t')$, and under $\mathbf{H}_0 : R\gamma^0 = 0$, where R is a $J \times J$ diagonal matrix with the first diagonal element being 0 and other diagonal elements being 1, we have

$$\sqrt{n}R\tilde{\gamma} \xrightarrow{d} N(0, \sigma_v^2 R Q_{\text{uu}}^{-1} R'),$$

where $\tilde{\gamma}$ is the OLS estimator and $\sigma_v^2 = E(v_t^2)$. This implies $R\tilde{\gamma} = O_P(n^{-1/2})$, which vanishes to zero in probability at rate $n^{-1/2}$. It is this term that yields the asymptotic χ_J^2 distribution for $(n-J-1)\tilde{R}^2$, which is asymptotically equivalent to the test statistic

$$\sqrt{n}(R\tilde{\gamma})'[s_v^2 R\hat{Q}_{\text{uu}}^{-1}R']^{-1}\sqrt{n}R\tilde{\gamma}.$$

Now suppose we replace ε_t^2 with e_t^2, and consider the auxiliary regression

$$e_t^2 = U_t'\gamma^0 + \tilde{v}_t.$$

Denote the OLS estimator by $\hat{\gamma}$. We decompose

$$\begin{aligned} e_t^2 &= \left[\varepsilon_t - X_t'(\hat{\beta} - \beta^o)\right]^2 \\ &= \varepsilon_t^2 + (\hat{\beta} - \beta^o)'X_tX_t'(\hat{\beta} - \beta^o) - 2(\hat{\beta} - \beta^o)'X_t\varepsilon_t \\ &= \gamma'U_t + \tilde{v}_t. \end{aligned}$$

Thus, $\hat{\gamma}$ can be written as follows:

$$\hat{\gamma} = \tilde{\gamma} + \hat{\delta} + \hat{\eta},$$

where $\tilde{\gamma}$ is the OLS estimator of the infeasible auxiliary regression, $\hat{\delta}$ is the effect of the second term, and $\hat{\eta}$ is the effect of the third term. For the third term, $X_t\varepsilon_t$ is uncorrelated with U_t given $E(\varepsilon_t|X_t) = 0$. Therefore, this term, after scaled by the factor $\hat{\beta} - \beta^o$ that itself vanishes to zero in probability at the rate $n^{-1/2}$, will vanish to zero in probability at a rate n^{-1}, that is, $\hat{\eta} = O_P(n^{-1})$. This is expected to have negligible impact on the asymptotic distribution of the test statistic. For the second term, X_tX_t' is perfectly correlated with U_t. However, it is scaled by a factor of $||\hat{\beta} - \beta^o||^2$ rather than by $||\hat{\beta} - \beta^o||$ only. As a consequence, the regression coefficient of $(\hat{\beta} - \beta^o)'X_tX_t'(\hat{\beta} - \beta^o)$ on U_t will also vanish to zero at rate n^{-1}, that is, $\hat{\delta} = O_P(n^{-1})$. Therefore, it also has negligible impact on the asymptotic distribution of $(n-J-1)R^2$.

Question: How to test conditional homoskedasticity if $E(\varepsilon_t^4|X_t)$ is not a constant (i.e., $E(\varepsilon_t^4|X_t) \neq \mu_4$ for some μ_4 under $\mathbf{H}_0$)? This corresponds to the case when v_t displays conditional heteroskedasticity.

Question: Suppose White's (1980) test rejects the null hypothesis of conditional homoskedasticity, one can then conclude that there exists evidence of conditional heteroskedasticity. What conclusion can one reach if White's test fails to reject $\mathbf{H}_0$: $E(\varepsilon_t^2|X_t) = \sigma^2$?

Because White (1980) considers a quadratic alternative to test $\mathbf{H}_0$, it may have no power against some conditional heteroskedastic alternatives for which $E(\varepsilon_t^2|X_t)$ does not depend on the quadratic form of X_t but depends on cubic or higher order polynomials of X_t. Thus, when White's test fails to reject $\mathbf{H}_0$, one can only say that we find no evidence against $\mathbf{H}_0$.

However, when White's test fails to reject $\mathbf{H}_0$, we have

$$E(\varepsilon_t^2 X_t X_t') = \sigma^2 E(X_t X_t') = \sigma^2 Q$$

even if $\mathbf{H}_0$ is false. Therefore, one can use the conventional variance-covariance matrix estimator $s^2(X'X)^{-1}$ for $\hat{\beta}$. Indeed, the main motivation for White's (1980) test for conditional heteroskedasticity is whether the heteroskedasticity-consistent variance-covariance matrix of $\hat{\beta}$ has to be used, not really whether conditional heteroskedasticity exists. For this purpose, it suffices to regress ε_t^2 or e_t^2 on the quadratic form of X_t. This can be seen from the decomposition

$$V = E(X_t X_t' \varepsilon_t^2) = \sigma^2 Q + \text{cov}(X_t X_t', \varepsilon_t^2),$$

which indicates that $V = \sigma^2 Q$ if and only if ε_t^2 is uncorrelated with $X_t X_t'$.

The validity of White's test procedure and associated interpretations is built upon the assumption that the linear regression model is correctly specified for the conditional mean $E(Y_t|X_t)$.Suppose the linear regression model is not correctly specified, i.e., $E(Y_t|X_t) \neq X_t'\beta$ for all β. Then the OLS $\hat{\beta}$ will converge to $\beta^* = [E(X_t X_t')]^{-1} E(X_t Y_t)$, the best linear least squares approximation coefficient, and $E(Y_t|X_t) \neq X_t'\beta^*$. In this case, the estimated residual

$$\begin{aligned} e_t &= Y_t - X_t'\hat{\beta} \\ &= \varepsilon_t + [E(Y_t|X_t) - X_t'\beta^*] + X_t'(\beta^* - \hat{\beta}), \end{aligned}$$

where $\varepsilon_t = Y_t - E(Y_t|X_t)$ is the true disturbance with $E(\varepsilon_t|X_t) = 0$, the estimation error $X_t'(\beta^* - \hat{\beta})$ vanishes to 0 as $n \to \infty$, but the approximation error $E(Y_t|X_t) - X_t'\beta^*$ never disappears. In other words, when the linear regression model is misspecified for $E(Y_t|X_t)$, the estimated residual e_t will contain not only the true disturbance but also the approximation error which is a function of X_t. This will result in a spurious conditional heteroskedasticity when White's test is used. Therefore, before using White's test or any other tests for conditional heteroskedasticity, it is important to first check whether the linear regression model is correctly specified. For tests of correct specification of a linear regression model, see Hausman's test in Chapter 7 and other specification tests mentioned there.

4.8 Empirical Applications

4.9 Conclusion

In this chapter, within the context of i.i.d. observations, we have relaxed some key assumptions of the classical linear regression model. In particular, we do not assume conditional normality for ε_t and allow for conditional heteroskedasticity. Because the exact finite sample distribution of the OLS is generally unknown, we have relied on asymptotic analysis. It is found that for large samples, the results of the OLS estimator $\hat{\beta}$ and related test statistics (e.g., t-test statistic and F-test statistic) are still applicable under conditional homoskedasticity. Under conditional heteroskedasticity, however, the statistic properties of $\hat{\beta}$ are different from those of $\hat{\beta}$ under conditional homoskedasticity, and as a consequence, the conventional t-test and F-test are invalid even when the sample size $n \to \infty$. One has to use White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator for the OLS estimator $\hat{\beta}$ and use it to construct robust test statistics. A direct test for conditional heteroskedasticity, due to White (1980), is described.

The asymptotic theory provides convenient inference procedures in practice. However, the finite sample distribution of $\hat{\beta}$ may be different from its asymptotic distribution. How well the approximation of the asymptotic distribution for the unknown finite sample distribution depends on the data generating process and the sample size of the data. In econometrics, simulation studies have been used to examine how well asymptotic theory can approximate the finite sample distributions of econometric estimators or related statistics. They are the nearest approach that econometricians can make to the laboratory experiments of the physical sciences and are a very useful way of reinforcing or checking the theoretical results. Alternatively, resampling methods called bootstrap have been proposed in econometrics to approximate the finite sample distributions of econometric estimators or related statistics by simulating data on a computer. In this book, we focus on asymptotic theory.

EXERCISES

4.1. Suppose Assumptions 3.1, 3.3 and 3.5 hold. Show (a) s^2 converges in probability to σ^2, and (b) s converges in probability to σ.

4.2. Let $Z_1, ..., Z_n$ be a random sample from a population with mean μ and variance σ^2. Show that

$$E\left[\frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma}\right] = 0 \text{ and } Var\left[\frac{\sqrt{n}(\bar{Z}_n - \mu)}{\sigma}\right] = 1.$$

4.3. Suppose a sequence of random variables $\{Z_n, n = 1, 2, ...\}$ is defined as

$$\begin{array}{cccc} Z_n & & \frac{1}{n} & n \\ P_{Z_n} & 1 - \frac{1}{n} & \frac{1}{n} & \end{array}$$

(a) Does Z_n converges in mean squares to 0? Give your reasoning clearly.

(b) Does Z_n converges in probability to 0? Give your reasoning clearly.

4.4. Let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define $Z(s) = s$ for all $s \in [0, 1]$. Also, for $n = 1, 2, ...$, define a sequence of random variables

$$Z_n(s) = \begin{cases} s + s^n & \text{if } s \in [0, 1 - n^{-1}] \\ s + 1 & \text{if } s \in (1 - n^{-1}, 1]. \end{cases}$$

(a) Does Z_n converge in quadratic mean to Z?

(a) Does Z_n converge in probability to Z?

(b) Does Z_n converge almost surely to Z?

4.5. Suppose $g(\cdot)$ is a real-valued continuous function, and $\{Z_n, n = 1, 2, ...\}$ is a sequence of real-valued random variables which converges in probability to random variable Z. Show $g(Z_n) \xrightarrow{p} g(Z)$.

4.6. Suppose a stochastic process $\{Y_t, X_t'\}_{t=1}^{'n}$ satisfies the following assumptions:

Assumption 1.1 [Linearity] $\{Y_t, X_t'\}_{t=1}^{'n}$ is an i.i.d. process with

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

for some unknown parameter β^o and some unobservable disturbance ε_t;

Assumption 1.2 [i.i.d.] The $K \times K$ matrix $E(X_t X_t') = Q$ is nonsingular and finite;

Assumption 1.3 [conditional heteroskedasticity]:
(i) $E(X_t \varepsilon_t) = 0$;
(ii) $E(\varepsilon_t^2 | X_t) \neq \sigma^2$;
(iii) $E(X_{jt}^4) \leq C$ for all $0 \leq j \leq k$, and $E(\varepsilon_t^4) \leq C$ for some $C < \infty$.

(a) Show that $\hat{\beta} \xrightarrow{p} \beta^o$?
(b) Show that $\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \Omega)$, where $\Omega = Q^{-1} V Q^{-1}$, and $V = E(X_t X_t' \varepsilon_t^2)$.
(c) Show that the asymptotic variance estimator

$$\hat{\Omega} = \hat{Q}^{-1} \hat{V} \hat{Q}^{-1} \xrightarrow{p} \Omega,$$

where $\hat{Q} = n^{-1} \sum_{t=1}^n X_t X_t'$ and $\hat{V} = n^{-1} \sum_{t=1}^n X_t X_t' e_t^2$. This is called White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator.
(d) Consider a test for hypothesis $\mathbf{H}_0 : R\beta^o = r$. Do we have $J \cdot F \xrightarrow{d} \chi_J^2$, where

$$F = \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1} R']^{-1} (R\hat{\beta} - r)/J}{s^2}$$

is the usual F-test statistic? If it holds, give the reasoning. If it does not, could you provide an alternative test statistic that converges in distribution to χ_J^2.

4.7. Put $Q = E(X_t X_t'), V = E(\varepsilon_t^2 X_t X_t')$ and $\sigma^2 = E(\varepsilon_t^2)$. Suppose there exists conditional heteroskedasticity, and $\text{cov}(\varepsilon_t^2, X_t X_t') = V - \sigma^2 Q$ is positive semi-definite, i.e, $\sigma^2(X_t)$ is positively correlated with $X_t X_t'$. Show that $Q^{-1} V Q^{-1} - \sigma^2 Q^{-1}$ is positive semi-definite.

4.8. Suppose the following assumptions hold:

Assumption 2.1: $\{Y_t, X_t'\}_{t=1}^{\prime n}$ is an i.i.d. random sample with

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

for some unknown parameter β^o and unobservable random disturbance ε_t.

Assumption 2.2: $E(\varepsilon_t | X_t) = 0$ a.s.

Assumption 2.3:
(i) $W_t = W(X_t)$ is a positive function of X_t;
(ii) The $K \times K$ matrix $E(X_t W_t X_t') = Q_w$ is finite and nonsingular.
(iii) $E(W_t^8) \leq C < \infty$, $E(X_{jt}^8) \leq C < \infty$ for all $0 \leq j \leq k$, and $E(\varepsilon_t^4) \leq C$;

Assumption 2.4: $V_w = E(W_t^2 X_t X_t' \varepsilon_t^2)$ is finite and nonsingular.

We consider the so-called weighted least squares (WLS) estimator for β^o :

$$\hat{\beta}_w = \left(n^{-1} \sum_{t=1}^{n} X_t W_t X_t' \right)^{-1} n^{-1} \sum_{t=1}^{n} X_t W_t Y_t.$$

(a) Show that $\hat{\beta}_w$ is the solution to the following problem

$$\min_{\beta} \sum_{t=1}^{n} W_t (Y_t - X_t'\beta)^2.$$

(b) Show that $\hat{\beta}_w$ is consistent for β^o;
(c) Show that $\sqrt{n}(\hat{\beta}_w - \beta^o) \xrightarrow{d} N(0, \Omega_w)$ for some $K \times K$ finite and positive definite matrix Ω_w. Obtain the expression of Ω_w under (i) conditional homoskedasticity $E(\varepsilon_t^2 | X_t) = \sigma^2$ a.s. and (ii) conditional heteroskedasticity $E(\varepsilon_t^2 | X_t) \neq \sigma^2$.
(d) Propose an estimator $\hat{\Omega}_w$ for Ω_w, and show that $\hat{\Omega}_w$ is consistent for Ω_w under conditional homoskedasticity and conditional heteroskedasticity respectively.
(e) Construct a test statistic for $\mathbf{H}_0 : R\beta^o = r$, where R is a $J \times K$ matrix and r is a $J \times 1$ vector under conditional homoskedasticity and under conditional heteroskedasticity respectively. Derive the asymptotic distribution of the test statistic under $\mathbf{H}_0$ in each case.
(f) Suppose $E(\varepsilon_t^2 | X_t) = \sigma^2(X_t)$ is
known, and we set $W_t = \sigma^{-1}(X_t)$. Construct a test statistic for $\mathbf{H}_0 : R\beta^o = r$, where R is a $J \times K$ matrix and r is a $J \times 1$ vector. Derive the asymptotic distribution of the test statistic under $\mathbf{H}_0$.

4.9. Consider the problem of testing conditional homoskedasticity ($\mathbf{H}_0 : E(\varepsilon_t^2 | X_t) = \sigma^2$) for a linear regression model

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

where X_t is a $K \times 1$ vector consisting of an intercept and explanatory variables. To test conditional homoskedasticity, we consider the auxiliary regression

$$\begin{aligned} \varepsilon_t^2 &= \text{vech}(X_t X_t')'\gamma + v_t \\ &= U_t'\gamma + v_t, \end{aligned}$$

Show that under $\mathbf{H}_0 : E(\varepsilon_t^2 | X_t) = \sigma^2$, (a) $E(v_t | X_t) = 0$, and (b) $E(v_t^2 | X_t) = \sigma_v^2$ if and only if $E(\varepsilon_t^4 | X_t) = \mu_4$ for some constant μ_4.

4.10. Consider the problem of testing conditional homoskedasticity ($\mathbf{H}_0 : E(\varepsilon_t^2|X_t) = \sigma^2$) for a linear regression model

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

where X_t is a $K \times 1$ vector consisting of an intercept and explanatory variables. To test conditional homoskedasticity, we consider the auxiliary regression

$$\begin{aligned} \varepsilon_t^2 &= \text{vech}(X_t X_t')'\gamma + v_t \\ &= U_t'\gamma + v_t. \end{aligned}$$

Suppose Assumptions 4.1, 4.2, 4.3, 4.4, 4.7 hold, and $E(\varepsilon_t^4|X_t) \neq \mu_4$. That is, $E(\varepsilon_t^4|X_t)$ is a function of X_t.

(a) Show $\text{var}(v_t|X_t) \neq \sigma_v^2$ under $\mathbf{H}_0$. That is, the disturbance v_t in the auxiliary regression model displays conditional heteroskedasticity.

(b) Suppose ε_t is directly observable. Construct an asymptotically valid test for the null hypothesis $\mathbf{H}_0$ of conditional homoskedasticity of ε_t. Justify your reasoning and test statistic.

CHAPTER 5 LINEAR REGRESSION MODELS WITH DEPENDENT OBSERVATIONS

Abstract: In this chapter, we will show that the asymptotic theory for linear regression models with i.i.d. observations carries over to linear time series regression models with martingale difference sequence disturbances. Some basic concepts in time series analysis are introduced, and some tests for serial correlation are described.

Key words: Dynamic regression model, Ergodicity, Martingale difference sequence, Random walk, Serial correlation, Static regression model, Stationarity, Time series, Unit root, White noise.

Motivation

The asymptotic theory developed in Chapter 4 is applicable for cross-sectional data (due to the i.i.d. random sample assumption). What happens if we have time series data? Could the asymptotic theory for linear regression models with i.i.d. observations be applicable to linear regression models with time series observations?

Consider a simple regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, \\ \{\varepsilon_t\} &\sim \text{i.i.d.} N(0, \sigma^2). \end{aligned}$$

Here, $X_t = (1, Y_{t-1})'$. This is called an autoregression model, which violates the i.i.d. assumption for $\{Y_t, X_t'\}_{t=1}^n$ in Chapter 4. Here, we have

$$E(\varepsilon_t | X_t) = 0 \text{ a.s.}$$

but we no longer have

$$\begin{aligned} E(\varepsilon_t | X) &= E(\varepsilon_t | X_1, X_2, ..., X_n) \\ &= 0 \text{ a.s.} \end{aligned}$$

because X_{t+j} contains ε_t when $j > 0$. Hence, Assumption 3.2 (strict exogeneity) fails.

In general, the i.i.d. assumption for $\{Y_t, X_t'\}_{t=1}^n$ in Chapter 4 rules out time series data. Most economic and financial data are time series observations.

Question: Under what conditions will the asymptotic theory developed in Chapter 4 carry over to linear regression models with dependent observations?

5.1 Introduction to Time Series Analysis

To establish asymptotic theory for linear regression models with time series observations, we need to first introduce some basic concepts in time series.

Question: What is a time series process?

A time series process can be stochastic or deterministic. In this book, we only consider stochastic time series processes, which is consistent with the fundamental axiom of modern econometrics discussed in Chapter 1.

Definition 5.1 [Stochastic Time Series Process]: *A stochastic time series* $\{Z_t\}$ *is a sequence of random variables or random vectors indexed by time* $t \in \{..., 0, 1, 2, ...\}$ *and governed by some probability law* (Ω, F, P)*, where* Ω *is the sample space,* F *is a* σ*-field, and* P *is a probability measure, with* $P : F \to [0, 1]$.

Remarks:

More precisely, we can write $Z_t = Z(t, \cdot)$, and its realization $z_t = Z(t, \omega)$, where $\omega \in \Omega$ is a basic outcome in sample space Ω.

For each ω, we can obtain a sample path $z_t = Z(t, \omega)$ of the process $\{Z_t\}$ as a deterministic function of time t. Different ω's will give different sample paths.

The dynamics of $\{Z_t\}$ is completely determined by the *transition probability* of Z_t; that is, the *conditional probability* of Z_t given its past history $I_{t-1} = \{Z_{t-1}, Z_{t-2}, ...\}$.

Time Series Random sample: Consider a subset (or a segment) of a time series process $\{Z_t\}$ for $t = 1, \cdots, n$. This is called a time series random sample of size n, denoted as

$$Z^n = \{Z_1, \cdots, Z_n\}'.$$

Any realization of this random sample is called a data set, denoted as

$$z^n = \{z_1, \cdots, z_n\}'.$$

This corresponds to the occurrence of some specific outcome $\omega \in \Omega$. In theory, a random sample Z^n can generate many data sets, each corresponding to a specific $\omega \in \Omega$. In reality, however, one only observes a data set for any random sample of the economic process, due to the nonexperimental nature of the economic system.

Question: Why can the dynamics of $\{Z_t\}$ be completely captured by its conditional probability distribution?

Consider the random sample Z^n. It is well-known from basic statistics courses that the joint probability distribution of the random sample Z^n,

$$f_{Z^n}(z^n) = f_{Z_1,Z_2,...,Z_n}(z_1, z_2, ..., z_n), \qquad z^n \in \mathbb{R}^n,$$

completely captures all the sample information contained in Z^n. With $f_{Z^n}(z^n)$, we can, in theory, obtain the sampling distribution of any statistic (e.g., sample mean estimator, sample variance estimator, confidence interval estimator) that is a function of Z^n.

Now, by sequential partition (repeating the multiplication rule $P(A \cap B) = P(A|B)P(B)$ for any event A and B), we can write

$$f_{Z^n}(z^n) = \prod_{t=1}^{n} f_{Z_t|I_{t-1}}(z_t|I_{t-1}),$$

where by convention, for $t = 1$, $f(z_1|I_0) = f(z_1)$, the marginal density of Z_1. Thus, the conditional density function $f_{Z_t|I_{t-1}}(z|I_{t-1})$ completely describes the joint probability of the random sample Z^n.

Example 1: Let Z_t be the US Gross Domestic Product (GDP) in quarter t. Then the quarterly records of U.S. GDP from the first quarter of 1961 to the last quarter of 2001 constitute a time series data set, denoted as $z^n = (z_1, \cdots, z_n)'$, with $n = 164$.

Example 2: Let Z_t be the S&P 500 closing price index at day t. Then the daily records of S & P 500 index from July 2, 1962 to December 31, 2001 constitute a time series data set, denoted as $z^n = (z_1, \cdots, z_n)'$, with $n = 9987$.

Here is a fundamental feature of economic time series: each random variable Z_t only has one observed realization z_t in practice. It is impossible to obtain more realizations for each economic variable Z_t, due to the nonexperimental nature of an economic system. In order to "aggregate" realizations from different random variables $\{Z_t\}_{t=1}^n$, we need to impose stationarity—a concept of stability for certain aspects of the probability law $f_{Z_t|I_{t-1}}(z_t|I_{t-1})$. For example, we may need to assume:
(i) The marginal probability of each Z_t shares some common features (e.g., the same mean, the same variance).
(ii) The relationship (joint distribution) between Z_t and I_{t-1} is time-invariant in certain aspects (e.g., $\text{cov}(Z_t, Z_{t-j}) = \gamma(j)$ does not depend on time t; it only depends on the time distance j).

With these assumptions, observations from different random variables $\{Z_t\}$ can be viewed to contain some common features of the data generating process, so that one can conduct statistical inference by pooling them together.

Stationarity

A stochastic time series $\{Z_t\}$ can be stationary or nonstationary. There are at least two notions for stationarity. The first is strict stationarity.

Definition 5.2 [Strict Stationarity]: *A stochastic time series process $\{Z_t\}$ is strictly stationary if for any admissible $t_1, t_2, \cdots, t_m$, the joint probability distribution of $\{Z_{t_1}, Z_{t_2}, \cdots, Z_{t_m}\}$ is the same as the joint distribution of $\{Z_{t_1+k}, Z_{t_2+k}, \cdots, Z_{t_m+k}\}$ for all integers k. That is,*

$$f_{Z_{t_1} Z_{t_2} \cdots Z_{t_m}}(z_1, \cdots, z_m) = f_{Z_{t_1+k} Z_{t_2+k} \cdots Z_{t_m+k}}(z_1, \cdots, z_m).$$

Remarks:

If Z_t is strictly stationary, the conditional probability of Z_t given I_{t-1} will have a time-invariant functional form. In other words, the probabilistic structure of a completely stationary process is invariant under a shift of the time origin.

Strict stationarity is also called "complete stationarity", because it characterizes the time-invariance property of the entire joint probability distribution of the process $\{Z_t\}$.

No moment condition on $\{Z_t\}$ is needed when defining strict stationarity. Thus, a strictly stationary process may not have finite moments (e.g., $\text{var}(Z_t) = \infty$). However, if moments (e.g., $E(Z_t)$) and cross-moments ($e.g., E(Z_t Z_{t-j})$) of $\{Z_t\}$ exist, then they are time-invariant when $\{Z_t\}$ is strictly stationary.

Any measurable transformation of a strictly stationary process is still strictly stationary.

Strict stationarity implies identical distribution for each of the Z_t. Thus, although a strictly stationary time series data are realizations from different random variables, they can be viewed as realizations from the same (marginal) population distribution.

Example 3: Suppose $\{Z_t\}$ is an i.i.d. Cauchy $(0, 1)$ sequence with marginal pdf

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

Note that Z_t has no moment. Consider $\{Z_{t_1}, \cdots, Z_{t_m}\}$. Because the joint distribution

$$f_{Z_{t_1} Z_{t_2} \cdots Z_{t_m}}(z_1, \cdots, z_m) = \prod_{j=1}^{m} f(z_j)$$

is time-invariant, $\{Z_t\}$ is strictly stationary.

We now introduce another concept of stationarity based on the time-invariance property of the joint moments of $\{Z_{t_1}, Z_{t_2}, ..., Z_{t_m}\}$.

Definition 5.3 [N**-th order stationarity**]: *The time series process* $\{Z_t\}$ *is said to be stationary up to order* N *if, for any admissible* $t_1, t_2, \cdots, t_m$, *and any* k, *all the joint moments up to order* N *of* $\{Z_{t_1}, Z_{t_2}, \cdots, Z_{t_m}\}$ *exist and equal to the corresponding joint moments up to order* N *of* $\{Z_{t_1+k}, \cdots, Z_{t_m+k}\}$. *That is,*

$$E\left[(Z_{t_1})^{n_1} \cdots (Z_{t_m})^{n_m}\right] = E\left[(Z_{t_1+k})^{n_1} \cdots (Z_{t_m+k})^{n_m}\right],$$

for any k *and all nonnegative integers* $n_1, \cdots, n_m$ *satisfying* $\sum_{j=1}^{m} n_j \leq N$.

Remarks:

Setting $n_2 = n_3 = \cdots = n_m = 0$, we have

$$E\left[(Z_t)^{n_1}\right] = E\left[(Z_0)^{n_1}\right] \text{ for all } t.$$

On the other hand, for $n_1 + n_2 \leq N$, we have the pairwise joint product moment

$$\begin{aligned} E\left[(Z_t)^{n_1}(Z_{t-j})^{n_2}\right] &= E\left[(Z_0)^{n_1}(Z_{-j})^{n_2}\right] \\ &= \text{ function of } j, \end{aligned}$$

where j is called a lag order.

We now consider a special case: $N = 2$. This yields a concept called weak stationarity.

Definition 5.4 [Weak Stationarity] *A stochastic time series process* $\{Z_t\}$ *is weakly stationary if*

(i) $E(Z_t) = \mu$ *for all* t;

(ii) $var(Z_t) = \sigma^2 < \infty$ *for all* t;

(iii) $cov(Z_t, Z_{t-j}) = \gamma(j)$ *is only a function of lag order* j *for all* t.

Remarks:

Strict stationarity is defined in terms of the "time invariance" property of the entire distribution of $\{Z_t\}$, while weak-stationarity is defined in terms of the "time-invariance" property in the first two moments (means, variances and covariances) of $\{Z_t\}$. Suppose all moments of $\{Z_t\}$ exist. Then it is possible that the first two moments are time-invariant but the higher order moments are time-varying. In other words, a process $\{Z_t\}$ can be weakly stationary but not strictly stationary. However, Example 1 shows that a process can be strictly stationary but not weakly stationary, because the first two moments simply do not exist.

Weak stationarity is also called "covariance-stationarity", or "2nd order stationarity" because it is based on the time-invariance property of the first two moments. It does not require identical distribution for each of the Z_t. The higher order moments of Z_t can be different for different t's.

Question: Which, strict or weak stationarity, is more restrictive?

We consider two cases:
(i) If $E(Z_t^2) < \infty$, then strict stationarity implies weak stationarity.
(ii) However, if $E(Z_t^2) = \infty$, strict stationarity does not imply weak stationarity. In other words, a time series process can be strictly stationary but not weakly stationary.

Example 4: An i.i.d. Cauchy$(0, 1)$ process is strictly stationary but not weakly stationary.

A special but important weakly stationary time series is a process with zero autocorrelations.

Definition 5.5 [White Noise]: *A time series process $\{Z_t\}$ is a white noise (or serially uncorrelated) process if*
(i) $E(Z_t) = 0$.
(ii) $var(Z_t) = \sigma^2$,
(iii) $cov(Z_t, Z_{t-j}) = \gamma(j) = 0$ *for all* $j > 0$.

Remarks:

Later we will explain why such a process is called a white noise (WN) process. WN is a basic building block for linear time series modeling.

When $\{Z_t\}$ is a white noise and $\{Z_t\}$ is a Gaussian process (i.e., any finite set $(Z_{t_1}, Z_{t_2}, ..., Z_{t_m})$ of $\{Z_t\}$ has a joint normal distribution), we call $\{Z_t\}$ is a Gaussian white noise. For a Gaussian white noise process, $\{Z_t\}$ is an i.i.d. sequence.

Example 5: A first order autoregressive (AR(1)) process

$$\begin{aligned} Z_t &= \alpha Z_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim \text{white noise } (0, \sigma^2) \end{aligned}$$

is weakly stationary if $|\alpha| < 1$ (Z_t is a unit root process if $\alpha = 1$) because $Z_t =$

$\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j}$, and

$$\begin{aligned} E(Z_t) &= 0, \\ \text{var}(Z_t) &= \frac{\sigma^2}{1-\alpha^2}, \\ \gamma(j) &= \frac{\sigma^2}{1-\alpha^2}\alpha^{|j|}, \qquad j = 0, \pm 1, \pm 2, \ldots . \end{aligned}$$

Here, ε_t may be interpreted as a random shock or an innovation that derives the movement of the process $\{Z_t\}$ over time.

More generally, $\{Z_t\}$ is an AR(p) process if

$$\begin{aligned} Z_t &= \alpha_0 + \sum_{j=1}^{p} \alpha_j Z_{t-j} + \varepsilon_t, \\ \varepsilon_t &\sim \text{White noise } (0, \sigma^2). \end{aligned}$$

Example 6: $\{Z_t\}$ is a q-th order moving-average process (MA(q)) if

$$\begin{aligned} Z_t &= \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j} + \varepsilon_t, \\ \{\varepsilon_t\} &\sim \text{White noise } (0, \sigma^2). \end{aligned}$$

This is a weakly stationary process. For an MA(q) process, we have $\gamma(j) = 0$ for all $|j| > q$.

Example 7: $\{Z_t\}$ is an autoregressive-moving average (ARMA) process of orders (p, q) if

$$\begin{aligned} Z_t &= \alpha_0 + \sum_{j=1}^{p} \alpha_j Z_{t-j} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t, \\ \{\varepsilon_t\} &\sim \text{white noise } (0, \sigma^2). \end{aligned}$$

ARMA models include AR models and MA models as special cases. An estimation method for ARMA models can be found in Chapter 9. In practice, the orders of (p, q) can be selected according to the AIC or BIC criterion.

Under rather mild regularity conditions, a zero-mean weakly stationary process can be represented by an MA(∞) process

$$\begin{aligned} Z_t &= \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \\ \varepsilon_t &\sim \text{WN}(0, \sigma^2), \end{aligned}$$

where $\Sigma_{j=1}^{\infty}\alpha_j^2 < \infty$. This is called Wold's decomposition. The partial derivative

$$\frac{\partial Z_{t+j}}{\partial \varepsilon_t} = \alpha_j, j = 0, 1, ...$$

is called the impulse response function of the time series process $\{Z_t\}$ with respect to a random shock ε_t. This function characterizes the impact of a random shock ε_t on the immediate and subsequent observations $\{Z_{t+j}, j \geq 0\}$. For a weakly stationary process, the impact of any shock on a future Z_{t+j} will always diminish to zero as the lag order $j \to \infty$, because $\alpha_j \to 0$. The ultimate cumulative impact of ε_t on the process $\{Z_t\}$ is the sum $\Sigma_{j=0}^{\infty}\alpha_j$.

The function $\gamma(j) = \text{cov}(Z_t, Z_{t-j})$ is called the autocovariance function of the weakly stationary process $\{Z_t\}$, where j is a lag order. It characterizes the (linear) serial dependence of Z_t on its own lagged variable Z_{t-j}. Note that $\gamma(j) = \gamma(-j)$ for all integers j.

The normalized function $\rho(j) = \gamma(j)/\gamma(0)$ is called the autocorrelation function of $\{Z_t\}$. It has the property that $|\rho(j)| \leq 1$. The plot of $\rho(j)$ as a function of j is called the autocorrelogram of the time series process $\{Z_t\}$. It can be used to judge which linear time series model (e.g., AR, MA, or ARMA) should be used to fit a particular time series data set.

We now consider the Fourier transform of the autocovariance function $\gamma(j)$.

Definition 5.6 [Spectral Density Function] *The Fourier transform of* $\gamma(j)$

$$h(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma(j)e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

where $i = \sqrt{-1}$, *is called the power spectral density of process* $\{Z_t\}$.

The normalized version

$$f(\omega) = \frac{h(\omega)}{\gamma(0)} = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\rho(j)e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

is called the standardized spectral density of $\{Z_t\}$.

Question: What are the properties of $f(\omega)$?

It can be shown that (i) $f(\omega)$ is real-valued, and $f(\omega) \geq 0$; (ii) $\int_{-\pi}^{\pi} f(\omega)d\omega = 1$; (iii) $f(-\omega) = f(\omega)$.

The spectral density $h(\omega)$ is widely used in economic analysis. For example, it can be used to search for business cycles. Specifically, a frequency ω_0 corresponding to a special peak is closely associated with a business cycle with periodicity $T_0 = 2\pi/\omega_0$. Intuitively, time series can be decomposed as the sum of many cyclical components with different frequencies ω, and $h(\omega)$ is the strength or magnitude of the component with frequency ω. When $h(\omega)$ has a peak at ω_0, it means that the cyclical component with frequency ω_0 or periodicity $T_0 = 2\pi/\omega_0$ dominates all other frequencies. Consequently, the whole time series behaves as mainly having a cycle with periodicity T_0.

The functions $h(\omega)$ and $\gamma(j)$ are Fourier transforms of each other. Thus, they contain the same information on serial dependence in $\{Z_t\}$. In time series analysis, the use of $\gamma(j)$ is called the time domain analysis, and the use of $h(\omega)$ is called the frequency domain analysis. Which tool to use depends on the convenience of the user. In some applications, the use of $\gamma(j)$ is simpler and more intuitive, while in other applications, the use of $h(\omega)$ is more enlightening. This is exactly the same as the case that it is more convenient to use Chinese in China, while it is more convenient to use English in U.S.

Example 8: Hamilton, James (1994, *Time Series Analysis*): Business cycles of U.S. industrial production

Example 9: Steven Durlauf (1990, *Journal of Monetary Economics*): Income tax rate changes

Reference: Sargent, T. (1987): *Macroeconomic Theory*, 2nd Edition. Academic Press: Orlando, U.S.A.

For a serially uncorrelated sequence, the spectral density $h(\omega)$ is flat as a function of frequency ω :

$$\begin{aligned} h(\omega) &= \frac{1}{2\pi}\gamma(0) \\ &= \frac{1}{2\pi}\sigma^2 \text{ for all } \omega \in [-\pi, \pi]. \end{aligned}$$

This is analogous to the power (or energy) spectral density of a physical white color light. It is for this reason that we call a serially uncorrelated time series a white noise process.

Intuitively, a white color light can be decomposed via a lens as the sum of equal magnitude components of different frequencies. That is, a white color light has a flat physical spectral density function.

It is important to point out that a white noise may not be i.i.d., as is illustrated by the example below:

Example 10: Consider an autoregressive conditional heteroskedastic (ARCH) process

$$\begin{aligned} Z_t &= \varepsilon_t h_t^{1/2}, \\ h_t &= \alpha_0 + \alpha_1 Z_{t-1}^2, \\ \varepsilon_t &\sim \text{ i.i.d.(0,1).} \end{aligned}$$

This is first proposed by Engle (1982) and it has been widely used to model volatility in economics and finance. We have $E(Z_t|I_{t-1}) = 0$ and $\text{var}(Z_t|I_{t-1}) = h_t$, where $I_{t-1} = \{Z_{t-1}, Z_{t-2}, ...\}$ is the information set containing all past history of Z_t.

It can be shown that

$$\begin{aligned} E(Z_t) &= 0, \\ \text{cov}(Z_t, Z_{t-j}) &= 0 \text{ for } j > 0, \\ \text{var}(Z_t) &= \frac{\alpha_0}{1-\alpha_1}. \end{aligned}$$

When $\alpha_1 < 1$, $\{Z_t\}$ is a stationary white noise. But it is not weakly stationary if $\alpha_1 = 1$, because $\text{var}(Z_t) = \infty$. In both cases, $\{Z_t\}$ is strictly stationary (e.g., Nelson 1990, *Journal of Econometrics*).

Although $\{Z_t\}$ is a white noise, it is not an i.i.d. sequence because the correlation in $\{Z_t^2\}$ is $\text{corr}(Z_t^2, Z_{t-j}^2) = \alpha_1^{|j|}$ for $j = 0, 1, 2, ...$ In other words, an ARCH process is uncorrelated in level but is autocorrelated in squares.

Nonstationarity

Usually, we call $\{Z_t\}$ a nonstationary time series when it is not covariance-stationary. In time series econometrics, there have been two types of nonstationary processes that display similar sample paths when the sample size is not large but have quite different implications. We first discuss a nonstationary process called trend-stationary process.

Example 11: $\{Z_t\}$ is called a trend-stationary process if

$$Z_t = \alpha_0 + \alpha_1 t + \varepsilon_t,$$

where ε_t is a weakly stationary process with mean 0 and variance σ^2. To see why $\{Z_t\}$ is not weakly stationary, we consider a simplest case where $\{\varepsilon_t\}$ is i.i.d. $(0, \sigma^2)$. Then

$$\begin{aligned} E(Z_t) &= \alpha_0 + \alpha_1 t, \\ var(Z_t) &= \sigma^2, \\ \text{cov}(Z_t, Z_{t-j}) &= 0. \end{aligned}$$

Question: What happens if $\Delta Z_t = Z_t - Z_{t-1}$?

More generally, a trend-stationary time series process can be defined as follows:

$$Z_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j t^j + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a weakly stationary process. The reason that $\{Z_t\}$ is called trend-stationary is that it will become weakly stationary after the deterministic trend is removed.

Next, we discuss the second type of nonstationary process called difference-stationary process. Again, we start with a special case:

Example 12: $\{Z_t\}$ is a random walk with a drift if

$$Z_t = \alpha_0 + Z_{t-1} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is i.i.d. $(0, \sigma^2)$. For simplicity, we assume $Z_0 = 0$. Then

$$\begin{aligned} E(Z_t) &= \alpha_0 t, \\ \text{var}(Z_t) &= \sigma^2 t, \\ \text{cov}(Z_t, Z_{t-j}) &= \sigma^2 (t-j). \end{aligned}$$

Note that for any given j,

$$\text{corr}(Z_t, Z_{t-j}) = \sqrt{\frac{t-j}{t}} \to 1 \text{ as } t \to \infty,$$

which implies that the impact of an infinite past event on today's behavior never dies out. Indeed, this can be seen more clearly if we write

$$Z_t = Z_0 + \alpha_0 t + \sum_{j=0}^{t-1} \varepsilon_{t-j}.$$

Note that $\{Z_t\}$ has a deterministic linear time trend but with an increasing variance over time. The impulse response function $\partial Z_{t+j}/\partial \varepsilon_t = 1$ for all $j \geq 0$, which never dies off to zero as $j \to \infty$.

There is another nonstationary process called martingale process which is closely related to a random walk.

Definition 5.7 [Martingale] *A time series process $\{Z_t\}$ is a martingale with drift if*

$$Z_t = \alpha + Z_{t-1} + \varepsilon_t,$$

and $\{\varepsilon_t\}$ *satisfies*

$$E(\varepsilon_t|I_{t-1}) = 0 \text{ a.s.},$$

where I_{t-1} *is the* σ*-field generated by* $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots\}$*. We call that* $\{\varepsilon_t\}$ *is a martingale difference sequence (MDS).*

Question: Why is ε_t called an MDS?

Because ε_t is the difference of a martingale process. That is, $\varepsilon_t = Z_t - Z_{t-1}$.

Example 13 [Martingale and Efficient Market Hypothesis]: Suppose a stock log-price $\ln P_t$ follows a martingale process, i.e.,

$$\ln P_t = \ln P_{t-1} + \varepsilon_t,$$

where $E(\varepsilon_t|I_{t-1}) = 0$. Then $\varepsilon_t = \ln P_t - \ln P_{t-1} \approx \frac{P_t - P_{t-1}}{P_{t-1}}$ is the stock relative price change or stock return (if no dividend) from time $t-1$ to time t, which can be viewed as the proxy for the new information arrival from time $t-1$ to time t that derives the stock price change in the same period. For this reason, ε_t is also called an innovation sequence. The MDS property of ε_t implies that the price change ε_t is unpredictable using the past information available at time $t-1$, and the market is called informationally efficient. Thus, the best predictor for the stock price at time t using the information available at time $t-1$ is P_{t-1}, that is, $E(P_t|I_{t-1}) = P_{t-1}$.

Question: What is the relationship between a random walk and a martingale?

A random walk is a martingale because IID with zero mean implies $E(\varepsilon_t|I_{t-1}) = E(\varepsilon_t) = 0$. However, the converse is not true.

Example 14: Reconsider an ARCH(1) process

$$\begin{aligned} \varepsilon_t &= h_t^{1/2} z_t, \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2, \\ \{z_t\} &\sim \text{i.i.d.(0,1)}. \end{aligned}$$

where $\alpha_0, \alpha_1 > 0$. It follows that

$$\begin{aligned} E(\varepsilon_t|I_{t-1}) &= 0, \\ \text{var}(\varepsilon_t|I_{t-1}) &= h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2, \end{aligned}$$

where I_{t-1} denotes the information available at time $t-1$. Clearly $\{\varepsilon_t\}$ is MDS but not IID, because its conditional variance h_t is time-varying (depending on the past information set I_{t-1}).

Since the only condition for MDS is $E(\varepsilon_t|I_{t-1}) = 0$ a.s., an MDS need not be strictly stationary or weakly stationary. However, if it is assumed that $\text{var}(\varepsilon_t) = \sigma^2$ exists, then an MDS is weakly stationary.

When the variance $E(\varepsilon_t^2)$ exists, we have the following directional relationships:

$$\text{IID (with } \mu = 0) \Longrightarrow \text{ MDS} \Longrightarrow \text{ WHITE NOISE.}$$

Lemma 5.1: *If $\{\varepsilon_t\}$ is an MDS with $E(\varepsilon_t^2) = \sigma^2 < \infty$, then $\{\varepsilon_t\}$ is a white noise.*

Proof: By the law of iterated expectations, we have

$$E(\varepsilon_t) = E[E(\varepsilon_t|I_{t-1})] = 0,$$

and for any $j > 0$,

$$\begin{aligned} \text{cov}(\varepsilon_t, \varepsilon_{t-j}) &= E(\varepsilon_t\varepsilon_{t-j}) - E(\varepsilon_t)E(\varepsilon_{t-j}) \\ &= E[E(\varepsilon_t\varepsilon_{t-j}|I_{t-1})] \\ &= E[E(\varepsilon_t|I_{t-1})\varepsilon_{t-j}] \\ &= E(0 \cdot \varepsilon_{t-j}) \\ &= 0. \end{aligned}$$

This implies that MDS, together with $\text{var}(\varepsilon_t) = \sigma^2$, is a white noise.

However, a white noise does not imply a MDS.

Example 15: A nonlinear MA process

$$\begin{aligned} \varepsilon_t &= \alpha z_{t-1}z_{t-2} + z_t, \\ \{z_t\} &\sim i.i.d.(0,1). \end{aligned}$$

Then it can be shown that $\{\varepsilon_t\}$ is a white noise but not MDS, because $\text{cov}(\varepsilon_t, \varepsilon_{t-j}) = 0$ for all $j > 0$ but

$$E(\varepsilon_t|I_{t-1}) = \alpha z_{t-1}z_{t-2} \neq 0.$$

Question: When will the concepts of IID, MDS and White noise coincide?

When $\{\varepsilon_t\}$ is a stationary Gaussian process. A time series is a stationary Gaussian process if $\{\varepsilon_{t_1}, \varepsilon_{t_2}, ..., \varepsilon_{t_m}\}$ is multivariate normally distributed for any admissible sets of integers $\{t_1, t_2, ..., t_m\}$. Unfortunately, an important stylized fact for economic and financial time series is that they are typically non-Gaussian. Therefore, it is important to emphasize the difference among the concepts of IID, MDS and White Noise in time series econometrics.

When var(ε_t) exists, both random walk and martingale processes are special cases of the so-called unit root process which is defined below.

Definition 5.8 [Unit root or difference stationary process]: $\{Z_t\}$ *is a unit root process if*

$$\begin{aligned} Z_t &= \alpha_0 + Z_{t-1} + \varepsilon_t, \\ & \quad \{\varepsilon_t\} \text{ is covariance-stationary } (0, \sigma^2). \end{aligned}$$

The process $\{Z_t\}$ is called a unit root process because its autoregressive coefficient is unity. It is also called a difference-stationary process because its first difference,

$$\Delta Z_t = Z_t - Z_{t-1} = \alpha_0 + \varepsilon_t,$$

becomes weakly stationary. In fact, the first difference of a linear trend-stationary process $Z_t = \alpha_0 + \alpha_1 t + \varepsilon_t$ is also weakly stationary:

$$\Delta Z_t = \alpha_1 + \varepsilon_t - \varepsilon_{t-1}.$$

The inverse of differencing is "integrating". For the difference-stationary process $\{Z_t\}$, we can write it as the integral of the weakly stationary process $\{\varepsilon_t\}$ in the sense that

$$Z_t = \alpha_0 t + Z_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j},$$

where Z_0 is the starting value of the process $\{Z_t\}$. This is analogous to differentiation and integration in calculus which are inverses of each other. For this reason, $\{Z_t\}$ is also called an Integrated process of order 1, denoted as $I(1)$.Obviously, a random walk and a martingale process are $I(1)$ processes if the variance of the innovation ε_t is finite.

We will assume strict stationarity in most cases in the present and subsequent chapters. This implies that some economic variables have to be transformed before used in $Y_t = X_t'\beta^o + \varepsilon_t$. Otherwise, the asymptotic theory developed here cannot be applied. Indeed, a different asymptotic theory should be developed for unit root processes (see, e.g., Hamilton (1994), *Time Series Analysis*).

In macroeconomics, it is important to check whether a nonstationary macroeconomic time series is trend-stationary or difference-stationary. If it is a unit root process, then a shock to the economy will never die out to zero as time evolves. In contrast, a random shock to a trend-stationary process will die out to zero eventually.

Question: Why has the unit root econometrics been so popular in econometrics?

It was found in empirical studies (e.g., Nelson and Plosser (1982, *Journal of Monetary Economics*)) that most macroeconomic time series display unit root properties.

Ergodicity

Next, we introduce a concept of asymptotic independence.

Question: Consider the following time series

$$\begin{aligned} Z^n &= (Z_1, Z_2, \cdots, Z_n)' \\ &= (W, W, \cdots, W)', \end{aligned}$$

where W is a random variable that does not depend on time index t. Obviously, the stationarity condition holds. However, any realization of this random sample Z^n will be

$$z^n = (w, w, \cdots, w)',$$

i.e., it will contain the same realization w for all n observations (no new information as n increases). In order to avoid this, we need to impose a condition called ergodicity that assumes that $(Z_t, \cdots, Z_{t+k})$ and $(Z_{m+t}, \cdots, Z_{m+t+l})$ are asymptotically independent when their time distance $m \to \infty$.

Statistically speaking, independence or little correlation generates new or more information as the sample size n increases. Recall that X and Y are independent if and only if

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$$

for any measurable functions $f(\cdot)$ and $g(\cdot)$. We now extend this definition to define ergodicity.

Definition 5.9 [Ergodicity]*: A strictly stationary process $\{Z_t\}$ is said to be ergodic if for any two bounded functions $f : R^{k+1} \to R$ and $g : \mathbb{R}^{l+1} \to \mathbb{R}$,*

$$\begin{aligned} &\lim_{m\to\infty} \left| E\left[f(Z_t, \cdots, Z_{t+k}) g(Z_{m+t}, \cdots, Z_{m+t+l}) \right] \right| \\ = \;& \left| E\left[f(Z_t, \cdots, Z_{t+k}) \right] \right| \cdot \left| E\left[g(Z_{m+t}, \cdots, Z_{m+t+l}) \right] \right|. \end{aligned}$$

Remarks:

Clearly, ergodicity is a concept of asymptotic independence. A strictly stationary process that is ergodic is called ergodic stationary. If $\{Z_t\}$ is ergodic stationary, then $\{f(Z_t)\}$ is also ergodic stationary for any measurable function $f(\cdot)$.

Theorem 5.2 [WLLN for Ergodic Stationary Random Samples]: *Let* $\{Z_t\}$ *be an ergodic stationary process with* $E(Z_t) = \mu$ *and* $E|Z_t| < \infty$. *Then the sample mean*

$$\bar{Z}_n = n^{-1} \sum_{t=1}^{n} Z_t \xrightarrow{p} \mu \text{ as } n \to \infty.$$

Question: Why do we need to assume ergodicity?

Consider a counter example which does not satisfy the ergodicity condition: $Z_t = W$ for all t. Then $\bar{Z}_n = W$, a random variable which will not converge to μ as $n \to \infty$.

Next, we state a CLT for ergodic stationary MDS random samples.

Theorem 5.3 [Central Limit Theorem for Ergodic Stationary MDS]: *Suppose* $\{Z_t\}$ *is a stationary ergodic MDS process, with* $var(Z_t) \equiv E(Z_t Z_t') = V$ *finite, symmetric and positive definite. Then as* $n \to \infty$,

$$\sqrt{n}\bar{Z}_n = n^{-1/2} \sum_{t=1}^{n} Z_t \xrightarrow{d} N(0, V)$$

or equivalently,

$$V^{-1/2}\sqrt{n}\bar{Z}_n \xrightarrow{d} N(0, I).$$

Question: Is $\text{avar}(\sqrt{n}\bar{Z}_n) = V = \text{var}(Z_t)$? That is, is the asymptotic variance of $\sqrt{n}\bar{Z}_n$ coincides with the individual variance $\text{var}(Z_t)$.
To check this, we have

$$\begin{aligned}
\text{var}(\sqrt{n}\bar{Z}_n) &= E[\sqrt{n}\bar{Z}_n \sqrt{n}\bar{Z}_n'] \\
&= E\left[\left(n^{-1/2}\sum_{t=1}^{n} Z_t\right)\left(n^{-1/2}\sum_{s=1}^{n} Z_s\right)'\right] \\
&= n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} E(Z_t Z_s') \\
[E(Z_t Z_s) &= 0 \text{ for } t \neq s, \text{ by the LIE}] \\
&= n^{-1}\sum_{t=1}^{n} E(Z_t Z_t') \\
&= E(Z_t Z_t') \\
&= V.
\end{aligned}$$

Here, the MDS property plays a crucial rule in simplifying the asymptotic variance of $\sqrt{n}\bar{Z}_n$ because it implies $\text{cov}(Z_t, Z_s) = 0$ for all $t \neq s$.MDS is one of the most important concepts in modern economics, particularly in macroeconomics, finance, and econometrics. For example, rational expectations theory can be characterized by an expectational error being an MDS.

5.2 Framework and Assumptions

With the basic time series concepts and analytic tools introduced above, we can now develop an asymptotic theory for linear regression models with time series observations. We first state the assumptions that allow for time series observations.

Assumption 5.1 [Ergodic stationarity]: The stochastic process $\{Y_t, X_t'\}_{t=1}'^n$ is jointly stationary and ergodic.

Assumption 5.2 [Linearity]:

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

where β^o is a $K \times 1$ unknown parameter vector, and ε_t is the unobservable disturbance.

Assumption 5.3 [Correct Model Specification]: $E(\varepsilon_t|X_t) = 0$ a.s. with $E(\varepsilon_t^2) = \sigma^2 < \infty$.

Assumption 5.4 [Nonsingularity]: The $K \times K$ matrix

$$Q = E(X_t X_t')$$

is finite and nonsingular.

Assumption 5.5 [MDS]: $\{X_t\varepsilon_t\}$ is an MDS process with respect to the σ-field generated by $\{X_s\varepsilon_s, s < t\}$ and the $K \times K$ matrix $V \equiv \text{var}(X_t\varepsilon_t) = E(X_t X_t'\varepsilon_t^2)$ is finite and positive definite.

Remarks:

In Assumption 5.1, the ergodic stationary process $Z_t = \{Y_t, X_t'\}_{t=1}'^n$ can be independent or serially dependent across different time periods. we thus allow for time series observations from a stationary stochastic process.

It is important to emphasize that the asymptotic theory to be developed below and in subsequent chapters is not applicable to nonstationary time series. A problem associated with nonstationary time series is the so-called spurious regression or spurious correlation problem. If the dependent variable Y_t and the regressors X_t display similar

trending behaviors over time, one is likely to obtain seemly highly "significant" regression coefficients and high values for R^2, even if they do not have any causal relationship. Such results are completely spurious. In fact, the OLS estimator for nonstationary time series regression model does not follow the asymptotic theory to be developed below. A different asymptotic theory for nonstationary time series regression models has to be used (see, e.g., Hamiltion 1994). Using the correct asymptotic theory, the seemingly highly "significant" regression coefficient estimators would become insignificant in the spurious regression models.

Unlike the i.i.d. case, where $E(\varepsilon_t|X_t) = 0$ is equivalent to the strict exogeneity condition that

$$E(\varepsilon_t|X) = E(\varepsilon_t|X_1, ..., X_t, ..., X_n) = 0,$$

the condition $E(\varepsilon_t|X_t) = 0$ is weaker than $E(\varepsilon_t|X) = 0$ in a time series context. In other words, it is possible that $E(\varepsilon_t|X_t) = 0$ but $E(\varepsilon_t|X) \neq 0$. Assumption 5.3 allows for the inclusion of predetermined variables in X_t, the lagged dependent variables Y_{t-1}, Y_{t-2}, etc.

For example, suppose $X_t = (1, Y_{t-1})'$. Then we obtain an AR(1) model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t, \qquad t = 2, \cdots, n. \\ \{\varepsilon_t\} &\sim \text{MDS}(0, \sigma^2). \end{aligned}$$

Then $E(\varepsilon_t|X_t) = 0$ holds if $E(\varepsilon_t|I_{t-1}) = 0$, namely if $\{\varepsilon_t\}$ is an MDS, where I_{t-1} is the sigma-field generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$. However, we generally have $E(\varepsilon_t|X) \neq 0$ because $E(\varepsilon_t X_{t+1}) \neq 0$.

When X_t contains an intercept the MDS condition for $\{X_t\varepsilon_t\}$ in Assumption 5.5 implies that $E(\varepsilon_t|I_{t-1}) = 0$; that is, $\{\varepsilon_t\}$ is an MDS, where $I_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$.

Question: When can an MDS disturbance ε_t arise in economics and finance?

Example 1: Rational Expectations Economics

Recall the dynamic asset pricing model under a rational expectations framework in Chapter 1. The behavior of the economic agent is characterized by the Euler equation:

$$\begin{aligned} E\left[\beta \frac{u'(C_t)}{u'(C_{t-1})} R_t \middle| I_{t-1}\right] &= 1 \text{ or} \\ E[M_t R_t | I_{t-1}] &= 1, \end{aligned}$$

where β is the time discount factor of the representative economic agent, C_t is the consumption, R_t is the asset gross return, and M_t is the stochastic discount factor defined as follows:

$$\begin{aligned} M_t &= \beta \frac{u'(C_t)}{u'(C_{t-1})} \\ &= \beta + \frac{u''(C_{t-1})}{u'(C_{t-1})} \Delta C_t + \text{ higher order} \\ &\sim \text{ risk adjustment factor.} \end{aligned}$$

Using the formula that $\text{cov}(X_t, Y_t|I_{t-1}) = E(X_t Y_t|I_{t-1}) - E(X_t|I_{t-1})E(Y_t|I_{t-1})$ and re-arranging, we can write the Euler equation as

$$E(M_t|I_{t-1})E(R_t|I_{t-1}) + \text{cov}(M_t, R_t|I_{t-1}) = 1.$$

It follows that

$$\begin{aligned} E(R_t|I_{t-1}) &= \frac{1}{E(M_t|I_{t-1})} + \frac{\text{cov}(M_t, R_t|I_{t-1})}{\text{var}(M_t|I_{t-1})} \cdot \frac{-\text{var}(M_t|I_{t-1})}{E(M_t|I_{t-1})} \\ &= \alpha_t + \beta_t \lambda_t, \end{aligned}$$

where $\alpha_t = \alpha(I_{t-1})$ is the riskfree interest rate, $\lambda_t = \lambda(I_{t-1})$ is the market risk, and $\beta_t = \beta(I_{t-1})$ is the price of market risk, or the so-called investment beta factor.

Equivalently, we can write a regression equation for the asset return

$$\begin{aligned} R_t &= \alpha_t + \beta_t \lambda_t + \varepsilon_t, \text{ where} \\ E(\varepsilon_t|I_{t-1}) &= 0. \end{aligned}$$

A conventional CAPM usually assumes $\alpha_{t-1} = \alpha, \beta_t = \beta$ and use some proxies for λ_t.

Like in Chapter 4, no normality on $\{\varepsilon_t\}$ is imposed. Furthermore, no conditional homoskedasticity is imposed. We now allow that $\text{var}(\varepsilon_t|X_t)$ is a function of X_t. Because X_t may contain lagged $Y_{t-1}, Y_{t-2}, \cdots$, $\text{var}(\varepsilon_t|X_t)$ may change over time (e.g., volatility clustering). Volatility clustering is a well-known financial phenomenon where a large volatility today tends to be followed by another large volatility tomorrow, and a small volatility today tends to be followed by another small volatility tomorrow.

Although Assumptions 5.1–5.5 allow for temporal dependences between observations, we will still obtain the same asymptotic properties for the OLS estimator and related test procedures as in the i.i.d. case. Put it differently, all the large sample properties for the OLS and related tests established under the i.i.d. assumption in Chapter 4 remain applicable to time series observations with the stationary MDS assumptions for $\{X_t \varepsilon_t\}$. We now show that this is indeed the case in subsequent sections.

5.3 Consistency of OLS

We first investigate the consistency of OLS $\hat{\beta}$. Recall the OLS estimator

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= \hat{Q}^{-1}n^{-1}\sum_{t=1}^{n} X_t Y_t,\end{aligned}$$

where, as before,

$$\hat{Q} = n^{-1}\sum_{t=1}^{n} X_t X_t'.$$

Substituting $Y_t = X_t'\beta^o + \varepsilon_t$ from Assumption 5.2, we have

$$\hat{\beta} - \beta^o = \hat{Q}^{-1}n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t.$$

Theorem 5.4: *Suppose Assumptions 5.1–5.5 hold. Then*

$$\hat{\beta} - \beta^o \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof: Because $\{X_t\}$ is ergodic stationary, $\{X_t X_t'\}$ is also ergodic stationary. Thus, given Assumption 5.4, which implies $E|X_{it}X_{jt}| \leq C < \infty$ for $0 \leq i, j \leq k$ and for some constant C, we have

$$\hat{Q} \xrightarrow{p} E(X_t X_t') = Q$$

by the WLLN for ergodic stationary processes. Because Q^{-1} exists, by continuity we have

$$\hat{Q}^{-1} \xrightarrow{p} Q^{-1} \text{ as } n \to \infty.$$

Next, we consider $n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t$. Because $\{Y_t, X_t'\}_{t=1}^{n}{}'$ is ergodic stationary, $\varepsilon_t = Y_t - X_t'\beta^o$ is ergodic stationary, and so is $X_t\varepsilon_t$. In addition,

$$E|X_{jt}\varepsilon_t| \leq \left[E(X_{jt}^2)E(\varepsilon_t^2)\right]^{1/2} \leq C < \infty \text{ for } 0 \leq j \leq k$$

by the Cauchy-Schwarz inequality and Assumptions 5.3 and 5.4. It follows that

$$n^{-1}\sum_{t=1}^{n} X_t\varepsilon_t \xrightarrow{p} E(X_t\varepsilon_t) = 0$$

by the WLLN for ergodic stationary processes, where

$$\begin{aligned}E(X_t\varepsilon_t) &= E[E(X_t\varepsilon_t|X_t)] \\ &= E[X_t E(\varepsilon_t|X_t)] \\ &= E(X_t \cdot 0) \\ &= 0\end{aligned}$$

by the law of iterated expectations and Assumption 5.3. Therefore, we have

$$\hat{\beta} - \beta^o = \hat{Q}^{-1} n^{-1} \sum_{t=1}^{n} X_t \varepsilon_t \xrightarrow{p} Q^{-1} \cdot 0 = 0.$$

This completes the proof.

5.4 Asymptotic Normality of OLS

Next, we derive the asymptotic distribution of $\hat{\beta}$.

Theorem 5.5: *Suppose Assumptions 5.1–5.5 hold. Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, Q^{-1} V Q^{-1}) \text{ as } n \to \infty.$$

Proof: Recall

$$\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}^{-1} n^{-\frac{1}{2}} \sum_{t=1}^{n} X_t \varepsilon_t.$$

First, we consider the second term

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} X_t \varepsilon_t.$$

Because $\{Y_t, X_t'\}_{t=1}^{\prime n}$ is stationary ergodic, $X_t \varepsilon_t$ is also stationary ergodic. Also, $\{X_t \varepsilon_t\}$ is a MDS with $\text{var}(X_t \varepsilon_t) = E(X_t X_t' \varepsilon_t^2) = V$ being finite and positive definite (Assumption 5.5). By the CLT of stationary ergodic MDS processes, we have

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} X_t \varepsilon_t \xrightarrow{d} N(0, V).$$

Moreover, $\hat{Q}^{-1} \xrightarrow{p} Q^{-1}$, as shown earlier. It follows from the Slutsky theorem that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta^o) &= \hat{Q}^{-1} n^{-\frac{1}{2}} \sum_{t=1}^{n} X_t \varepsilon_t \\ &\xrightarrow{d} Q^{-1} N(0, V) \\ &\sim N(0, Q^{-1} V Q^{-1}). \end{aligned}$$

This completes the proof.

The asymptotic distribution of $\hat{\beta}$ under Assumptions 5.1–5.5 is exactly the same as that of $\hat{\beta}$ in Chapter 4. In particular, the asymptotic mean of $\sqrt{n}(\hat{\beta} - \beta^o)$ is 0, and the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is $Q^{-1} V Q^{-1}$; we denote

$$\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1} V Q^{-1}.$$

Special Case: Conditional Homoskedasticity

The asymptotic variance of $\sqrt{n}\hat{\beta}$ can be simplified if there exists conditional homoskedasticity.

Assumption 5.6: $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.

This assumption rules out the possibility that the conditional variance of ε_t changes with X_t. For low-frequency macroeconomic time series, this might be a reasonable assumption. For high-frequency financial time series, however, this assumption will be rather restrictive.

Theorem 5.6: *Suppose Assumptions 5.1–5.6 hold. Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

Proof: Under Assumption 5.6, we can simplify

$$\begin{aligned} V &= E(X_t X_t' \varepsilon_t^2) \\ &= E[E(X_t X_t' \varepsilon_t^2 | X_t)] \\ &= E[X_t X_t' E(\varepsilon_t^2 | X_t)] \\ &= \sigma^2 E(X_t X_t') \\ &= \sigma^2 Q. \end{aligned}$$

The desired results follow immediately from the previous theorem. This completes the proof.

Under conditional homoskedasticity, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is

$$\begin{aligned} \text{avar}(\sqrt{n}\hat{\beta}) &= Q^{-1} V Q^{-1} \\ &= \sigma^2 Q^{-1}. \end{aligned}$$

This is rather convenient to estimate.

5.5 Asymptotic Variance Estimator for OLS

To construct confidence interval estimators or hypothesis test statistics, we need to estimate the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$, namely $\text{avar}(\sqrt{n}\hat{\beta})$. We consider consistent estimation for $\text{avar}(\sqrt{n}\hat{\beta})$ under conditional homoskedasticity and conditional heteroskedasticity respectively.

Case I: Conditional Homoskedasticity

Under this case, the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ is

$$\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1} = \sigma^2 Q^{-1}.$$

It suffices to have consistent estimators for σ^2 and Q respectively.

Question: How to estimate Q?

Lemma 5.7: *Suppose Assumptions 5.1 and 5.3 hold. Then*

$$\hat{Q} \xrightarrow{p} Q \text{ as } n \to \infty.$$

Question: How to estimate σ^2?

To estimate the residual sample variance estimator, we have

$$s^2 = \frac{e'e}{n-K}.$$

Theorem 5.8 [Consistent Estimator for σ^2]: *Under Assumptions 5.1-5.5, as $n \to \infty$,*

$$s^2 \xrightarrow{p} \sigma^2.$$

Proof: The proof is analogous to the proof of Theorem 4.15 in Chapter 4. We have

$$\begin{aligned} s^2 &= \frac{1}{n-K}\sum_{t=1}^{n} e_t^2 \\ &= (n-K)^{-1}\sum_{t=1}^{n}\varepsilon_t^2 \\ &\quad +(\hat{\beta}-\beta^o)'\left(\frac{1}{n-K}\sum_{t=1}^{n}X_tX_t'\right)(\hat{\beta}-\beta^o) \\ &\quad -2(\hat{\beta}-\beta^o)'\frac{1}{n-K}\sum_{t=1}^{n}X_t\varepsilon_t \\ &\xrightarrow{p} \sigma^2 + 0\cdot Q\cdot 0 - 2\cdot 0\cdot 0 = \sigma^2 \end{aligned}$$

given that K is a fixed number, where we have made use of the WLLN for ergodic stationary processes in several places. This completes the proof.

We can then estimate $\text{avar}(\sqrt{n}\hat{\beta}) = \sigma^2 Q^{-1}$ by $s^2\hat{Q}^{-1}$.

Theorem 5.9: [Asymptotic Variance Estimator of $\hat{\beta}$]: *Under Assumptions 5.1-5.4, we can consistently estimate the asymptotic variance avar$(\sqrt{n}\hat{\beta})$ by*

$$s^2\hat{Q}^{-1} \xrightarrow{p} \sigma^2 Q^{-1}.$$

This implies that the variance estimator of $\hat{\beta}$ is calculated as

$$s^2\hat{Q}^{-1}/n = s^2(\mathbf{X}'\mathbf{X})^{-1},$$

which is the same as in the classical linear regression case.

Case II: Conditional Heteroskedasticity

In this case,

$$\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1}$$

cannot be further simplified.

Question: How to estimate $Q^{-1}VQ^{-1}$?

Question: It is straightforward to estimate Q by $\hat{Q}$. How to estimate $V = E(X_tX_t'\varepsilon_t^2)$?

We can use its sample analog

$$\hat{V} = n^{-1}\sum_{t=1}^{n} X_tX_t'e_t^2.$$

To ensure consistency of $\hat{V}$ for V, we impose the following moment condition:

Assumption 5.7: $E(X_{jt}^4) < \infty$ for $0 \leq j \leq k$ and $E(\varepsilon_t^4) < \infty$.
Lemma 5.10: *Suppose Assumptions 5.1–5.5 and 5.7 hold. Then*

$$\hat{V} \xrightarrow{p} V \text{ as } n \to \infty.$$

Proof: The proof is analogous to the proof of Lemma 4.17 in Chapter 4. Because $e_t = \varepsilon_t - (\hat{\beta} - \beta^o)'X_t$, we have

$$\begin{aligned}
\hat{V} &= n^{-1}\sum_{t=1}^{n} X_tX_t'\varepsilon_t^2 \\
&\quad + n^{-1}\sum_{t=1}^{n} X_tX_t'[(\hat{\beta} - \beta^o)'X_tX_t'(\hat{\beta} - \beta^o)] \\
&\quad - 2n^{-1}\sum_{t=1}^{n} X_tX_t'[\varepsilon_tX_t'(\hat{\beta} - \beta^o)] \\
&\xrightarrow{p} V + 0 - 2 \cdot 0,
\end{aligned}$$

where for the first term, we have

$$n^{-1}\sum_{t=1}^{n} X_t X_t' \varepsilon_t^2 \xrightarrow{p} E(X_t X_t' \varepsilon_t^2) = V$$

by the WLLN for ergodic stationary processes and Assumption 5.5. For the second term, it suffices to show that for any combination (i, j, l, m), where $0 \leq i, j, l, m \leq k$,

$$\begin{aligned} & n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}[(\hat{\beta}-\beta^o)'X_t X_t'(\hat{\beta}-\beta^o)] \\ = & \sum_{l=0}^{k}\sum_{m=0}^{k}(\hat{\beta}_l-\beta_l^o)(\hat{\beta}_m-\beta_m^o)\left(n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}X_{mt}\right) \\ \xrightarrow{p} & \ 0, \end{aligned}$$

which follows from $\hat{\beta}-\beta^o \xrightarrow{p} 0$ and $n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}X_{mt} \xrightarrow{p} E(X_{it}X_{jt}X_{lt}X_{mt}) = O(1)$ by the WLLN and Assumption 5.7.

For the last term, it suffices to show

$$\begin{aligned} & n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}[\varepsilon_t X_t'(\hat{\beta}-\beta^o)] \\ = & \sum_{l=0}^{k}(\hat{\beta}_l-\beta_l^o)\left(n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}\varepsilon_t\right) \\ \xrightarrow{p} & \ 0, \end{aligned}$$

which follows from $\hat{\beta}-\beta^o \xrightarrow{p} 0$, $n^{-1}\sum_{t=1}^{n} X_{it}X_{jt}X_{lt}\varepsilon_t \xrightarrow{p} E(X_{it}X_{jt}X_{lt}\varepsilon_t) = 0$ by the WLLN for ergodic stationary processes, the law of iterated expectations, and $E(\varepsilon_t|X_t) = 0$ a.s.

We have proved the following result.

Theorem 5.11 [Asymptotic variance estimator for $\sqrt{n}(\hat{\beta}-\beta^o)$]: *Under Assumptions 5.1–5.5 and 5.7, we can estimate avar$(\sqrt{n}\hat{\beta})$ by*

$$\hat{Q}^{-1}\hat{V}\hat{Q}^{-1} \xrightarrow{p} Q^{-1}VQ^{-1}.$$

The variance estimator $\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}$ is the so-called White's heteroskedasticity-consistent variance-covariance matrix of estimator $\sqrt{n}(\hat{\beta}-\beta^o)$ in a linear time series regression model with MDS disturbances.

5.6 Hypothesis Testing

Question: How to construct a test for the null hypothesis

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

where R is a $J \times K$ constant matrix, and r is a $J \times 1$ constant vector?

Because

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

we have under $\mathbf{H}_0$,

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

When $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s., we have $V = \sigma^2 Q$, and so

$$R\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2 RQ^{-1}R').$$

The test statistics differ in two cases. We first construct a test under conditional homoskedasticity.

Case I: Conditional Homoskedasticity

When $J = 1$, we can use the conventional t-test statistic for large sample inference.

Theorem 5.12 [t-test]: *Suppose Assumptions 5.1-5.6 hold. Then under* $\mathbf{H}_0$ with $J = 1$,

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0, 1)$$

as $n \to \infty$.

Proof: Given $R\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2 RQ^{-1}R')$, $R\beta^o = r$ under $\mathbf{H}_0$, and $J = 1$, we have

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{\sigma^2 RQ^{-1}R'}} \xrightarrow{d} N(0, 1).$$

By the Slutsky theorem and $\hat{Q} = \mathbf{X}'\mathbf{X}/n$, we obtain

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{s^2 R\hat{Q}^{-1}R'}} \xrightarrow{d} N(0, 1).$$

This ratio is the conventional t-test statistic we examined in Chapter 3, namely:

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{s^2 R\hat{Q}^{-1}R'}} = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} = T.$$

For $J > 1$, we can consider an asymptotic χ^2 test that is based on the conventional F-statistic.

Theorem 5.13 [Asymptotic χ^2 Test]: *Suppose Assumptions 5.1-5.6 hold. Then under* $\mathbf{H}_0$,

$$J \cdot F \xrightarrow{d} \chi^2_J$$

as $n \to \infty$.

Proof: We write

$$R\hat{\beta} - r = R(\hat{\beta} - \beta^o) + R\beta^o - r.$$

Under $\mathbf{H}_0 : R\beta^o = r$, we have

$$\begin{aligned} \sqrt{n}(R\hat{\beta} - r) &= R\sqrt{n}(\hat{\beta} - \beta^o) \\ &\xrightarrow{d} N(0, \sigma^2 RQ^{-1}R'). \end{aligned}$$

It follows that the quadratic form

$$\sqrt{n}(R\hat{\beta} - r)'[\sigma^2 RQ^{-1}R']^{-1}\sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

Also, because $s^2\hat{Q}^{-1} \xrightarrow{p} \sigma^2 Q^{-1}$, we have the Wald test statistic

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

by the Slutsky theorem. This can be written equivalently as follows:

$$W = \frac{(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r)}{s^2} \xrightarrow{d} \chi^2_J,$$

namely

$$W = J \cdot F \xrightarrow{d} \chi^2_J,$$

where F is the conventional F-test statistic derived in Chapter 3.

Remarks:

We cannot use the F distribution for a finite sample size n, but we can still compute the F-statistic and the appropriate test statistic is J times the F-statistic, which is asymptotically χ^2_J as $n \to \infty$. That is,

$$J \cdot F = \frac{(\tilde{e}'\tilde{e} - e'e)}{e'e/(n-K)} \xrightarrow{d} \chi^2_J.$$

Put it differently, the classical F-test is still approximately applicable under Assumptions 5.1–5.6 for a large n.

We now give two examples that are not covered under the assumptions of classical linear regression models.

Example 1 [Testing for Granger Causality]: Consider a bivariate time series $\{Y_t, X_t\}$, where t is the time index, $I_{t-1}^{(Y)} = \{Y_{t-1}, ..., Y_1\}$ and $I_{t-1}^{(X)} = \{X_{t-1}, ..., X_1\}$. For example, Y_t is the GDP growth, and X_t is the money supply growth. We say that X_t does not Granger-cause Y_t in conditional mean with respect to $I_{t-1} = \{I_{t-1}^{(Y)}, I_{t-1}^{(X)}\}$ if

$$E(Y_t|I_{t-1}^{(Y)}, I_{t-1}^{(X)}) = E(Y_t|I_{t-1}^{(Y)}).$$

In other words, the lagged variables of X_t have no impact on the level of Y_t.

Granger causality is defined in terms of incremental predictability rather than the real cause-effect relationship. From an econometric point of view, it is a test of omitted variables in a time series context. It is first introduced by Granger (1969).

Question: How to test Granger causality?

We consider two approaches to testing Granger causality. The first test is proposed by Granger (1969). Consider now a linear regression model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} \\ &\quad + \beta_{p+1} X_{t-1} + \cdots + \beta_{p+q} X_{t-q} + \varepsilon_t. \end{aligned}$$

Under non-Granger causality, we have

$$\mathbf{H}_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0.$$

The F-test statistic

$$F \sim F_{q,n-(p+q+1)}.$$

The classical regression theory of Chapter 3 (Assumption 3.2: $E(\varepsilon_t|\mathbf{X}) = 0$) rules out this application, because it is a dynamic regression model. However, we have justified in this chapter that under H_0,

$$q \cdot F \xrightarrow{d} \chi_q^2$$

as $n \to \infty$ under conditional homoskedasticity even for a linear dynamic regression model.

There is another well-known test for Granger causality proposed by Sims (1980), which is based on the fact that the future cannot cause the present in any notion of causality. To test whether $\{X_t\}$ Granger-causes $\{Y_t\}$, we consider the following linear regression model

$$X_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{j=1}^{J} \beta_j Y_{t+j} + \sum_{j=1}^{q} \gamma_j Y_{t-j} + \varepsilon_t.$$

Here, the dependent variable is X_t rather than Y_t. If $\{X_t\}$ Granger-causes $\{Y_t\}$, we expect some relationship between the current X_t and the future values of Y_t. Note that nonzero values for any of $\{\beta_j\}_{j=1}^{J}$ cannot be interpreted as causality from the future values of Y_t to the current X_t, simply because the future cannot cause the present. Nonzero values of any β_j must imply that there exists causality from current X_t to future values of Y_t. Therefore, we test the null hypothesis

$$\mathbf{H}_0 : \beta_j = 0 \text{ for } 1 \leq j \leq J.$$

Let F be the associated F-test statistic. Then under H_0,

$$J \cdot F \xrightarrow{d} \chi^2_J$$

as $n \to \infty$ under conditional homoskedasticity.

Example 2 [Wage Determination]: Consider the wage function

$$\begin{aligned} W_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 U_t \\ &\quad + \beta_4 V_t + \beta_5 W_{t-1} + \varepsilon_t, \end{aligned}$$

where W_t = wage, P_t = price, U_t = unemployment, and V_t = unfilled vacancies. We will test the null hypothesis

$$\mathbf{H}_0 : \beta_1 + \beta_2 = 0, \beta_3 + \beta_4 = 0, \text{ and } \beta_5 = 1.$$

Question: What is the economic interpretation of the null hypothesis $\mathbf{H}_0$?

Under $\mathbf{H}_0$, we have the restricted wage equation:

$$\Delta W_t = \beta_0 + \beta_1 \Delta P_t + \beta_4 D_t + \varepsilon_t,$$

where $\Delta W_t = W_t - W_{t-1}$ is the wage growth rate, $\Delta P_t = P_t - P_{t-1}$ is the inflation rate, and $D_t = V_t - U_t$ is an index for job market situation (excess job supply). This implies that the wage increase depends on the inflation rate and the excess labor supply.

Under $\mathbf{H}_0$, we have

$$3F \xrightarrow{d} \chi^2_3.$$

A Special Case: Testing for Joint Significance of All Economic Variables

Theorem 5.14 [$(n-K)R^2$ **Test**]: *Suppose Assumption 5.1-5.6 hold, and we are interested in testing the null hypothesis that*

$$\mathbf{H}_0 : \beta_1^o = \beta_2^o = \cdots = \beta_k^o = 0,$$

where the β_j^o, $1 \leq j \leq k$, *are the slope coefficients in the linear regression model* $Y_t = X_t'\beta^o + \varepsilon_t$.

Let R^2 *be the coefficient of determination from the unrestricted regression model*

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

Then under $\mathbf{H}_0$,

$$(n-K)R^2 \xrightarrow{d} \chi^2_k.$$

Proof: First, note that as shown earlier, we have in this case,

$$F = \frac{R^2/k}{(1-R^2)/(n-K)}.$$

Here, we have $J = k$, and under $\mathbf{H}_0$,

$$k \cdot F = \frac{(n-K)R^2}{1-R^2} \xrightarrow{d} \chi^2_k.$$

This implies that $k \cdot F$ is bounded in probability; that is,

$$\frac{(n-K)R^2}{1-R^2} = O_P(1).$$

Consequently, given that k is fixed (i.e., does not grow with the sample size n), we have

$$R^2/(1-R^2) \xrightarrow{p} 0$$

or equivalently,

$$R^2 \xrightarrow{p} 0.$$

Therefore, $1 - R^2 \xrightarrow{p} 1$. By the Slutsky theorem, we have

$$\begin{aligned}(n-K)R^2 &= \frac{(n-K)R^2}{1-R^2} \cdot (1-R^2) \\ &\xrightarrow{d} \chi^2_k.\end{aligned}$$

This completes the proof. ■

Example 3 [Efficient Market Hypothesis]: Suppose Y_t is the exchange rate return in period t, and I_{t-1} is the information available at time $t-1$. Then a classical version of the efficient market hypothesis (EMH) can be stated as follows:

$$E(Y_t|I_{t-1}) = E(Y_t)$$

To check whether exchange rate changes are unpredictable using the past history of exchange rate changes, we specify a linear regression model:

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

where

$$X_t = (1, Y_{t-1}, ..., Y_{t-k})'.$$

Under EMH, we have

$$\mathbf{H}_0 : \beta_j^o = 0 \text{ for all } j = 1, ..., k.$$

If the alternative

$$\mathbb{H}_A : \beta_j^o \neq 0 \text{ at least for some } j \in \{1, ..., k\}$$

holds, then exchange rate changes are predictable using the past information.

Remarks:

What is the appropriate interpretation if $\mathbf{H}_0$ is not rejected? Note that there exists a gap between the efficiency hypothesis and $\mathbf{H}_0$, because the linear regression model is just one of many ways to check EMH. Thus, $\mathbf{H}_0$ is not rejected, at most we can only say that no evidence against the efficiency hypothesis is found. We should not conclude that EMH holds.

In using $k \cdot F$ or $(n-K)R^2$ statistic to test EMH, although the normality assumption is not needed for this result, we still require conditional homoskedasticity, which rules out autoregressive conditional heteroskedasticity (ARCH) in the dynamic time series regression framework. ARCH effects arise in high-frequency financial time series processes.

Case II: Conditional Heteroskedasticity

Next, we construct hypothesis tests for $\mathbf{H}_0$ under conditional heteroskedasticity. Recall that under $\mathbf{H}_0$,

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

where $V = E(X_t X_t' \varepsilon_t^2)$.

For $J = 1$, we have

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{RQ^{-1}VQ^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Because $\hat{Q} \xrightarrow{p} Q$ and $\hat{V} \xrightarrow{p} V$, where $\hat{V} = \mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}/n$, we have by the Slutsky theorem that the robust t-test statistic

$$T_r = \frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Theorem 5.15 [Robust t-Test Under Conditional Heteroskedasticity] *Suppose Assumptions 5.1–5.5 and 5.7 hold. Then under* $\mathbf{H}_0$ with $J = 1$, *as* $n \to \infty$, the robust t-test statistic

$$T_r = \frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1).$$

For $J > 1$, the quadratic form

$$\begin{aligned}&\sqrt{n}(R\hat{\beta} - r)'[RQ^{-1}VQ^{-1}R']^{-1}\sqrt{n}(R\hat{\beta} - r) \\ &\xrightarrow{d} \chi_J^2\end{aligned}$$

under $\mathbf{H}_0$. Given $\hat{Q} \xrightarrow{p} Q$ and $\hat{V} \xrightarrow{p} V$, where $\hat{V} = \mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}/n$, we have a robust Wald test statistic

$$\begin{aligned}W &= n(R\hat{\beta} - r)'[R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R']^{-1}(R\hat{\beta} - r) \\ &\xrightarrow{d} \chi_J^2\end{aligned}$$

by the Slutsky theorem. We can equivalently write

$$W = (R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi_J^2.$$

Theorem 5.16 [Robust Wald Test Under Conditional Heteroskedasticity] *Suppose Assumptions 5.1–5.5 and 5.7 hold. Then under* $\mathbf{H}_0$, *as* $n \to \infty$,

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

Remarks:

Under conditional heteroskedasticity, $J \cdot F$ and $(n-K)R^2$ cannot be used even when $n \to \infty$.

On the other hand, although the general form of the test statistic W developed here can be used no matter whether there exists conditional homoskedasticity, W may perform poorly in small samples (i.e., the asymptotic χ^2_J approximation may be poor in small samples, or Type I errors are large). Thus, if one has information that the error term is conditionally homoskedastic, one should use the test statistics derived under conditional homoskedasticity, which will perform better in small sample sizes. Because of this reason, it is important to test whether conditional homoskedasticity holds in a time series context.

5.7 Testing for Conditional Heteroskedasticity and Autoregressive Conditional Heteroskedasticity

Question: How to test conditional heteroskedasticity in a time series regression context?

Question: Can we still use White's (1980) test for conditional heteroskedasticity?

Yes. Although White's (1980) test is developed under the independence assumption, it is still applicable to a time series linear regression model when $\{X_t\varepsilon_t\}$ is an MDS process. Thus, the test procedure to implement White's (1980) test as is discussed in Chapter 4 can be used here.

In the time series econometrics, there is an alternative approach to testing conditional heteroskedasticity in an autoregressive time series context. This is Engle's (1982, *Econometrica*) Lagrange Multiplier test for autoregressive conditional heteroskedasticity (ARCH) in $\{\varepsilon_t\}$.

Consider the regression model

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t, \\ \varepsilon_t &= \sigma_t z_t, \\ \{z_t\} &\sim i.i.d.(0,1). \end{aligned}$$

The null hypothesis

$$\mathbf{H}_0 : \sigma_t^2 = \sigma^2 \text{ for some } \sigma^2 > 0.$$

where $I_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$.

Here, to allow for a possibly time-varying conditional variance of the regression disturbance ε_t given I_{t-1}, ε_t is formulated as the product between a random shock z_t and $\sigma_t = \sigma(I_{t-1})$. When the random shock series $\{z_t\}$ is i.i.d.$(0, 1)$, we have

$$\begin{aligned} \text{var}(\varepsilon_t | I_{t-1}) &= E(z_t^2 \sigma_t^2 | I_{t-1}) \\ &= \sigma_t^2 E(z_t^2 | I_{t-1}) \\ &= \sigma_t^2. \end{aligned}$$

That is, σ_t^2 is the conditional variance of ε_t given I_{t-1}. The null hypothesis $\mathbf{H}_0$ says that the conditional variance of ε_t given I_{t-1} does not change over time.

The alternative hypothesis to $\mathbf{H}_0$ is that σ_t^2 is a function of I_{t-1}, so it changes over time. In particular, we consider the following auxiliary regression for ε_t^2 :

$$\varepsilon_t^2 = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + v_t,$$

where $E(v_t | I_{t-1}) = 0$ a.s. This is called an ARCH(q) process in Engle (1982). ARCH models can capture a well-known empirical styles fact called volatility clustering in financial markets, that is, a high volatility today tends to be followed by another large volatility tomorrow, and a small volatility today tends to be followed by another small volatility tomorrow, and such patterns alternate over time. To see this more clearly, we consider an ARCH(1) model where

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2,$$

where, to ensure nonnegativity of σ_t^2, both α_0 and α_1 are required to be nonnegative parameters. Suppose $\alpha_1 > 0$. Then if ε_{t-1} is an unusually large deviation from its expectation of 0 so that ε_{t-1}^2 is large, then the conditional variance of ε_t is larger than usual. Therefore, ε_t is expected to have an unusually large deviation from its mean of 0, with either direction. Similarly, if ε_{t-1}^2 is usually small. then σ_t^2 is small, and ε_t^2 is expected to be small as well. Because of this behavior, volatility clustering arises.

In addition to volatility clustering, the ARCH(1) model can also generate heavy tails for ε_t even when the random shock z_t is i.i.d.$N(0,1)$. This can be seen from its kurtosis

$$\begin{aligned} K &= \frac{E(\varepsilon_t^4)}{[E(\varepsilon_t^2)]^2} \\ &= \frac{E(z_t^4)(1-\alpha_1^2)}{(1-3\alpha_1^2)} \\ &> 3 \end{aligned}$$

given $\alpha_1 > 0$.

With an ARCH modeling framework, all autoregressive coefficients $\alpha_j, 1 \leq j \leq q$, are identically zero when $\mathbf{H}_0$ holds. Thus, we can test $\mathbf{H}_0$ by checking whether all $\alpha_j, 1 \leq j \leq q$, are jointly zero. If $\alpha_j \neq 0$ for some $1 \leq j \leq q$, then there exists autocorrelation in $\{\varepsilon_t^2\}$ and $\mathbf{H}_0$ is false.

Observe that with $\varepsilon_t = \sigma_t z_t$ and $\{z_t\}$ is i.i.d.(0,1), the disturbance v_t in the auxiliary autoregression model is an i.i.d. sequence under $\mathbf{H}_0$, which implies that $E(v_t^2|I_{t-1}) = \sigma_v^2$, that is, $\{v_t\}$ is conditionally homoskedastic. Thus, when $\mathbf{H}_0$ holds, we have

$$(n-q-1)\tilde{R}^2 \xrightarrow{d} \chi_q^2,$$

where $\tilde{R}^2$ is the centered R^2 from the auxiliary regression.

The auxiliary regression for ε_t^2, unfortunately, is infeasible because ε_t is not observable. However, we can replace ε_t by the estimated residual e_t and consider the regression

$$e_t^2 = \alpha_0 + \sum_{j=1}^{q} \alpha_j e_{t-j}^2 + \tilde{v}_t.$$

Then we have

$$(n-q-1)R^2 \xrightarrow{d} \chi_q^2.$$

Note that the replacement of ε_t by e_t has no impact on the asymptotic distribution of the test statistic, for the same reason as in White's (1980) direct test for conditional heteroskedasticity. See Chapter 4 for more discussions.

Remarks:

The existence of ARCH effect for $\{\varepsilon_t\}$ does not automatically imply that we have to use White's heteroskedasticity-consistent variance-covariance matrix $Q^{-1}VQ^{-1}$ for the OLS estimator $\hat{\beta}$. Suppose $Y_t = X_t'\beta^o + \varepsilon_t$ is a static time series model such that the

two time series $\{X_t\}$ and $\{\varepsilon_t\}$ are independent of each other, and $\{\varepsilon_t\}$ displays ARCH effect, i.e.,

$$var(\varepsilon_t|I_{t-1}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2$$

with at least some $\alpha_j \neq 0$. Then Assumption 5.6 still holds because $\text{var}(\varepsilon_t|X_t) = \text{var}(\varepsilon_t) = \sigma^2$ given the assumption that $\{X_t\}$ and $\{\varepsilon_t\}$ are independent. In this case, we have $\text{avar}(\sqrt{n}\hat{\beta}) = \sigma^2 Q^{-1}$.

Next, suppose $Y_t = X_t'\beta^o + \varepsilon_t$ is a dynamic time series regression model such that X_t contains some lagged dependent variables (say Y_{t-1}). Then if $\{\varepsilon_t\}$ displays ARCH effect, Assumption 5.6 may fail because we may have $E(\varepsilon_t^2|X_t) \neq \sigma^2$, which generally occurs when X_t and $\{\varepsilon_{t-j}^2, j = 1, ..., p\}$ are not independent. In this case, we have to use $\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1}$.

5.8 Testing for Serial Correlation

Question: Why is it important to test serial correlation for $\{\varepsilon_t\}$?

We first provide some motivation for doing so. Recall that under Assumptions 5.1–5.5,

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

where V =var$(X_t\varepsilon_t)$. Among other things, this implies that the asymptotic variance of $n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t$ is the same as the variance of $X_t\varepsilon_t$. This follows from the MDS assumption for $\{X_t\varepsilon_t\}$:

$$\begin{aligned}
& \text{var}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) \\
= \; & n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} E(X_t\varepsilon_t X_s'\varepsilon_s) \\
= \; & n^{-1}\sum_{t=1}^{n} E(X_t X_t'\varepsilon_t^2) \\
= \; & E(X_t X_t'\varepsilon_t^2) \\
= \; & V.
\end{aligned}$$

This result will not generally hold if the MDS property for $\{X_t\varepsilon_t\}$ is violated.

Question: How to check $E(X_t\varepsilon_t|I_{t-1}) = 0$, where I_{t-1} is the σ-field generated by $\{X_s\varepsilon_s, s < t\}$?

When X_t contains the intercept, we have that $\{\varepsilon_t\}$ is MDS with respect to the σ-field generated by $\{\varepsilon_s, s < t\}$, which implies that $\{\varepsilon_t\}$ is serially uncorrelated (or is a white noise).

If $\{\varepsilon_t\}$ is serially correlated, then $\{X_t\varepsilon_t\}$ will not be MDS, and consequently we will generally have $\text{var}(n^{-1/2}\sum_{t=1}^n X_t\varepsilon_t) \neq V$. Therefore, serial uncorrelatedness is an important necessary condition for the validity of $\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1}$ with $V = E(X_tX_t'\varepsilon_t^2)$.

On the other hand, let us revisit the correct model specification condition that

$$E(\varepsilon_t|X_t) = 0 \text{ a.s.}$$

in a time series context. Note that this condition does not necessarily imply that $\{\varepsilon_t\}$ or $\{X_t\varepsilon_t\}$ is MDS in a time series context.

To see this, consider the case when $Y_t = X_t'\beta^o + \varepsilon_t$ is a static regression model (i.e., when $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent, or at least when $\text{cov}(X_t, \varepsilon_s) = 0$ for all t, s), it is possible that $E(\varepsilon_t|X_t) = 0$ but $\{\varepsilon_t\}$ is serially correlated. An example is that $\{\varepsilon_t\}$ is an AR(1) process but $\{\varepsilon_t\}$ and $\{X_t\}$ are mutually independent. In this case, serial dependence in $\{\varepsilon_t\}$ does not cause inconsistency of OLS $\hat{\beta}$ to β^o, but we no longer have $\text{var}(n^{-1/2}\Sigma_{t=1}^n X_t\varepsilon_t) = V = E(X_tX_t'\varepsilon_t^2)$. In other words, the MDS property for $\{\varepsilon_t\}$ is crucial for $\text{var}(n^{-1/2}\Sigma_{t=1}^n X_t\varepsilon_t) = V$ in a static regression model, although it is not needed to ensure $E(\varepsilon_t|X_t) = 0$. For a static regression model, the regressors X_t are usually called exogenous variables. In particular, if $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent, then X_t is called strictly exogenous.

On the other hand, when $Y_t = X_t'\beta^o + \varepsilon_t$ is a dynamic model (i.e., when X_t includes lagged dependent variables such as $\{Y_{t-1}, \cdots, Y_{t-k}\}$ so that X_t and ε_{t-j} are generally not independent for $j > 0$), the correct model specification condition

$$E(\varepsilon_t|X_t) = 0 \text{ a.s.}$$

holds when $\{\varepsilon_t\}$ is MDS. If $\{\varepsilon_t\}$ is not an MDS, the condition that $E(\varepsilon_t|X_t) = 0$ a.s. generally does not hold. To see this, we consider, for example, an AR(1) model

$$\begin{aligned} Y_t &= \beta_0^o + \beta_1^o Y_{t-1} + \varepsilon_t \\ &= X_t'\beta^o + \varepsilon_t. \end{aligned}$$

Suppose $\{\varepsilon_t\}$ is an MA(1) process. Then $E(X_t\varepsilon_t) \neq 0$, and so $E(\varepsilon_t|X_t) \neq 0$. Thus, to ensure correct specification ($E(Y_t|X_t) = X_t'\beta^o$ a.s.) of a dynamic regression model in a

time series context, it is important to check MDS for $\{\varepsilon_t\}$. In this case, tests for MDS can be viewed as specification tests for dynamic regression models.

In time series econometrics such as rational expectations econometrics, correct model specification usually requires that ε_t be MDS:

$$E(\varepsilon_t|I_{t-1}) = 0 \text{ a.s.}$$

where I_{t-1} is the information set available to the economic agent at time $t-1$. In this content, X_t is usually a subset of I_{t-1}, namely $X_t \in I_{t-1}$. Thus both Assumptions 5.3 and 5.5 hold simultaneously:

$$\begin{aligned} E(\varepsilon_t|X_t) &= E[E(\varepsilon_t|I_{t-1})|X_t] = 0 \text{ a.s.} \\ &\quad \text{and} \\ E(X_t\varepsilon_t|I_{t-1}) &= X_t E(\varepsilon_t|I_{t-1}) = 0 \text{ a.s.} \end{aligned}$$

because X_t belongs to I_{t-1}.

To check the MDS property of $\{\varepsilon_t\}$, one may check whether there exists serial correlation in $\{\varepsilon_t\}$. Evidence of serial correlation in $\{\varepsilon_t\}$ will indicate that $\{\varepsilon_t\}$ is not MDS. The existence of serial correlation may be due to various sources of model misspecification. For example, it may be that in the linear regression model, an important explanatory variable is missing (omitted variables), or that the functional relationship is nonlinear (functional form misspecification), or that lagged dependent variables or lagged explanatory variables should be included as regressors (neglected dynamics or dynamic misspecification). Therefore, tests for serial correlation can also be viewed as a model specification check in a dynamic time series regression context.

Question: How to check serial dependence in $\{\varepsilon_t\}$?

We now introduce a number of tests for serial correlation of the disturbance $\{\varepsilon_t\}$ in a linear regression model.

Breusch and Godfrey's Lagrange Multiplier Test for Serial Correlation

The null hypothesis

$$\mathbf{H}_0 : E(\varepsilon_t|I_{t-1}) = 0,$$

where ε_t is the regression error in the linear regression model

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

$I_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$, and $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.

Below, following the vast literature, we will first assume conditional homoskedasticity in testing serial correlation for $\{\varepsilon_t\}$. Thus, this method is not suitable for high-frequency financial time series, where volatility clustering has been well-documented. Extensions to conditional heteroskedasticity will be discussed later.

First, suppose ε_t is observed, and we consider the auxiliary regression model (an AR(p))

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + u_t, \quad t = p+1, \cdots, n,$$

where $\{u_t\}$ is MDS. Under $\mathbf{H}_0$, we have $\alpha_j = 0$ for $1 \leq j \leq p$. Thus, we can test $\mathbf{H}_0$ by checking whether the α_j are jointly equal to 0. Assuming $E(\varepsilon_t^2|X_t) = \sigma^2$ (which implies $E(u_t^2|X_t) = \sigma^2$ under $\mathbf{H}_0$), then we can run an OLS regression and obtain

$$(n-2p)\tilde{R}_{uc}^2 \xrightarrow{d} \chi_p^2,$$

where $\tilde{R}_{uc}^2$ is the uncentered R^2 in the auxiliary regression (note that there is no intercept), and p is the number of the regressors. The reason that we use $(n-2p)\tilde{R}_{uc}^2$ is that t begins from $p+1$.

Unfortunately, ε_t is not observable. However, we can replace ε_t with the estimated residual $e_t = Y_t - X_t'\hat{\beta}$. Unlike White's (1980) test for heteroskedasticity of unknown form, this replacement will generally change the asymptotic χ_p^2 distribution for $(n-2p)R_{uc}^2$ here. To remove the impact of the estimation error $X_t'(\hat{\beta} - \beta^o)$, we have to modify the auxiliary regression as follows:

$$\begin{aligned} e_t &= \sum_{j=1}^{K} \gamma_j X_{jt} + \sum_{j=1}^{p} \alpha_j e_{t-j} + u_t \\ &= \gamma' X_t + \sum_{j=1}^{p} \alpha_j e_{t-j} + u_t, \quad t = p+1, \cdots, n, \end{aligned}$$

where X_t contains the intercept. The inclusion of the regressors X_t in the auxiliary regression will purge the impact of the estimation error $X_t'(\hat{\beta} - \beta^o)$ of the test statistic, because X_t and $X_t'(\hat{\beta} - \beta^o)$ are perfectly correlated. Therefore, the resulting statistic

$$(n-2p-K)R^2 \xrightarrow{d} \chi_p^2,$$

under $\mathbf{H}_0$, where R^2 is the centered squared multi-correlation coefficient in the feasible auxiliary regression model.

Question: Why should X_t be generally included in the auxiliary regression?

When we replace ε_t by $e_t = \varepsilon_t - X_t'(\hat{\beta} - \beta^o)$, the estimation error $X_t'(\hat{\beta} - \beta^o)$ will have nontrivial impact on the asymptotic distribution of a test statistic for $\mathbf{H}_0$, because X_t may be correlated with ε_{t-j} at least for some lag order $j > 0$ (this occurs when the regression model is dynamic). To remove the impact of $X_t'(\hat{\beta} - \beta^o)$, we have to add the regressor X_t in the auxiliary regression, which is perfectly correlated with the estimation error $X_t'(\hat{\beta} - \beta^o)$, and thus can extract its impact. This can be proven rigorously but we do not attempt to do so here, because it would be very tedious and offer no much new insight than the above intuition. Below, we provide a heuristic explanation.

First, we consider the infeasible auxiliary autoregression. Under the null hypothesis of no serial correlation, the OLS estimator

$$\sqrt{n}(\tilde{\alpha} - \alpha^0) = \sqrt{n}\tilde{\alpha}$$

converges to an asymptotic normal distribution, which implies $\tilde{\alpha} = O_P(n^{-1/2})$ vanishes in probability at a rate of $n^{-1/2}$. The test statistic $n\tilde{R}^2_{uc}$ is asymptotically equivalent to a quadratic form in $\sqrt{n}\tilde{\alpha}$ which follows an asymptotic χ^2_p distribution. In other words, the asymptotic distribution of $n\tilde{R}^2_{uc}$ is determined by the asymptotic distribution of $\sqrt{n}\tilde{\alpha}$.

Now, suppose we replace ε_t by $e_t = \varepsilon_t - (\hat{\beta} - \beta^o)'X_t$, and consider the feasible autoregression

$$e_t = \sum_{j=1}^{p} \alpha_j e_{t-j} + v_t.$$

Suppose the OLS estimator is $\hat{\alpha}$. We can then decompose

$$\hat{\alpha} = \tilde{\alpha} + \hat{\delta} + \text{ reminder term},$$

where $\tilde{\alpha}$, as discussed above, is the OLS estimator of regressing ε_t on $\varepsilon_{t-1}, ..., \varepsilon_{t-p}$, and $\hat{\delta}$ is the OLS estimator of regressing $(\hat{\beta} - \beta^o)'X_t$ on $\varepsilon_{t-1}, ..., \varepsilon_{t-p}$. For a dynamic regression model, the regressor X_t contains lagged dependent variables and so $E(X_t\varepsilon_{t-j})$ is likely nonzero for some $j \in \{1, ..., p\}$. It follows that $\hat{\delta}$ will converge to zero at the same rate as $\tilde{\alpha} - \alpha^0$, which is $n^{-1/2}$. Because $\hat{\delta} \xrightarrow{p} 0$ at the same rate as $\tilde{\alpha}$, $\hat{\delta}$ will have impact on the asymptotic distribution of nR^2_{uc}, where R^2_{uc} is the uncentered R^2 in the auxiliary autoregression. To remove the impact of $\hat{\delta}$, we need to include X_t as additional regressors in the auxiliary regression.

Question: When do we need not include X_t in the auxiliary regression?

Answer: When we have a static regression model, $\text{cov}(X_t, \varepsilon_s) = 0$ for all t, s (so $E(X_t\varepsilon_{t-j}) = 0$ for all $j = 1, ..., p$), the estimation error $X_t'(\hat{\beta} - \beta^o)$ has no impact on the asymptotic distribution of nR_{uc}^2. It follows that we do not need to include X_t in the auxiliary autoregression. In other words, we can test serial correlation for $\{\varepsilon_t\}$ by running the following auxiliary regression model

$$e_t = \sum_{j=1}^{p} \alpha_j e_{t-j} + u_t.$$

The resulting nR_{uc}^2 is asymptotically χ_p^2 under the null hypothesis of no serial correlation.

Question: Suppose we have a static regression model, and we include X_t in the auxiliary regression in testing serial correlation of $\{\varepsilon_t\}$. What will happen?

For a static regression model, whether X_t is included in the auxiliary regression has no impact on the asymptotic χ_p^2 distribution of $(n-2p)R_{uc}^2$ or $(n-2p)R^2$ under the null hypothesis of no serial correlation in $\{\varepsilon_t\}$. Thus, we will still obtain an asymptotic valid test statistic $(n-2p)R^2$ under $\mathbf{H}_0$. In fact, the size performance of the test can be better in finite samples. However, the test may be less powerful than the test without including X_t, because X_t may take away some serial correlation in $\{\varepsilon_t\}$ under the alternative to $\mathbf{H}_0$.

Question: What happens if we include an intercept in the auxiliary regression

$$e_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j e_{t-j} + u_t,$$

where e_t is the OLS residual from a static regression model.

With the inclusion of the intercept here, we can then use $(n-2p)R^2$ to test serial correlation in $\{\varepsilon_t\}$, which is more convenient to compute than $(n-2p)R_{uc}^2$. (Most statistical software report R^2 but not R_{uc}^2.) Under $\mathbf{H}_0$, $(n-2p)R^2 \xrightarrow{d} \chi_p^2$. However, the inclusion of the intercept α_0 may have some adverse impact on the power of the test in small samples, because there is an additional parameter to estimate.

As discussed at the beginning of this section, a test for serial correlation can be viewed as a specification test for dynamic regression models in a time series context, because existence of serial correlation in the estimated model residual $\{e_t\}$ will generally indicate misspecification of a dynamic regression model.

On the other hand, for static regression models with time series observations, it is possible that a static regression model $Y_t = X_t'\beta^o + \varepsilon_t$ is correctly specified in the sense that $E(\varepsilon_t|X_t) = 0$ but $\{\varepsilon_t\}$ displays serial correlation. In this case, existence of serial correlation in $\{\varepsilon_t\}$ does not affect the consistency of the OLS estimator $\hat{\beta}$ but affects the asymptotic variance and therefore the efficiency of the OLS estimator $\hat{\beta}$. However, since ε_t is unobservable, one has to use the estimated residual e_t in testing for serial correlation in a static regression model in the same way as in a dynamic regression model. Because the estimated residual

$$\begin{aligned} e_t &= Y_t - X_t'\hat{\beta} \\ &= \varepsilon_t + [E(Y_t|X_t) - X_t'\beta^*] + X_t'(\beta^* - \hat{\beta}), \end{aligned}$$

it contains the true disturbance $\varepsilon_t = Y_t - E(Y_t|X_t)$ and model approximation error $E(Y_t|X_t) - X_t'\beta^*$, where $\beta^* = [E(X_tX_t')]^{-1}E(X_tY_t)$ is the best linear least squares approximation coefficient which the OLS $\hat{\beta}$ always converges to as $n \to \infty$. If the linear regression model is misspecified for $E(Y_t|X_t)$, then the approximation error $E(Y_t|X_t) - X_t'\beta^*$ will never vanish to zero and this term can cause serial correlation in e_t if X_t is a time series process. Thus, when one finds that there exists serial correlation in the estimated residuals $\{e_t\}$ of a static linear regression model, it is also likely due to the misspecification of the static regression model. In this case, the OLS estimator $\hat{\beta}$ is generally not consistent. Therefore, one has to first check correct specification of a static regression model in order to give correct interpretation of any documented serial correlation in the estimated residuals.

In the development of tests for serial correlation in regression disturbances, there have been two very popular tests that have historical importance. One is the Durbin-Watson test and the other is Durbin's h test. The Durbin-Watson test is the first formal procedure developed for testing first order serial correlation

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \qquad \{u_t\} \sim i.i.d.\left(0, \sigma^2\right),$$

using the OLS residuals $\{e_t\}_{t=1}^n$ in a static linear regression model $Y_t = X_t'\beta^o + \varepsilon_t$. Durbin and Watson (1950,1951) propose a test statistic

$$d = \frac{\Sigma_{t=2}^n \left(e_t - e_{t-1}\right)^2}{\Sigma_{t=1}^n e_t^2}.$$

Durbin and Watson present tables of bounds at the 0.05, 0.025 and 0.01 significance levels of the d statistic for static regressions with an intercept. Against the one-sided alternative that $\rho > 0$, if d is less than the lower bound d_L, the null hypothesis that

$\rho = 0$ is rejected; if d is greater than the upper bound d_U, the null hypothesis is accepted. Otherwise, the test is equivocal. Against the one-sided alternative that $\rho < 0$, $4 - d$ can be used to replace d in the above procedure.

The Durbin-Watson test has been extended to test for lag 4 autocorrelation by Wallis (1972) and for autocorrelation at any lag by Vinod (1973).

The Durbin-Watson d test is not applicable to dynamic linear regression models, because parameter estimation uncertainty in the OLS estimator $\hat{\beta}$ will have nontrivial impact on the distribution of d. Durbin (1970) developed the so-called h test for first-order autocorrelation in $\{\varepsilon_t\}$ that takes into account parameter estimation uncertainty in $\hat{\beta}$. Consider a simple dynamic linear regression model

$$Y_t = \beta_0^o + \beta_1^o Y_{t-1} + \beta_2^o X_t + \varepsilon_t,$$

where X_t is strictly exogenous. Durbin's h statistic is defined as:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n \cdot \hat{\text{var}}(\hat{\beta}_1)}},$$

where $\hat{\text{var}}(\hat{\beta}_1)$ is an estimator for the asymptotic variance of $\hat{\beta}_1$, $\hat{\rho}$ is the OLS estimator from regressing e_t on e_{t-1} (in fact, $\hat{\rho} \approx 1 - d/2$). Durbin (1970) shows that $h \xrightarrow{d} N(0,1)$ as $n \to \infty$ under null hypothesis that $\rho = 0$. In fact, Durbin's h test is asymptotically equivalent to the Lagrange multiplier test introduced above.

The Box-Pierce Portmanteau Test

Define the sample autocovariance function

$$\hat{\gamma}(j) = n^{-1} \sum_{t=j+1}^{n} (e_t - \bar{e})(e_{t-j} - \bar{e}),$$

where $\bar{e} = n^{-1} \sum_{t=1}^{n} e_t$ (this is zero when X_t contains an intercept). The Box-Pierce portmanteau test statistic is defined as

$$Q(p) = n \sum_{j=1}^{p} \hat{\rho}^2(j),$$

where the sample aucorrelation function

$$\hat{\rho}(j) = \hat{\gamma}(j)/\hat{\gamma}(0).$$

When $\{e_t\}$ is a directly observed data or is the estimated residual from a static regression model, we can show

$$Q(p) \xrightarrow{d} \chi^2_p$$

under the null hypothesis of no serial correlation.

On the other hand, when e_t is an estimated residual from an ARMA(r, s) model

$$Y_t = \alpha_0 + \sum_{j=1}^{r} \alpha_j Y_{t-j} + \sum_{j=1}^{s} \beta_j \varepsilon_{t-j} + \varepsilon_t,$$

then

$$Q(p) \xrightarrow{d} \chi^2_{p-(r+s)}$$

where $p > r + s$. See Box and Pierce (1970).

To improve small sample performance of the $Q(p)$ test, Ljung and Box (1978) propose a modified $Q(p)$ test statistic:

$$Q^*(p) \equiv n(n+2) \sum_{j=1}^{p} (n-j)^{-1} \hat{\rho}^2(j) \xrightarrow{d} \chi^2_{p-(r+q)}.$$

The modification matches the first two moments of $Q^*(p)$ with those of the χ^2 distribution. This improves the size in small samples, although not the power of the test.

When $\{e_t\}$ is an estimated residual from a dynamic regression model with regressors including both lagged dependent variables and exogenous variables, then the asymptotic distribution of $Q(p)$ is generally unknown (Breusch and Pagan 1980). One solution is to modify the $Q(p)$ test statistic as follows:

$$\hat{Q}(p) \equiv n\hat{\rho}'(I - \hat{\Phi})^{-1}\hat{\rho} \xrightarrow{d} \chi^2_p \text{ as } n \to \infty,$$

where $\hat{\rho} = [\hat{\rho}(1), \cdots, \hat{\rho}(p)]'$, and $\hat{\Phi}$ captures the impact caused by nonzero correlation between $\{X_t\}$ and $\{\varepsilon_{t-j}, 1 \leq j \leq p\}$. See Hayashi (2000, Section 2.10) for more discussion and the expression of $\hat{\Phi}$.

Like the $(n-p)R^2$ test, the $Q(p)$ test also assumes conditional homoskedasticity. In fact, it can be shown to be asymptotically equivalent to the $(n-p)R^2$ test statistic when e_t is the estimated residual of a static regression model.

The Kernel-Based Test for Serial Correlation

Hong (1996, *Econometrica*)

Let $k : R \to [-1, 1]$ be a symmetric function that is continuous at all points except a finite number of points on R, with $k(0) = 1$ and $\int_{-\infty}^{\infty} k^2(z)dz < \infty$.

Examples of $k(\cdot)$:

(i) The truncated kernel

$$k(z) = \mathbf{1}(|z| \leq 1).$$

(ii) The Bartlett kernel

$$k(z) = (1 - |z|)\mathbf{1}(|z| \leq 1).$$

(iii) The Daniell kernel

$$k(z) = \frac{\sin(\pi z)}{\pi z}, \qquad z \in \mathbf{R},$$

Here, $\mathbf{1}(|z| \leq 1)$ is the indicator function that takes value 1 if $|z| \leq 1$ and 0 otherwise.

Define a test statistic

$$M(p) = \left[n \sum_{j=1}^{n-1} k^2(j/p)\hat{\rho}^2(j) - C(p) \right] / \sqrt{D(p)},$$

where $\hat{\rho}(j)$ is the sample autocorrelation function,

$$\begin{aligned} C(p) &= \sum_{j=1}^{n-1} k^2(j/p), \\ D(p) &= 2 \sum_{j=1}^{n-2} k^4(j/p). \end{aligned}$$

Under the null hypothesis of no serial correlation with conditional homoskedasticity, it can be shown that

$$M(p) \xrightarrow{p} N(0, 1)$$

as $p = p(n) \to \infty, p/n \to 0$. This holds no matter whether e_t is the estimated residual from a static regression model or a dynamic regression model.

To appreciate why $M(p) \xrightarrow{d} N(0, 1)$, we consider a special case of using the truncated kernel $k(z) = \mathbf{1}(|z| \leq 1)$, which assigns an equal weight to each of the first p lags. In this case, $M(p)$ becomes

$$M_T(p) = \frac{n \sum_{j=1}^{p} \hat{\rho}^2(j) - p}{\sqrt{2p}}.$$

This can be viewed as a generalized version of the Box-Pierce test. In other words, the Box-Pierce test can be viewed as a kernel-based test with the choice of the truncated kernel.

For a static regression model, we have $n\Sigma_{j=1}^{p}\hat{\rho}^2(j) \xrightarrow{d} \chi_p^2$ under the null hypothesis of no serial correlation. When p is large, we can obtain a normal approximation for χ_p^2 by substracting its mean p and dividing by its standard deviation $\sqrt{2p}$:

$$\frac{\chi_p^2 - p}{\sqrt{2p}} \xrightarrow{d} N(0,1) \text{ as } p \to \infty.$$

In fact, when $p \to \infty$ as $n \to \infty$, we have the same asymptotic result even when the regression model is dynamic.

Question: Why is it not needed to correct for the impact of the estimation error contained in e_t even when the regression model is dynamic?

Answer: The estimation error indeed does have some impact but such impact becomes asymptotically negligible when p grows to infinity as $n \to \infty$. In contrast, the Box-Pierce portmanteau test has some problem because it uses a fixed lag order p (i.e., p is fixed when $n \to \infty$.)

Question: What is the advantage of using a kernel function?

For a weakly stationary process $\{\varepsilon_t\}$, the autocorrelation function $\rho(j)$ typically decays to zero as j increases. Consequently, it is more powerful if one can discount higher order lags rather than treat all lags equally. This can be achieved by using a downward weighting kernel function such as the Bartlett kernel and the Daniell kernel. Hong (1996) shows that the Daniell kernel gives a most powerful test among a class of kernel functions.

Testing Serial Correlation Under Conditional Heteroskedasticity

We have been testing serial correlation under conditional homoskedasticity. All aforementioned tests assume conditional homoskedasticity or even *i.i.d.* on $\{\varepsilon_t\}$ under the null hypothesis of no serial correlation, which rules out high frequency financial time series, which has been documented to have persistent volatility clustering. To test serial correlation under conditional heteroskedasticity, we need to use different procedures because the F-test and $(n-p)R^2$ are no longer valid.

Question: Under what conditions will conditional homoskedasticity be a reasonable assumption? And under what conditions will it not be a reasonable assumption?

Answer: It is a reasonable assumption for low-frequency macroeconomic time series. It is not a reasonable assumption for high-frequency financial time series.

Question: How to construct a test for serial correlation under conditional heteroskedasticity?

Wooldridge's (1991) Robust Test

Some effort has been devoted to robustifying tests for serial correlation. Wooldridge (1990,1991) proposes some regression-based new procedures to test serial correlation that are robust to conditional heteroskedasticity. Specifically, Wooldridge (1990,1991) proposes a two-stage procedure to robustify the nR^2 test for serial correlation in estimated residuals $\{e_t\}$ of a linear regression model (2.1):

- Step 1: Regress $(e_{t-1}, \cdots, e_{t-p})$ on X_t and save the estimated $p \times 1$ residual vector $\hat{v}_t$;
- Step 2: Regress 1 on $\hat{v}_t e_t$ and obtain SSR, the sum of squared residuals;
- Step 3: Compare the $n - SSR$ statistic with the asymptotic χ^2_p distribution.

The first auxiliary regression purges the impact of parameter estimation uncertainty in the OLS estimator $\hat{\beta}$ and the second auxiliary regression delivers a test statistic robust to conditional heteroskedasticity of unknown form.

The Robust Kernel-based Test

Hong and Lee (2006) have recently robustified Hong's (1996) spectral density-based consistent test for serial correlation of unknown form:

$$\hat{M} \equiv \left[n^{-1} \sum_{j=1}^{n-1} k^2(j/p) \hat{\gamma}^2(j) - \hat{C}(p) \right] / \sqrt{\hat{D}(p)},$$

where the centering and scaling factors

$$\begin{aligned} \hat{C}(p) &\equiv \hat{\gamma}^2(0) \sum_{j=1}^{n-1} k^2(j/p) + \sum_{j=1}^{n-1} k^2(j/p) \hat{\gamma}_{22}(j), \\ \hat{D}(p) &\equiv 2\hat{\gamma}^4(0) \sum_{j=1}^{n-2} k^4(j/p) + 4\hat{\gamma}^2(0) \sum_{j=1}^{n-2} k^4(j/p) \hat{\gamma}_{22}(j) \\ &\quad + 2 \sum_{j=1}^{n-2} \sum_{l=1}^{n-2} k^2(j/p) k^2(l/p) \hat{C}(0,j,l)^2, \end{aligned}$$

with

$$\hat{\gamma}_{22}(j) \equiv n^{-1} \sum_{t=j+1}^{n-1} [e_t^2 - \hat{\gamma}(0)][e_{t-j}^2 - \hat{\gamma}(0)]$$

and

$$\hat{C}(0, j, l) \equiv n^{-1} \sum_{t=\max(j,l)+1}^{n} [e_t^2 - \hat{\gamma}(0)] e_{t-j} e_{t-l}.$$

Intuitively, the centering and scaling factors have taken into account possible volatility clustering and asymmetric features of volatility dynamics, so the $\hat{M}$ test is robust to these effects. It allows for various volatility processes, including GARCH models, Nelson's (1991) EGARCH, and Glosten *et al.*'s (1993) Threshold GARCH models.

5.9 Conclusion

In this chapter, after introducing some basic concepts in time series analysis, we show that the asymptotic theory established under the i.i.d. assumption in Chapter 4 carries over to linear ergodic stationary time series regression models with MDS disturbances. The MDS assumption for the regression disturbances plays a key role here. For a static linear regression model, the MDS assumption is crucial for the validity of White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator. For a dynamic linear regression model, the MDS assumption is crucial for correct model specification for the conditional mean $E(Y_t|I_{t-1})$.

To check the validity of the MDS assumption, one can test serial correlation in the regression disturbances. We introduce a number of tests for serial correlation and discuss the difference in testing serial correlation between a static regression model and a dynamic regression model.

EXERCISES

5.1. **(a)** Suppose that using the Lagrange Multiplier test, one finds that there exists serial correlation in $\{\varepsilon_t\}$. Can we conclude that $\{\varepsilon_t\}$ is not a martingale difference sequence $(m.d.s)$? Give your reasoning.

(b) Suppose one finds that there exists no serial correlation in $\{\varepsilon_t\}$. Can we conclude that $\{\varepsilon_t\}$ is a $m.d.s.$? Give your reasoning. [Hint: Consider a process $\varepsilon_t = z_{t-1}z_{t-2} + z_t$, where $z_t \sim i.i.d.(0, \sigma^2)$.]

5.2. Suppose $\{Z_t\}$ is a zero-mean weakly stationary process with spectral density function $h(\omega)$ and normalized spectral density function $f(\omega)$. Show that:

(a) $f(\omega)$ is real-valued for all $\omega \in [-\pi, \pi]$;

(b) $f(\omega)$ is a symmetric function, i.e., $f(-\omega) = f(\omega)$;

(c) $\int_{-\pi}^{\pi} f(\omega)d\omega = 1$;

(d) $f(\omega) \geq 0$ for all $\omega \in [-\pi, \pi]$. [Hint: Consider the limit of $E|n^{-1/2}\sum_{t=1}^{n} Z_t e^{it\omega}|^2$, the variance of the complex-valued random variable $n^{-1/2}\sum_{t=1}^{n} Z_t e^{it\omega}$.

5.3. Suppose a time series linear regression model

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

where the disturbance ε_t is directly observable, satisfies Assumptions 5.1–5.3. This class of models contains both static regression models and dynamic regression models.

(a) Does the condition $E(\varepsilon_t|X_t) = 0$ imply that $\{\varepsilon_t\}$ is a white noise? Explain.

(b) If $\{\varepsilon_t\}$ is MDS, does it imply $E(\varepsilon_t|X_t) = 0$? Explain.

(c) If $\{\varepsilon_t\}$ is serially correlated, does it necessarily imply $E(\varepsilon_t|X_t) \neq 0$, i.e., the linear regression model is misspecified for $E(Y_t|X_t)$? Explain.

5.4. Suppose that in a linear regression model

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

the disturbance ε_t is directly observable. We are interested in testing the null hypothesis $\mathbf{H}_0$ that $\{\varepsilon_t\}$ is serially uncorrelated. Suppose Assumptions 5.1–5.6 hold.

(a) Consider the auxiliary regression

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + u_t, \ t = p+1, ..., n.$$

Let $\tilde{R}_{uc}^2$ is the uncentered R^2 from the OLS estimation of this auxiliary regression. Show that $(n-2p)\tilde{R}_{uc} \xrightarrow{d} \chi_p^2$ as $n \to \infty$ under $\mathbf{H}_0$.

(b) Now consider another auxiliary regression

$$\varepsilon_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + u_t, t = p+1, ..., n.$$

Let $\tilde{R}^2$ be the centered R^2 from this auxiliary regression model. Show that $(n-2p)\tilde{R}^2 \xrightarrow{d} \chi^2_p$ as $n \to \infty$ under $\mathbf{H}_0$.

(c) Which test statistic, $(n-2p)\tilde{R}^2_{uc}$ or $(n-2p)\tilde{R}^2$, performs better in finite samples? Give your heuristic reasoning.

5.5. Suppose that in a linear regression model

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

the disturbance ε_t is directly observable. We are interested in testing the null hypothesis $\mathbf{H}_0$ that $\{\varepsilon_t\}$ is serially uncorrelated. Suppose Assumptions 5.1–5.5 hold, and $E(\varepsilon_t^2|X_t) \neq \sigma^2$.

(a) Consider the auxiliary regression

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + u_t, \ t = p+1, ..., n.$$

Construct an asymptotically valid test statistic for the null hypothesis that there exists no serial correlation in $\{\varepsilon_t\}$.

5.6. Suppose ε_t follows an ARCH(1) process

$$\begin{aligned} \varepsilon_t &= z_t\sigma_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2, \\ \{z_t\} &\sim i.i.d.N(0,1) \end{aligned}$$

(a) Show $E(\varepsilon_t|I_{t-1}) = 0$ and $\text{cov}(\varepsilon_t, \varepsilon_{t-j}) = 0$ for all $j > 0$, where $I_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$.
(b) Show $\text{cov}(\varepsilon_t^2, \varepsilon_{t-1}^2) = \alpha_1$.
(c) Show the kurtosis of ε_t is given by

$$\begin{aligned} K &= \frac{E(\varepsilon_t^4)}{[E(\varepsilon_t^2)]^2} = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2} \\ &> 3 \text{ if } \alpha_1 > 0. \end{aligned}$$

5.7. Suppose a time series linear regression model

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

where the disturbance ε_t is directly observable, satisfies Assumptions 5.1–5.5. Both static and dynamic regression models are covered.

Suppose there exists autoregressive conditional heteroskedasticity (ARCH) for $\{\varepsilon_t\}$, namely,

$$E(\varepsilon_t^2|I_{t-1}) = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2,$$

where I_{t-1} is the sigma-field generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$. Does this imply that one has to use the asymptotic variance formula $Q^{-1}VQ^{-1}$ for avar$(\sqrt{n}\hat{\beta})$? Explain.

5.8. Suppose a time series linear regression model

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

where the disturbance ε_t is directly observable, satisfies Assumptions 5.1–5.5, and the two time series $\{X_t\}$ and $\{\varepsilon_t\}$ are independent of each other.

Suppose there exists autoregressive conditional heteroskedasticity for $\{\varepsilon_t\}$, namely,

$$E(\varepsilon_t^2|I_{t-1}) = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2,$$

where I_{t-1} is the sigma-field generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$.

What is the form of avar$(\sqrt{n}\hat{\beta})$, where $\hat{\beta}$ is the OLS estimator?

5.9. Suppose a dynamic time series linear regression model

$$\begin{aligned} Y_t &= \beta_0^o + \beta_1^o Y_{t-1} + \varepsilon_t \\ &= X_t'\beta^o + \varepsilon_t \end{aligned}$$

satisfies Assumptions 5.1–5.5. Suppose further there exists autoregressive conditional heteroskedasticity for $\{\varepsilon_t\}$ in form of the following:

$$E(\varepsilon_t^2|I_{t-1}) = \alpha_0 + \alpha_1 Y_{t-1}^2.$$

What is the form of avar$(\sqrt{n}\hat{\beta})$, where $\hat{\beta}$ is the OLS estimator?

5.10. Suppose a time series linear regression model

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

satisfies Assumptions 5.1, 5.2 and 5.4, the two time series $\{X_t\}$ and $\{\varepsilon_t\}$ are independent of each other, and $E(\varepsilon_t) = 0$. Suppose further that there exist serial correlation in $\{\varepsilon_t\}$.

(a) Does the presence of serial correlation in $\{\varepsilon_t\}$ affect the consistency of $\hat{\beta}$ for β^o? Explain.

(b) Does the presence of serial correlation in $\{\varepsilon_t\}$ affect the form of asymptotic variance $\text{avar}(\sqrt{n}\hat{\beta}) = Q^{-1}VQ^{-1}$, where $V = \lim_{n\to\infty}\text{var}(n^{-1/2}\Sigma_{t=1}^{n} X_t\varepsilon_t)$? In particular, do we still have $V = E(X_tX_t'\varepsilon_t^2)$? Explain.

5.11. Suppose a dynamic time series linear regression model

$$\begin{aligned} Y_t &= \beta_0^o + \beta_1^o Y_{t-1} + \varepsilon_t \\ &= X_t'\beta^o + \varepsilon_t, \end{aligned}$$

where $X_t = (1, Y_{t-1})'$, satisfies Assumptions 5.1, 5.2 and 5.4. Suppose further $\{\varepsilon_t\}$ follows an MA(1) process:

$$\varepsilon_t = \rho v_{t-1} + v_t,$$

where $\{v_t\}$ is i.i.d.$(0, \sigma_v^2)$. Thus, there exists first order serial correlation in $\{\varepsilon_t\}$.

Is the OLS estimator $\hat{\beta}$ consistent for β^o? Explain.

CHAPTER 6. LINEAR REGRESSION MODELS UNDER CONDITIONAL HETEROSKEDASTICITY AND AUTOCORRELATION

Abstract: When the regression disturbance $\{\varepsilon_t\}$ displays serial correlation, the asymptotic results in Chapter 5 are no longer applicable, because the asymptotic variance of the OLS estimator will depend on serial correlation in $\{X_t\varepsilon_t\}$. In this chapter, we introduce a method to estimate the asymptotic variance of the OLS estimator in the presence of heteroskedasticity and autocorrelation, and then develop test procedures based on it. Some empirical applications are considered.

Key words: Heteroskedasticity and Autocorrelation (HAC) consistent variance-covariance matrix, Kernel function, Long-run variance-covariance matrix, Newey-West estimator, Nonparametric estimation, Spectral density matrix.

Motivation

In Chapter 5, we assumed that $\{X_t\varepsilon_t\}$ is an MDS. In many economic applications, there may exist serial correlation in the regression error $\{\varepsilon_t\}$. As a consequence, $\{X_t\varepsilon_t\}$ is generally no longer an MDS. We now provide a few examples where $\{\varepsilon_t\}$ is serially correlated.

Example 1 [Testing a zero population mean]: Suppose the daily stock return $\{Y_t\}$ is a stationary ergodic process with $E(Y_t) = \mu$. We are interested in testing the null hypothesis

$$\mathbf{H}_0 : \mu = 0$$

versus the alternative hypothesis

$$\mathbb{H}_A : \mu \neq 0.$$

A test for $\mathbf{H}_0$ can be based on the sample mean

$$\bar{Y}_n = n^{-1}\sum_{t=1}^{n} Y_t.$$

By a suitable CLT (White (1999)), the sampling distribution of the sample mean $\bar{Y}_n$ scaled by $\sqrt{n}$

$$\sqrt{n}\bar{Y}_n \xrightarrow{d} N(0, V),$$

where the asymptotic variance of the sample mean

$$V \equiv avar\left(\sqrt{n}\bar{Y}_n\right).$$

Because

$$\begin{aligned}\text{var}(\sqrt{n}\bar{Y}_n) &= n^{-1}\sum_{t=1}^{n}\text{var}\left(Y_t\right) \\ &\quad +2n^{-1}\sum_{t=2}^{n-1}\sum_{j=1}^{t-1}\text{cov}(Y_t, Y_{t-j}),\end{aligned}$$

serial correlation in $\{Y_t\}$ is expected to affect the asymptotic variance of $\sqrt{n}\bar{Y}_n$. Thus, unlike in Chapter 5, avar$(\sqrt{n}\bar{Y}_n)$ is no longer equal to var(Y_t).

Suppose there exists a variance-covariance estimator $\hat{V}$ such that $\hat{V} \xrightarrow{p} V$. Then, by the Slutsky theorem, we can construct a test statistic which is asymptotically N(0,1) under $\mathbf{H}_0$:

$$\frac{\sqrt{n}\bar{Y}_n}{\sqrt{\hat{V}}} \xrightarrow{d} N(0,1).$$

Example 2 [Unbiasedness Hypothesis]: Consider the following linear regression model

$$S_{t+\tau} = \alpha + \beta F_t(\tau) + \varepsilon_{t+\tau},$$

where $S_{t+\tau}$ is the spot foreign exchange rate at time $t+\tau$, $F_t(\tau)$ is the forward exchange rate (with maturity $\tau > 0$) at time t, and the disturbance $\varepsilon_{t+\tau}$ is not observable. Forward currency contracts are agreements to exchange, in the future, fixed amounts of two currencies at prices set today. No money changes hand over until the contract expires or is offset.

It has been a longstanding controversy on whether the current forward rate $F_t(\tau)$, as opposed to the current spot rate S_t, is a better predictor of the future spot rate $S_{t+\tau}$. The unbiasedness hypothesis states that the forward exchange rate (with maturity τ) at time t is the optimal predictor for the spot exchange rate at time $t+\tau$, namely,

$$E(S_{t+\tau}|I_t) = F_t(\tau) \text{ a.s.},$$

where I_t is the information set available at time t. This implies

$$\mathbf{H}_0 : \alpha = 0, \beta = 1,$$

and

$$E(\varepsilon_{t+\tau}|I_t) = 0 \text{ a.s., } t = 1, 2, ...$$

However, with $\tau > 1$, we generally do not have $E(\varepsilon_{t+j}|I_t) = 0$ a.s. for $1 \leq j \leq \tau - 1$. Consequently, there exists serial correlation in $\{\varepsilon_t\}$ up to $\tau - 1$ lags under H_0.

Example 3 [Long Horizon Return Predictability]: There has been much interest in regressions of asset returns, measured over various horizons, on various forecasting variables. The latter include ratios of price to dividends or earnings various interest rate measures such as the yield spread between long and short term rates, and the quality yield spread between low and high-grade corporate bonds, and the short term interest rate.

Consider a regression

$$Y_{t+h,h} = \beta_0 + \beta_1 r_t + \beta_2 (d_t - p_t) + \varepsilon_{t+h,h}$$

where $Y_{t+h,h}$ is the cumulative return over the holding period from time t to time $t + h$, namely,

$$Y_{t+h,h} = \sum_{j=1}^{h} R_{t+j},$$

where R_{t+j} is an asset return in period $t + j$, r_t is the short term interest rate in time t, and $d_t - p_t$ is the log dividend-price ratio, which is expected to be a good proxy for market expectations of future stock return, because $d_t - p_t$ is equal to the expectation of the sum of all discounted future returns and dividend growth rates. In the empirical finance, there has been an interest in investigating how the predictability of asset returns by various forecasting variables depends on time horizon h. For example, it is expected that $d_t - p_t$ is a better proxy for expectations of long horizon returns than for expectations of short horizon returns. When monthly data is used and $h > 1$, there exists an overlapping for observations on $Y_{t+h,h}$. As a result, the regression disturbance $\varepsilon_{t+h,h}$ is expected to display serial correlation up to lag order $h - 1$.

Example 4 [Relationship between GDP and Money Supply]: Consider the linear macroeconomic regression model

$$Y_t = \alpha + \beta M_t + \varepsilon_t,$$

where Y_t is GDP at time t, M_t is the money supply at time t, and ε_t is an unobservable disturbance such that $E(\varepsilon_t|M_t) = 0$ but there may exist strong serial correlation of unknown form in $\{\varepsilon_t\}$.

Question: What happens to the OLS estimator $\hat{\beta}$ if the disturbance $\{\varepsilon_t\}$ displays conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s. fails) *and/or* autocorrelation (i.e., $\mathrm{cov}(\varepsilon_t, \varepsilon_{t-j}) \neq 0$ for some $j > 0$)? In particular,

- Is the OLS estimator $\hat{\beta}$ consistent for β^o?
- Is $\hat{\beta}$ asymptotically most efficient?
- Is $\hat{\beta}$, after properly scaled, asymptotically normal?
- Are the t-test and F-test statistics are applicable for large sample inference?

6.1 Framework and Assumptions

We now state the set of assumptions which allow for serial correlation and conditional heteroskedasticity of unknown form.

Assumption 6.1 [Ergodic Stationarity]: $\{Y_t, X_t'\}_{t=1}^{'n}$ is a stationary ergodic process.

Assumption 6.2 [Linearity]:

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

where β^o is a $K \times 1$ unknown parameter and ε_t is the unobservable disturbance.

Assumption 6.3 [Correct Model Specification]: $E(\varepsilon_t|X_t) = 0$ a.s.

Assumption 6.4 [Nonsingularity]: The $K \times K$ matrix

$$Q = E(X_t X_t')$$

is finite and nonsingular.

Assumption 6.5 [Long-run Variance]: (i) For $j = 0, \pm 1, ...$, put the $K \times K$ matrix

$$\begin{aligned} \Gamma(j) &= \text{cov}(X_t\varepsilon_t, X_{t-j}\varepsilon_{t-j}) \\ &= E[X_t\varepsilon_t\varepsilon_{t-j}X_{t-j}']. \end{aligned}$$

Then $\Sigma_{j=-\infty}^{\infty}||\Gamma(j)|| < \infty$, where $||A|| = \Sigma_{i=1}^{K}\Sigma_{j=1}^{K}|A_{(i,j)}|$ for any $K \times K$ matrix, and the long-run variance-covariance matrix

$$V = \sum_{j=-\infty}^{\infty} \Gamma(j)$$

is p.d.

(ii) The conditional expectation

$$E(X_t\varepsilon_t|X_{t-j}\varepsilon_{t-j}, X_{t-j-1}\varepsilon_{t-j-1}, ...) \overset{q.m.}{\to} 0 \text{ as } j \to \infty;$$

(iii) $\sum_{j=0}^{\infty}[E(r_j'r_j)]^{1/2} < \infty$, where

$$\begin{aligned} r_j &= E(X_t\varepsilon_t|X_{t-j}\varepsilon_{t-j}, X_{t-j-1}\varepsilon_{t-j-1}, ...) \\ &\quad -E(X_t\varepsilon_t|X_{t-j-1}\varepsilon_{t-j-1}, X_{t-j-2}\varepsilon_{t-j-2}, ...). \end{aligned}$$

Remarks:

Assumptions 6.1–6.4 have been assumed in Chapter 5 but Assumption 6.5 is new. Assumption 6.5(i) allows for both conditional heteroskedasticity and autocorrelation of unknown form in $\{\varepsilon_t\}$, and no normality assumption is imposed on $\{\varepsilon_t\}$.

We do not assume that $\{X_t\varepsilon_t\}$ is an MDS, although $E(X_t\varepsilon_t) = 0$ as implied by $E(\varepsilon_t|X_t) = 0$ a.s. Note that $E(\varepsilon_t|X_t) = 0$ a.s. does not necessarily imply that $\{X_t\varepsilon_t\}$ is MDS in a time series context. See the aforementioned examples for which $\{X_t\varepsilon_t\}$ is not MDS.

Assumptions 6.5(ii, iii) imply that the serial dependence of $X_t\varepsilon_t$ on its past history in term of mean and variance respectively vanishes to zero as the lag order $j \to \infty$. Intuitively, Assumption 6.5(iii) may be viewed as the net effect of $X_{t-j}\varepsilon_{t-j}$ on the conditional mean of $X_t\varepsilon_t$. It assumes that $E(r_j'r_j) \to 0$ as $j \to \infty$.

6.2 Long-run Variance Estimation

Question: Why are we interested in V?

Recall that for the OLS estimator $\hat{\beta}$, we have

$$\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}^{-1}n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t.$$

Suppose the CLT holds for $\{X_t\varepsilon_t\}$. That is, suppose

$$n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t \xrightarrow{d} N(0, V),$$

where V is an asymptotic variance, namely

$$\begin{aligned} V &\equiv \text{avar}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) \\ &= \lim_{n\to\infty}\text{var}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right). \end{aligned}$$

Then, by the Slutsky theorem, we have

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

under suitable regularity conditions.

Put

$$g_t = X_t\varepsilon_t.$$

Note that $E(g_t) = 0$ given $E(\varepsilon_t|X_t) = 0$ and the law of iterated expectations. Because $\{g_t\}$ is not an MDS, it may be serially correlated. Thus, the autocovariance function $\Gamma(j) = \text{cov}(g_t, g_{t-j})$ may not be zero at least for some lag order $j > 0$.

Now we consider the variance

$$\begin{aligned}
\text{var}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) &= \text{var}\left(n^{-1/2}\sum_{t=1}^{n} g_t\right) \\
&= E\left[\left(n^{-1/2}\sum_{t=1}^{n} g_t\right)\left(n^{-1/2}\sum_{s=1}^{n} g_s\right)'\right] \\
&= n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} E(g_t g_s') \\
&= n^{-1}\sum_{t=1}^{n} E(g_t g_t') \\
&\quad + n^{-1}\sum_{t=2}^{n}\sum_{s=1}^{t-1} E(g_t g_s') \\
&\quad + n^{-1}\sum_{t=1}^{n-1}\sum_{s=t+1}^{n} E(g_t g_s') \\
&= n^{-1}\sum_{t=1}^{n} E(g_t g_t') \\
&\quad + \sum_{j=1}^{n-1} n^{-1}\sum_{t=j+1}^{n} E(g_t g_{t-j}') \\
&\quad + \sum_{j=-(n-1)}^{-1} n^{-1}\sum_{t=1}^{n+j} E(g_t g_{t-j}') \\
&= \sum_{j=-(n-1)}^{n-1} (1 - |j|/n)\Gamma(j) \\
&\to \sum_{j=-\infty}^{\infty} \Gamma(j) \text{ as } n \to \infty
\end{aligned}$$

by dominated convergence. Therefore, we have $V = \sum_{j=-\infty}^{\infty} \Gamma(j)$.

In contrast, when $\{g_t\}$ is MDS, we have

$$\begin{aligned} V &\equiv \text{avar}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) \\ &= E(g_t g_t') \\ &= E(X_t X_t' \varepsilon_t^2) \\ &= \Gamma(0) \end{aligned}$$

when $\{g_t\}$ is MDS.

When $\text{cov}(g_t, g_{t-j})$ is p.s.d. for all $j > 0$, the difference $\Sigma_{j=-\infty}^{\infty}\Gamma(j) - \Gamma(0)$ is a p.s.d matrix. Intuitively, when $\Gamma(j)$ is p.s.d., a large deviation of g_t from its mean will tend to be followed by another large deviation. As a result, $V - \Gamma(0)$ is p.s.d.

To explore the link between the long-run variance V and the spectral density matrix of $\{X_t\varepsilon_t\}$, which is crucial for consistent estimation of V, we now extend the concept of the spectral density of a univariate time series to a multivariate time series context.

Definition 6.1 [Spectral Density Matrix] *Suppose $\{g_t = X_t\varepsilon_t\}$ is a $K \times 1$ weakly stationary process with $E(g_t) = 0$ and autocovariance function $\Gamma(j) \equiv cov(g_t, g_{t-j}) = E(g_t g_{t-j}')$, which is a $K \times K$ matrix. Suppose*

$$\sum_{j=-\infty}^{\infty} ||\Gamma(j)|| < \infty.$$

Then the Fourier transform of the autocovariance function $\Gamma(j)$ exists and is given by

$$H(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma(j)\exp(-ij\omega), \qquad \omega \in [-\pi, \pi],$$

where $i = \sqrt{-1}$. The $K \times K$ matrix-valued function $H(\omega)$ is called the spectral density matrix of the weakly stationary time series vector-valued process $\{g_t\}$.

Remarks:

The inverse Fourier transform of the spectral density matrix is

$$\Gamma(j) = \int_{-\pi}^{\pi} H(\omega)e^{ij\omega}d\omega.$$

Both $H(\omega)$ and $\Gamma(j)$ are Fourier transforms of each other. They contain the same amount of information on serial dependence of the process $\{g_t = X_t\varepsilon_t\}$. The spectral distribution function $H(\omega)$ is useful to identify business cycles (see Sargent 1987, *Dynamic Marcoeconomics*, 2nd Edition). For example, if g_t is the GDP growth rate at time t, then $H(\omega)$ can be used to identify business cycle of the economy.

When $\omega = 0$, then the long-run variance-covariance matrix

$$V = 2\pi H(0) = \sum_{j=-\infty}^{\infty} \Gamma(j).$$

That is, the long-run variance V is 2π times the spectral density matrix of the time series process $\{g_t\}$ at frequency zero. As will be seen below, this link provides a basis for consistent nonparametric estimation of V.

Question: What are the elements of the $K \times K$ matrix $\Gamma(j)$?

Recall that $g_t = (g_{0t}, g_{1t}, ..., g_{kt})'$, where $g_{lt} = X_{lt}\varepsilon_t$ for $0 \leq l \leq k$. Then the $(l+1, m+1)$-th element of $\Gamma(j)$ is

$$\begin{aligned} [\Gamma(j)]_{(l+1,m+1)} &= \Gamma_{lm}(j) \\ &= \text{cov}[g_{lt}, g_{m(t-j)}] \\ &= \text{cov}[X_{lt}\varepsilon_t, X_{m(t-j)}\varepsilon_{(t-j)}], \end{aligned}$$

which is the cross-covariance between $X_{lt}\varepsilon_t$ and $X_{m(t-j)}\varepsilon_{(t-j)}$. We note that

$$\Gamma_{lm}(j) \neq \Gamma_{lm}(-j),$$

because g_t is a vector, not a scalar. Instead, we have

$$\Gamma(j) = \Gamma(-j)',$$

which implies $\Gamma_{lm}(j) = \Gamma_{ml}(-j)$.

Question: What is the $(l+1, m+1)$-th element of $H(\omega)$ when $l \neq m$? The function

$$H_{lm}(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \Gamma_{lm}(j) e^{-ij\omega}$$

is called the cross-spectral density between $\{g_{lt}\}$ and $\{g_{mt}\}$. The cross-spectrum is very useful in investigating the comovements between different economic time series. The popular concept of Granger causality was first defined using the cross-spectrum (see Granger 1969, *Econometrica*). In general, $H_{lm}(\omega)$ is complex-valued.

Question: How to estimate V?

Recall the important identity:

$$V = 2\pi H(0) = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

where $\Gamma(j) = \text{cov}(g_t, g_{t-j})$. The long-run variance V is 2π times $H(0)$, the spectral density matrix at frequency zero. This provides the basis to use a nonparametric approach to estimating V.

A possible naive estimation method:

Given a random sample $\{Y_t, X_t'\}_{t=1}^n$, we can obtain the estimated OLS residual e_t from the linear regression model $Y_t = X_t'\beta^o + \varepsilon_t$. Because

$$V = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

we first consider a naive estimator

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

where the sample autocovariance function

$$\hat{\Gamma}(j) = \begin{cases} n^{-1}\sum_{t=j+1}^{n} X_t e_t X_{t-j}' e_{t-j}, & j = 0, 1, ..., n-1, \\ n^{-1}\sum_{t=1-j}^{n} X_t e_t X_{t-j}' e_{t-j}, & j = -1, -2, ..., -(n-1). \end{cases}$$

There is no need to subtract the same mean from $X_t e_t$ and $X_{t-j} e_{t-j}$ because $\mathbf{X}'e = \Sigma_{t=1}^n X_t e_t = 0$. Also, note that the summation over lag orders in $\hat{V}$ extends to the maximum lag order $n-1$ for the sample autocovariance function $\hat{\Gamma}(j)$. Unfortunately, although $\hat{\Gamma}(j)$ is consistent for $\Gamma(j)$ for each given j as $n \to \infty$, the estimator $\hat{V}$ is not consistent for V.

Question: Why?

There are too many estimated terms in the summation over lag orders. In fact, there are n estimated parameters $\{\hat{\Gamma}(j)\}_{j=0}^{n-1}$ in $\hat{V}$. The asymptotic variance of the estimator $\hat{V}$ defined above is proportional to the ratio of the number of estimated autocovariance matrices $\{\hat{\Gamma}(j)\}$ to the sample size n, which will not vanish to zero if the number of estimated covariances is the same as or close to the sample size n.

Nonparametric Kernel Estimation

The above explanation motivates us to consider the following truncated sum

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

where p is a positive integer. If p is fixed (i.e., p does not grow when the sample size n increases), however, we expect

$$\hat{V} \xrightarrow{p} \sum_{j=-p}^{p} \Gamma(j) \neq 2\pi H(0) = V,$$

because the resulting bias

$$2\pi H(0) - \sum_{j=-p}^{p} \Gamma(j) = \sum_{|j|>p} \Gamma(j)$$

will never vanish to zero as $n \to \infty$ when p is fixed. Hence, we should let p grows to infinity as $n \to \infty$; that is, let $p = p(n) \to \infty$ as $n \to \infty$. The bias will then vanish to zero as $n \to \infty$. However, we cannot let p grow as fast as the sample size n. Otherwise, the variance of $\hat{V}$ will never vanish to zero. Therefore, to ensure consistency of $\hat{V}$ to V, we should balance the bias and the variance of $\hat{V}$ properly. This requires using a truncated variance estimator

$$\hat{V} = \sum_{j=-p_n}^{p_n} \hat{\Gamma}(j),$$

where $p_n \to \infty, p_n/n \to 0$. An example $p_n = n^{1/3}$.

Although this estimator is consistent for V, it may not be positive semi-definite for all n. To ensure that it is always positive semi-definite, we can use a weighted average estimator

$$\hat{V} = \sum_{j=-p_n}^{p_n} k(j/p_n)\hat{\Gamma}(j)$$

where the weighting function $k(\cdot)$ is called a kernel function. An example of such kernels is the Bartlett kernel

$$k(z) = (1 - |z|)\mathbf{1}(|z| \leq 1),$$

where $\mathbf{1}(\cdot)$ is the indicator function, which takes value 1 if the condition inside holds, and takes value 0 if the condition inside does not hold. Newey and West (1987, *Econometrica*; 1994, *Review of Economic Studies*) first used this kernel function to estimate V in econometrics. The truncated variance estimator $\hat{V}$ can be viewed as a kernel-based

estimator with the use of the truncated kernel $k(z) = \mathbf{1}(|z| \leq 1)$, which assigns equal weighting to each of the first p_n lags.

Most kernels are downward-weighting in the sense that $k(z) \to 0$ as $|z| \to \infty$. The use of a downward weighting kernel may enhance estimation efficiency of V because when $\Sigma_{j=-\infty}^{\infty} ||\Gamma(j)|| < \infty$, we have $\Gamma(j) \to 0$ as $j \to \infty$, and so it is more efficient to assign a larger weight to a lower order j and a smaller weight to a higher order j.

In fact, we can consider a more general form of estimator for V :

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

where $k(\cdot)$ may have unbounded support. Although the lag order j sums up from $1-n$ to $n-1$, the variance of the estimator $\hat{V}$ still vanishes to zero, provided $p_n \to \infty, p_n/n \to 0$, and $k(\cdot)$ discounts higher order lags as $j \to \infty$. An example of $k(\cdot)$ that has unbounded support is the Quadratic-Spectral kernel:

$$k(z) = \frac{3}{(\pi z)^2}\left\{\frac{\sin(\pi z)}{\pi z} - \cos(\pi z)\right\}, \quad -\infty < z < \infty.$$

Andrews (1991, *Econometrica*) uses it to estimate for V. This kernel also delivers a p.s.d. matrix. Moreover, it minimizes the asymptotic MSE of the estimator $\hat{V}$ over a class of kernel functions.

Under certain regularity conditions on random sample $\{Y_t, X_t'\}_{t=1}^{'n}$, kernel function $k(\cdot)$, and lag order p_n (Newey and West 1987, Andrews 1991), we have

$$\hat{V} \xrightarrow{p} V$$

provided $p_n \to \infty, p_n/n \to 0$. Intuitively, although the summation over lag orders in $\hat{V}$ extends to the maximum lag order $n-1$, the lag orders that are much larger than p_n are expected to have negligible contributions to $\hat{V}$, given that $k(\cdot)$ discounts higher order lags. As a consequence, we have $\hat{V} \xrightarrow{p} V$. There are many rules to satisfy $p_n \to \infty, p_n/n \to 0$. Andrews (1991) and Newey and West (1994) discuss data-driven methods to choose p_n.

Question: What are the regularity conditions on $k(\cdot)$?

Assumption on the kernel function: $k : \mathbb{R} \to [-1, 1]$ is symmetric about 0, and is continuous at all points except a finite number of points on $\mathbb{R}$, with $k(0) = 1$ and $\int_{-\infty}^{\infty} k^2(z)dz < \infty$.

At point 0, $k(\cdot)$ attains the maximal value, and the fact that $k(\cdot)$ is square-integrable implies $k(z) \to 0$ as $|z| \to \infty$.

For derivations of asymptotic variance and asymptotic bias of the long-run variance estimator $\hat{V}$, see Newey and West (1987) and Andrews (1991).

6.3 Consistency of OLS

When there exists conditional heteroskedasticity and autocorrelation of unknown form in $\{\varepsilon_t\}$, it is very difficult, if not impossible, to use the GLS estimation. Instead, the OLS estimator $\hat{\beta}$ is convenient to use in practice. We now investigate the asymptotic properties of the OLS $\hat{\beta}$ when there exist conditional heteroskedasticity and autocorrelation of unknown form.

Theorem 6.1: *Suppose Assumptions 6.1–6.5(i) hold. Then*

$$\hat{\beta} \xrightarrow{p} \beta^o \text{ as } n \to \infty.$$

Proof: Recall that we have

$$\hat{\beta} - \beta^o = \hat{Q}^{-1} n^{-1} \sum_{t=1}^{n} X_t \varepsilon_t.$$

By Assumptions 6.1, 6.2 and 6.4 and the WLLN for stationary ergodic processes, we have

$$\hat{Q} \xrightarrow{p} Q \text{ and } \hat{Q}^{-1} \xrightarrow{p} Q^{-1}.$$

Similarly, by Assumptions 6.1–6.3 and 6.5(i), we have

$$n^{-1} \sum_{t=1}^{n} X_t \varepsilon_t \xrightarrow{p} E(X_t \varepsilon_t) = 0$$

using the WLLN for ergodic stationary processes, where $E(X_t \varepsilon_t) = 0$ given Assumption 6.2 ($E(\varepsilon_t | X_t) = 0$ a.s.) and LIE.

6.4 Asymptotic Normality of OLS

Next, we derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

Theorem 6.2: *Suppose Assumptions 6.1–6.5 hold. Then*

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

where $V = \sum_{j=-\infty}^{\infty} \Gamma(j)$ *is as in Assumption 6.5.*

The proof of this theorem calls for the use of a new CLT.

Lemma 6.3 [CLT for Zero Mean Ergodic Stationary Processes (White 1984, Theorem 5.15)]: *Suppose* $\{Z_t\}$ *is a stationary ergodic process with*

(i) $E(Z_t) = 0$;

(ii) $V = \sum_{j=-\infty}^{\infty} \Gamma(j)$ *is finite and nonsingular, where* $\Gamma(j) = E(Z_t Z_{t-j}')$;

(iii) $E(Z_t | Z_{t-j}, Z_{t-j-1}, ...) \overset{q.m.}{\to} 0$;

(iv) $\sum_{j=0}^{\infty} [E(r_j' r_j)]^{1/2} < \infty$, *where*

$$r_j = E(Z_t | Z_{t-j}, Z_{t-j-1}, ...) - E(Z_t | Z_{t-j-1}, Z_{t-j-2}, ...).$$

Then as $n \to \infty$,

$$n^{1/2} \bar{Z}_n = n^{-1/2} \sum_{t=1}^{n} Z_t \xrightarrow{d} N(0, V).$$

We now use this CLT to derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

Proof: Recall that

$$\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}^{-1} n^{-1/2} \sum_{t=1}^{n} X_t \varepsilon_t.$$

By Assumptions 6.1–6.3 and 6.5 and the CLT for stationary ergodic processes, we have

$$n^{-1/2} \sum_{t=1}^{n} X_t \varepsilon_t \xrightarrow{d} N(0, V),$$

where $V = \sum_{j=-\infty}^{\infty} \Gamma(j)$ is as in Assumption 6.5. Also, $\hat{Q} \xrightarrow{p} Q$ and $\hat{Q}^{-1} \xrightarrow{p} Q^{-1}$ by Assumption 6.4 and the WLLN for ergodic stationary processes. We then have by the Slutsky theorem

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

6.5 Hypothesis Testing

We now consider testing the null hypothesis

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

where R is a nonstochastic $J \times K$ matrix, and r is a $J \times 1$ nonstochastic vector.

When there exists autocorrelation in $\{X_t \varepsilon_t\}$, there is no need (and in fact there is no way) to consider the cases of conditional homoskedasticity and conditional heteroskedasticity separately (why?).

Corollary 6.4: *Suppose Assumptions 6.1–6.5 hold. Then under* $\mathbf{H}_0$, *as* $n \to \infty$,

$$\sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} N(0, R Q^{-1} V Q^{-1} R').$$

We directly assume a consistent estimator $\hat{V}$ for V.

Assumption 6.6: $\hat{V} \xrightarrow{p} V$.

When there exists serial correlation of unknown form, we can estimate V using the nonparametric kernel estimator $\hat{V}$, as described in Section 6.3. In some special scenarios, we may have $\Gamma(j) = 0$ for all $j > p_0$, where p_0 is a fixed lag order. An example of this case is Example 2 in Section 6.1. In this case, we can use the following estimator

$$\hat{V} = \sum_{j=-p_0}^{p_0} \hat{\Gamma}(j).$$

It can be shown that $\hat{V} \xrightarrow{p} V$ in this case.

For the case where $J = 1$, a robust t-type test statistic

$$\frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0, 1),$$

where the convergence to $N(0, 1)$ in distribution holds under $\mathbf{H}_0$.

Question: Why is it called a "robust" t-type test?

This statistic has used the asymptotic variance estimator that is robust to conditional heteroskedasticity and autocorrelation of unknown form.

For the case where $J > 1$, we consider a "robust" Wald test.

Theorem 6.5: *Under Assumptions 6.1–6.6, we have the Wald test statistic*

$$\hat{W} = n^{-1}(R\hat{\beta} - r)'[R(\mathbf{X}'\mathbf{X})^{-1}\hat{V}(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_J$$

as $n \to \infty$ *under* $\mathbf{H}_0 : R\beta^o = r$.

Proof: Because

$$\sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} N(0, RQ^{-1}VQ^{-1}R'),$$

we have the quadratic form

$$\sqrt{n}(R\hat{\beta} - r)' \left(RQ^{-1}VQ^{-1}R'\right)^{-1} \sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

By the Slutsky theorem, we have the Wald test statistic

$$\hat{W} = n(R\hat{\beta} - r)' \left(R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'\right)^{-1} (R\hat{\beta} - r) \xrightarrow{d} \chi^2_J.$$

Using the expression of $\hat{Q} = \mathbf{X}'\mathbf{X}/n$, we have an equivalent expression for $\hat{W}$:

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

Remarks:

The standard t-statistic and F-statistic cannot be used when there exists autocorrelation and conditional heteroskedasticity in $\{X_t\varepsilon_t\}$.

Question: Can we use this Wald test when $\Gamma(j) = 0$ for all nonzero j?

Yes. But this is not a good test statistic because it may perform poorly in finite samples. In particular, it usually overrejects the correct null hypothesis $\mathbf{H}_0$ in finite samples even if $\Gamma(j) = 0$ for all $j \neq 0$. In the case where $\Gamma(j) = 0$ for all $j \neq 0$, a better estimator to use is

$$\begin{aligned}\hat{V} &= \hat{\Gamma}(0) \\ &= n^{-1}\sum_{t=1}^{n} X_t e_t e_t X_t' \\ &= \mathbf{X}'\mathrm{D}(e)\mathrm{D}(e)'\mathbf{X}/n.\end{aligned}$$

This is essentially White's heteroskedasticity consistent variance estimator (also see Chapter 5).

Question: Why do the robust t- and Wald tests tend to overreject $\mathbf{H}_0$ in the presence of HAC?

We use the robust t-test as an example. Recall $\hat{V}$ is an estimator for $H(0)$ up to a factor of 2π. When there exists strong positive serial correlation in $\{\varepsilon_t\}$, as is the case of economic time series, $H(\omega)$ will display a peak or mode at frequency zero. The kernel estimator, which is a local averaging estimator, always tends to underestimate $H(0)$, because it has an asymptotic negative bias. Consequently, the robust t-statistic tends to be a larger statistic value, because it is the ratio of $R\hat{\beta} - r$ to the square root of a variance estimator which tends to be smaller than the true variance.

Simulation Evidence

6.6 Testing Whether Long-run Variance Estimation Is Needed

Because of the notorious poor performance of the robust t- and W tests even when $\Gamma(j) = 0$ for all $j \neq 0$, it is very important to test whether we really have to use a long-run variance estimator.

Question: How to test whether we need to use the long-run variance-covariance matrix estimator? That is, how to test whether the null hypothesis that

$$\mathbf{H}_0 : 2\pi H(0) \equiv \sum_{j=-\infty}^{\infty} \Gamma(j) = \Gamma(0)?$$

The null hypothesis $\mathbf{H}_0$ can be equivalently written as follows:

$$\mathbf{H}_0 : \sum_{j=1}^{\infty} \Gamma(j) = 0.$$

It can arise from two cases:

(i) $\Gamma(j) = 0$ for all $j \neq 0$.

(ii) $\Gamma(j) \neq 0$ for some $j \neq 0$, but $\sum_{j=1}^{\infty} \Gamma(j) = 0$. For simplicity, we will consider the first case only. Case (ii) is pathological, although it could occur in practice.

We now provide a test for $\mathbf{H}_0$ under case (i). See Hong (1997) in a related univariate context.

To test the null hypothesis that $\Sigma_{j=1}^{\infty}\Gamma(j) = 0$, we can use a consistent estimator $\hat{A}$ (say) for $\Sigma_{j=1}^{\infty}\Gamma(j)$ and then check whether $\hat{A}$ is close to a zero matrix. Any significant difference of $\hat{A}$ from zero will indicate the violation of the null hypothesis, and thus a long-run variance estimator is needed.

To estimate $\Sigma_{j=1}^{\infty}\Gamma(j)$ consistently, we can use a nonparametric kernel estimator

$$\hat{A} = \sum_{j=1}^{n-1} k(j/p_n)\text{vech}[\hat{\Gamma}(j)],$$

where $p_n = p(n) \to \infty$ at a suitable rate as $n \to \infty$. We shall derive the asymptotic distribution of $\hat{A}$ (with suitable scaling) under the assumption that $\{g_t = X_t\varepsilon_t\}$ is MDS, which implies the null hypothesis H_o that $\Sigma_{j=1}^{\infty}\Gamma(j) = 0$. First, we consider the case when $\{g_t = X_t\varepsilon_t\}$ is autoregressively conditionally homoskedastic, namely, $\text{var}(g_t|I_{t-1}) = \text{var}(g_t)$, where $I_{t-1} = \{g_{t-1}, g_{t-2}, ...\}$. In this case, we can show

$$\left[p\int_0^{\infty} k^2(z)dz\right]^{-1/2} \text{vech}^{-1}\left[\Gamma(0)\right]\sqrt{n}\hat{A} \xrightarrow{d} N(0, I_{K(K+1)/2}).$$

We can then construct a test statistic

$$\begin{aligned}\hat{M} &= \left[p\int_0^{\infty} k^2(z)dz\right]^{-1} n\hat{A}'\text{vech}^{-2}\left[\hat{\Gamma}(0)\right]\hat{A} \\ &\xrightarrow{d} \chi^2_{K(K+1)/2}.\end{aligned}$$

Next, we consider the case when $\{g_t = X_t\varepsilon_t\}$ is autoregressively conditionally heteroskedastic, namely $\text{var}(g_t|I_{t-1}) \neq \text{var}(g_t)$. In this case, the test statistic is

$$\hat{M} = \hat{A}'\hat{B}^{-1}\hat{A},$$

where

$$\begin{aligned}\hat{B} &= \sum_{j=1}^{n-1}\sum_{l=1}^{n-1} k(j/p)k(l/p)\hat{C}(j,l),\\ \hat{C}(j,l) &= \frac{1}{n}\sum_{t=1+\max(j,l)}^{n-1} \text{vech}(\hat{g}_t\hat{g}'_{t-j})\text{vech}'(\hat{g}_t\hat{g}'_{t-l}),\end{aligned}$$

with $\hat{g}_t = X_t e_t$. Under the assumption that $\{g_t = X_t\varepsilon_t\}$ is an MDS, we have

$$\hat{M} \xrightarrow{d} \chi^2_{K(K+1)/2}.$$

This test is robust to autoregressive conditional heteroskedasticity of unknown form for $\{g_t = X_t\varepsilon_t\}$.

A Related Test: Variance Ratio Test

In fact, the above test is closely related to a variance ratio test that is popular in financial econometrics. Extending an idea of Cochrane (1988), Lo and MacKinlay (1988) first rigorously present an asymptotic theory for a variance ratio test for the MDS hypothesis of asset returns $\{Y_t\}$. Recall that $\sum_{j=1}^p Y_{t-j}$ is the cumulative asset return over a total of p periods. Then under the MDS hypothesis, which implies $\gamma(j) \equiv \text{cov}(Y_t, Y_{t-j}) = 0$ for all $j > 0$, one has

$$\frac{\text{var}\left(\sum_{j=1}^p Y_{t-j}\right)}{p \cdot \text{var}(Y_t)} = \frac{p\gamma(0) + 2p\sum_{j=1}^p (1 - j/p)\gamma(j)}{p\gamma(0)} = 1.$$

This unity property of the variance ratio can be used to test the MDS hypothesis because any departure from unity is evidence against the MDS hypothesis.

The variance ratio test is essentially based on the statistic

$$\text{VR}_o \equiv \sqrt{n/p}\sum_{j=1}^{p}(1 - j/p)\hat{\rho}(j) = \frac{\pi}{2}\sqrt{n/p}\left[\hat{f}(0) - \frac{1}{2\pi}\right],$$

where

$$\hat{f}(0) = \frac{1}{2\pi}\sum_{j=-p}^{p}\left(1 - \frac{|j|}{p}\right)\hat{\rho}(j)$$

is a kernel-based normalized spectral density estimator at frequency 0, with the Bartlett kernel $K(z) = (1-|z|)\mathbf{1}(|z| \leq 1)$ and a lag order equal to p. This, the variance ratio test is the same as checking whether the long-run variance is equal to the individual variance $\gamma(0)$. Because VR_o is based on a spectral density estimator of frequency 0, and because of this, it is particularly powerful against long memory processes, whose spectral density at frequency 0 is infinity (see Robinson 1994, for discussion on long memory processes).

Under the MDS hypothesis with conditional homoskedasticity for $\{Y_t\}$, Lo and MacKinlay (1988) show that for any fixed p,

$$\text{VR}_o \xrightarrow{d} N[0, 2(2p-1)(p-1)/3p] \text{ as } n \to \infty.$$

When $\{Y_t\}$ displays conditional heteroskedasticity, Lo and MacKinlay (1988) also consider a heteroskedasticity-consistent variance ratio test:

$$\text{VR} \equiv \sqrt{n/p} \sum_{j=1}^{p} (1 - j/p)\hat{\gamma}(j)/\sqrt{\hat{\gamma}_2(j)},$$

where $\hat{\gamma}_2(j)$ is a consistent estimator for the asymptotic variance of $\hat{\gamma}(j)$ under conditional heteroskedasticity. Lo and MacKinlay (1988) assume a fourth order cumulant condition that

$$E\left[(Y_t - \mu)^2 (Y_{t-j} - \mu)(Y_{t-l} - \mu)\right] = 0, \qquad j, l > 0, j \neq l.$$

Intuitively, this condition ensures that the sample autocovariances at different lags are asymptotically uncorrelated; that is, $\text{cov}[\sqrt{n}\hat{\gamma}(j), \sqrt{n}\hat{\gamma}(l)] \to 0$ for all $j \neq l$. As a result, the heteroskedasticity-consistent VR has the same asymptotic distribution as VR_o. However, the condition in the above equation rules out many important volatility processes, such as EGARCH and Threshold GARCH models. Moreover, the variance ratio test only exploits the implication of the MDS hypothesis on the spectral density at frequency 0; it does not check the spectral density at nonzero frequencies. As a result, it is not consistent against serial correlation of unknown form. See Durlauf (1991) for more discussion.

6.7 A Classical Ornut-Cochrane Procedure

Long-run variance estimators are necessary for statistical inference of the OLS estimation in a linear regression model when there exists serial correlation of *unknown form*. If serial correlation in the regression error has a known special pattern, then simpler statistical inference procedures are possible. One example is the classical Ornut-Cochrane

procedure. Consider a linear regression model with serially correlated errors:

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

where $E(\varepsilon_t|X_t) = 0$ but $\{\varepsilon_t\}$ follows an AR(p) process

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + v_t, \{v_t\} \sim \text{i.i.d.}(0, \sigma^2).$$

The OLS estimator $\hat{\beta}$ is consistent for β^o given $E(X_t\varepsilon_t) = 0$ but its asymptotic variance depends on serial correlation in $\{\varepsilon_t\}$. We can consider the following transformed linear regression model

$$\begin{aligned} Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j} &= \left(X_t - \sum_{j=1}^{p} \alpha_j X_{t-j}\right)' \beta^o \\ &\quad + \left(\varepsilon_t - \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}\right) \\ &= \left(X_t - \sum_{j=1}^{p} \alpha_j X_{t-j}\right)' \beta^o + v_t. \end{aligned}$$

We can write it as follows:

$$Y_t^* = X_t^{*\prime}\beta^o + v_t,$$

where

$$\begin{aligned} Y_t^* &= Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j}, \\ X_t^* &= X_t - \sum_{j=1}^{p} \alpha_j X_{t-j}. \end{aligned}$$

The OLS estimator $\tilde{\beta}$ of this transformed regression will be consistent for β^o and asymptotically normal:

$$\sqrt{n}(\tilde{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma_v^2 Q_{x^*x^*}^{-1}),$$

where $Q_{x^*x^*} = E(X_t^* X_t^{*\prime})$. Moreover it is asymptotically BLUE. However, the OLS estimator $\tilde{\beta}$ is infeasible, because (Y_t^*, X_t^*) is not available due to the unknown parameters $\{\alpha_j\}_{j=1}^p$. As a solution, one can use a feasible two-step procedure:

- Step 1: Regress

$$Y_t = X_t'\beta^o + \varepsilon_t, t = 1, ..., n,$$

Y_t on X_t, and obtain the estimated OLS residual $e_t = Y_t - X_t'\hat{\beta}$;

- Step 2: Regress an AR(p) model

$$e_t = \sum_{j=1}^{p} \alpha_j e_{t-j} + \tilde{v}_t, t = p+1, ..., n,$$

obtain the OLS estimators $\{\hat{\alpha}_j\}_{j=1}^{p}$;

- Step 3: Regress the transformed model

$$\hat{Y}_t^* = \hat{X}_t^{*\prime}\beta^o + v_t^*, t = p+1, ..., n,$$

where $\hat{Y}_t^*$ and $\hat{X}_t^*$ are defined in the same way as Y_t and X_t respectively, with $\{\hat{\alpha}_j\}_{j=1}^{p}$ replacing $\{\alpha_j\}_{j=1}^{p}$. The resulting OLS estimator is denoted as $\tilde{\beta}_a$.

It can be shown that the adaptive feasible OLS estimator $\tilde{\beta}_a$ has the same asymptotic properties as the infeasible OLS estimator $\tilde{\beta}$. In other words, the sampling error resulting from the first step estimation has no impact on the asymptotic properties of the OLS estimator in the second step. The asymptotic variance estimator of $\tilde{\beta}_a$ is given by

$$\hat{s}_v^2 \hat{Q}_{x^*x^*}^{-1},$$

where

$$\begin{aligned} \hat{s}_v^2 &= \frac{1}{n-K}\sum_{t=1}^{n} \hat{v}_t^{*2}, \\ \hat{Q}_{x^*x^*} &= \frac{1}{n}\sum_{t=1}^{n} \hat{X}_t^* \hat{X}_t^{*\prime}, \end{aligned}$$

with $\hat{v}_t = \hat{Y}_t^* - \hat{X}_t^{*\prime}\tilde{\beta}_a$. The t-test statistic which is asymptotically $N(0,1)$ and the $J \cdot F$-test statistic which is asymptotically χ_J^2 from the last stage regression are applicable when the sample size n is large.

The estimator $\tilde{\beta}_a$ is essentially the adaptive feasible GLS estimator described in Chapter 3, and it is asymptotically BLUE. This estimation method is therefore asymptotically more efficient than the robust test procedures developed in Section 6, but it is based on the assumption that the AR(p) process for the disturbance $\{\varepsilon_t\}$ is known. The robust test procedures in Section 6 are applicable when $\{\varepsilon_t\}$ has conditional heteroskedasticity and serial correlation of unknown form.

6.8 Empirical Applications

6.9 Conclusion

In this chapter, we have first discussed some motivating economic examples where a long-run variance estimator is needed. Then we discussed consistent estimation of a long-run variance-covariance matrix by a nonparametric kernel method. The asymptotic properties of the OLS estimator are investigated, which calls for the use of a new CLT because $\{X_t\varepsilon_t\}$ is not a MDS. Robust t- and Wald test statistics that are valid under conditional heteroskedasticity and autocorrelation of unknown form are then derived. When there exists serial correlation of unknown form, there is no need (and no way) to separate the cases of conditional homoskedasticity and conditional heteroskedasticity. Because robust t- and Wald tests have very poor finite sample performances even if $\{X_t\varepsilon_t\}$ is a MDS, it is desirable to first check whether we really need a long-run variance estimator. We provide such a test. Finally, some empirical applications are considered. We also introduce a classical estimation method called Ornut-Ochrance procedure when it is known that the regression disburbance follows an AR process with a known order.

Long-run variances have been also widely used in nonstationary time series econometrics such as in unit root and cointegration (e.g., Phillips 1987).

EXERCISES

6.1. Suppose Assumptions 6.1–6.3 and 6.5(i) hold. Show

$$\begin{aligned} \text{avar}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) &\equiv \lim_{n\to\infty} \text{var}\left(n^{-1/2}\sum_{t=1}^{n} X_t\varepsilon_t\right) \\ &= \sum_{j=-\infty}^{\infty} \Gamma(j). \end{aligned}$$

6.2. Suppose $\Gamma(j) = 0$ for all $j > p_0$, where p_0 is a fixed lag order. An example of this case is Example 2 in Section 6.1. In this case, the long-run variance $V = \sum_{j=-p_0}^{p_0} \Gamma(j)$ and we can estimate it by using the following estimator

$$\hat{V} = \sum_{j=-p_0}^{p_0} \hat{\Gamma}(j).$$

where $\hat{\Gamma}(j)$ is defined as in Section 6.1. Show that for each given j, $\hat{\Gamma}(j) \xrightarrow{p} \Gamma(j)$ as $n \to \infty$.

Given that p_0 is a fixed interger, an important implication that $\hat{\Gamma}(j) \xrightarrow{p} \Gamma(j)$ for each given j as $n \to \infty$ is that $\hat{V} \xrightarrow{p} V$ as $n \to \infty$.

6.3. Suppose $\{Y_t\}$ is a stationary time series process with the following spectral density function exists:

$$h(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \gamma(j)e^{-ij\omega}.$$

Show that

$$\text{var}\left(\sum_{j=1}^{p} Y_{t-j}\right) \to 2\pi h(0) \text{ as } p \to \infty.$$

6.4. Suppose $\{Y_t\}$ is a weakly stationary process with $\gamma(j) = \text{cov}(Y_t, Y_{t-j})$.

(a) Find an example of $\{Y_t\}$ such that $\sum_{j=1}^{\infty} \gamma(j) = 0$ but there exists at least one $j > 0$, such that $\gamma(j) \neq 0$.

(b) Can the variance ration test detect the time series process in part (a) with a high probability.

CHAPTER 7 INSTRUMENTAL VARIABLES REGRESSION

Abstract: In this chapter we first discuss possibilities that the condition $E(\varepsilon_t|X_t) = 0$ *a.s.* may fail, which will generally render inconsistent the OLS estimator for the true model parameters. We then introduce a consistent two-stage least squares (2SLS) estimator, investigating its statistical properties and providing intuitions for the nature of the 2SLS estimator. Hypothesis tests are constructed. We consider various test procedures corresponding to the cases for which the disturbance is an MDS with conditional homoskedasticity, an MDS with conditional heteroskedasticity, and a non-MDS process, respectively. The latter case will require consistent estimation of a long-run variance-covariance matrix. It is important to emphasize that the t-test and F-test obtained from the second stage regression estimation cannot be used even for large samples. Finally, we consider some empirical applications and conclude this chapter by presenting a brief summary of the comprehensive econometric theory for linear regression models developed in Chapters 2–7.

Key Words: Endogeneity, Instrumental variables, Hausman's test, 2SLS.

Motivation

In all previous chapters, we always assumed that $E(\varepsilon_t|X_t) = 0$ holds even when there exist conditional heteroskedasticity and autocorrelation.

Questions: When may the condition $E(\varepsilon_t|X_t) = 0$ fail? And, what will happen to the OLS estimator $\hat{\beta}$ if $E(\varepsilon_t|X_t) = 0$ fails?

There are at least three possibilities where $E(\varepsilon_t|X_t) = 0$ may fail. The first is model misspecification (e.g., functional form misspecification or omitted variables). The second is the existence of measurement errors in regressors (also called errors in variables). The third is the estimation of a subset of a simultaneous equation system. We will consider the last two possibilities in this chapter. For the first case (i.e., model misspecification), it may not be meaningful to discuss consistent estimation of the parameters in a misspecified regression model.

Some Motivating Examples

We first provide some examples in which $E(\varepsilon_t|X_t) \neq 0$.

Example 1 [Errors of Measurements or Errors in Variables]:

Often, economic data measure concepts that differ somewhat from those of economic theory. It is therefore important to take into account errors of measurements. This is

usually called errors in variables in econometrics. Consider a data generating process (DGP)

$$Y_t^* = \beta_0^o + \beta_1^o X_t^* + u_t, \tag{7.1}$$

where X_t^* is the income, Y_t^* is the consumption, and $\{u_t\}$ is i.i.d. $(0, \sigma_u^2)$ and is independent of $\{X_t^*\}$.

Suppose both X_t^* and Y_t^* are not observable. The observed variables X_t and Y_t contain measurement errors in the sense that

$$\begin{aligned} X_t &= X_t^* + v_t, & (7.2)\\ Y_t &= Y_t^* + w_t, & (7.3) \end{aligned}$$

where $\{v_t\}$ and $\{w_t\}$ are measurement errors independent of $\{X_t^*\}$ and $\{Y_t^*\}$, such that $\{v_t\} \sim i.i.d.\ (0, \sigma_v^2)$ and $\{w_t\} \sim i.i.d.\ (0, \sigma_w^2)$. We assume that the series $\{v_t\}, \{w_t\}$ and $\{u_t\}$ are all mutually independent of each other.

Because we only observe (X_t, Y_t), we are forced to estimate the following regression model

$$Y_t = \beta_0^o + \beta_1^o X_t + \varepsilon_t, \tag{7.4}$$

where ε_t is some unobservable disturbance.

Clearly, the disturbance ε_t is different from the original (true) disturbance u_t. Although the linear regression model is correctly specified, we no longer have $E(\varepsilon_t | X_t) = 0$ due to the existence of the measurement errors. This is explained below.

Question: If we use the OLS estimator $\hat{\beta}$ to estimate this model, is $\hat{\beta}$ consistent for β^o?

From the general regression analysis in Chapter 2, we have known that the key for the consistency of the OLS estimator $\hat{\beta}$ for β^o is to check if $E(X_t \varepsilon_t) = 0$. From Eqs. $(7.1) - (7.3)$, we have

$$\begin{aligned} Y_t &= Y_t^* + w_t \\ &= (\beta_0^o + \beta_1^o X_t^* + u_t) + w_t \\ X_t &= X_t^* + v_t. \end{aligned}$$

Therefore, from Eq. (7.4), we obtain

$$\begin{aligned} \varepsilon_t &= Y_t - \beta_0^o - \beta_1^o X_t \\ &= [\beta_0^o + \beta_1^o X_t^* + u_t + w_t] - \beta_0^o - \beta_1^o (X_t^* + v_t) \\ &= u_t + w_t - \beta_1^o v_t. \end{aligned}$$

The regression error ε_t contains the true disturbance u_t and a linear combination of measurement errors.

Now, the expectation

$$\begin{aligned} E(X_t\varepsilon_t) &= E[(X_t^* + v_t)\varepsilon_t] \\ &= E(X_t^*\varepsilon_t) + E(v_t\varepsilon_t) \\ &= 0 - \beta_1^o E(v_t^2) \\ &= -\beta_1^o\sigma_v^2 \\ &\neq 0. \end{aligned}$$

Consequently, by the WLLN, the OLS estimator

$$\begin{aligned} \hat{\beta} - \beta^o &= \hat{Q}_{xx}^{-1} n^{-1} \sum_{t=1}^{n} X_t\varepsilon_t \\ &\xrightarrow{p} Q_{xx}^{-1} E(X_t\varepsilon_t) \\ &= -\beta_1^o\sigma_v^2 Q_{xx}^{-1} \neq 0. \end{aligned}$$

In other words, $\hat{\beta}$ is not consistent for β^o due to the existence of the measurement errors in regressors $\{X_t\}$.

Question: What is the effect of the measurement errors $\{w_t\}$ in the dependent variable Y_t?

Example 2 [Errors of Measurements in Dependent Variable]: Now we consider a data generating process (DGP) given by

$$Y_t^* = \beta_0^o + \beta_1^o X_t^* + u_t,$$

where X_t^* is the income, Y_t^* is the consumption, and $\{u_t\}$ is i.i.d. $(0, \sigma_u^2)$ and is independent of $\{X_t^*\}$.

Suppose X_t^* is now observed, and Y_t^* is still not observable, such that

$$\begin{aligned} X_t &= X_t^*, \\ Y_t &= Y_t^* + w_t, \end{aligned}$$

where $\{w_t\}$ is i.i.d.$(0, \sigma_w^2)$ measurement errors independent of $\{X_t^*\}$ and $\{Y_t^*\}$. We assume that the two series $\{w_t\}$ and $\{u_t\}$ are mutually independent.

Because we only observe (X_t, Y_t), we are forced to estimate the following model

$$Y_t = \beta_0^o + \beta_1^o X_t + \varepsilon_t.$$

Question: If we use the OLS estimator $\hat{\beta}$ to estimate this model, is $\hat{\beta}$ consistent for β^o?

Answer: Yes! The measurement errors in Y_t do not cause any trouble for consistent estimation of β^o.

The measurement error in Y_t can be regarded as part of the true regression disturbance. It increases the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$, that is, the existence of measurement errors in Y_t renders the estimation of β^o less precise.

Example 3 [Errors in Expectations] Consider a linear regression model

$$Y_t = \beta_0 + \beta_1 X_t^* + \varepsilon_t,$$

where X_t^* is the economic agent's conditional expectation of X_t at time $t-1$, and $\{\varepsilon_t\}$ is an i.i.d.$(0, \sigma^2)$ sequence with $E(\varepsilon_t | X_t^*) = 0$. The conditional expectation X_t^* is a latent variable. When the economic agent follows rational expectations, then $X_t^* = E(X_t | I_{t-1})$ and we have

$$X_t = X_t^* + v_t,$$

where

$$E(v_t | I_{t-1}) = 0,$$

where I_{t-1} is the information available to the economic agent at time $t-1$. Assume that two error series $\{\varepsilon_t\}$ and $\{v_t\}$ are independent of each other.

We can consider the following regression model

$$Y_t = \beta_0^o + \beta_1^o X_t + u_t,$$

where the error term

$$u_t = \varepsilon_t - \beta_1^o v_t.$$

Since

$$\begin{aligned} E(X_t u_t) &= E[(X_t^* + v_t)(\varepsilon_t - \beta_1^o v_t)] \\ &= -\beta_1^o \sigma_v^2 \\ &\neq 0 \end{aligned}$$

provided $\beta_1^o \neq 0$, the OLS estimator is not consistent for β_1^o.

Example 4 [Endogeneity due to Omitted variables] Consider an earning data generating process

$$Y_t = X_t' \beta^o + \gamma A_t + u_t,$$

where Y_t is the earning, X_t is a vector consisting of working experience and schooling, and A_t is ability which is unobservable, and the disturbance u_t satisfies the condition

that $E(u_t|X_t, A_t) = 0$. Because one does not observe A_t, one is forced to consider the regression model

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

and is interested in knowing β^o, the marginal effect of schooling and working experience. However, we have $E(X_t\varepsilon_t) \neq 0$ because A_t is usually correlated with X_t.

Example 5 [Production-Bonus Causality; Groves, Hong, McMillan and Naughton 1994]: Consider a production function data generating process

$$\ln(Y_t) = \beta_0^o + \beta_1^o \ln(L_t) + \beta_2^o \ln(K_t) + \beta_3^o B_t + \varepsilon_t,$$

where Y_t, L_t, K_t are the output, labor and capital stock, B_t is the proportion of bonus out of total pay, and t is a time index. Without loss of generality, we assume that

$$\begin{aligned} E(\varepsilon_t) &= 0, \\ E[\ln(L_t)\varepsilon_t] &= 0, \\ E[\ln(K_t)\varepsilon_t] &= 0. \end{aligned}$$

Economic theory suggests that the use of bonus in addition to basic wage will provide a stronger incentive for workers to work harder in a transitional economy. This theory can be tested by checking if $\beta_3^o = 0$. However, the test procedure is complicated because there exists a possibility that when a firm is more productive, it will pay more bonus to workers regardless of the effort of its workers. In this case, the OLS estimator $\hat{\beta}_3$ cannot consistently estimate β_3^o and cannot be used to test the null hypothesis.

Why?

To reflect the fact that a more productive firm pays more bonus to its workers, we can assume a structural equation for bonus:

$$B_t = \alpha_0^0 + \alpha_1^0 \ln(Y_t) + w_t \tag{7.5}$$

where $\alpha_1^0 > 0$, and $\{w_t\}$ is an i.i.d. $(0, \sigma_w^2)$ sequence that is independent of $\{Y_t\}$. For simplicity, we assume that $\{w_t\}$ is independent of $\{\varepsilon_t\}$.

Put $X_t = [1, \ln(L_t), \ln(K_t), B_t]'$. Now, from Eq. (7.5) and then Eq. (7.4), we have

$$\begin{aligned} E(B_t\varepsilon_t) &= E[(\alpha_0^o + \alpha_1^o \ln(Y_t) + w_t)\varepsilon_t] \\ &= \alpha_1^o E[\ln(Y_t)\varepsilon_t] \\ &= \alpha_1^o \beta_3^o E(B_t\varepsilon_t) + \alpha_1^o E(\varepsilon_t^2). \end{aligned}$$

It follows that

$$E(B_t\varepsilon_t) = \frac{\alpha_1^o}{1-\alpha_1^o\beta_3^o}\sigma^2 \neq 0,$$

where $\sigma^2 = \text{var}(\varepsilon_t)$. Consequently, the OLS estimator $\hat{\beta}_3$ is inconsistent for β_3^o due to the existence of the causality from productivity $\ln(Y_t)$ to bonus B_t.

The bias of the OLS estimator for β_3^o in the above model is usually called the simultaneous equation bias because it arises from the fact that productivity function is but one of two relationships that hold simultaneously. It is a common phenomena in economics. It is the rule rather than the exception for economic relationships to be embedded in a simultaneous system of equations. We now consider two more examples with simultaneous equation bias.

Example 6 [Simultaneous Equation Bias] We consider the following simple model of national income determination:

$$\begin{aligned} C_t &= \beta_0^o + \beta_1^o I_t + \varepsilon_t, \quad (7.6) \\ I_t &= C_t + D_t, \quad (7.7) \end{aligned}$$

where I_t is the income, C_t is the consumption expenditure, and D_t is the non-consumption expenditure. The variables I_t and C_t are called endogenous variables, as they can be determined by the two-equation model. The variable D_t is called an exogenous variable, because it is determined outside the model (or the system considered). We assume that $\{D_t\}$ and $\{\varepsilon_t\}$ are mutually independent, and $\{\varepsilon_t\}$ is i.i.d. $(0, \sigma^2)$.

Question: If the OLS estimator $\hat{\beta}$ is applied to the first equation, is it consistent for β^o?

To answer this question, we have from Eq. (7.7)

$$\begin{aligned} E(I_t\varepsilon_t) &= E[(C_t + D_t)\varepsilon_t] \\ &= E(C_t\varepsilon_t) + E(D_t\varepsilon_t) \\ &= \beta_1^o E(I_t\varepsilon_t) + E(\varepsilon_t^2) + 0. \end{aligned}$$

It follows that

$$E(I_t\varepsilon_t) = \frac{1}{1-\beta_1^o}\sigma^2 \neq 0.$$

Thus, $\hat{\beta}$ is not consistent for β^o.

In fact, this bias problem can also be seen from the so-called reduced form model.

Question: What is the reduced form?

Solving for Eqs. (7.6) and (7.7) simultaneously, we can obtain the "reduced forms" that express endogenous variables in terms of exogenous variables and disturbances:

$$\begin{aligned} C_t &= \frac{\beta_0^o}{1-\beta_1^o} + \frac{\beta_1^o}{1-\beta_1^o} D_t + \frac{1}{1-\beta_1^o}\varepsilon_t, \\ I_t &= \frac{\beta_0^o}{1-\beta_1^o} + \frac{1}{1-\beta_1^o} D_t + \frac{1}{1-\beta_1^o}\varepsilon_t. \end{aligned}$$

Obviously, I_t is positively correlated with ε_t (i.e., $E(I_t\varepsilon_t) \neq 0$). Thus, the OLS estimator for the regression of C_t on I_t in Eq. (7.6) will not be consistent for β_1^o, the parameter for marginal propensity to consume. Generally speaking, the OLS estimator for the reduced form is consistent for new parameters, which are functions of original parameters.

Example 7 [Wage-Price Spiral Model] Consider the system of equations

$$W_t = \beta_0^o + \beta_1^o P_t + \beta_2^o D_t + \varepsilon_t, \tag{7.8}$$

$$P_t = \alpha_0^o + \alpha_1^o W_t + v_t, \tag{7.9}$$

where W_t, P_t, D_t are the wage, price, and excess demand in the labor market respectively. Eq. (7.8) describes the mechanism of how wage is determined. In particular, wage depends on price and excess demand for labor. Eq. (7.9) describes how price depends on wage (or income).

Suppose D_t is an exogenous variable, with $E(\varepsilon_t|D_t) = 0$. There are two endogenous variables, W_t and P_t, in the system of equations (7.8) and (7.9).

Question: Will W_t be correlated with v_t? And, will P_t be correlated with ε_t?

To answer these questions, we first obtain the reduced form equations:

$$\begin{aligned} W_t &= \frac{\beta_0^o + \beta_1^o\alpha_0}{1-\beta_1^o\alpha_1^o} + \frac{\beta_1^o}{1-\beta_1^o\alpha_1^o} D_t + \frac{\varepsilon_t + \beta_1^o v_t}{1-\beta_1^o\alpha_1^o}, \\ P_t &= \frac{\alpha_0^o}{1-\beta_1^o\alpha_1^o} + \frac{\alpha_1^o\beta_2^o}{1-\beta_1^o\alpha_1^o} D_t + \frac{\alpha_1^o\varepsilon_t + v_t}{1-\beta_1^o\alpha_1^o}. \end{aligned}$$

Conditional on the exogenous variable D_t, both W_t and P_t are correlated with ε_t and v_t. As a consequence, both the OLS estimator for β_1^o in Eq. (7.8) and the OLS estimator for α_1^o in Eq. (7.9) will be inconsistent.

In this chapter, we will consider a method called two-stage least squares estimation to obtain consistent estimators for the unknown parameters in all above examples except for the parameter β_2^o in Eq. (7.8) of Example 7. No methods can deliver a consistent estimator for β_2^o in Eq. (7.8) because it is not identifiable. This is the so-called identification problem of the simultaneous equations.

A Digression: Identification Problem in Simultaneous Equation Models

To see why there is no way to obtain a consistent estimator for β_2^o in Eq. (7.8), from Eq. (7.9), we can write

$$W_t = -\frac{\alpha_1^o}{\alpha_2^o} + \frac{1}{\alpha_2^o} P_t - \frac{v_t}{\alpha_2^o}. \tag{7.10}$$

Let a and b be two arbitrary constants. We multiply Eq. (7.8) with a, and multiply Eq. (7.10) with b, and add them together:

$$(a+b)W_t = a\beta_1^o - \frac{b\alpha_1}{\alpha_2} + (a\beta_2^o + \frac{b}{\alpha_2})P_t + a\beta_3^o D_t + (a\varepsilon_t - \frac{b}{\alpha_2} v_t),$$

or

$$W_t = \left[\frac{a\beta_1^o}{a+b} - \frac{b\alpha_1^o}{(a+b)\alpha_2^o}\right] + \frac{1}{a+b}(a\beta_2^o + \frac{b}{\alpha_2^o})P_t + \frac{a\beta_3^o}{a+b} D_t + \frac{1}{a+b}(a\varepsilon_t - \frac{b}{\alpha_2^o} v_t). \tag{7.11}$$

This new equation, (7.11), is a combination of the original wage equation (7.8) and the price equation (7.9). It is of the same statistical form as Eq. (7.8). Since a and b are arbitrary, there is an infinite number of parameters that can satisfy Eq. (11) and they are all indistinguishable from Eq. (7.8). Consequently, if we use OLS to run regression of W_t on P_t and D_t, or more generally, use any other method to estimate the equation (7.8) or (7.11), there is no way to know which model, either Eq. (7.8) or Eq. (7.11), is being estimated. Therefore, there is no way to estimate β_2^o. This is the so-called identification problem with simultaneous equation models. To avoid such identification problems in simultaneous equations, certain conditions are required to make the system of simultaneous equations identifiable. For example, if an extra variable, say money supply growth rate, is added in the price equation in (7.9), we then obtain

$$P_t = \alpha_0^o + \alpha_1^o W_t + \alpha_2^o M_t + v_t, \tag{7.12}$$

then the system of equations (7.8) and (7.12) is identifiable provided $\alpha_2^o \neq 0$, and so the parameters in Eqs. (7.8) and (7.12) can be consistently estimated. [Question: Check why the system of equations (7.8) and (7.12) is identifiable.]

We note that for the system of equations (7.8) and (7.9), although Eq. (7.8) cannot be consistently estimated by any method, Eq. (7.9) can still be consistently estimated using the method proposed below. For an identifiable system of simultaneous equations with simultaneous equation bias, we can use various methods to estimate them consistently, including 2SLS, the generalized method of moments and the maximum likelihood or quasi-maximum likelihood estimation methods. These methods will be introduced below and in subsequent chapters.

7.1 Framework and Assumptions

We now provide a set of regularity conditions for our formal analysis in this chapter.

Assumption 7.1 [Ergodic Stationarity]: $\{Y_t, X_t', Z_t'\}_{t=1}'^n$ is an ergodic stationary stochastic process, where X_t is a $K \times 1$ vector, Z_t is a $l \times 1$ vector, and $l \geq K$.

Assumption 7.2 [Linearity]:

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

for some unknown parameter β^o and some unobservable disturbance ε_t;

Assumption 7.3 [Nonsingularity]: The $K \times K$ matrix

$$Q_{xx} = E(X_t X_t')$$

is nonsingular and finite;

Assumption 7.4 [IV Conditions]:

(i) $E(X_t\varepsilon_t) \neq 0$;

(ii) $E(Z_t\varepsilon_t) = 0$;

(iii) The $l \times l$ matrix

$$Q_{zz} = E(Z_t Z_t')$$

is finite and nonsingular, and the $l \times K$ matrix

$$Q_{zx} = E(Z_t X_t')$$

is finite and of full rank.

Assumption 7.5 [CLT]: $n^{-1/2}\sum_{t=1}^n Z_t\varepsilon_t \xrightarrow{d} N(0, V)$ for some $K \times K$ symmetric matrix $V \equiv \text{avar}(n^{-1/2}\sum_{t=1}^n Z_t\varepsilon_t)$ finite and nonsingular.

Remarks:

Assumption 7.1 allows for i.i.d. and stationary time series observations.

Assumption 7.5 directly assumes that the CLT holds. This is often called a "high level assumption." It covers three cases: IID, MDS and non-MDS for $\{X_t\varepsilon_t\}$, respectively. For an IID or MDS sequence $\{Z_t\varepsilon_t\}$, we have $V = \text{var}(Z_t\varepsilon_t) = E(Z_t Z_t'\varepsilon_t^2)$. For a non-MDS process $\{Z_t\varepsilon_t\}$, $V = \Sigma_{j=-\infty}^{\infty}\text{cov}(Z_t\varepsilon_t, Z_{t-j}\varepsilon_{t-j})$ is a long-run variance-covariance matrix.

The random vector Z_t that satisfies Assumption 7.4 is called instruments. The condition that $l \geq K$ in Assumption 7.1 implies that the number of instruments Z_t is larger than or at least equal to the number of regressors X_t.

Question: Why is the condition of $l \geq K$ required?

Question: How to choose instruments Z_t in practice?

First of all, one should analyze which explanatory variables in X_t are endogenous or exogenous. If an explanatory variable is exogenous, then this variable should be included in Z_t, the set of instruments. For example, the constant term should always be included, because a constant is uncorrelated with any random variables. All other exogenous variables in X_t should also be included in the set of Z_t. If k_0 of K regressors are endogenous, one should find at least k_0 additional instruments.

Most importantly, we should choose an instrument vector Z_t which is closely related to X_t as much as possible. As we will see below, the strength of the correlation between Z_t and X_t affects the magnitude of the asymptotic variance of the 2SLS estimator for β_0 which we will propose, although it does not affect the consistency provided the correlation between Z_t and X_t is not zero.

In time series regression models, it is often reasonable to assume that lagged variables of X_t are not correlated with ε_t. Therefore, we can use lagged values of X_t, for example, X_{t-1}, as an instrument. This instrument is expected to be highly correlated with X_t if $\{X_t\}$ is a time series process. In light of this, we can choose the set of instruments $Z_t = (1, \ln L_t, \ln K_t, B_{t-1})'$ in estimating Eq.(7.4) in Example 5, choose $Z_t = (1, D_t, I_{t-1})'$ in estimating Eq.(7.6) in Example 6, choose $Z_t = (1, D_t, P_{t-1})'$ in estimating Eq.(7.8) in Example 7. For examples with measurement errors or expectational errors, where $E(X_t\varepsilon_t) \neq 0$ due to the presence of measurement errors or expectational errors, we can choose $Z_t = X_{t-1}$ if the measurement errors or expectational errors in X_t are serially uncorrelated (check this!). The expectational errors in X_t are MDS and so are uncorrelated in Example 3 when the economic agent has rational expectations.

7.2 Two-Stage Least Squares (2SLS) Estimation

Question: Because $E(\varepsilon_t|X_t) \neq 0$, the OLS estimator $\hat{\beta}$ is not consistent for β^o. How to obtain consistent estimators for β^o in situations similar to the examples described in Section 7.1?

We now introduce a two-stage least squares (2SLS) procedure, which can consistently estimate the true parameter β^o. The 2SLS procedure can be described as follows:

Stage 1: Regress X_t on Z_t via OLS and save the predicted value $\hat{X}_t$.

Here, the artificial linear regression model is

$$X_t = \gamma' Z_t + v_t, \qquad t = 1, ..., n,$$

where γ is a $l \times K$ parameter matrix, and v_t is a $K \times 1$ regression error. From the result in Chapter 2, we have $E(Z_t v_t) = 0$ if and only if γ is the best LS approximation coefficient, i.e., if and only if

$$\gamma = [E(Z_t Z_t')]^{-1} E(Z_t X_t').$$

In matrix form, we can write

$$\mathbf{X} = \mathbf{Z}\gamma + v,$$

where $\mathbf{X}$ is a $n \times K$ matrix, $\mathbf{Z}$ is a $n \times l$ matrix, γ is a $l \times K$ matrix, and v is a $n \times K$ matrix.

The OLS estimator for γ is

$$\begin{aligned} \hat{\gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \\ &= \left(n^{-1} \sum_{t=1}^{n} Z_t Z_t' \right)^{-1} n^{-1} \sum_{t=1}^{n} Z_t X_t'. \end{aligned}$$

The predicted value or the sample projection of X_t on Z_t is

$$\hat{X}_t = \hat{\gamma}' Z_t$$

or in matrix form

$$\hat{\mathbf{X}} = \mathbf{Z}\hat{\gamma} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

Stage 2: Use the predicted value $\hat{X}_t$ as regressors for Y_t. Regress Y_t on $\hat{X}_t$, and the resulting OLS estimator is called the 2SLS estimator, denoted as $\hat{\beta}_{2sls}$.

Question: Why use the fitted value $\hat{X}_t = \hat{\gamma}' Z_t$ as regressors?

We first consider

$$X_t = \gamma' Z_t + v_t,$$

where γ is the best linear LS approximation coefficient, and so v_t is orthogonal to Z_t in the sense $E(Z_t v_t') = 0$. Because $E(Z_t \varepsilon_t) = 0$, the population projection $\gamma' Z_t$ is orthogonal to ε. In general, $v_t = X_t - \gamma' Z_t$, which is orthogonal to Z_t, is correlated with ε_t. In other words, the auxiliary regression in stage 1 decomposes X_t into two components: $\gamma' Z_t$ and v_t, where $\gamma' Z_t$ is orthogonal to ε_t, and v_t is correlated with ε_t.

Since the best linear LS approximation coefficient γ is unknown, we have to replace it with $\hat{\gamma}$. The fitted value $\hat{X}_t = \hat{\gamma}' Z_t$ is the (sample) projection X_t onto Z_t. The regression of X_t on Z_t purges the component of X_t that is correlated with ε_t so that the projection $\hat{X}_t$ is approximately orthogonal to ε_t given that Z_t is orthogonal to ε_t. (The word "approximately" is used here because $\hat{\gamma}$ is an estimator of γ and thus contains some estimation error.)

The regression model in the second stage can be written as

$$Y_t = \hat{X}_t' \beta^o + \tilde{\varepsilon}_t$$

or in matrix form

$$Y = \hat{\mathbf{X}} \beta^o + \tilde{\varepsilon}.$$

Note that the disturbance $\tilde{\varepsilon}_t$ is not ε_t because $\hat{X}_t$ is not X_t.

Using $\hat{\mathbf{X}} = \mathbf{Z}\hat{\gamma} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$, we can write the second stage OLS estimator, namely the 2SLS estimator as follows:

$$\begin{aligned}
\hat{\beta}_{2sls} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'Y \\
&= [(\mathbf{Z}\hat{\gamma})'(\mathbf{Z}\hat{\gamma})]^{-1}(\mathbf{Z}\hat{\gamma})'Y \\
&= \left\{[\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]'[\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]\right\}^{-1}[\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]'Y \\
&= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y \\
&= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y \\
&= \left[\frac{\mathbf{X}'\mathbf{Z}}{n}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\frac{\mathbf{Z}'\mathbf{X}}{n}\right]^{-1}\frac{\mathbf{X}'\mathbf{Z}}{n}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\frac{\mathbf{Z}'Y}{n}.
\end{aligned}$$

Using the expression $Y = \mathbf{X}\beta^o + \varepsilon$ from Assumption 7.2, we have

$$\begin{aligned}
\hat{\beta}_{2sls} - \beta^o &= \left[\frac{\mathbf{X}'\mathbf{Z}}{n}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\frac{\mathbf{Z}'\mathbf{X}}{n}\right]^{-1}\frac{\mathbf{X}'\mathbf{Z}}{n}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\frac{\mathbf{Z}'\varepsilon}{n} \\
&= \left[\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right]^{-1}\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\frac{Z'\varepsilon}{n},
\end{aligned}$$

where

$$\begin{aligned}
\hat{Q}_{zz} &= \frac{\mathbf{Z}'\mathbf{Z}}{n} = n^{-1}\sum_{t=1}^{n} Z_t Z_t', \\
\hat{Q}_{xz} &= \frac{\mathbf{X}'\mathbf{Z}}{n} = n^{-1}\sum_{t=1}^{n} X_t Z_t', \\
\hat{Q}_{zx} &= \frac{\mathbf{Z}'\mathbf{X}}{n} = n^{-1}\sum_{t=1}^{n} Z_t X_t' = \hat{Q}_{xz}'.
\end{aligned}$$

Question: What are the statistical properties of $\hat{\beta}_{2sls}$?

7.3 Consistency of 2SLS

By the WLLN for a stationary ergodic process, we have

$$\hat{Q}_{zz} \xrightarrow{p} Q_{zz}, \quad l \times l$$
$$\hat{Q}_{xz} \xrightarrow{p} Q_{xz}, \quad K \times l,$$
$$\frac{Z'\varepsilon}{n} \xrightarrow{p} E(Z_t\varepsilon_t) = 0, \quad l \times 1.$$

Also, $Q_{xz}Q_{zz}^{-1}Q_{zx}$ is a $K \times K$ symmetric and nonsingular matrix because Q_{xz} is of full rank, Q_{zz} is nonsingular, and $l \geq K$. It follows from continuity that

$$\left[\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right]^{-1} \xrightarrow{p} [Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}.$$

Consequently, we have

$$\hat{\beta}_{2sls} - \beta^o \xrightarrow{p} [Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}Q_{xz}Q_{zz}^{-1} \cdot 0 = 0.$$

We now state this consistency result in the following theorem.

Theorem 7.1 [Consistency of 2SLS]: *Under Assumptions 7.1-7.4, as* $n \to \infty$,

$$\hat{\beta}_{2sls} \xrightarrow{p} \beta^o.$$

To provide intuition why the 2SLS estimator $\hat{\beta}_{2sls}$ is consistent for β^o, we consider

$$Y_t = X_t'\beta^o + \varepsilon_t.$$

The OLS estimator $\hat{\beta}$ is not consistent for β^o because $E(X_t\varepsilon_t) \neq 0$. Suppose we decompose the regressor X_t into two terms:

$$X_t = \tilde{X}_t + v_t,$$

where one $\tilde{X}_t = \gamma' Z_t$ is a projection of X_t on Z_t and so it is orthogonal to ε_t. The other component, $v_t = X_t - \tilde{X}_t$, is generally correlated with ε_t. Then consistent estimation for β^o is possible if we observe v_t and run the following augmented regression

$$\begin{aligned} Y_t &= X_t'\beta^o + \varepsilon_t \\ &= \tilde{X}_t'\beta^o + (v_t'\beta^o + \varepsilon_t) \\ &= \tilde{X}_t'\beta^o + u_t, \end{aligned}$$

where $u_t = v_t'\beta^o + \varepsilon_t$ is the disturbance when regressing Y_t on $\tilde{X}_t$. Because

$$\begin{aligned} E(\tilde{X}_t u_t) &= \gamma' E(Z_t u_t) \\ &= \gamma' E(Z_t v_t')\beta^o + \gamma' E(Z_t \varepsilon_t) \\ &= 0, \end{aligned}$$

the OLS estimator of regressing Y_t on $\tilde{X}_t$ would be consistent for β^o.

However, $\tilde{X}_t = \gamma' Z_t$ is not observable, so we need to use a proxy, i.e., $\hat{X}_t = \hat{\gamma}' Z_t$, where $\hat{\gamma}$ is the OLS estimator of regressing X_t on Z_t. This results in the 2SLS estimator $\hat{\beta}_{2sls}$. The estimation error of $\hat{\gamma}$ does not affect the consistency of the 2SLS estimator $\hat{\beta}$.

7.4 Asymptotic Normality of 2SLS

We now derive the asymptotic distribution of $\hat{\beta}_{2sls}$. Write

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{2sls} - \beta^o) &= \left[\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right]^{-1} \hat{Q}_{xz}\hat{Q}_{zz}^{-1}\frac{\mathbf{Z}'\varepsilon}{\sqrt{n}}. \\ &= \hat{A} \cdot \frac{\mathbf{Z}'\varepsilon}{\sqrt{n}}, \end{aligned}$$

where the $K \times l$ matrix

$$\hat{A} = \left[\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right]^{-1} \hat{Q}_{xz}\hat{Q}_{zz}^{-1}.$$

By the CLT assumption (Assumption 7.5), we have

$$\frac{\mathbf{Z}'\varepsilon}{\sqrt{n}} = n^{-\frac{1}{2}} \sum_{t=1}^{n} Z_t \varepsilon_t \xrightarrow{d} N(0, V) \sim G,$$

where V is a finite and nonsingular $l \times l$ matrix, and we denote the random vector $G \sim N(0, V)$. Then by the Slutsky theorem, we have

$$\begin{aligned} & \sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} \left(Q_{xz}Q_{zz}^{-1}Q_{zx}\right)^{-1} Q_{xz}Q_{zz}^{-1} \cdot N(0, V) \\ \sim\ & N(0, AVA') \\ \sim\ & N(0, \Omega), \end{aligned}$$

where $A = (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}$. The asymptotic variance of $\sqrt{n}(\hat{\beta}_{2sls} - \beta^o)$

$$\begin{aligned} \text{avar}(\sqrt{n}\hat{\beta}_{2sls}) &= \Omega \\ &= AVA' \\ &= \left\{[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}Q_{xz}Q_{zz}^{-1}\right\} V \left\{[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}Q_{xz}Q_{zz}^{-1}\right\}' \\ &= [Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}Q_{xz}Q_{zz}^{-1}VQ_{zz}^{-1}Q_{zx}[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}. \end{aligned}$$

Theorem 7.2 [Asymptotic Normality of 2SLS]: *Under Assumptions 7.1-7.5, as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, \Omega).$$

The estimation of V depends on whether $\{Z_t\varepsilon_t\}$ is an MDS. We first consider the case where $\{Z_t\varepsilon_t\}$ is an MDS process. In this case, $V = E(Z_tZ_t'\varepsilon_t^2)$ and so we need not estimate the long-run variance-covariance matrix.

Case I: $\{Z_t\varepsilon_t\}$ is a Stationary Ergodic MDS

Assumption 7.6 [MDS]: *(i)* $\{Z_t\varepsilon_t\}$ *is an MDS; (ii)* $\text{var}(Z_t\varepsilon_t) = E(Z_tZ_t'\varepsilon_t^2)$ *is finite and nonsingular.*

Corollary 7.3: *Under Assumptions 7.1–7.4 and 7.6, we have as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, \Omega),$$

where Ω *is defined as above with* $V = E(Z_tZ_t'\varepsilon_t^2)$.

There is no need to estimate a long-run variance-covariance matrix but Ω involves consistent estimation of the heteroskedasticity-consistent variance-covariance matrix V.

When $\{Z_t\varepsilon_t\}$ is an MDS with conditional homoskedasticity, the asymptotic variance Ω can be greatly simplified.

Special Case: Conditional Homoskedasticity

Assumption 7.7 [Conditional Homoskedasticity]: $E(\varepsilon_t^2|Z_t) = \sigma^2$ a.s.

Note that the conditional expectation in Assumption 7.7 is conditional on Z_t, not on X_t.

Under this assumption, by the law of iterated expectations, we obtain

$$\begin{aligned} V &= E(Z_tZ_t'\varepsilon_t^2) \\ &= E[Z_tZ_t'E(\varepsilon_t^2|Z_t)] \\ &= \sigma^2 E(Z_tZ_t') \\ &= \sigma^2 Q_{zz}. \end{aligned}$$

It follows that

$$\begin{aligned} \Omega &= (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}\sigma^2 Q_{zz}Q_{zz}^{-1}Q_{zx}(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1} \\ &= \sigma^2(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}. \end{aligned}$$

Corollary 7.4 [Asymptotic Normality of 2SLS under MDS with Conditional Homoskedasticity] *Under Assumptions 7.1–7.4,7.6 and 7.7, we have as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \sigma^2 [Q_{xz} Q_{zz}^{-1} Q_{zx}]^{-1}.$$

Case II: $\{Z_t \varepsilon_t\}$ is a Stationary Ergodic non-MDS

In this general case, we have

$$V \equiv \text{avar}\left(n^{-1/2} \sum_{t=1}^{n} Z_t \varepsilon_t \right) = \sum_{j=-\infty}^{\infty} \Gamma(j)$$

where $\Gamma(j) = \text{cov}(Z_t \varepsilon_t, Z_{t-j} \varepsilon_{t-j})$. We need to use a long-run variance-covariance matrix estimator for V. When $\{Z_t \varepsilon_t\}$ is not an MDS, there is no need (and in fact there is no way) to consider conditional homoskedasticity and conditional heteroskedasticity separately.

7.5 Interpretation and Estimation of the 2SLS Asymptotic Variance

The asymptotic variance Ω of $\hat{\beta}_{2sls}$ is so complicated that it will be highly desirable if we can find an interpretation to help understand its structure. What is the nature of $\hat{\beta}_{2sls}$? What is Ω?

Let us revisit the second stage regression model

$$Y_t = \hat{X}_t' \beta^o + \tilde{\varepsilon}_t,$$

where the regressor

$$\hat{X}_t = \hat{\gamma}' Z_t$$

is the sample projection of X_t on Z_t, and the disturbance $\tilde{\varepsilon}_t = Y_t - \hat{X}_t' \beta^o$. Note that $\tilde{\varepsilon}_t \neq \varepsilon_t$ because $\hat{X}_t \neq X_t$. Given $Y_t = X_t' \beta^o + \varepsilon_t$ from Assumption 7.2, we have

$$\begin{aligned} \tilde{\varepsilon}_t &= Y_t - \hat{X}_t' \beta^o \\ &= \varepsilon_t + (X_t - \hat{X}_t)' \beta^o \\ &= \varepsilon_t + \hat{v}_t' \beta^o, \end{aligned}$$

where ε_t is the true disturbance and $\hat{v}_t \equiv X_t - \hat{X}_t = X_t - \hat{\gamma}' Z_t$. Since $\hat{v}_t$ is the estimated residual from the first OLS regression

$$\mathbf{X} = \mathbf{Z}\gamma + v,$$

we have the following FOC holds:

$$\mathbf{Z}'(\mathbf{X} - \hat{\mathbf{X}}) = \mathbf{Z}'\hat{v} = 0.$$

It follows that the 2SLS estimator

$$\begin{aligned}
\hat{\beta}_{2sls} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'Y \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\hat{\mathbf{X}}\beta^o + \tilde{\varepsilon}) \\
&= \beta^o + (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'[\varepsilon + \hat{v}\beta^o] \\
&= \beta^o + (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\varepsilon
\end{aligned}$$

because $\hat{\mathbf{X}}'\hat{v} = 0$ (why?). Therefore, the asymptotic properties of $\hat{\beta}_{2sls}$ are determined by

$$\begin{aligned}
\hat{\beta}_{2sls} - \beta^o &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\varepsilon \\
&= \left(\frac{\hat{\mathbf{X}}'\hat{\mathbf{X}}}{n}\right)^{-1}\frac{\hat{\mathbf{X}}'\varepsilon}{n}.
\end{aligned}$$

In other words, the estimated residual $\hat{v} = \mathbf{X} - \hat{\mathbf{X}}$ from the first stage regression has no impact on the statistical properties of $\hat{\beta}_{2sls}$, although it is a component of $\tilde{\varepsilon}_t$. Thus, when analyzing the asymptotic properties of $\hat{\beta}_{2sls}$, we can proceed as if we were estimating $Y = \hat{\mathbf{X}}\beta^o + \varepsilon$ by OLS.

Next, recall that we have

$$\begin{aligned}
\hat{\mathbf{X}} &= \mathbf{Z}\hat{\gamma}, \\
\hat{\gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \\
\xrightarrow{p} Q_{zz}^{-1}Q_{zx} &= \gamma
\end{aligned}$$

By the WLLN, the sample projection $\hat{X}_t$ "converges" to the population projection $\tilde{X}_t \equiv \gamma' Z_t$ as $n \to \infty$. That is, $\hat{X}_t$ will become arbitrarily close to $\tilde{X}_t$ as $n \to \infty$. In fact, the estimation error of $\hat{\gamma}$ in the first stage has no impact on the asymptotic properties of $\hat{\beta}_{2sls}$.

Thus, we can consider the following artificial regression model

$$Y_t = \tilde{X}_t'\beta^o + \varepsilon_t,$$

whose infeasible OLS estimator

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y.$$

As we will show below, the asymptotic properties of $\hat{\beta}_{2sls}$ are the same as those of the infeasible OLS estimator $\tilde{\beta}$. This helps a lot in understanding the variance-covariance structure of $\hat{\beta}_{2sls}$. It is important to emphasize that the equation in (7.13) is not derived from other equations. It is just a convenient way to understand the nature of $\hat{\beta}_{2sls}$.

We now show that the asymptotic properties of $\hat{\beta}_{2sls}$ are the same as the asymptotic properties of $\tilde{\beta}$. For the asymptotic normality, observe that

$$\begin{aligned}\sqrt{n}(\tilde{\beta}-\beta^o) &= \hat{Q}_{\tilde{x}\tilde{x}}^{-1}\frac{\tilde{X}'\varepsilon}{\sqrt{n}} \\ &\xrightarrow{d} Q_{\tilde{x}\tilde{x}}^{-1}\cdot N(0,\tilde{V}) \\ &\sim N(0,Q_{\tilde{x}\tilde{x}}^{-1}\tilde{V}Q_{\tilde{x}\tilde{x}}^{-1})\end{aligned}$$

using the asymptotic theory in Chapters 5 and 6, where

$$\begin{aligned}Q_{\tilde{x}\tilde{x}} &\equiv E(\tilde{X}_t\tilde{X}_t'), \\ \tilde{V} &\equiv \text{avar}\left(n^{-1/2}\sum_{t=1}^{n}\tilde{X}_t\varepsilon_t\right).\end{aligned}$$

We first consider the case where $\{Z_t\varepsilon_t\}$ is MDS with conditional homoskedasticity.

Case I: MDS with Conditional Homoskedasticity

Suppose $\{\tilde{X}_t\varepsilon_t\}$ is MDS, and $E(\varepsilon_t^2|\tilde{X}_t)=\sigma^2$ a.s. Then we have

$$\begin{aligned}\tilde{V} &= E(\tilde{X}_t\tilde{X}_t'\varepsilon_t^2) \\ &= \sigma^2 Q_{\tilde{x}\tilde{x}}\end{aligned}$$

by the law of iterated expectations (LIE). It follows that

$$\sqrt{n}(\tilde{\beta}-\beta^o)\xrightarrow{d}N(0,\sigma^2Q_{\tilde{x}\tilde{x}}^{-1}).$$

Because $\tilde{X}_t=\gamma'Z_t$, $\gamma=Q_{zz}^{-1}Q_{zx}$, we have

$$\begin{aligned}Q_{\tilde{x}\tilde{x}} &= E(\tilde{X}_t\tilde{X}_t') \\ &= \gamma'E(Z_tZ_t')\gamma \\ &= \gamma'Q_{zz}\gamma \\ &= Q_{xz}Q_{zz}^{-1}Q_{zz}Q_{zz}^{-1}Q_{zx} \\ &= Q_{xz}Q_{zz}^{-1}Q_{zx}.\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma^2 Q_{\tilde{x}\tilde{x}}^{-1} &= \sigma^2 \left(Q_{xz} Q_{zz}^{-1} Q_{zx}\right)^{-1} \\ &= \Omega \equiv \text{avar}(\sqrt{n}\hat{\beta}_{2sls}).\end{aligned}$$

This implies that the asymptotic distribution of $\tilde{\beta}$ is indeed the same as the asymptotic distribution of $\hat{\beta}_{2sls}$ under the MDS with conditional homoskedasticity.

The asymptotic variance formula

$$\text{avar}(\sqrt{n}\hat{\beta}_{2sls}) = \sigma^2 Q_{\tilde{x}\tilde{x}}^{-1} = \sigma^2(\gamma' Q_{zz}\gamma)^{-1}$$

indicates that the asymptotic variance of $\sqrt{n}\hat{\beta}_{2sls}$ will be large if the correlation between Z_t and X_t, as measured by γ, is weak. Thus, more precise estimation of β^o will be obtained if one chooses the instrument vector Z_t such that Z_t is highly correlated with X_t.

Question: How to estimate Ω under the MDS disturbances with conditional homoskedasticity?

Consider the asymptotic variance estimator

$$\begin{aligned}\hat{\Omega} &= \hat{s}^2 \hat{Q}_{\hat{x}\hat{x}}^{-1} \\ &= \hat{s}^2 \left(\hat{Q}_{xz} \hat{Q}_{zz}^{-1} \hat{Q}_{zx}\right)^{-1}\end{aligned}$$

where $\hat{s}^2 = \hat{e}'\hat{e}/(n-K)$, $\hat{e} = Y - \mathbf{X}\hat{\beta}_{2sls}$,

$$\hat{Q}_{\hat{x}\hat{x}} = n^{-1}\sum_{t=1}^{n} \hat{X}_t \hat{X}_t'$$

and $\hat{X}_t = \hat{\gamma}' Z_t$ is the sample projection of X_t on Z_t. Note that we have to use $\hat{X}_t$ rather than $\tilde{X}_t$ because $\tilde{X}_t = \gamma' Z_t$ is unknown.

It should be emphasized that $\hat{e}$ is not the estimated residual from the second stage regression (i.e., not from the regression of Y on $\hat{X}$). This implies that even under conditional homoskedasticity, the conventional t-statistic in the second stage regression does not converge to $N(0,1)$ in distribution, and $J \cdot \hat{F}$ does not converge to χ_J^2 where $\hat{F}$ is the F-statistic in the second stage regression.

To show $\hat{\Omega} \xrightarrow{p} \Omega$, we shall show (i) $\hat{Q}_{\hat{x}\hat{x}}^{-1} \xrightarrow{p} Q_{\tilde{x}\tilde{x}}^{-1}$ and (ii) $\hat{s}^2 \xrightarrow{p} \sigma^2$.

We first show (i). There are two methods for proving this.

Method 1: We shall show $\hat{Q}_{\hat{x}\hat{x}}^{-1} \xrightarrow{p} Q_{\tilde{x}\tilde{x}}^{-1}$. Because $\hat{X}_t = \hat{\gamma}' Z_t$ and $\hat{\gamma} \xrightarrow{p} \gamma$, we have

$$\begin{aligned}
\hat{Q}_{\hat{x}\hat{x}} &= n^{-1} \sum_{t=1}^{n} \hat{X}_t \hat{X}_t' \\
&= \hat{\gamma}' \left(n^{-1} \sum_{t=1}^{n} Z_t Z_t' \right) \hat{\gamma} \\
&= \hat{\gamma}' \hat{Q}_{zz} \hat{\gamma} \\
&\xrightarrow{p} \gamma' Q_{zz} \gamma \\
&= E[(\gamma' Z_t)(Z_t' \gamma)] \\
&= E(\tilde{X}_t \tilde{X}_t') \\
&= Q_{\tilde{x}\tilde{x}}.
\end{aligned}$$

Method 2: We shall show $(\hat{Q}_{xz} \hat{Q}_{zz}^{-1} \hat{Q}_{zx})^{-1} \xrightarrow{p} (Q_{xz} Q_{zz}^{-1} Q_{zx})^{-1}$, which follows immediately from $\hat{Q}_{xz} \xrightarrow{p} Q_{xz}$ and $\hat{Q}_{zz} \xrightarrow{p} Q_{zz}$ by the WLLN. This method is more straightforward but is less intuitive than the first method.

Next, we shall show (ii) $\hat{s}^2 \xrightarrow{p} \sigma^2$. We decompose

$$\begin{aligned}
\hat{s}^2 &= \frac{\hat{e}'\hat{e}}{n-K} \\
&= \frac{1}{n-K} \sum_{t=1}^{n} (Y_t - X_t' \hat{\beta}_{2sls})^2 \\
&= \frac{1}{n-K} \sum_{t=1}^{n} [\varepsilon_t - X_t' (\hat{\beta}_{2sls} - \beta^o)]^2 \\
&= \frac{1}{n-K} \sum_{t=1}^{n} \varepsilon_t^2 \\
&\quad + (\hat{\beta}_{2sls} - \beta^o)' \frac{1}{n-K} \sum_{t=1}^{n} X_t X_t' (\hat{\beta}_{2sls} - \beta^o) \\
&\quad - 2(\hat{\beta}_{2sls} - \beta^o)' \frac{1}{n-K} \sum_{t=1}^{n} X_t \varepsilon_t \\
&\xrightarrow{p} \quad \sigma^2 + 0 \cdot Q_{xx} \cdot 0 - 2 \cdot 0 \cdot E(X_t \varepsilon_t) \\
&= \sigma^2.
\end{aligned}$$

Note that although $E(X_t \varepsilon_t) \neq 0$, the last term still vanishes to zero in probability, because $\hat{\beta}_{2sls} - \beta^o \xrightarrow{p} 0$.

Question: What happens if we use $s^2 = e'e/(n-K)$, where $e = Y - \hat{\mathbf{X}}\hat{\beta}_{2sls}$ is the estimated residual from the second stage regression? Do we still have $s^2 \xrightarrow{p} \sigma^2$?

We have proved the following theorem.

Theorem 7.5 [Consistency of $\hat{\Omega}$ under MDS with Conditional Homoskedasticity]: *Under Assumptions 7.1 – 7.4, 7.6 and 7.7, we have as $n \to \infty$,*

$$\hat{\Omega} = \hat{s}^2 \hat{Q}_{\hat{x}\hat{x}}^{-1} \xrightarrow{p} \Omega = \sigma^2 Q_{\tilde{x}\tilde{x}}^{-1} = \sigma^2 \left(Q_{xz} Q_{zz}^{-1} Q_{zx}\right)^{-1}.$$

Case II: $\{Z_t\varepsilon_t\}$ is an MDS with Conditional Heteroskedasticity

When there exists conditional heteroskedasticity but $\{Z_t\varepsilon_t\}$ is still an MDS, the infeasible OLS estimator $\tilde{\beta}$ in the artificial regression

$$Y = \tilde{X}\beta^o + \varepsilon$$

has the following asymptotic distribution:

$$\sqrt{n}(\tilde{\beta} - \beta^o) \xrightarrow{d} N(0, Q_{\tilde{x}\tilde{x}}^{-1} \tilde{V} Q_{\tilde{x}\tilde{x}}^{-1}),$$

where

$$\tilde{V} = E(\tilde{X}_t \tilde{X}_t' \varepsilon_t^2).$$

Given $\tilde{X}_t = \gamma' Z_t$, $\gamma = Q_{zz}^{-1} Q_{zx}$, $Q_{\tilde{x}\tilde{x}} = \gamma' Q_{zz} \gamma$, and $\tilde{V} = \gamma' E(Z_t Z_t' \varepsilon_t^2)\gamma = \gamma'\ V\gamma$, where $V = E(Z_t Z_t' \varepsilon_t^2)$ under the MDS assumption with conditional heteroskedasticity, we have

$$\begin{aligned}
\text{avar}(\sqrt{n}\tilde{\beta}) &= Q_{\tilde{x}\tilde{x}}^{-1} \tilde{V} Q_{\tilde{x}\tilde{x}}^{-1} \\
&= [E(\tilde{X}_t \tilde{X}_t')]^{-1} E[\tilde{X}_t \tilde{X}_t' \varepsilon_t^2] [E(\tilde{X}_t \tilde{X}_t')]^{-1} \\
&= [\gamma' E(Z_t Z_t')\gamma]^{-1} \gamma' E(Z_t Z_t' \varepsilon_t^2)\gamma [\gamma' E(Z_t Z_t')\gamma]^{-1} \\
&= \left(Q_{xz} Q_{zz}^{-1} Q_{zx}\right)^{-1} Q_{xz} Q_{zz}^{-1} V Q_{zz}^{-1} Q_{zx} (Q_{xz} Q_{zz}^{-1} Q_{zx})^{-1} \\
&= \Omega \equiv \text{ avar}(\sqrt{n}\hat{\beta}_{2sls}).
\end{aligned}$$

This implies that the asymptotic distribution of the infeasible OLS estimator $\tilde{\beta}$ is the same as the asymptotic distribution of $\hat{\beta}_{2sls}$ under MDS with conditional heteroskedasticity. Therefore, the estimator for Ω is

$$\hat{\Omega} = \hat{Q}_{\hat{x}\hat{x}}^{-1} \hat{V}_{\hat{x}\hat{x}} \hat{Q}_{\hat{x}\hat{x}}^{-1},$$

where

$$\begin{aligned}\hat{V}_{\hat{x}\hat{x}} &= n^{-1}\sum_{t=1}^{n}\hat{X}_t\hat{X}_t'\hat{e}_t^2 \\ &= \hat{\gamma}'\left(n^{-1}\sum_{t=1}^{n}Z_tZ_t'\hat{e}_t^2\right)\hat{\gamma},\end{aligned}$$

where $\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \hat{Q}_{zz}^{-1}\hat{Q}_{zx}$ and $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. This is a White's (1980) heteroskedasticity-consistent variance-covariance matrix estimator for $\hat{\beta}_{2sls}$.

Now, put

$$\hat{V} \equiv n^{-1}\sum_{t=1}^{n}Z_tZ_t'\hat{e}_t^2,$$

Then

$$\hat{\Omega} = [\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}]^{-1}\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{V}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}[\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}]^{-1},$$

where (please check it!)

$$\begin{aligned}\hat{V} &= n^{-1}\sum_{t=1}^{n}Z_tZ_t'\hat{e}_t^2 \\ \xrightarrow{p} V &= E(Z_tZ_t'\varepsilon_t^2)\end{aligned}$$

under suitable regularity conditions.

Question: How to show $\hat{\Omega} \xrightarrow{p} \Omega$ under MDS with conditional heteroskedasticity?

Again, there are two methods to show $\hat{\Omega} \xrightarrow{p} \Omega$ here.

Method 1: We shall show $\hat{Q}_{\hat{x}\hat{x}} \xrightarrow{p} Q_{\tilde{x}\tilde{x}}$ and $\hat{V}_{\hat{x}\hat{x}} \xrightarrow{p} \tilde{V}$. The fact that $\hat{Q}_{\hat{x}\hat{x}} \xrightarrow{p} Q_{\tilde{x}\tilde{x}}$ has been shown earlier in the case of conditional homoskedasticity. To show $\hat{V}_{\hat{x}\hat{x}} \xrightarrow{p} \tilde{V}$, we write

$$\begin{aligned}\hat{V}_{\hat{x}\hat{x}} &= n^{-1}\sum_{t=1}^{n}\hat{X}_t\hat{X}_t\hat{e}_t^2 \\ &= \hat{\gamma}'\left(n^{-1}\sum_{t=1}^{n}Z_tZ_t'\hat{e}_t^2\right)\hat{\gamma} \\ &= \hat{\gamma}'\hat{V}\hat{\gamma}.\end{aligned}$$

Because $\hat{\gamma} \xrightarrow{p} \gamma$, and following the consistency proof for $n^{-1}\sum_{t=1}^{n}X_tX_t'e_t^2$ in Chapter 4, we can show (please verify!) that

$$\hat{V} = n^{-1}\sum_{t=1}^{n}Z_tZ_t'\hat{e}_t^2 \xrightarrow{p} E(Z_tZ_t'\varepsilon_t^2) = V,$$

under the following additional moment condition:

Assumption 7.8: *(i)* $E(Z_{jt}^4) < \infty$ *for all* $0 \leq j \leq l$; *and (ii)* $E(\varepsilon_t^4) < \infty$.

It follows that

$$\begin{aligned} & \hat{V}_{\hat{x}\hat{x}} \xrightarrow{p} \gamma' E(Z_t Z_t' \varepsilon_t^2)\gamma \\ = \; & E(\tilde{X}_t \tilde{X}_t' \varepsilon_t^2) \\ = \; & \tilde{V}. \end{aligned}$$

This and $\hat{Q}_{\hat{x}\hat{x}} \xrightarrow{p} Q_{\tilde{x}\tilde{x}}$ imply $\hat{\Omega} \xrightarrow{p} \Omega$.

Method 2: Given that

$$\hat{\Omega} = \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1} \hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{V}\hat{Q}_{zz}^{-1}\hat{Q}_{zx} \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1},$$

it suffices to show $\hat{Q}_{xz} \xrightarrow{p} Q_{xz}, \hat{Q}_{zz} \xrightarrow{p} Q_{zz}$ and $\hat{V} \xrightarrow{p} V$. The first two results immediately follow by the WLLN. The last result follows by using a similar reasoning of the consistency proof for $n^{-1}\sum_{t=1}^n X_t X_t' e_t^2$ in Chapter 4 or 5.

We now summarize the result derived above.

Theorem 7.6 [Consistency of $\hat{\Omega}$ under MDS with Conditional Heteroskedasticity]: *Under Assumptions 7.1-7.4, 7.6 and 7.8, we have as* $n \to \infty$,

$$\begin{aligned} \hat{\Omega} &= \hat{Q}_{\hat{x}\hat{x}}^{-1}\hat{V}_{\hat{x}\hat{x}}\hat{Q}_{\hat{x}\hat{x}}^{-1} \xrightarrow{p} \Omega = Q_{\tilde{x}\tilde{x}}^{-1}\tilde{V}Q_{\tilde{x}\tilde{x}}^{-1} \\ &= (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}VQ_{zz}^{-1}Q_{zx}(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}. \end{aligned}$$

where $\tilde{V} = E(\tilde{X}_t\tilde{X}_t'\varepsilon_t^2)$ and $V = E(Z_t Z_t'\varepsilon_t^2)$.

Case III: $\{Z_t\varepsilon_t\}$ is a Stationary Ergodic non-MDS

Finally, we consider a general case where $\{Z_t\varepsilon_t\}$ is not an MDS, which may arise as in the examples discussed in Chapter 6.

In this case, we have $\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, \Omega)$ as $n \to \infty$, where

$$\begin{aligned} \Omega &= Q_{\tilde{x}\tilde{x}}^{-1}\tilde{V}Q_{\tilde{x}\tilde{x}}^{-1} \\ &= (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}VQ_{zz}^{-1}Q_{zx}(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}, \end{aligned}$$

with

$$\begin{aligned}\tilde{V} &= \sum_{j=-\infty}^{\infty}\tilde{\Gamma}(j), \qquad \tilde{\Gamma}(j) = \text{cov}(\tilde{X}_t\varepsilon_t, \tilde{X}_{t-j}\varepsilon_{t-j}),\\ V &= \sum_{j=-\infty}^{\infty}\Gamma(j), \qquad \Gamma(j) = \text{cov}(Z_t\varepsilon_t, Z_{t-j}\varepsilon_{t-j}).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\text{avar}(\sqrt{n}\tilde{\beta}) &= Q_{\tilde{x}\tilde{x}}^{-1}V_{\tilde{x}\tilde{x}}Q_{\tilde{x}\tilde{x}}^{-1}\\ &= (\gamma'Q_{xx}\gamma)^{-1}\gamma'V\gamma(\gamma'Q_{xx}\gamma)^{-1}\\ &= \Omega \equiv \text{avar}(\sqrt{n}\hat{\beta}_{2sls}).\end{aligned}$$

Thus, the asymptotic variance of $\sqrt{n}\hat{\beta}_{2sls}$ is the same as the asymptotic variance of $\tilde{\beta}$ under this general case.

Question: How to estimate Ω?
Answer: Use a long-run variance-covariance matrix estimator for V or $\tilde{V}$.

We directly assume that we have a consistent estimator $\hat{V}$ for V.

Assumption 7.9: $\hat{V} \xrightarrow{p} V \equiv \Sigma_{j=-\infty}^{\infty}\Gamma(j)$, where $\Gamma(j) = \text{cov}(Z_t\varepsilon_t, Z_{t-j}\varepsilon_{t-j})$.

Question: How to estimate $\tilde{V} = \Sigma_{j=-\infty}^{\infty}\tilde{\Gamma}(j)$?

Recall that $\tilde{\Gamma}(j) = \gamma'\Gamma(j)\gamma$. A consistent estimator for $\tilde{V}$ can be given by

$$\hat{\gamma}'\hat{V}\hat{\gamma} \xrightarrow{p} \tilde{V}.$$

Theorem 7.7 [Consistency of $\hat{\Omega}$ under Non-MDS]: *Under Assumptions 7.1-7.4, and 7.9, we have as $n \to \infty$,*

$$\begin{aligned}\hat{\Omega} &= \hat{Q}_{\hat{x}\hat{x}}^{-1}\hat{V}_{\hat{x}\hat{x}}\hat{Q}_{\hat{x}\hat{x}}^{-1}\\ &= (\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{V}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}\\ \xrightarrow{p} \Omega &= Q_{\tilde{x}\tilde{x}}^{-1}\tilde{V}Q_{\tilde{x}\tilde{x}}^{-1},\end{aligned}$$

where $\hat{V}_{\hat{x}\hat{x}} = \hat{\gamma}\hat{V}\hat{\gamma}'$ and

$$\Omega = (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}VQ_{zz}^{-1}Q_{zx}(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}.$$

With a consistent estimator of Ω, we can develop various confidence interval estimators and various tests for the null hypothesis $\mathbf{H}_0 : R\beta^o = r$. We will consider the latter now.

7.6 Hypothesis Testing

Now, consider the null hypothesis of interest

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

where R is a $J \times K$ nonstochastic matrix, and r is a $J \times 1$ nonstochastic vector. The test statistics will differ depending on whether $\{Z_t\varepsilon_t\}$ is an MDS, and whether $\{\varepsilon_t\}$ is conditionally homoskedastic when $\{Z_t\varepsilon_t\}$ is an MDS. For space, here we do not present the results on t-type test statistics when $J = 1$.

Case I: $\{Z_t\varepsilon_t\}$ is an MDS with Conditional Homoskedasticity

Theorem 7.8 [Hypothesis Testing]: *Put $\hat{e} \equiv Y - \mathbf{X}\hat{\beta}_{2sls}$. Then under Assumptions 7.1-7.4, 7.6 and 7.7, the Wald test statistic*

$$\hat{W} = \frac{n(R\hat{\beta}_{2sls} - r)'[R(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}R']^{-1}(R\hat{\beta}_{2sls} - r)}{\hat{e}'\hat{e}/(n-K)} \xrightarrow{d} \chi^2_J$$

as $n \to \infty$,*under* $\mathbf{H}_0$.

Proof: The result follows immediately from the asymptotic normality theorem for $\sqrt{n}(\hat{\beta}_{2sls} - \beta^o)$, $\mathbf{H}_0$ (which implies $\sqrt{n}(R\hat{\beta}_{2sls} - r) = R\sqrt{n}(\hat{\beta}_{2sls} - \beta^o)$), the consistent asymptotic variance estimation theorem, and the Slutsky theorem.

Remarks:

Question: Is $\hat{W}/J$ the F-statistic from the second stage regression?

Answer: No, because $\hat{e}$ is not the estimated residual from the second stage regression.

Question: Do we still have

$$\hat{F} = \frac{(e_r'e_r - e_u'e_u)/J}{e_u'e_u/(n-K)},$$

where e_r and e_u are estimated residuals from the restricted and unrestricted regression models in the second stage regression respectively?

Answer: No. (Why?)

Case II: $\{Z_t\varepsilon_t\}$ is a Stationary Ergodic MDS with Conditional Heteroskedasticity

Theorem 7.9 [Hypothesis Testing]: *Under Assumptions 7.1-7.4, 7.6 and 7.8, the Wald test statistic*

$$\hat{W} \equiv n(R\hat{\beta}_{2sls} - r)'[R\hat{Q}_{\hat{x}\hat{x}}^{-1}\hat{V}_{\hat{x}\hat{x}}\hat{Q}_{\hat{x}\hat{x}}^{-1}R']^{-1}(R\hat{\beta}_{2sls} - r) \xrightarrow{d} \chi_J^2$$

under $\mathbf{H}_0$, *where* $\hat{V}_{\hat{x}\hat{x}} = n^{-1}\Sigma_{t=1}^n \hat{X}_t\hat{X}_t'\hat{e}_t^2$ *and* $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$.

Question: Suppose there exists conditional homoskedasticity but we use $\hat{W}$ above. Is $\hat{W}$ an asymptically valid procedure in this case?
Answer: Yes, $\hat{W}$ is asymptotically valid. However, the finite sample performance of $\hat{W}$ will be generally less satisfactory than the test statistic in Case I.

Case III: $\{Z_t\varepsilon_t\}$ is a Stationary ergodic non-MDS

When $\{Z_t\varepsilon_t\}$ is non-MDS, we can still construct a Wald test which is robust to conditional heteroskedasticity and autocorrelation, as is stated below.

Theorem 7.10 [Hypothesis Testing]: *Under Assumptions 7.1-7.5 and 7.9, the Wald test statistic*

$$\hat{W} = n(R\hat{\beta}_{2sls} - r)'[R\hat{Q}_{\hat{x}\hat{x}}^{-1}\hat{V}_{\hat{x}\hat{x}}\hat{Q}_{\hat{x}\hat{x}}^{-1}R']^{-1}(R\hat{\beta}_{2sls} - r) \xrightarrow{d} \chi_J^2$$

under $\mathbf{H}_0$, *where* $\hat{V}_{\hat{x}\hat{x}} = \hat{\gamma}'\hat{V}\hat{\gamma}, \hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ and $\hat{V}$ *is a long-run variance-covariance estimator for* $V = \Sigma_{j=-\infty}^{\infty}\Gamma(j)$ *with* $\Gamma(j) = \text{cov}(Z_t\varepsilon_t, Z_{t-j}\varepsilon_{t-j})$.

7.7 Hausman's Test

When there exists endogeneity so that $E(X_t\varepsilon_t) \neq 0$, the OLS estimator $\hat{\beta}$ is inconsistent for β^o. Instead, the 2SLS estimator $\hat{\beta}_{2sls}$ should be used, which involves the choice of the instrumental vector Z_t which in turn affects the efficiency of $\hat{\beta}_{2sls}$. In practice, it is not uncommon that practitioners are not sure whether there exists endogeneity. In this section, we introduce Hausman's (1978) test for endogeneity. The null hypothesis of interest is:

$$\mathbf{H}_0 : E(\varepsilon_t|X_t) = 0.$$

If this null hypothesis is rejected, one has to use the 2SLS estimator $\hat{\beta}_{2sls}$ provided that one can find a set of instruments Z_t that satisfies Assumption 7.4.

For simplicity, we impose the following conditions.

Assumption 7.10: (i) $\{(X_t', Z_t')'\varepsilon_t\}$ is an MDS process; and (ii) $E(\varepsilon_t^2|X_t, Z_t) = \sigma^2$ a.s.

Assumptions 7.10 is made for simplicity. They could be relaxed to be a non-MDS process with conditional heteroskedasticity but Hausman's (1978) test statistic to be introduced below should be generalized.

Question: How to test the conditional homoskedasticity assumption that $E(\varepsilon_t^2|X_t, Z_t) = \sigma^2$?

Answer: Put $\hat{e}_t = Y_t - \hat{X}_t'\hat{\beta}_{2sls}$. (Question: Can we use $e_t = Y_t - X_t'\hat{\beta}_{2sls}$?) Then run an auxiliary regression of $\hat{e}_t^2$ on vech(U_t), where $U_t = (X_t', Z_t')'$, a $(K + l) \times 1$ vector. Then under the condition that $E(\varepsilon_t^4|X_t, Z_t) = \mu_4$ is a constant, we have $nR^2 \xrightarrow{d} \chi_J^2$ under the null hypothesis of conditional homoskedasticity, where $J = (K + l)(K + l + 1)/2 - 1$.

The basic idea of Hausman's test is under $\mathbf{H}_0 : E(\varepsilon_t|X_t) = 0$, both the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ and the 2SLS estimator $\hat{\beta}_{2sls}$ are consistent for β^o. They converge to the same limit β^o but it can be shown that $\hat{\beta}$ is an asymptotically efficient estimator while $\hat{\beta}_{2sls}$ is not. Under the alternatives to H_0, $\hat{\beta}_{2sls}$ remains to be consistent for β^o but $\hat{\beta}$ is not. Hausman (1978) considers a test for H_0 based on the difference between the two estimators:

$$\hat{\beta}_{2sls} - \hat{\beta},$$

which converges to zero under H_0 but generally to a nonzero constant under the alternatives to H_0, giving the test its power against H_0 when the sample size n is sufficiently large.

To construct Hausman's (1978) test statistic, we need to derive the asymptotic distribution of $\hat{\beta}_{2sls} - \hat{\beta}$ For this purpose, we first state a lemma.

Lemma 7.11: Suppose $\hat{A} \xrightarrow{p} A$ and $\hat{B} = O_P(1)$. Then $(\hat{A} - A)\hat{B} \xrightarrow{p} 0$.

We first consider the OLS $\hat{\beta}$. Note that

$$\sqrt{n}(\hat{\beta} - \beta^o) = \hat{Q}_{xx}^{-1} n^{-1/2} \sum_{t=1}^{n} X_t\varepsilon_t$$

where $\hat{Q}_{xx}^{-1} \xrightarrow{p} Q_{xx}^{-1}$ and

$$n^{-1/2} \sum_{t=1}^{n} X_t\varepsilon_t \xrightarrow{d} N(0, \sigma^2 Q_{xx})$$

as $n \to \infty$ (see Chapter 5). It follows that $n^{-1/2}\sum_{t=1}^n X_t\varepsilon_t = O_P(1)$, and by Lemma 7.11, we have

$$\sqrt{n}(\hat{\beta} - \beta^o) = Q_{xx}^{-1} n^{-1/2} \sum_{t=1}^n X_t\varepsilon_t + o_P(1).$$

Similarly, we can obtain

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) &= \hat{A} n^{-1/2} \sum_{t=1}^n Z_t\varepsilon_t \\ &= A n^{-1/2} \sum_{t=1}^n Z_t\varepsilon_t + o_P(1),\end{aligned}$$

where $\hat{A} = (\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}\hat{Q}_{xz}\hat{Q}_{zz} \xrightarrow{p} A = (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}$ and $n^{-1/2}\sum_{t=1}^n Z_t\varepsilon_t \xrightarrow{d} N(0, \sigma^2 Q_{zz})$ (see Corollary 7.4). It follows that

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{2sls} - \hat{\beta}) &= n^{-1/2} \sum_{t=1}^n \left[(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}Z_t - Q_{xx}^{-1}X_t\right]\varepsilon_t + o_P(1) \\ &\xrightarrow{d} N(0, \sigma^2(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1} - \sigma^2 Q_{xx}^{-1})\end{aligned}$$

by the CLT for the stationary ergodic MDS process and Assumption 7.10. Therefore, under the null hypothesis $\mathbf{H}_0$, the quadratic form

$$\begin{aligned}H &= \frac{n(\hat{\beta}_{2sls} - \hat{\beta})' \left[(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1} - \hat{Q}_{xx}^{-1}\right]^{-1} (\hat{\beta}_{2sls} - \hat{\beta})}{s^2} \\ &\xrightarrow{d} \chi_K^2\end{aligned}$$

as $n \to \infty$ by the Slutsky theorem, where $s^2 = e'e/n$ is the residual variance estimator based on the OLS residual $e = Y - X\hat{\beta}$. This is called Hausman's test statistic.

Question: Can we replace the residual variance estimator s^2 by $\hat{s}^2 = \hat{e}'\hat{e}/n$, where $\hat{e} = Y - X\hat{\beta}_{2sls}$?

Theorem 7.12 [Hausman's Test for Endogeneity] *Suppose Assumptions 7.1–7.4, 7.10 and* $\mathbf{H}_0$ *hold, and* $Q_{xx} - Q_{xz}Q_{zz}^{-1}Q_{zx}$ *is strictly positive definite. Then as* $n \to \infty$,

$$H \xrightarrow{d} \chi_K^2.$$

Remarks:

We note that in the above Theorem,

$$\begin{aligned}\text{avar}[\sqrt{n}(\hat{\beta}_{2sls} - \hat{\beta})] &= \sigma^2(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1} - \sigma^2 Q_{xx}^{-1} \\ &= \text{avar}(\sqrt{n}\hat{\beta}_{2sls}) - \text{avar}(\sqrt{n}\hat{\beta}).\end{aligned}$$

This simple asymptotic variance-covariance structure is made possible under Assumption 7.10. Suppose there exists conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t, Z_t) \neq \sigma^2$). Then we no longer have the above simple variance-covariance structure for avar$[\sqrt{n}(\hat{\beta} - \hat{\beta}_{2sls})]$.

The variance-covariance $(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1} - Q_{xx}^{-1}$ may become singular when its rank $J < K$. In this case, we have to modify the Hausman's test statistic by using the generalized inverse of the variance estimator:

$$H = \frac{n(\hat{\beta}_{2sls} - \hat{\beta})' \left[(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1} - \hat{Q}_{xx}^{-1}\right]^{-} (\hat{\beta}_{2sls} - \hat{\beta})}{s^2}$$

Note that now $H \xrightarrow{d} \chi_J^2$ under $\mathbf{H}_0$ where $J < K$.

Question: What is the generalized inverse A^- of matrix A?

Question: How to modify the Hausman's test statistic so that it remains asymptotically χ_K^2 when there exists conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t, Z_t) \neq \sigma^2$) but $\{(X_t', Z_t')'\varepsilon_t\}$ is still an MDS process?

In fact, Hausman's (1978) test is a general approach to testing model specification, not merely whether endogeneity exists. For example, it can be used to test whether a fixed effect panel regression model or a random effect panel regression model should be used. In Hausman (1978), two estimators are compared, one of which is asymptotically efficient under the null hypothesis but inconsistent under the alternative and another of which is asymptotically inefficient but consistent under the alternative hypothesis. This approach was extended by White (1981) to compare any two different estimators either of which need not be asymptotically most efficient. The methods of Hausman and White were further extended by Newey (1985), Tauchen (1985) and White (1990) to construct moment-based tests for model specification.

Hausman's test is used to check whether $E(\varepsilon_t|X_t) = 0$. Suppose this condition fails, one has to choose an instrumental vector Z_t that satisfies Assumption 7.4. When we choose a set of variables Z_t, how can we check the validity of Z_t as instruments? In particular, how to check whether $E(\varepsilon_t|Z_t) = 0$? For this purpose, we will consider a so-called overidentification test, which will be introduced in Chapter 8.

7.8 Empirical Applications

Application I: Incentives in Chinese State-owned Enterprises

Groves, Hong, McMillan and Naughton (1994, Quaterly Journal of Economics)

Application II: The Consumption Function

Campbell and Mankiw (1989, 1991)
The consumption function

$$\begin{aligned} \Delta C_t &= \mu + \lambda \Delta Y_t + \varepsilon_t, \\ \Delta Y_t &= Z_t'\delta + v_t, \end{aligned}$$

where ΔY_t is income growth, ΔC_t is consumption growth.

7.9 Conclusion

In this chapter, we discuss the possibilities that the condition of $E(\varepsilon_t|X_t) = 0$ may fail in practice, which will render inconsistent the OLS estimator for the true model parameters. With the use of instrumental variables, we introduce a consistent two-stage least squares (2SLS) estimator. We investigate the statistical properties of the 2SLS estimator and provide some interpretations that can enhance deeper understanding of the nature of the 2SLS estimator. We discuss how to construct consistent estimators for the asymptotic variance of the 2SLS estimator under various scenarios, including MDS with conditional homoskedasticity, MDS with conditional heteroskedasticity, and non-MDS possibly with conditional heteroskedasticity. For the latter, consistent estimation for the long-run variance covariance matrix is needed. With these consistent asymptotic variance estimators, various hypothesis test procedures are proposed. It is important to emphasize that the conventional t-test and F-test cannot be used even for large samples. Finally, some empirical applications that employ 2SLS are considered.

In fact, the 2SLS procedure is one of several approaches to consistent estimation of model parameters when the condition of $E(\varepsilon_t|X_t) = 0$ fails. There are alternative estimation procedures that also yield consistent estimators. For example, suppose the correlation between X_t and ε_t is caused by the omitted variables problem, namely

$$\varepsilon_t = g(W_t) + u_t,$$

when $E(u_t|X_t, W_t) = 0$ and W_t is an omitted variable which is correlated with X_t. This delivers a partially linear regression model

$$Y_t = X_t'\beta^o + g(W_t) + u_t.$$

Because $E(Y_t|W_t) = E(X_t|W_t)'\beta^o + g(W_t)$, we obtain

$$Y_t - E(Y_t|W_t) = [X_t - E(X_t|W_t)]'\beta^o + u_t$$

or

$$Y_t^* = X_t^{*\prime}\beta^o + u_t,$$

where $Y_t^* = Y_t - E(Y_t|W_t)$ and $X_t^* = X_t - E(X_t|W_t)$. Because $E(X_t^* u_t) = 0$, the OLS estimator $\tilde{\beta}^*$ of regressing Y_t^* on X_t^* would be consistent for β^o. However, (Y_t^*, X_t^*) are not observable, so $\tilde{\beta}^*$ is infeasible. Nevertheless, one can first estimate $E(Y_t|W_t)$ and $E(X_t|W_t)$ nonparametrically, and then obtain a feasible OLS estimator which will be consistent for the true model parameter (e.g., Robinson 1988). Specifically, let $\hat{m}_Y(W_t)$ and $\hat{m}_X(W_t)$ be consistent nonparametric estimators for $E(Y_t|W_t)$ and $E(X_t|W_t)$ respectively. Then we can obtain a feasible OLS estimator

$$\tilde{\beta}_a^* = \left[\sum_{t=1}^n \hat{X}_t^* \hat{X}_t^{*\prime}\right]^{-1} \sum_{t=1}^n \hat{X}_t^* \hat{Y}_t^*,$$

where $\hat{X}_t^* = X_t - \hat{m}_X(W_t)$ and $\hat{Y}_t^* = Y_t - \hat{m}_Y(W_t)$. It can be shown that $\tilde{\beta}_a^* \xrightarrow{p} \beta^o$ and

$$\sqrt{n}(\tilde{\beta}_a^* - \beta^o) \xrightarrow{d} N(0, Q^{*-1}V^*Q^{*-1}),$$

where $Q^* = E(X_t^* X_t^{*\prime})$ and $V^* = \text{var}(n^{-1/2}\Sigma_{t=1}^n X_t^* u_t)$. The first stage nonparametric estimation has no impact on the asymptotic properties of the feasible OLS estimator $\tilde{\beta}_a$.

Another method to consistently estimate the true model parameters is to make use of panel data. A panel data is a collection of observations for a total of n cross-sectional units and each of these units has T time series observations over the same time period. This is called a balanced panel data. In contrast, an unbalanced panel data is a collection of observations for a total of n cross-sectional units and each unit may have different lengths of time series observations with some common overlapping time periods.

With a balanced panel data, we have

$$\begin{aligned} Y_{it} &= X_{it}'\beta^o + \varepsilon_{it} \\ &= X_{it}'\beta^o + \alpha_i + u_{it}, \end{aligned}$$

where α_i is called individual-specific effect and u_{it} is called idiosyncratic disturbance such that $E(u_{it}|X_{it}, \alpha_i) = 0$. When α_i is correlated with X_{it}, which may be caused by omitted variables which do not change over time, the panel data model is called a fixed effect panel data model. When α_i is uncorrelated with X_{it}, the panel data model is called a random effect panel data model. Here, we consider a fixed effect panel data

model with strictly exogenous variables X_{it}. Because ε_{it} is correlated with X_{it}, the OLS estimator of regressing Y_{it} on X_{it} is not consistent for β^o. However, one can consider the demeaned model

$$Y_{it} - \dot{Y}_{i.} = (X_{it} - \dot{X}_{i.})'\beta^o + (\varepsilon_{it} - \dot{\varepsilon}_{i.}),$$

where $\dot{Y}_{i.} = T^{-1}\Sigma_{t=1}^{T} Y_{it}$ and similarly for $\dot{X}_{i.}$ and $\dot{\varepsilon}_{i.}$. The demeaning procedure removes the unobservable individual-specific effect and as a result, the OLS estimator for the demeaned model, which is called the within estimator in the panel data literature, will be consistent for the true model parameter β^o. (It should be noted that for a dynamic panel data model where X_{it} is not strictly exogenous, the within estimator is not consistent for β^o when the number of the time periods T is fixed. Different estimation methods have to be used.) See Hsiao (2002) for detailed discussion of panel data econometric models.

Chapters 2 to 7 present a relatively comprehensive econometric theory for linear regression models often encountered in economics and finance. We start with a general regression analysis, discussing the interpretation of a linear regression model, which depends on whether the linear regression model is correctly specified. After discussing the classical linear regression model in Chapter 3, Chapters 4 to 7 discuss various extensions and generalizations when some assumptions in the classical linear regression model are violated. In particular, we consider the scenarios under which the results for classical linear regression models are approximately applicable for large samples. The key condition here are conditional homokedasticity and serial uncorrelatedness in the regression disturbance. When there exists conditional heteroskedasticity or serial correlation in the regression disturbance, the results for classical linear regression models are no longer applicable; we provide robust asymptotically valid procedures under these scenarios.

The asymptotic theory developed for linear regression models in Chapters 4–7 can be easily extended to more complicated, nonlinear models. For example, consider a nonlinear regression model

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

where $E(\varepsilon_t|X_t) = 0$ a.s. The nonlinear least squares estimator solves the minimization of the sum of squared residual problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{n} [Y_t - g(X_t, \beta)]^2 .$$

The first order condition is

$$D(\hat{\beta})'e = 0,$$

where $D(\beta)$ is a $n \times K$ matrix, with the t-th row being $\partial g(X_t, \beta)/\partial\beta$. Although one generally does not have a closed form expression for $\hat{\beta}$, all asymptotic theory and procedures in Chapters 4–7 are applicable to the nonlinear least squares estimator if one replaces X_t by $(\partial/\partial\beta)g(X_t, \beta)$. See also the discussion in Chapters 8 and 9.

The asymptotic theory in Chapters 4–7 however, cannot be directly applied to some popular nonlinear models. Examples of such nonlinear models are

- Rational Expectations Model:

$$E\left[m(Z_t, \beta^o)\right] = 0;$$

- Conditional Variance Model:

$$Y_t = g(X_t, \beta^o) + \sigma(X_t, \beta^o)u_t,$$

  where $g(X_t, \beta)$ is a parametric model for $E(Y_t|X_t), \sigma^2(X_t, \beta)$ is a parametric model for $\text{var}(Y_t|X_t)$, and $\{u_t\}$ is i.i.d.$(0, 1)$;

- Conditional probability model of Y_t given X_t :

$$f(y|X_t, \beta).$$

These nonlinear models are not models for conditional mean or regression; they also model other characteristics of the conditional distribution of Y_t given X_t. For these models, we need to develop new estimation methods and new asymptotic theory, which we will turn to in subsequent chapters.

One important part that we do not discuss in Chapters 2–7 is model specification testing. Chapter 2 emphasizes the importance of correct model specification for the validity of economic interpretation of model parameters. How to check whether a linear regression model is correctly specified for conditional mean $E(Y_t|X_t)$? This is called model specification testing. Some popular specification tests in econometrics are Hausman's (1978) test and White's (1981) test which compares two parameter estimators for the same model parameter. Also, see Hong and White's (1995) specification test using a nonparametric series regression approach.

EXERCISES

7.1. Consider the following simple Keynes national income qmodel

$$\begin{aligned} C_t &= \beta_1^o + \beta_2^o(Y_t - T_t) + \varepsilon_t, && (1.1)\\ T_t &= \gamma_1^o + \gamma_2^o Y_t + v_t, && (1.2)\\ Y_t &= C_t + G_t, && (1.3) \end{aligned}$$

where C_t, Y_t, T_t, G_t are the consumption, income, tax, and government spending respectively, and $\{\varepsilon_t\}$ and $\{v_t\}$ are i.i.d. $(0, \sigma_\varepsilon^2)$ and $(0, \sigma_v^2)$ respectively. Model (1.1) is a consumption function which we are interested in, (1.2) is a tax function, and (1.3) is an income identity.

(a) Can the OLS estimator $\hat{\beta}$ of model (1.1) give consistent estimation for the marginal propensity to consume? Explain.

(b) Suppose G_t is an exogenous variable (i.e., G_t does not depend on both C_t and Y_t). Can G_t be used as a valid instrumental variable? If yes, describe a 2SLS procedure. If not, explain.

(c) Suppose the government has to maintain a budget balance such that

$$G_t = T_t + w_t, \tag{1.4}$$

where $\{w_t\}$ is i.i.d. $(0, \sigma_w^2)$. Could G_t be used as a valid instrumental variable? If yes, describe a 2SLS procedure. If not, explain.

7.2. Consider the data generating process

$$Y_t = X_t'\beta^o + \varepsilon_t, \tag{2.1}$$

where $X_t = (1, X_{1t})'$,

$$\begin{aligned} X_{1t} &= v_t + u_t, && (2.2)\\ \varepsilon_t &= w_t + u_t. && (2.3) \end{aligned}$$

where $\{v_t\}, \{u_t\}$ and $\{w_t\}$ are all i.i.d. $N(0,1)$, and they are mutually independent.

(a) Is the OLS estimator $\hat{\beta}$ consistent for β^o? Explain.

(b) Suppose that $Z_{1t} = w_t - \varepsilon_t$. Is $Z_t = (1, Z_{1t})'$ a valid instrumental vector? Explain.

(c) Find an instrumental vector and the asymptotic distribution of $\hat{\beta}_{2sls}$ using this instrumental vector. [Note you need to find $\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, V)$ for some V, where the expression of V should be given.]

(d) To test the hypothesis

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

where R is a $J \times 2$ matrix, and r is a $J \times 1$ vector. Suppose that $\tilde{F}$ is the F-statistic in the second stage regression of 2SLS. Could we use $J \cdot \tilde{F}$ as an asymptotic χ^2_J test? Explain.

7.3. Consider the following demand-supply system:

$$\begin{aligned} Y_t &= \alpha_0^o + \alpha_1^o P_t + \alpha_2^o S_t + \varepsilon_t, \\ Y_t &= \beta_0^o + \beta_1^o P_t + \beta_2^o C_t + v_t, \end{aligned}$$

where the first equation is a model for the demand of certain good, where Y_t is the quantity demanded for the good, P_t is the price of the good, S_t is the price of a substitute, and ε_t is a shock to the demand. The second equation is a model for the supply of the good, where Y_t is the quantity supplied, C_t is the cost of production, and v_t is a shock to the supply. Suppose S_t and C_t are exogenous variables, $\{\varepsilon_t\}$ is i.i.d.$(0, \sigma_\varepsilon^2)$ and $\{v_t\}$ is i.i.d.$(0, \sigma_v^2)$, and two series $\{\varepsilon_t\}$ and $\{v_t\}$ are independent of each other. We have also assumed that the market is always clear so the quantity demanded is equal to the quantity supplied.

(a) Suppose we use a 2SLS estimator to estimate the demand model with the instruments $Z_t = (S_t, C_t)'$. Describe the 2SLS procedure. Is the resulting 2SLS $\hat{\alpha}_{2sls}$ consistent for $\alpha^o = (\alpha_0^o, \alpha_1^o, \alpha_2^o)'$? Explain.

(b) Suppose we use a 2SLS estimator to estimate the supply equation with instruments $Z_t = (S_t, C_t)'$. Describe the 2SLS procedure. Is the resulting 2SLS $\hat{\beta}_{2sls}$ consistent for $\beta^o = (\beta_0^o, \beta_1^o, \beta_2^o)'$? Explain.

(c) Suppose $\{\varepsilon_t\}$ and $\{v_t\}$ are contemporaneously correlated, namely, $E(\varepsilon_t v_t) \neq 0$. This can occur when there is a common shock to both the demand and supply of the good. Does this affect the conclusions in part (a) and part (b). Explain.

7.4. Show that under Assumptions 7.1-7.4, $\hat{\beta}_{2sls} \xrightarrow{p} \beta^o$ as $n \to \infty$.

7.5. Suppose Assumptions 7.1-7.5 hold.

(a) Show that $\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) \xrightarrow{d} N(0, \Omega)$ as $n \to \infty$, where

$$\Omega = [Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}Q_{xz}Q_{zz}^{-1}VQ_{zz}^{-1}Q_{zx}[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1},$$

and V is given in Assumption 7.5;

(b) If in addition that $\{Z_t\varepsilon_t\}$ is an ergodic stationary MDS process with $E(\varepsilon_t^2|Z_t) = \sigma^2$. Show that

$$\Omega = \sigma^2[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}.$$

7.6. Suppose Assumptions 7.1 – 7.4, 7.6 and 7.7 hold.

(a) Define

$$\hat{s}^2 = \frac{\hat{e}'\hat{e}}{n}$$

where $\hat{e} = Y - X\hat{\beta}_{2sls}$. Show $\hat{s}^2 \xrightarrow{p} \sigma^2 = \text{var}(\varepsilon_t)$ as $n \to \infty$.

(b) Define

$$s^2 = \frac{e'e}{n},$$

where $e = Y - \hat{X}\hat{\beta}_{2sls}$ is the estimated residual from the second stage regression of Y_t on $\hat{X}_t = \hat{\gamma}'Z_t$. Show that s^2 is not a consistent estimator for σ^2.

7.7. [2SLS Hypothesis Testing] Suppose Assumptions 7.1-7.5 hold. Define a F-statistic

$$F = \frac{n(R\hat{\beta}_{2sls} - r)'[R\hat{Q}_{\hat{x}\hat{x}}^{-1}R']^{-1}(R\hat{\beta}_{2sls} - r)/J}{e'e/(n-K)},$$

where $e_t = Y_t - \hat{X}_t'\hat{\beta}_{2sls}$ is the estimated residual from the second stage regression of Y_t on $\hat{X}_t$. Does $J \cdot F \xrightarrow{d} \chi_J^2$ under the null hypothesis $\mathbf{H}_0 : R\beta^o = \gamma$? If yes, give your reasoning. If not, provide a modification so that the modified test statistic converges to χ_J^2 under $\mathbf{H}_0$.

7.8. Let

$$\hat{V} = \frac{1}{n}\sum_{t=1}^{n} Z_t Z_t' \hat{e}_t^2,$$

where $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. Show $\hat{V} \xrightarrow{p} V$ under Assumptions 7.1–7.8.

7.9. Suppose the following assumptions hold:

Assumption 3.1 [Linearity]: $\{Y_t, X_t'\}_{t=1}'^n$ is a stationary ergodic process with

$$Y_t = X_t'\beta^o + \varepsilon_t, \qquad t = 1, ..., n,$$

for some unknown parameter β^o and some unobservable disturbance ε_t;

Assumption 3.2 [Nonsingularity] The $K \times K$ matrix

$$Q_{xx} = E(X_t X_t')$$

is nonsingular and finite;

Assumption 3.3 [Orthogonality]

(i) $E(X_t\varepsilon_t) = 0$;

(ii) $E(Z_t\varepsilon_t) = 0$, where Z_t is a $l \times 1$ random vector, with $l \geq K$;

(iii) The $l \times l$ matrix

$$Q_{zz} = E(Z_t Z_t')$$

is finite and nonsingular, and the $l \times K$ matrix

$$Q_{xz} = E(Z_t X_t')$$

is finite and of full rank;

Assumption 3.4: $\{(X_t', Z_t')'\varepsilon_t\}$ is an martingale difference sequence.
Assumption 3.5: $E(\varepsilon_t^2|X_t, Z_t) = \sigma^2$ a.s.

Under these assumptions, both OLS

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

and 2SLS

$$\hat{\beta}_{2sls} = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y$$

are consistent for β^o.

(a) Show that $\hat{\beta}$ is a special 2SLS estimator $\hat{\beta}_{2sls}$ with some proper choice of instrumental vector Z_t.

(b) Which estimator, $\hat{\beta}$ or $\hat{\beta}_{2sls}$, is more asymptotically efficient? [Hint: if $\sqrt{n}(\hat{\beta}_1 - \beta^o) \xrightarrow{d} N(0, \Omega_1)$ and $\sqrt{n}(\hat{\beta}_2 - \beta^o) \xrightarrow{d} N(0, \Omega_2)$, then $\hat{\beta}_1$ is asymptotically more efficient than $\hat{\beta}_2$ if and only if $\Omega_2 - \Omega_1$ or $\Omega_1^{-1} - \Omega_2^{-1}$ is positive semi-definite.]

7.10. Consider the linear regression model

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

where $E(X_t\varepsilon_t) \neq 0$. Our purpose is to find a consistent estimation procedure for β^o.

First, consider the artificial regression

$$X_t = \gamma' Z_t + v_t,$$

where X_t is the regressor vector, Z_t is the instrumental vector, $\gamma = [E(Z_t Z_t')]^{-1}E(Z_t X_t')$ is the best linear LS approximation coefficient, and v_t is the $K \times 1$ regression error.

Now, suppose instead of decomposing X_t, we decompose the regression error ε_t as follows:

$$\varepsilon_t = v_t'\rho^0 + u_t,$$

where $\rho^0 = [E(v_t v_t')]^{-1} E(v_t \varepsilon_t)$ is the best linear LS approximation coefficient.

Now, assuming that v_t is observable, we consider the augmented linear regression model

$$Y_t = X_t'\beta^o + v_t'\rho^0 + u_t.$$

Show $E[(X_t', v_t)' u_t] = 0$. One important implication of this orthogonality condition is that if v_t is observable then the OLS estimator of regressing Y_t on X_t and v_t will be consistent for $(\beta^o, \rho^o)'$.

7.11. In practice, v_t is unobservable. However, it can be estimated by the estimated residual

$$\hat{v}_t = X_t - \hat{\gamma}' Z_t = X_t - \hat{X}_t.$$

We now consider the following feasible augmented linear regression model

$$Y_t = X_t'\beta^o + \hat{v}_t'\rho + \tilde{u}_t,$$

and we denote the resulting OLS estimator as $\hat{\alpha} = (\hat{\beta}', \hat{\rho}')'$, where $\hat{\beta}$ is the OLS estimator for β^o and $\hat{\rho}$ is the OLS estimator for ρ.

Show $\hat{\beta} = \hat{\beta}_{2sls}$.[Hint: The following decomposition may be useful: Suppose

$$A = \begin{bmatrix} B & C' \\ C & D \end{bmatrix}$$

is a nonsingular square matrix, where B is $k_1 \times k_1, C$ is $k_2 \times k_1$ and D is $k_2 \times k_2$. Then

$$A^{-1} = \begin{bmatrix} B^{-1}(I + C'E^{-1}CB^{-1}) & -B'^{-1}C'E^{-1} \\ -E^{-1}CB^{-1} & E^{-1} \end{bmatrix},$$

where $E = D - CB^{-1}C'$.]

7.12. Suppose $\hat{Y}$ is a $n \times 1$ vector of the fitted values of regressing Y_t on Z_t, and $\hat{X}$ is a $n \times K$ matrix of fitted values of regressing X_t on Z_t. Show that $\hat{\beta}_{2sls}$ is equal to the OLS estimator of regressing $\hat{Y}$ on $\hat{X}$.

7.13 [Hausman's Test] Suppose Assumptions 3.1, 3.2, 3.3(ii, iii), 3.4 and 3.5 in Problem 7.8 hold. A test for the null hypothesis $\mathbf{H}_0 : E(X_t \varepsilon_t) = 0$ can be constructed by comparing $\hat{\beta}$ and $\hat{\beta}_{2sls}$, because they will converge in probability to the same limit β^o under H_0 and to different limits under the alternatives to $\mathbf{H}_0$. Assume $\mathbf{H}_0$ holds.

(a) Show that

$$\sqrt{n}(\hat{\beta} - \beta^o) - Q_{xx}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \varepsilon_t \xrightarrow{p} 0$$

or equivalently

$$\sqrt{n}(\hat{\beta} - \beta^o) = Q_{xx}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \varepsilon_t + o_P(1),$$

where $Q_{xx} = E(X_t X_t')$. [Hint: If $\hat{A} \xrightarrow{p} A$ and $\hat{B} = O_P(1)$, then $\hat{A}\hat{B} - A\hat{B} \xrightarrow{p} 0$ or $\hat{A}\hat{B} = A\hat{B} + o_P(1)$.]

(b) Show that

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta^o) = Q_{\tilde{x}\tilde{x}}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{X}_t \varepsilon_t + o_P(1),$$

where $Q_{\tilde{x}\tilde{x}} = E(X_t X_t')$, $\tilde{X}_t = \gamma' Z_t$ and $\gamma = [E(Z_t Z_t')]^{-1} E(Z_t X_t)$.

(e) Show that

$$\sqrt{n}(\hat{\beta}_{2sls} - \hat{\beta}) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{ Q_{xx}^{-1} X_t - Q_{\tilde{x}\tilde{x}}^{-1} \tilde{X}_t \right\} \varepsilon_t + o_P(1).$$

(d) The asymptotic distribution of $\sqrt{n}(\hat{\beta}_{2sls} - \hat{\beta})$ is determined by the leading term only in part (c). Find its asymptotic distribution.

(e) Construct an asymptotically χ^2 test statistic. What is the degree of freedom of the asymptotic χ^2 distribution? Assume that $Q_{xx} - Q_{\tilde{x}\tilde{x}}$ is strictly positive definite.

7.14. Suppose Assumptions 3.1, 3.2, 3.3(ii, iii) and 3.4 in Problem 7.8 hold, $E(X_{jt}^4) < \infty$ for $1 \leq j \leq K$, $E(Z_{jt}^4) < \infty$ for $1 \leq j \leq l$, and $E(\varepsilon_t^4) < \infty$. Construct a Hausman's test statistic for $\mathbf{H}_0 : E(\varepsilon_t | X_t) = 0$ and derive its asymptotic distribution under $\mathbf{H}_0$.

CHAPTER 8 GENERALIZED METHOD OF MOMENTS ESTIMATION

Abstract: Many economic theories and hypotheses have implications on and only on a moment condition or a set of moment conditions. A popular method to estimate model parameters contained in the moment condition is the Generalized Method of Moments (GMM). In this chapter, we first provide some economic examples for the moment condition, and define the GMM estimator. We then establish the consistency and asymptotic normality of the GMM estimator. Since the asymptotic variance of a GMM estimator depends on the choice of a weighting matrix, we introduce an asymptotically optimal two-stage GMM estimator with a suitable choice of a weighting matrix. With the construction of a consistent asymptotic variance estimator, we then propose an asymptotically χ^2 Wald test statistic for the hypothesis of interest, and a model specification test for the moment condition.

Key words: CAPM, GMM, IV Estimation, Model specification test, Moment condition, Moment matching, Optimal estimation, Overidentification, Rational expectations.

8.1 Introduction to the Method of Moments Estimation (MME)

To motivate the generalized method of moments (GMM) estimation, we first consider a traditional method in statistics which is called the method of moments estimation (MME).

MME Procedure: Suppose $f(y, \beta^o)$ is the probability density function (pdf) or the probability mass function (pmf) of a univariate random variable Y_t.

Question: How to estimate the unknown parameter β^o using a realization of the random sample $\{Y_t\}_{t=1}^n$?

The basic idea of MME is to match the sample moments with the population moments obtained under the probability distributional model. Specifically, MME can be implemented as follows:

Step 1: Compute population moments $\mu_k(\beta^o) \equiv E(Y_t^k)$ under the model density $f(y, \beta^o)$.

For example, for $k = 1, 2$, we have

$$\begin{aligned} E(Y_t) &= \int_{-\infty}^{\infty} y f(y, \beta^o) dy = \mu_1(\beta^o) \\ E(Y_t^2) &= \int_{-\infty}^{\infty} y^2 f(y, \beta^o) dy \\ &= \sigma^2(\beta^o) + \mu_1^2(\beta^o), \end{aligned}$$

where $\sigma^2(\beta^o)$ is the variance of Y_t.

Step 2: Compute the sample moments from the random sample $Y^n = (Y_1, ..., Y_n)'$:

For example, for $k = 1, 2$, we have

$$\begin{aligned} \hat{m}_1 &= \bar{Y}_n \xrightarrow{p} \mu(\beta^o) \\ \hat{m}_2 &= n^{-1} \sum_{t=1}^{n} Y_t^2 \\ \xrightarrow{p} \quad E(Y_t^2) &= \sigma^2(\beta^o) + \mu_1^2(\beta^o), \end{aligned}$$

where $\sigma^2(\beta^o) = \mu_2(\beta^o) - \mu_1^2(\beta^o)$, and the weak convergence follows from the WLLN.

Step 3: Match the sample moments with the corresponding population moments evaluated at some parameter value $\hat{\beta}$:

For example, for $k = 1, 2$, we set

$$\begin{aligned} \hat{m}_1 &= \mu(\hat{\beta}), \\ \hat{m}_2 &= \sigma^2(\hat{\beta}) + \mu^2(\hat{\beta}). \end{aligned}$$

Step 4: Solve for the system of equations. The solution $\hat{\beta}$ is called the method of moment estimator for β^o.

Remarks: In general, if β is a $K \times 1$ parameter vector, we need K equations of matching moments.

Question: Is MME consistent for β^o?

Answer: Because $\mu_k(\hat{\beta}) = \hat{m}_k \xrightarrow{p} \mu_k(\beta^o)$ by the WLLN, we expect that $\hat{\beta} \xrightarrow{p} \beta^o$ as $n \to \infty$.

We now illustrate MME by two simple examples.

Example 1: Suppose the random sample $\{Y_t\}_{t=1}^n \sim$ i.i.d. EXP(λ). Find an estimator for λ using the method of moment estimation.

Solution: In our application, $\beta = \lambda$. Because the exponential pdf

$$f(y, \lambda) = \lambda e^{-\lambda y} \text{ for } y > 0,$$

it can be shown that

$$\begin{aligned}\mu(\lambda) &= E(Y_t) = \int_0^\infty y f(y,\lambda) dy \\ &= \int_0^\infty y\lambda e^{-\lambda y} dy \\ &= \frac{1}{\lambda}.\end{aligned}$$

On the other hand, the first sample moment is the sample mean:

$$\hat{m}_1 = \bar{Y}_n.$$

Matching the sample mean with the population mean evaluated at $\hat{\lambda}$:

$$\hat{m}_1 = \mu(\hat{\lambda}) = \frac{1}{\hat{\lambda}},$$

we obtain the method of moment estimator

$$\hat{\lambda} = \frac{1}{\hat{m}_1} = \frac{1}{\bar{Y}_n}.$$

Example 2: Suppose the random sample $\{Y_t\}_{t=1}^n \sim$ i.i.d.$N(\mu, \sigma^2)$. Find MME for $\beta^o = (\mu, \sigma^2)'$.

Solution: The first two population moments are

$$\begin{aligned}E(Y_t) &= \mu, \\ E(Y_t^2) &= \sigma^2 + \mu^2.\end{aligned}$$

The first two sample moments are

$$\begin{aligned}\hat{m}_1 &= \bar{Y}_n, \\ \hat{m}_2 &= \frac{1}{n}\sum_{t=1}^n Y_t^2.\end{aligned}$$

Matching the first two moments, we have

$$\begin{aligned}\bar{Y}_n &= \hat{\mu}, \\ \frac{1}{n}\sum_{t=1}^n Y_t^2 &= \hat{\sigma}^2 + \hat{\mu}^2.\end{aligned}$$

It follows that the MME

$$\begin{aligned}\hat{\mu} &= \bar{Y}_n, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - \bar{Y}_n^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y}_n)^2.\end{aligned}$$

It is well-known that $\hat{\mu} \xrightarrow{p} \mu$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$.

8.2 Generalized Method of Moments (GMM) Estimation

Suppose β is a $K \times 1$ unknown parameter vector, and there exists a $l \times 1$ moment function $m_t(\beta)$ such that

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

where sub-index t denotes that $m_t(\beta)$ is a function of both β and some random variables indexed by t. For example, we may have

$$m_t(\beta) = X_t(Y_t - X_t'\beta)$$

in the OLS estimation, or

$$m_t(\beta) = Z_t(Y_t - X_t'\beta)$$

in the 2SLS estimation, or more generally in the instrumental variable (IV) estimation, where Z_t is a $l \times 1$ instrument vector.

If $l = K$, that is, if the number of moment conditions is the same as the number of unknown parameters, the model $E[m_t(\beta^o)] = 0$ is called exactly identified. If $l > K$, that is, if the number of moment conditions is more than the number of unknown parameters, the model is called overidentified.

The moment condition $E[m_t(\beta^o)] = 0$ may follow from economic and financial theory (e.g. rational expectations and correct asset pricing). We now illustrate this by the following example.

Example 1 [Capital Asset Pricing Model (CAPM)]: Define Y_t as an $L \times 1$ vector of excess returns for L assets (or portfolios of assets) in period t. For these L assets, the excess returns can be described using the excess-return market model:

$$\begin{aligned}Y_t &= \beta_0^o + \beta_1^o R_{mt} + \varepsilon_t \\ &= \beta^{o\prime} X_t + \varepsilon_t,\end{aligned}$$

where $X_t = (1, R_{mt})'$ is a bivariate vector, R_{mt} is the excess market portfolio return, β^o is a $2 \times L$ parameter matrix, and ε_t is an $L \times 1$ disturbance, with $E(\varepsilon_t | X_t) = 0$.

Define the $l \times 1$ moment function

$$m_t(\beta) = X_t \otimes (Y_t - \beta' X_t),$$

where $l = 2L$ and $\otimes$ denotes the Kronecker product. When CAPM holds, we have

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

These $l \times 1$ moment conditions form a basis to estimate and test the CAPM.

In fact, for any measurable function $h : R^2 \to R^l$, CAPM implies

$$E[h(X_t)(Y_t - \beta' X_t)] = 0.$$

This can also be used to estimate the CAPM model.

Question: How to choose the instruments $h(X_t)$?

Example 2 [Hansen and Singleton (1982, Econometrica) Dynamic Capital Asset Pricing Model]:

Suppose a representative economic agent has a constant relative risk aversion utility over his lifetime

$$U = \sum_{t=0}^{n} \delta^t u(C_t) = \sum_{t=0}^{n} \delta^t \frac{C_t^\gamma - 1}{\gamma},$$

where $u(\cdot)$ is the time-invariant utility function of the economic agent in each time period (here we assume $u(c) = (c^\gamma - 1)/\gamma$), δ is the agent's time discount factor, γ is the economic agent's risk aversion parameter, and C_t is the consumption during period t. Let the information available to the agent at time $t-1$ be represented by the sigma-algebra I_{t-1}– in the sense that any variable whose value is known at time $t-1$ is presumed to be I_{t-1}-measurable, and let

$$R_t = \frac{P_t}{P_{t-1}} = 1 + \frac{P_t - P_{t-1}}{P_{t-1}}$$

be the gross return to an asset acquired at time $t-1$ at the price of P_{t-1} (we assume no dividend on the asset). The agent's optimization problem is to

$$\max_{\{C_t\}} E(U)$$

subject to the intertemporal budget constraint

$$C_t + P_t q_t = Y_t + P_t q_{t-1},$$

where q_t is the quantity of the asset purchased at time t and Y_t is the agent's labor income during period t. Define the marginal rate of intertemporal substitution

$$\text{MRS}_t(\gamma) = \frac{\frac{\partial u(C_t)}{\partial C_t}}{\frac{\partial u(C_{t-1})}{\partial C_{t-1}}} = \left(\frac{C_t}{C_{t-1}}\right)^{\gamma-1}.$$

The first order conditions of the agent optimization problem are characterized by the Euler equation:

$$E\left[\delta^o \text{MRS}_t(\gamma^o) R_t | I_{t-1}\right] = 1 \text{ for some } \beta^o = (\delta^o, \gamma^o)'.$$

That is, the marginal rate of intertemporal substitution discounts gross returns to unity.

Remarks: Any dynamic asset pricing model is equivalent to a specification of MRS_t.

We may write the Euler equation as follows:

$$E\left[\{\delta^o \text{MRS}_t(\gamma^o) R_t - 1\} | I_{t-1}\right] = 0.$$

Thus, one may view that $\{\delta \text{MRS}_t(\gamma) R_t - 1\}$ is a generalized model residual which has the MDS property when evaluated at the true structural parameters $\beta^o = (\delta^o, \gamma^o)'$.

Question: How to estimate the unknown parameter β^o in an asset pricing model?

More generally, how to estimate β^o from any linear or nonlinear econometric model which can be formulated as a set of moment conditions? Note that the joint distribution of the random sample is not given or implied by economic theory; only a set of conditional moments is given.

From the Euler equation, we can induce the following conditional moment restrictions:

$$\begin{aligned} E\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right) &= 0, \\ E\left[\frac{C_{t-1}}{C_{t-2}}\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right)\right] &= 0, \\ E\left[R_{t-1}\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right)\right] &= 0. \end{aligned}$$

Therefore, we can consider the 3×1 sample moments

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

where

$$m_t(\beta) = [\delta \text{MRS}_t(\gamma) R_t - 1] \left(1, \frac{C_{t-1}}{C_{t-2}}, R_{t-1}\right)'$$

can serve as the basis for estimation. The elements of the vector

$$Z_t \equiv \left(1, \frac{C_{t-1}}{C_{t-2}}, R_{t-1}\right)'$$

are called instrumental variables which are a subset of information set I_{t-1}.

We now define the GMM estimator.

Definition 8.1 [GMM Estimator] The generalized method of moments (GMM) estimator is

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where

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

is a $l \times 1$ sample moment vector, $\hat{W}$ is a $l \times l$ symmetric nonsingular matrix which is possibly data-dependent, and β is a $K \times 1$ unknown parameter vector, and Θ is a K-dimensional parameter space. Here, we assume $l \geq K$, i.e., the number of moments may be larger than or at least equal to the number of parameters.

Question: Why do we require $l \geq K$ in GMM estimation?

Question: Why is the GMM estimator $\hat{\beta}$ not defined by setting the $l \times 1$ sample moments to zero jointly, namely

$$\hat{m}(\hat{\beta}) = 0?$$

Remarks: When $l > K$, i.e., when the number of equations is larger than the number of unknown parameters, we generally cannot find a $\hat{\beta}$ such that $\hat{m}(\hat{\beta}) = 0$. However, we can find a $\hat{\beta}$ which makes $\hat{m}(\hat{\beta})$ as close to a $l \times 1$ zero vector as possible by minimizing the quadratic form

$$\hat{m}(\beta)' \hat{m}(\beta) = \sum_{i=1}^{l} \hat{m}_i^2(\beta),$$

where $\hat{m}_i(\beta) = n^{-1} \sum_{t=1}^{n} m_{it}(\beta)$, $i = 1, ..., l$. Since each sample moment component $\hat{m}_i(\beta)$ has a different variance, and $\hat{m}_i(\beta)$ and $\hat{m}_j(\beta)$ may be correlated, we can introduce a weighting matrix $\hat{W}$ and choose $\hat{\beta}$ to minimize a weighted quadratic form in $\hat{m}(\hat{\beta})$, namely

$$\hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta).$$

Question: What is the role of $\hat{W}$?

When $\hat{W} = I$, an identity matrix, each of the l component sample moments is weighted equally. If $\hat{W} \neq I$, then the l sample moment components are weighted differently. A suitable choice of weighting matrix $\hat{W}$ can improve the efficiency of the resulting estimator. Here, a natural question is: what is the optimal weighting function for the choice of $\hat{W}$?

Intuitively, the sample moment components which have large sampling variations should be discounted. This is an idea similar to GLS, which discounts noisy observations by dividing by the conditional standard deviation of the disturbance term and differencing out serial correlations.

Special Case: Linear IV Estimation

Question: Does the GMM estimator have a closed form expression?

In general, when the moment function $m_t(\beta)$ is nonlinear in parameter β, there is no closed form solution for $\hat{\beta}$. However, there is an important special case where the GMM estimator $\hat{\beta}$ has a closed form. This is the case of so-called linear IV estimation where we have

$$m_t(\beta) = Z_t(Y_t - X_t'\beta)$$

and

$$E[Z_t(Y_t - X_t'\beta^o)] = 0 \text{ for some } \beta^o,$$

where Y_t is a scalar, X_t is a $K \times 1$ vector, and Z_t is $l \times 1$ vector, with $l \geq K$.

In this case, the GMM estimator, or more precisely, the linear IV estimator, $\hat{\beta}$, solves the following minimization problem:

$$\min_{\beta \in R^K} \hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta) = n^{-2} \min_{\beta \in R^K} (Y - \mathbf{X}\beta)'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\beta),$$

where

$$\hat{m}(\beta) = \frac{\mathbf{Z}'(Y - \mathbf{X}\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n} Z_t(Y_t - X_t'\beta).$$

The FOC is given by

$$\begin{aligned} & \frac{\partial}{\partial \beta}\left[(Y - \mathbf{X}\beta)'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\beta)\right]_{\beta=\hat{\beta}} \\ = \; & -2\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) = 0. \end{aligned}$$

It follows that

$$\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

When the $K \times l$ matrix $Q_{xz} = E(X_t Z_t')$ is of full rank of K, the $K \times K$ matrix $Q_{xz} W Q_{zx}$ is nonsingular. Therefore, $\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X}$ is not singular at least for large samples, and consequently the GMM estimator $\hat{\beta}$ has the closed form expression:

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

This is called a linear IV estimator because it estimates the parameter β^o in the linear model $Y_t = X_t'\beta^o + \varepsilon_t$ with $E(\varepsilon_t|Z_t) = 0$.

Interestingly, the 2SLS estimator $\hat{\beta}_{2sls}$ considered in Chapter 7 is a special case of the IV estimator by choosing

$$\hat{W} = \mathbf{Z}'\mathbf{Z}.$$

or more generally, by choosing $\hat{W} = c(\mathbf{Z}'\mathbf{Z})$ for any constant $c \neq 0$.

Question: Is the choice of $\hat{W} = \mathbf{Z}'\mathbf{Z}$ optimal? In other words, is the 2SLS estimator $\hat{\beta}_{2sls}$ asymptotically efficient in estimating β^o?

When $l = K$ such that $Q_{xz} = E(X_t Z_t')$ is nonsingular, the $K \times K$ matrix $\mathbf{X}'\mathbf{Z}$ is nonsingular at least for large samples. Consequently,

$$\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y.$$

Theorem 8.1: *Suppose $m_t(\beta) = Z_t(Y_t - X_t'\beta)$, where Y_t is a scalar, Z_t is a $l \times 1$ vector, X_t is $K \times 1$ vector, with $l \geq K$. Also, the $K \times l$ matrix $\mathbf{X}'\mathbf{Z}$ is of full rank K and the $l \times l$ weighting matrx $\hat{W}$ is nonsingular. Then the resulting GMM estimator $\hat{\beta}$ is called a linear IV estimator and has the closed form expression*

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

When $l = K$, and Q_{xz} is nonsignular,

$$\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y.$$

Note that the IV estimator $\hat{\beta}$ generally depends on the choice of instruments Z_t and weighting matrix $\hat{W}$. However, when $l = K$, the exact identification case, the IV estimator $\hat{\beta}$ does not depend on the choice of $\hat{W}$. This is because in this case the FOC that $\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) = 0$ becomes

$$\begin{aligned} \mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) &= 0 \\ (K \times n)(n \times 1) &= K \times 1 \end{aligned}$$

given $\mathbf{X}'\mathbf{Z}$ and $\hat{W}$ are nonsingular at least for large samples. Obviously, the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ is a special case of the linear IV estimator by choosing $Z_t = X_t$.

8.3 Consistency of GMM

Question: What are the statistical properties of GMM $\hat{\beta}$?

To investigate the asymptotic properties of the GMM estimator $\hat{\beta}$, we first provide a set of regularity conditions.

Assumption 8.1 [Compactness]: The parameter space Θ is compact (closed and bounded);

Assumption 8.2 [Uniform convergence]: (i) The moment function $m_t(\beta)$ is a measurable function of a random vector indexed by t for each $\beta \in \Theta$, and given each t, $m_t(\beta)$ is continuous in $\beta \in \Theta$; (ii) $\{m_t(\beta)\}$ is a stationary ergodic process; (iii) $\hat{m}(\beta)$ converges uniformly over Θ to $m(\beta) \equiv E[m_t(\beta)]$ in probability in the sense that

$$\sup_{\beta \in \Theta} ||\hat{m}(\beta) - m(\beta)|| \xrightarrow{p} 0,$$

where $||\cdot||$ is an Euclidean norm; (iv) $m(\beta)$ is continuous in $\beta \in \Theta$.

Assumption 8.3 [Identification]: There exists a unique parameter β^o in Θ such that $m(\beta^o) = 0$.

Assumption 8.4 [Weighting Matrix]: $\hat{W} \xrightarrow{p} W$, where W is a nonstochastic $l \times l$ symmetric, finite and nonsingular matrix.

Remarks:

Assumption 8.3 is an identification condition. If the moment condition $m(\beta^o) = 0$ is implied by economic theory, β^o can be viewed as the true model parameter value. Assumptions 8.1 and 8.3 imply that the true model parameter β^o lies inside the compact parameter space Θ. Compactness is sometimes restrictive, but it greatly simplifies our asymptotic analysis and is sometime necessary (as in the case of estimating GARCH models) where some parameters must be restricted to ensure a positive conditional variance estimator.

In many applications, the moment function $m_t(\beta)$ usually has the form

$$m_t(\beta) = h_t \varepsilon_t(\beta)$$

for some weighting function h_t and some error or generalized error term $\varepsilon_t(\beta)$. Assumption 8.2 allows but does not require such a multiplicative form for $m_t(\beta)$. Also, in Assumption 8.2, we

impose a uniform WLLN for $\hat{m}(\beta)$ over Θ. Intuitively, uniform convergence implies that the largest (or worse) deviation between $\hat{m}(\beta)$ and $m(\beta)$ over Θ vanishes to 0 in probability as $n \to \infty$.

Question: How to ensure uniform convergence in probability?

This can be achieved by a suitable uniform weak law of large numbers (UWLLN). For example, when $\{Y_t, X_t'\}_{t=1}'^n$ is i.i.d., we have the following:

Lemma 8.2 [Uniform Strong Law of Large Numbers for IID Processes (USLLN)]: *Let $\{Z_t, t = 1, 2, ...\}$ be an IID sequence of random $d \times 1$ vectors, with common cumulative distribution function F.*

Let Θ be a compact subset of R^K, and let $q : R^d \times \Theta \to R$ be a function such that $q(\cdot, \beta)$ is measurable for each $\beta \in \Theta$ and $q(z, \cdot)$ is continuous on Θ for each $z \in R^d$.

Suppose there exists a measurable function $D : R^d \to R^+$ such that $|q(z, \beta)| \leq D(z)$ for all $\beta \in \Theta$ and $z \in S$, where S is the support of Z_t and $E[D(Z_t)] < \infty$.

Then

(i) $Q(\beta) = E[q(Z_t, \beta)]$ is continuous on Θ;

(ii) $\sup_{\beta \in \Theta} |\hat{Q}(\beta) - Q(\beta)| \to 0$ a.s. as $n \to \infty$, where $\hat{Q}(\beta) = n^{-1} \sum_{t=1}^n q(Z_t, \beta)$.

Proof: See Jennrich (1969, Theorem 2).

A USLLN for stationary ergodic processes is following:

Lemma 8.3 [Uniform Strong Law of Large Numbers for Stationary Ergodic Processes {Ranga Rao (1962)}]: *Let (Ω, F, P) be a probability space, and let $T : \Omega \to \Omega$ be a one-to-one measure preserving transformation.*

Let Θ be a compact subset of R^K, and let $q : \Omega \times \Theta \to R$ be a function such that $q(\cdot, \beta)$ is measurable for each $\theta \in \Theta$ and $q(\omega, \cdot)$ is continuous on Θ for each $\omega \in \Omega$.

Suppose there exists a measurable function $D : \Omega \to R^+$ such that $|q(\omega, \beta)| \leq D(\omega)$ for all $\beta \in \Theta$ and $\omega \in \Omega$, and $E(D) = \int D dP < \infty$.

If for each $\beta \in \Theta$, $q_t(\theta) = q(T^t\omega, \beta)$ is ergodic, then

(i) $Q(\beta) = E[q_t(\beta)]$ is continuous on Θ;

(ii) $\sup_{\beta \in \Theta} |\hat{Q}(\beta) - Q(\beta)| \to 0$ a.s. as $n \to \infty$, where $\hat{Q}(\beta) = n^{-1} \sum_{t=1}^n q_t(\beta)$.

Proof: See Ranga Rao (1962).

Remarks: Uniform almost sure convergence implies uniform convergence in probability.

We first state the consistency result for the GMM estimator $\hat{\beta}$.

Theorem 8.4 [Consistency of the GMM Estimator]: *Suppose Assumptions 8.1–8.4 hold. Then* $\hat{\beta} \xrightarrow{p} \beta^o$.

To show this consistency theorem, we need the following extrema estimator lemma.

Lemma 8.5 [White, 1994, Consistency of Extrema Estimators]: *Let* $\hat{Q}(\beta)$ *be a stochastic real-valued function of* $\beta \in \Theta$, *and* $Q(\beta)$ *be a nonstochastic real-valued continuous function of* β, *where* Θ *is a compact parameter space. Suppose that for each* β, $\hat{Q}(\beta)$ *is a measurable function of the random sample with sample* n, *and for each* n, $\hat{Q}(\cdot)$ *is continuous in* $\beta \in \Theta$ *with probability one. Also suppose* $\hat{Q}(\beta) - Q(\beta) \xrightarrow{p} 0$ *uniformly in* $\beta \in \Theta$.

Let $\hat{\beta} = \arg\max_{\beta \in \Theta} \hat{Q}(\beta)$, *and* $\beta^o = \arg\max_{\beta \in \Theta} Q(\beta)$ *is the unique maximizer. Then* $\hat{\beta} - \beta^o \xrightarrow{p} 0$.

Remarks: This lemma continues to hold if we change all convergences in probability to almost sure convergences.

We now show the consistency of the GMM estimator $\hat{\beta}$ by applying the above lemma.

Proof: Put

$$\hat{Q}(\beta) = -\hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta)$$

and

$$Q(\beta) = -m(\beta)'W^{-1}m(\beta).$$

Then

$$\begin{aligned}
& \left|\hat{Q}(\beta) - Q(\beta)\right| \\
= & \left|\hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta) - m(\beta)'W^{-1}m(\beta)\right| \\
= & \left|[\hat{m}(\beta) - m(\beta) + m(\beta)]'\hat{W}^{-1}[\hat{m}(\beta) - m(\beta) + m(\beta)] - m(\beta)'W^{-1}m(\beta)\right| \\
\leq & \left|[\hat{m}(\beta) - m(\beta)]'\,\hat{W}^{-1}\,[\hat{m}(\beta) - m(\beta)]\right| \\
& +2\left|m(\beta)'\hat{W}^{-1}\,[\hat{m}(\beta) - m(\beta)]\right| \\
& +\left|m(\beta)'(\hat{W}^{-1} - W^{-1})m(\beta)\right|.
\end{aligned}$$

It follows from Assumptions 8.1, 8.2 and 8.4 that

$$\hat{Q}(\beta) \xrightarrow{p} Q(\beta)$$

uniformly over Θ, and $Q(\cdot) = m(\cdot)'W^{-1}m(\cdot)$ is continuous in β over Θ. Moreover, Assumption 8.3 implies that β^o is the unique minimizer of $Q(\beta)$ over Θ. It follows that $\hat{\beta} \xrightarrow{p} \beta^o$ by the extrema

estimator Lemma. Note that the proof of the consistency theorem does not require the existence of the FOC. This is made possible by using the extrema estimator lemma. This completes the proof of consistency.

8.4 Asymptotic Normality of GMM

To derive the asymptotic distribution of the GMM estimator, we impose two additional regularity conditions.

Assumption 8.5 [Interiorness]: $\beta^o \in int(\Theta)$.

Assumption 8.6 [CLT]:

(i) For each t, $m_t(\beta)$ is continuously differentiable with respect to $\beta \in \Theta$ with probability one.

(ii) As $n \to \infty$,

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

where $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$ is finite and p.d.

(iii) $\{\frac{\partial m_t(\beta)}{\partial \beta}\}$ obeys the uniform weak law of large numbers (UWLLN), i.e.,

$$\sup_{\beta \in \Theta} \left\| n^{-1} \sum_{t=1}^{n} \frac{\partial m_t(\beta)}{\partial \beta} - D(\beta) \right\| \xrightarrow{p} 0,$$

where the $l \times K$ matrix

$$\begin{aligned} D(\beta) &\equiv E\left[\frac{\partial m_t(\beta)}{\partial \beta}\right] \\ &= \frac{dm(\beta)}{d\beta} \end{aligned}$$

is continuous in $\beta \in \Theta$ and is of full rank K.

Remarks:

Question: Why do we need to assume that β^o is an interior point in Θ?

This is because we will have to use a Taylor series expansion. We need to make use of the FOC for GMM in order to derive the asymptotic distribution of $\hat{\beta}$.

In Assumption 8.6, we assume both CLT and UWLLN directly. These are called "high-level assumptions." They can be ensured by imposing more primitive conditions on the data generating processes (e.g., i.i.d. random samples or MDS random samples), and the moment and smoothness conditions of $m_t(\beta)$. Fore more discussion, see White (1994).

We now establish the asymptotic normality of the GMM estimator $\hat{\beta}$.

Theorem 8.6 [Asymptotic Normality]: *Suppose Assumptions 8.1–8.6 hold. Then as* $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta} - \beta^o\right) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = (D_o' W^{-1} D_o)^{-1} D_o' W^{-1} V_o W^{-1} D_o (D_o' W^{-1} D_o)^{-1},$$

and $D_o \equiv D(\beta^o) = \frac{\partial m(\beta^o)}{\partial \beta}$.

Proof: Because β^o is an interior element in Θ, and $\hat{\beta} \xrightarrow{p} \beta^o$ as $n \to \infty$, we have that $\hat{\beta}$ is an interior element of Θ with probability approaching one as $n \to \infty$.

For n sufficiently large, the first order conditions for the maximization of $\hat{Q}(\beta) = -\hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta)$ are

$$\begin{aligned} 0 &= \frac{d\hat{Q}(\beta)}{d\beta}\Big|_{\beta=\hat{\beta}} \\ &= -2\frac{d\hat{m}(\hat{\beta})}{d\beta'} \hat{W}^{-1} \hat{m}(\hat{\beta}). \\ 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'} \hat{W}^{-1} \sqrt{n}\hat{m}(\hat{\beta}). \\ K \times 1 &= (K \times l) \times (l \times l) \times (l \times 1) \end{aligned}$$

Note that $\hat{W}$ is not a function of β. Also, this FOC does not necessarily imply $\hat{m}(\hat{\beta}) = 0$. Instead, it only says that a set (with dimension $K \leq l$) of linear combinations of the l components in $\hat{m}(\hat{\beta})$ is equal to zero. Here, the $l \times K$ matrix $\frac{d\hat{m}(\beta)}{d\beta}$ is the gradient of the $l \times 1$ vector $\hat{m}(\hat{\beta})$ with respect to the $K \times 1$ vector β.

Using the Taylor series expansion around the true parameter value β^o, we have

$$\sqrt{n}\hat{m}(\hat{\beta}) = \sqrt{n}\hat{m}(\beta^o) + \frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o),$$

where $\bar{\beta} = \lambda\hat{\beta} + (1-\lambda)\beta^o$ lies between $\hat{\beta}$ and β^o, with $\lambda \in [0,1]$. Here, for notational simplicity, we have abused the notation in the expression of$\frac{d\hat{m}(\bar{\beta})}{d\beta}$. Precisely speaking, a different $\bar{\beta}$ is needed for each partial derivative of $\hat{m}(\cdot)$ with respect to each parameter β_i, $i = 1, ..., K$.

The first term in the above Taylor series expansion is contributed by the sampling randomness of the sample average of the moment functions evaluated at the true parameter β^o, and the second term is contributed by the randomness of parameter estimator $\hat{\beta} - \beta^o$.

It follows from FOC that

$$\begin{aligned} 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\hat{\beta}) \\ &= \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta}-\beta^o). \end{aligned}$$

Now let us show that $\frac{d\hat{m}(\hat{\beta})}{d\beta} \xrightarrow{p} D_o \equiv D(\beta^o)$. To show this, consider

$$\begin{aligned} & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D_0 \right\| \\ = & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D(\hat{\beta}) + D(\hat{\beta}) - D(\beta^o) \right\| \\ \leq & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D(\hat{\beta}) \right\| + \left\| D(\hat{\beta}) - D(\beta^o) \right\| \\ \leq & \sup_{\beta\in\Theta} \left\| \frac{d\hat{m}(\beta)}{d\beta} - D(\beta) \right\| + \left\| D(\hat{\beta}) - D(\beta^o) \right\| \\ & \xrightarrow{p} 0 \end{aligned}$$

by the triangle inequality and Assumption 8.6 (the UWLLN, the continuity of $D(\beta)$, and $\hat{\beta}-\beta^o \xrightarrow{p} 0$).

Similarly, because $\bar{\beta} = \lambda\hat{\beta} + (1-\lambda)\beta^o$ for $\lambda \in [0,1]$, we have

$$||\bar{\beta} - \beta^o|| = ||\lambda(\hat{\beta} - \beta^o)|| \leq ||\hat{\beta} - \beta^o|| \xrightarrow{p} 0.$$

It follows that

$$\frac{d\hat{m}(\bar{\beta})}{d\beta} \xrightarrow{p} D_o.$$

Then the $K \times K$ matrix

$$D_o' W^{-1} D_o$$

is nonsingular by Assumptions 8.4 and 8.6. Therefore, for n sufficiently large, the inverse

$$\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta} \right]^{-1}$$

exists and it converges in probability to $(D_o' W^{-1} D_o)^{-1}$. Therefore, when n is sufficiently large,

we have

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta^o) &= -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\beta^o)\\ &= \hat{A}\sqrt{n}\hat{m}(\beta^o),\end{aligned}$$

where

$$\hat{A} = -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}.$$

By Assumption 8.6(ii), the CLT for $\{m_t(\beta^o)\}$, we have

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

where $V_o \equiv \text{avar}[n^{-1/2}\Sigma_{t=1}^n m_t(\beta^o)]$. Moreover,

$$\begin{aligned}\hat{A} &= -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\\ &\xrightarrow{p} -\left(D_o'W^{-1}D_o\right)^{-1}D_o'W^{-1}\\ &\equiv A.\end{aligned}$$

It follows from the Slutsky theorem that

$$\sqrt{n}(\hat{\beta}-\beta^o) \xrightarrow{d} A\cdot N(0,V_o) \sim N(0,\Omega),$$

where

$$\begin{aligned}\Omega &= AV_oA'\\ &= (D_o'W^{-1}D_o)^{-1}D_o'W^{-1}V_oW^{-1}D_o(D_o'W^{-1}D_o)^{-1}.\end{aligned}$$

This completes the proof.

Remarks:

The structure of $\text{avar}(\sqrt{n}\hat{\beta})$ is very similar to that of $\text{avar}(\sqrt{n}\hat{\beta}_{2sls})$. In fact, as pointed out earlier, 2SLS is a special case of the GMM estimator with the choice of

$$\begin{aligned}m_t(\beta) &= Z_t(Y_t - X_t'\beta)\\ W &= E(Z_tZ_t') = Q_{zz}.\end{aligned}$$

Similarly, the OLS estimator is a special case of GMM with the choice of

$$\begin{aligned} m_t(\beta) &= X_t(Y_t - X_t'\beta), \\ W &= E(X_t X_t') = Q_{xx}. \end{aligned}$$

Most econometric estimators can be viewed as a special case of GMM, at least asymptotically. In other words, GMM provides a convenient unified framework to view most econometric estimators. See White (1994) for more discussion.

8.5 Asymptotic Efficiency of GMM

Question: There are many possible choices of $\hat{W}$. Is there any optimal choice for $\hat{W}$? If so, what is the optimal choice of $\hat{W}$?

The following theorem shows that the optimal choice of W is given by $W = V_o \equiv \text{var}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem 8.7 [Asymptotic Efficiency]: *Suppose Assumptions 8.4 and 8.6 hold. Define* $\Omega_o = (D_o' V_o^{-1} D_o)^{-1}$, *which is obtained from* Ω *by choosing* $W = V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. *Then*

$$\Omega - \Omega_o \text{ is p.s.d.}$$

for any finite, symmetric and nonsingular matrix W.

Proof: Observe that $\Omega - \Omega_o$ is p.s.d. if and only if $\Omega_o^{-1} - \Omega^{-1}$ is p.s.d. We therefore consider

$$\begin{aligned} & \Omega_o^{-1} - \Omega^{-1} \\ = \; & D_o' V_o^{-1} D_o - D_o' W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} D_o \\ = \; & D_o' V_o^{-1/2} [I - V_o^{1/2} W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} V_o^{1/2}] V_o^{-1/2} D_o \\ = \; & D_o' V_o^{-1/2} G V_o^{-1/2} D_o, \end{aligned}$$

where $V_o = V_o^{1/2} V_o^{1/2}$ for some symmetric and nonsingular matrix $V_o^{1/2}$, and

$$G \equiv I - V_o^{1/2} W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} V_o^{1/2}$$

is a symmetric idempotent matrix (i.e., $G = G'$ and $G^2 = G$). It follows that we have

$$\begin{aligned} \Omega_o^{-1} - \Omega^{-1} &= (D_o' V_o^{-1/2} G)(G V_o^{-1/2} D_o) \\ &= (G V_o^{-1/2} D_o)'(G V_o^{-1/2} D_o) \\ &= B'B \\ &\sim \text{ p.s.d. (why?)}, \end{aligned}$$

where $B = GV_o^{-1/2}D_o$ is a $l \times K$ matrix. This completes the proof.

Remarks:

The optimal choice of $W = V_o$ is not unique. The choice of $W = cV_o$ for any nonzero constant c is also optimal.

In practice, the matrix V_o is unavailable. However, we can use a feasible asymptotically optimal choice $\hat{W} = \tilde{V}$, a consistent estimator for $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Question: What is the intuition that $\hat{W} = \tilde{V}$ is an optimal weighting matrix?

Answer: $\hat{W} \xrightarrow{p} V_o$, and V_o is the variance-covariance matrix of the sample moments $\sqrt{n}\hat{m}(\beta^o)$. The use of $\hat{W}^{-1} \xrightarrow{p} V_o^{-1}$, therefore, downweighs the sample moments which have large sampling variations and differences out correlations between different components$\sqrt{n}\hat{m}_i(\beta^o)$ and $\sqrt{n}\hat{m}_j(\beta^o)$ for $i \neq j$, where $i, j = 1, ..., K$. This is similar in spirit to the GLS estimator in the linear regression model. It also corrects serial correlations between different sample moments when they exist.

Optimality of the 2SLS Estimator $\hat{\beta}_{2sls}$

As pointed out earlier, the 2SLS estimator $\hat{\beta}_{2sls}$ is a special case of the GMM estimator with $m_t(\beta) = Z_t(Y_t - X_t'\beta)$ and the the choice of weighting matrix $W = E(Z_tZ_t') = Q_{zz}$. Suppose $\{m_t(\beta^o)\}$ is an MDS and $E(\varepsilon_t^2|Z_t) = \sigma^2$, where $\varepsilon_t = Y_t - X_t'\beta^o$. Then

$$\begin{aligned} V_o &= \text{avar}[\sqrt{n}\hat{m}(\beta^o)] \\ &= E\left[m_t(\beta^o)m_t(\beta^o)'\right] \\ &= \sigma^2 Q_{zz} \end{aligned}$$

where the last equality follows from the law of iterated expectations and conditional homoskedasticity. Because $W = Q_{zz}$ is proportional to V_o, the 2SLS estimator $\hat{\beta}$ is asymptotically optimal in this case. In contrast, when $\{m_t(\beta^o)\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|Z_t) \neq \sigma^2$) or $\{m_t(\beta^o)\}$ is not an MDS, then the choice of $W = Q_{zz}$ does not deliver an asymptotically optimal 2SLS estimator. Instead, the GMM estimator with the choice of $W = V_o = E(Z_tZ_t'\varepsilon_t^2)$ is asymptotically optimal.

Two-Stage GMM Estimator

The previous theorem suggests that the following two-stage GMM estimator will be asymptotically optimal.

Step 1: Find a consistent preliminary estimator $\tilde{\beta}$:

$$\tilde{\beta} = \arg\min_{\beta\in\Theta} \hat{m}(\beta)'\tilde{W}^{-1}\hat{m}(\beta),$$

for some prespecified $\tilde{W}$ which converges in probability to some finite and p.d. matrix. For convenience, we can set $\tilde{W} = I$, an $l \times l$ identity matrix. This is not an optimal estimator, but it is a consistent estimator for β^o.

Step 2: Find a preliminary consistent estimator $\tilde{V}$ for $V_o \equiv$avar$[\sqrt{n}\hat{m}(\beta^o)]$, and choose $\hat{W} = \tilde{V}$.

The construction of $\tilde{V}$ differs in the following two cases, depending on whether $\{m_t(\beta^o)\}$ is an MDS:

Case (i): $\{m_t(\beta^o)\}$ is an ergodic stationary MDS process. In this case,

$$V_o \equiv \text{ avar}[\sqrt{n}\hat{m}(\beta^o)] = E[m_t(\beta^o)m_t(\beta^o)'].$$

The asymptotic variance estimator

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

will be consistent for

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

Question: How to show this?

Answer: We need to assume that $\{n^{-1}\sum_{t=1}^{n} m_t(\beta)m_t(\beta)' - E[m_t(\beta)m_t(\beta)']\}$ satisfies the uniform convergence:

$$\sup_{\beta\in\Theta}\left\| n^{-1}\sum_{t=1}^{n} m_t(\beta)m_t(\beta)' - E[m_t(\beta)m_t(\beta)']\right\| \xrightarrow{p} 0.$$

Also, we need to assume that $E[m_t(\beta)m_t(\beta)']$ is continuous in $\beta \in \Theta$.

Case (ii): $\{m_t(\beta^o)\}$ is not MDS. In this case, a long-run variance estimator for $V_o \equiv$ avar$[\sqrt{n}\hat{m}(\beta^o)]$ is needed:

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

where $k(\cdot)$ is a kernel function, $p = p(n)$ is a smoothing parameter,

$$\tilde{\Gamma}(j) = n^{-1}\sum_{t=j+1}^{n} m_t(\tilde{\beta})m_{t-j}(\tilde{\beta})' \text{ for } j \geq 0,$$

and $\tilde{\Gamma}(j) = \tilde{\Gamma}(-j)'$ if $j < 0$. Under regularity conditions, it can be shown that $\tilde{V}$ is consistent for the long-run variance

$$V_o = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

where $\Gamma(j) = \text{cov}[m_t(\beta^o), m_{t-j}(\beta^o)] = E[m_t(\beta^o)m_{t-j}(\beta^o)']$. See more discussion in Chapter 6.

Question: Why do not we need demean when defining $\tilde{\Gamma}(j)$?

Step 3: Find an asymptotically optimal estimator $\hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \tilde{V}^{-1} \hat{m}(\beta).$$

Remarks: The weighting matrix $\tilde{V}$ does not involve the unknown parameter β. It is a given (stochastic) weighting matrix. This two-stage GMM estimator $\hat{\beta}$ is asymptotically optimal because $\tilde{V} \xrightarrow{p} V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem 8.8 [Two-Stage Asymptotically Most Efficient GMM]: *Suppose Assumptions 8.1–8.3, 8.5 and 8.6 hold,* $\tilde{V} \xrightarrow{p} V$, *and* $\tilde{W} \xrightarrow{p} W$ *for some symmetric finite and positive definite matrix* W. *Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \Omega_o) \text{ as } n \to \infty,$$

where $\Omega_o = (D_o' V_o^{-1} D_o)^{-1}$.

Question: Why do we need the asymptotically two-stage GMM estimator?

First, most macroeconomic time series data sets are usually short, and second, the use of instruments Z_t is usually inefficient. These factors lead to a large estimation error so it is desirable to have an asymptotically efficient estimator.

Although the two-stage GMM procedure is asymptotically efficient, one may like to iterate the procedure further until the GMM parameter estimates and the values of the minimized objective function converge. This will eliminate any dependence of the GMM estimator on the choice of the initial weighting matrix $\tilde{W}$, and it may improve the finite sample performance of the GMM estimator when the number of parameters is large (e.g., Ferson and Foerster 1994).

8.6 Asymptotic Variance Estimator

To construct confidence interval estimators and conduct hypothesis tests, we need to estimate the asymptotic variance Ω_o of the optimal GMM estimator.

Question: How to estimate $\Omega_o \equiv (D_o' V_o^{-1} D_o)^{-1}$?

We need to estimate both D_o and V_o.

(i) To estimate $D_o = E[\frac{\partial m_t(\beta^o)}{\partial \beta}]$, we can use

$$\hat{D} = \frac{d\hat{m}(\hat{\beta})}{d\beta}.$$

We have shown earlier that

$$\hat{D} \xrightarrow{p} D_o.$$

(ii) To estimate V_o, we need to consider two cases—MDS and non-MDS separately:

Case I: $\{m_t(\beta^o)\}$ is ergodic stationary MDS. In this case,

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

A consistent variance estimator is

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

Assuming the UWLLN for $\{m_t(\beta)m_t(\beta)'\}$, we can show that $\hat{V}$ is consistent for

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

Case II: $\{m_t(\beta^o)\}$ is not MDS. In this case,

$$V_0 = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

where $\Gamma(j) = E[m_t(\beta^o)m_{t-j}(\beta^o)']$. A consistent variance estimator is

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

where $k(\cdot)$ is a kernel function, and

$$\hat{\Gamma}(j) = n^{-1}\sum_{t=j+1}^{n} m_t(\hat{\beta})m_{t-j}(\hat{\beta})' \text{ for } j \geq 0,$$

Under suitable conditions (e.g., Newey and West 1994, Andrews 1991), we can show

$$\hat{V} \xrightarrow{p} V_o$$

but the proof of this is beyond the scope of this course.

To cover both cases, we directly impose the following "high-level assumption":

Assumption 8.7: $\hat{V} - V_o \xrightarrow{p} 0$, where $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem 8.9 [Asymptotic Variance Estimator for the Optimal GMM Estimator]: *Suppose Assumptions 8.1–8.7 hold. Then*

$$\hat{\Omega}_o \equiv (\hat{D}'\hat{V}^{-1}\hat{D})^{-1} \xrightarrow{p} \Omega_o \text{ as } n \to \infty.$$

8.7 Hypothesis Testing

We now consider testing the hypothesis of interest

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

where $R(\cdot)$ is a $J \times 1$ continuously differentiable vector-valued function, $J \leq K$, and the $J \times K$ matrix $\frac{dR(\beta^o)}{d\beta} = R'(\beta^o)$ is of full rank J. Note that $R(\beta^o) = r$ covers both linear and nonlinear restrictions on model parameters. An example of nonlinear restriction on β^o is $\beta_1^o\beta_2^o = 1$.

Remarks: We need $J \leq K$. The number of restrictions is less than the number of parameters. We now allow hypotheses of both linear and nonlinear restrictions on β^o.

Question: How to construct a test statistic for $\mathbf{H}_0$?

The basic idea is to check whether $R(\hat{\beta}) - r$ is close to 0. By the Taylor series expansion and $R(\beta^o) = r$ under $\mathbf{H}_0$, we have

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

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

Because $R'(\bar{\beta}) \xrightarrow{p} R'(\beta^o)$ given continuity of $R'(\cdot)$ and $\bar{\beta} - \beta^o \xrightarrow{p} 0$, and

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

we have

$$\sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} N[0, R'(\beta^o)\Omega_o R'(\beta^o)'].$$

by the Slutsky theorem. It follows that the quadratic form

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

The Wald test statistic is then

$$W = n[R(\hat{\beta}) - r]'[R'(\hat{\beta})\hat{\Omega}_o R'(\hat{\beta})']^{-1}[R(\hat{\beta}) - r] \xrightarrow{d} \chi^2_J$$

where the convergence in distribution to χ^2_J follows from the Slutsky theorem.

When $J = 1$, we can have an asymptotically N(0,1) test statistic

$$T = \frac{\sqrt{n}[R(\hat{\beta}) - r)]}{\sqrt{R'(\hat{\beta})\hat{\Omega}_o R'(\hat{\beta})'}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Theorem 8.10 [Wald Test Statistic]: *Suppose Assumptions 8.1–8.7 hold. Then under* $\mathbf{H}_0$: $R(\beta^o) = r$, *we have*

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

Remarks: This can be used for hypothesis testing. This Wald test is built upon an asymptotically optimal GMM estimator. One could also construct a Wald test using a consistent but suboptimal GMM estimator (how?).

8.8 Model Specification Testing

As pointed out earlier, many dynamic economic theories can be formulated as a moment condition or a set of moment conditions. Thus, to test validity of an economic theory, one can check whether the related moment condition holds.

Question: How to test whether the econometric model as characterized by

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

is correctly specified?

Answer: We can check correct model specification by testing whether the above moment condition holds.

Question: How to check if the moment condition

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

holds?

Answer: Use the sample moment

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

and see if it is significantly different from zero (the value of the population moment evaluated at the true parameter value β^o). For this purpose, we need to know the asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$.

Consider the test statistic

$$\begin{aligned} \sqrt{n}\hat{m}(\hat{\beta}) &= \sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o) \end{aligned}$$

which follows from a first order Taylor series expansion, and $\bar{\beta}$ lies between $\hat{\beta}$ and β^o. The asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$ is contributed from two sources.

Recall that the two-stage GMM

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \tilde{V}^{-1} \hat{m}(\beta).$$

The FOC of the two-stage GMM estimation is given by

$$0 = \frac{d}{d\beta}\left[\hat{m}(\hat{\beta})' \tilde{V}^{-1} \hat{m}(\hat{\beta})\right].$$

It is very important to note that $\tilde{V}$ is not a function of β, so it has nothing to do with the differentiation with respect to β. We then have

$$\begin{aligned} 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'} \tilde{V}^{-1} \sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\hat{\beta})}{d\beta'} \tilde{V}^{-1} \frac{d\hat{m}(\bar{\beta})}{d\beta} \sqrt{n}(\hat{\beta} - \beta^o). \end{aligned}$$

It follows that for n sufficiently large, we have

$$\begin{aligned}
& \sqrt{n}(\hat{\beta} - \beta^o) \\
= & -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1} \\
& \times \frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\sqrt{n}\hat{m}(\beta^o).
\end{aligned}$$

Hence,

$$\begin{aligned}
& \tilde{V}^{-1/2}\sqrt{n}\hat{m}(\hat{\beta}) \\
= & \tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \\
& + \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o) \\
= & \left[I - \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1/2}\right]\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \\
= & \hat{\Pi}[\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)].
\end{aligned}$$

By the CLT for $\{m_t(\beta^o)\}$ and the Slutsky theorem, we have

$$\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \xrightarrow{d} N(0, I).$$

where I is a $l \times l$ identity matrix. Also, we have

$$\begin{aligned}
\hat{\Pi} &= I - \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1/2} \\
&\xrightarrow{p} I - V_o^{-1/2}D_o(D_o'V_o^{-1}D_o)^{-1}D_o'V_o^{-1/2} \\
&= \Pi,
\end{aligned}$$

where

$$\Pi = I - V_o^{-1/2}D_o(D_o'V_o^{-1}D_o)^{-1}D_o'V_o^{-1/2}$$

is a $l \times l$ symmetric matrix which is also idempotent (i.e., $\Pi^2 = \Pi$) with $\text{tr}(\Pi) = l - K$ (why? Use $\text{tr}(AB) = \text{tr}(BA)$!).

It follows that under correct model specification, we have

$$\begin{aligned} n[\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})] &= [\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)]'\hat{\Pi}^2[\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)] + o_P(1) \\ &\xrightarrow{d} G'\Pi G \\ &\sim \chi^2_{l-K} \end{aligned}$$

by the following lemma, where $G \sim N(0, I)$:

Lemma 8.11 [Quadratic Form in Normal Random Variables]: *If $v \sim N(0, I)$ and Π is an $l \times l$ symmetric and idempotent with rank $q \leq l$, then the quadratic form*

$$v'\Pi v \sim \chi^2_q.$$

Remarks: The adjustment of degrees of freedom from l to $l - K$ is due to the impact of the asymptotically optimal parameter estimator $\hat{\beta}$.

Theorem 8.12 [Overidentification Test] *Suppose Assumptions 8.1–8.6 hold, and $\tilde{V} \xrightarrow{p} V_o$ as $n \to \infty$. Then under the null hypothesis that $E[m_t(\beta^o)] = 0$ for some unknown β^o, the test statistic*

$$n \cdot \hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \xrightarrow{d} \chi^2_{l-K}.$$

Remarks: This is often called the J-test or the test for overidentification in the GMM literature, because it requires $l > K$. This test can be used to check if the model characterized as $E[m_t(\beta^o)] = 0$ is correctly specified.

It is important to note that the fact that

$$n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \to G'\Pi G$$

where Π is an idempotent matrix is due to the fact that $\hat{\beta}$ is an asymptotically optimal GMM estimator that minimizes the objective function $n\hat{m}(\beta)'\tilde{V}^{-1}\hat{m}(\beta)$. If a suboptimal GMM estimator is used, we would have no above result. Instead, we need to use a different asymptotic variance estimator to replace $\tilde{V}$ and obtain an asymptotically χ^2_l distribution under correct model specification. Because the critical value of χ^2_{l-K} is smaller than that of χ^2_l when $K > 0$, the use of the asymptotically optimal estimator $\hat{\beta}$ leads to an asymptotically more efficient test.

Remarks: When $l = K$, the exactly identified case, the moment conditions cannot be tested by the asymptotically optimal GMM $\hat{\beta}$, because $\hat{m}(\hat{\beta})$ will be identically zero, no matter whether $E[m(\beta^o)] = 0$.

Question: Why is the degree of freedom equal to $l - K$?

Answer: The adjustment of degrees of freedom (minus K) is due to the impact of the sampling variation of the asymptotically optimal GMM estimator. In other words, the use of an asymptotically optimal GMM estimator $\hat{\beta}$ instead of $\tilde{\beta}$ renders the degrees of freedom to change from l to $l-K$. Note that if $\hat{\beta}$ is not an asymptotically optimal GMM estimator, the asymptotic distribution of $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ will be changed.

Question: In the J test, why do we use the preliminary weighting matrix $\tilde{V}$, which is evaluated at a preliminary parameter estimator $\tilde{\beta}$? Why not use $\hat{V}$, a consistent estimator for V that is evaluated at the asymptotically optimal estimator $\hat{\beta}$?

Answer: With the preliminary matrix $\tilde{V}$, the J-test statistic is n times the minimum value of the objective function—the quadratic form in the second stage of GMM estimation. Thus, the value of the test statistic $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ is directly available as a by-product of the second stage GMM estimation. For this reason and for its asymptotic χ^2 distribution, the J-test is also called the minimum chi-square test.

Question: Can we use $\hat{V}$ to replace $\tilde{V}$ in the J-test statistic?

Answer: Yes. The test statistic $n\hat{m}(\hat{\beta})'\hat{V}^{-1}\hat{m}(\hat{\beta})$ is also asymptotically χ^2_{l-K} under correct model specification (please verify!), but this statistic is less convenient to compute than $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$, because the latter is the objective function of the second stage GMM estimation. This is analogous to the F-test statistic, which is based on the sums of squared residuals of linear regression models.

Question: Can we replace $\hat{\beta}$ by some suboptimal but consistent GMM estimator $\tilde{\beta}$, say?

Answer: No. We cannot obtain the asymptotically χ^2_{l-K} distribution. We need to replace $\tilde{V}$ in the $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ with a suitable asymptotic variance estimator and will obtain an asymptotic χ^2_l distribution. Note that $\text{avar}(\sqrt{n}\hat{\beta}) \neq \text{avar}(\sqrt{n}\tilde{\beta})$ if $\tilde{\beta}$ is a consistent but suboptimal estimator for β^o.

Testing for Validity of Instruments

In the linear IV estimation context, where

$$m_t(\beta) = Z_t(Y_t - X_t'\beta),$$

the overidentification test can be used to check the validity of the moment condition

$$\begin{aligned} E[m_t(\beta^o)] &= E[Z_t(Y_t - X_t'\beta^o)] \\ &= 0 \text{ for some } \beta^o. \end{aligned}$$

This is essentially to check whether Z_t is a valid instrument vector, that is, whether Z_t is orthogonal to $\varepsilon_t = Y_t - X_t'\beta^o$. Put $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. We can use the following test statistic

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n}$$

Note that the numerator

$$\hat{e}'Z(Z'Z)^{-1}Z'\hat{e} = n \cdot \hat{m}(\hat{\beta}_{2sls})'\hat{W}^{-1}\hat{m}(\hat{\beta}_{2sls})$$

is n times the value of the objective function of the GMM minimization with the choice of $\hat{W} = (Z'Z/n)$, which is an optimal choice when $\{m_t(\beta^o)\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|Z_t) = \sigma^2$). In this case,

$$\frac{\hat{e}'\hat{e}}{n}\frac{Z'Z}{n} \xrightarrow{p} \sigma^2 Q_{zz} = V_o.$$

It follows that the test statistic

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n} \xrightarrow{d} \chi^2_{l-K}$$

under the null hypothesis that $E(\varepsilon_t|Z_t) = 0$ for some β^o.

Corollary 8.13: *Suppose Assumptions 7.1–7.4, 7.6 and 7.7 hold, and $l > K$. Then under the null hypothesis that $E(\varepsilon_t|Z_t) = 0$, the test statistic*

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n} \xrightarrow{d} \chi^2_{l-K},$$

where $\hat{e} = Y - X\hat{\beta}_{2sls}$.

In fact, the overidentification test statistic is equal to nR^2_{uc}, where R^2_{uc} is the uncentered R^2 from the auxiliary regression

$$\hat{e}_t = \alpha' Z_t + w_t.$$

In fact, it can be shown that under the null hypothesis of $E(\varepsilon_t|Z_t) = 0$, nR^2_{uc} is asymptotically equivalent to nR^2 in the sense that $nR^2_{uc} = nR^2 + o_P(1)$, where R^2 is the uncentered R^2 of regressing $\hat{e}_t$ on Z_t.This provides a convenient way to calculate the test statistic. However, it is important to emphasize that this convenient procedure is asymptotically valid only when $E(\varepsilon_t^2|Z_t) = \sigma^2$.

8.9 Empirical Applications

8.10 Conclusion

Most economic and financial theories have implications on and only on a moment restriction

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

where $m_t(\beta)$ is a $l \times 1$ moment function. This moment condition can be used to estimate model parameter β^o via the so-called GMM estimation method. The GMM estimator is defined as:

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where

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

Under a set of regularity conditions, it can be shown that

$$\hat{\beta} \xrightarrow{p} \beta^o$$

and

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

where

$$\Omega = (D_o' W^{-1} D_o)^{-1} D_o' W^{-1} V_o W^{-1} D_o (D_o' W^{-1} D_o)^{-1}.$$

The asymptotic variance Ω of the GMM estimator $\hat{\beta}$ depends on the choice of weighting matrix W. An asymptotically most efficient GMM estimator is to choose $W = V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. In this case, the asymptotic variance of the GMM estimator is given by

$$\Omega_o = (D_o' V_o^{-1} D_o)^{-1}$$

which is a minimum variance. This is similar in spirit to the GLS estimator in a linear regression model. This suggests a two-stage asymptotically optimal GMM estimator $\hat{\beta}$: First, one can obtain a consistent but suboptimal GMM estimator $\tilde{\beta}$ by choosing some convenient weighting matrix $\tilde{W}$. Then one can use $\tilde{\beta}$ to construct a consistent estimator $\tilde{V}$ for V_o, and use it as a weighting matrix to obtain the second stage GMM estimator $\hat{\beta}$.

To construct confidence interval estimators and hypothesis tests, one has to obtain consistent asymptotic variance estimators for GMM estimators. A consistent asymptotic variance estimator for an asymptotically optimal GMM estimator is

$$\hat{\Omega}_o = (\hat{D}' \hat{V}^{-1} \hat{D})^{-1},$$

where

$$\hat{D} = n^{-1} \sum_{t=1}^{n} \frac{dm_t(\hat{\beta})}{d\beta},$$

and the construction of $\hat{V}$ depends on the properties of $\{m_t(\beta^o)\}$, particularly on whether $\{m_t(\beta^o)\}$ is an ergodic stationary MDS process.

Suppose a two-stage asymptotically optimal GMM estimator is used. Then the associated Wald test statistic for the null hypothesis

$$H_0 : R(\beta^o) = r.$$

is given by

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

The moment condition $E[m_t(\beta^o)] = 0$ also provides a basis to check whether an economic theory or economic model is correctly specified. This can be done by checking whether the sample moment $\hat{m}(\hat{\beta})$ is close to zero. A popular model specification test in the GMM framework is the J-test statistic

$$n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \xrightarrow{d} \chi^2_{l-K}$$

under correct model specification, where $\hat{\beta}$ is an asymptotically optimal GMM estimator (question: what will happen if a consistent but suboptimal GMM estimator is used). This is also called the overidentification test. The J-test statistic $n\hat{m}(\hat{\beta})\tilde{V}^{-1}\hat{m}(\hat{\beta})$ is rather convenient to compute, because it is the objective function of the GMM estimator.

GMM provides a convenient unified framework to view most econometric estimators. In other words, most econometric estimators can be viewed as a special case of the GMM framework with suitable choice of moment function and weighting matrix. In particular, the OLS and 2SLS estimators are special cases of the class of GMM estimators.

EXERCISES

8.1. A generalized method of moment (GMM) estimator is defined as

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where β is a $K \times 1$ vector, $\hat{W}$ is a possibly stochastic $l \times l$ symmetric and nonsingular matrix,

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

and $m_t(\beta)$ is a $l \times 1$ moment function of random vector Z_t, and $l \geq K$. We make the following assumptions:

Assumption 1.1: β^o is the unique solution to $E[m(Z_t, \beta^o)] = 0$, and β^o is an interior point in Θ.

Assumption 1.2: $\{Z_t\}$ is a stationary time series process and $m(Z_t, \beta^o)$ is a martingale difference sequence in the sense that

$$E\left[m(Z_t, \beta^o) | \, Z^{t-1} \right] = 0,$$

where $Z^{t-1} = \{Z_{t-1}, Z_{t-2}, ..., Z_1\}$ is the information available at time $t-1$..

Assumption 1.3: $m(Z_t, \beta)$ is continuously differentiable with respect to $\beta \in \Theta$ such that

$$\sup_{\beta \in \Theta} \| \hat{m}'(\beta) - m'(\beta) \| \xrightarrow{p} 0,$$

where $\hat{m}'(\beta) = \frac{d}{d\beta} \hat{m}(\beta)$ and $m'(\beta) = \frac{d}{d\beta} E[m(Z_t, \beta)] = E[\frac{\partial}{\partial \beta} m(Z_t, \beta)]$.

Assumption 1.4: $\sqrt{n} \hat{m}(\beta^o) \xrightarrow{d} N(0, V_o)$ for some finite and positive definite matrix V_o.

Assumption 1.5: $\hat{W} \xrightarrow{p} W$, where W is a finite and positive definite matrix.

From these assumptions, one can show that $\hat{\beta} \xrightarrow{p} \beta^o$, and this result can be used in answering the following questions in parts (a)–(d). Moreover, you can make additional assumptions if you feel appropriate and necessary.

(a) Find the expression of V_o in terms of $m(Z_t, \beta^o)$.
(b) Find the first order condition of the above GMM minimization problem.
(c) Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

(d) Find the optimal choice of $\hat{W}$. Explain why your choice of $\hat{W}$ is optimal.

8.2. (a) Show that the 2SLS $\hat{\beta}_{2sls}$ for the parameter β^o in the regression model $Y_t = X_t'\beta^o + \varepsilon_t$ is a special case of the GMM estimator with suitable choices of moment function $m_t(\beta)$ and weighting matrix $\hat{W}$;

(b) Assume that $\{Z_t\varepsilon_t\}$ is a stationary ergodic process and other regularity conditions hold. Compare the relative efficiency between an asymptotically optimal GMM estimator (with the optimal choice of the weighting matrix) and $\hat{\beta}_{2sls}$ under conditional homoskedasticity and conditional heteroskedasticity respectively.

8.3. Use a suboptimal GMM estimator $\hat{\beta}$ with a given weighting function $\hat{W} \xrightarrow{p} W$ to construct a Wald test statistic for the null hypothesis $\mathbf{H}_0 : R\beta^o = r$, and justify your reasoning. Assume all necessary regularity conditions hold.

8.4. Suppose that $\{m_t(\beta)\}$ is an ergodic stationary MDS process, where $m_t(\cdot)$ is continuous on a compact parameter set Θ, and $\{m_t(\beta)m_t(\beta)'\}$ follows a uniform weak law of large numbers, and $V_o = E[m_t(\beta^o)m_t(\beta^o)']$ is finite and nonsingular. Let $\hat{V} = n^{-1}\sum_{t=1}^n m_t(\hat{\beta})m_t(\hat{\beta})'$, where $\hat{\beta}$ is a consistent estimator of β^o. Show $\hat{V} \xrightarrow{p} V_o$.

8.5. Suppose $\hat{V}$ is a consistent estimator for $V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. Show that replacing $\tilde{V}$ by $\hat{V}$ has no impact on the asymptotic distribution of the overidentification test statistic, that is, show

$$n\hat{m}(\hat{\beta})\tilde{V}^{-1}\hat{m}(\hat{\beta}) - n\hat{m}(\hat{\beta})\hat{V}^{-1}\hat{m}(\hat{\beta}) \xrightarrow{p} 0.$$

Assume all necessary regularity conditions hold.

8.6. Suppose $\tilde{\beta}$ is a suboptimal but consistent GMM estimator. Could we simply replace $\hat{\beta}$ by $\tilde{\beta}$ and still obtain the asymptotic χ^2_{l-K} distribution for the overidentification test statistic? Give your reasoning. Assume all necessary regularity conditions hold.

8.7. Suppose Assumptions 7.1–7.4, 7.6 and 7.7 hold. To test the null hypothesis that $E(\varepsilon_t|Z_t) = 0$, where Z_t is a $l \times 1$ instrumental vector, one can consider the auxiliary regression

$$\hat{e}_t = \alpha' Z_t + w_t,$$

where $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. Show $nR^2_{uc} = nR^2 + o_P(1)$ as $n \to \infty$ under the null hypothesis. [*Hint: Recall the definitions of* R^2_{uc} *and* R^2 *in Chapter 3.*]

8.8 [Nonlinear Least Squares Estimation]. Consider a nonlinear regression model

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

where β^o is an unknown $K \times 1$ parameter vector and $E(\varepsilon_t|X_t) = 0$ a.s. Assume that $g(X_t, \cdot)$ is twice continuously differentiable with respect to β with the $K \times K$ matrices $E[\frac{\partial g(X_t,\beta)}{\partial \beta}\frac{\partial g(X_t,\beta)}{\partial \beta'}]$ and $E[\frac{\partial^2 g(X_t,\beta)}{\partial \beta \partial \beta'}]$ finite and nonsingular for all $\theta \in \Theta$.

The nonlinear least squares (NLS) estimator solves the minimization of the sum of squared residual problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{n} [Y_t - g(X_t, \beta)]^2 .$$

The first order condition is

$$D(\hat{\beta})'e = \sum_{t=1}^{n} \frac{\partial g(X_t, \hat{\beta})}{\partial \beta}[Y_t - g(X_t, \hat{\beta})] = 0,$$

where $D(\beta)$ is a $n \times K$ matrix, with the t-th row being $\frac{\partial}{\partial \beta} g(X_t, \beta)$. This FOC can be viewed as the FOC

$$\hat{m}(\hat{\beta}) = 0$$

for an GMM estimation with

$$m_t(\beta) = \frac{\partial g(X_t, \beta)}{\partial \beta}[Y_t - g(X_t, \beta)]$$

in an exact identification case ($l = K$). Generally, there exists no closed form expression for $\hat{\beta}$. Assume all necessary regularities conditions hold.

(a) Show that $\hat{\beta} \xrightarrow{p} \beta^o$ as $n \to \infty$.

(b) Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

(c) What is the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ if $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.)? Give your reasoning.

(d) What is the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ if $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t) \neq \sigma^2$ a.s.)? Give your reasoning.

(e) Suppose $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.). Construct a test for the null hypothesis $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix such that $R'(\beta^o) = \frac{\partial}{\partial \beta} R(\beta^o)$ is a $J \times L$ matrix with full rank $J \leq L$, and r is a $J \times 1$ nonstochastic vector.

8.9. [Nonlinear IV Estimation] Consider a nonlinear regression model

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

where $g(X_t, \cdot)$ is twice continuously differentiable with respect to β, $E(\varepsilon_t|X_t) \neq 0$ but $E(\varepsilon_t|Z_t) = 0$, where Y_t is a scalar, X_t is a $K \times 1$ vector and Z_t is a $l \times 1$ vector with $l \geq K$.

Suppose $\{Y_t, X_t', Z_t'\}_{t=1}^{\prime n}$ is a stationary ergodic process, and $\{Z_t\varepsilon_t\}$ is an MDS.

The unknown parameter β^o can be consistently estimated based on the moment condition

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

where $m_t(\beta) = Z_t[Y_t - g(X_t, \beta^o)]$. Suppose a nonlinear IV estimator solves the minimization problem

$$\hat{\beta} = \arg\min_{\beta} \hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta),$$

where $\hat{m}(\beta) = n^{-1}\sum_{t=1}^{n} Z_t[Y_t - g(X_t, \beta)]$, and $\hat{W} \xrightarrow{p} W$, a finite and positive definite matrix.

(a) Show $\hat{\beta} \xrightarrow{p} \beta^o$.

(b) Derive FOC.

(c) Derive the asymptotic distribution of $\hat{\beta}$. Discuss the cases of conditional homoskedasticity and conditional heteroskedasticity respectively.

(d) What is the optimal choice of W so that $\hat{\beta}$ is asymptotically most efficient?

(e) Construct a test for the null hypothesis that $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix with $R'(\beta^o)$ of full rank, r is a $J \times 1$ nonstochastic vector, and $J \leq K$.

(f) Suppose $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t) \neq \sigma^2$ a.s.). Construct a test for the null hypothesis $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix such that $R'(\beta^o) = \frac{\partial}{\partial \beta}R(\beta^o)$ is a $J \times L$ matrix with full rank $J \leq L$, and r is a $J \times 1$ nonstochastic vector.

8.10. Consider testing the hypothesis of interest $H_0 : R(\beta^o) = r$ under the GMM framework, where $R(\beta^o)$ is a $J \times K$ nonstochastic matrix, r is a $J \times 1$ nonstochastic vector, and $R'(\beta^o)$ is a $J \times K$ matrix with full rank J, where $J \leq K$. We can construct a Lagrangian multiplier test based on the Lagrangian multiplier $\hat{\lambda}^*$, where $\hat{\lambda}^*$ is the optimal solution of the following constrained GMM minimization problem:

$$(\hat{\beta}^*, \hat{\lambda}^*) = \arg\min_{\beta\in\Theta, \lambda\in R} \left[\hat{m}(\beta)'\tilde{V}^{-1}\hat{m}(\beta) + \lambda'[r - R(\beta)\right],$$

where $\tilde{V}$ is a preliminary consistent estimator for $V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$ that does not depend β. Construct the LM test statistic and derive its asymptotic distribution. Assume all regularity conditions hold.

CHAPTER 9 MAXIMUM LIKELIHOOD ESTIMATION AND QUASI-MAXIMUM LIKELIHOOD ESTIMATION

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.

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.

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

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

where the last equality follows from 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 Modeling] 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 modeling the conditional distribution of Y_t given $\Psi_t = (X_t', Z^{t-1'})'$.

Some Important Special Cases

Case I [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 II: [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, and $f_{X_t|Z^{t-1}}(\cdot|z^{t-1}, \gamma)$ 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 modeling 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 5 [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 $\varepsilon_t \geq c - X_t'\alpha^o$. 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 6 [Loan applications]: Only those successful loan applications will be recorded.

Example 7 [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 9.1 [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 9.1 [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 9.2 [(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 9.2 [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 Weierstrass 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

$$f_{Y_t|\Psi_t}(y|\Psi_t,\beta) = f(y|\Psi_t,\beta), \quad -\infty < y < \infty$$

We first provide a set of regularity conditions.

Assumption 9.1 [Parametric Distribution Model]: (i) $\{Z_t = (Y_t, X_t')'\}_{t=1}^n$ 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.,

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

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.

9.3.1 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 9.3 [Consistency of MLE/QMLE]: *Suppose Assumptions 9.1–9.4 hold. Then as* $n \to \infty$,

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

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 $\hat{\beta} \xrightarrow{p} \beta^*$ as $n \to \infty$.

Model Specification and Interpretation of β^*

Definition 9.3 [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 9.4: *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 from 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

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

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

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

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

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

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

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

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

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

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.

Remarks:

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.

Under Theorem 9.3, we have $\hat{\beta} \xrightarrow{p} \beta^*$ as $n \to \infty$. Furthermore, by correct specification for conditional distribution (i.e., Lemma 9.4), we know $\beta^* = \beta^o$, where β^o is the true model parameter. Thus, we have $\hat{\beta} \xrightarrow{p} \beta^o$ as $n \to \infty$.

This is essentially equivalent to the consistency in the linear regression context, in which, $\hat{\beta}_{OLS}$ always converges to β^* no matter whether the model is correctly specified. And only when the model we have coincides with the true model, we have $\beta^* = \beta^o$ and then $\hat{\beta}_{OLS}$ will converge to the true model parameter β^o.Otherwise, our estimation will be biased since $\hat{\beta}_{OLS}$ does not converge to β^o, as $n \to \infty$.

9.3.2 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 9.5 [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 9.6 [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,

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

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

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

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.,

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

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 $n \to \infty$,

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

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_o 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.

9.3.3 Asymptotic Distribution

Next, we derive the asymptotic normality of the MLE.

Theorem 9.7 [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 $\hat{\beta} - \beta^o \xrightarrow{p} 0$ as $n \to \infty$, 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 $\hat{\beta} - \beta^o \xrightarrow{p} 0$, we have

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

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

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

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.

9.3.4 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 $\hat{\beta} \xrightarrow{p} \beta^o$, we have $\hat{\Omega} \xrightarrow{p} -H_o^{-1}$.

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,

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

where

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

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

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

9.3.5 MLE-based 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 $J \leq K$, 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

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

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

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

By the Slutsky theorem, we have the Wald test statistic

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

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 9.8 [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 $\mathbf{H}_0$

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

as $n \to \infty$,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 9.9 [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 $LR = 2n[\hat{l}(\hat{\beta}) - \hat{l}(\tilde{\beta})]$.

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

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

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 $\text{tr}(ABC) = \text{tr}(BCA)$).

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.

Remarks:

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 $\mathbf{H}_0$ 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 9.10 [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 $\mathbf{H}_0$,

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

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^*, where $\beta^* = \arg\max_{\beta\in\Theta} l(\beta)$ is as in Assumption 9.4 when $f(y|\Psi_t,\beta)$ is misspecified?

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

It should be noted that in QMLE, we no longer have the following equality:

$$\beta^* = \beta^o$$

where β^* is as defined in Assumption 9.4 and β^o is the true model parameter.

Although it always holds that $\hat{\beta}_{QMLE} \xrightarrow{p} \beta^*$, as $n \to \infty$, we no longer have $\hat{\beta}_{QMLE} \xrightarrow{p} \beta^o$, as $n \to \infty$, given that the conditional probability distribution is misspecified.

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

Lemma 9.11: *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

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

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

Remarks:

The parameter value β^* minimizes the "distance" of $f(\cdot|\cdot,\beta^*)$ 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.

In other words, even though β^* does not equal to β^o element by element, we can have equality in some parameters of interests. For example, the first two elements (e.g. $\beta_0^* = \mu_0$ and $\beta_1^* = \sigma_0^2$, where μ_0 and σ_0^2 are the parameters we are interested in) in β^* could be equal to the corresponding elements in β^o, i.e., $\beta_0^* = \beta_0^o$ and $\beta_1^* = \beta_1^o$.Therefore, by using QMLE, we can have an inconsistent estimator $\hat{\beta}_{QMLE}$ in which $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent for the population mean and variance. See Example 1.

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 $E[S_t(\beta^*)|X_t] = 0$ 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 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:

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

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

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:

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

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

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

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

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

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^*, 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 $Y_t|\Psi_t \sim N(\mu_t(\Psi_t, \beta^*), \sigma^2(\Psi_t, \beta^*))$, 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^* 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's$\sqrt{(\nu-2)/\nu} \cdot t(\nu)$ 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 $\hat{\beta}$.

Remarks: The parameter β^* is not always consistently estimable by QMLE when the likelihood function is misspecified. In some cases, β^* 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 9.12: *Suppose Assumptions 9.4–9.6(i) hold. Then*

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

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

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

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

By differentiating, we have

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

Exchanging differentiation and integration yields the desired result:

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

This completes the proof. ■

Remarks:

No matter whether the conditional distributional model $f(y|\Psi_t, \beta)$ is correctly specified, the score function $S_t(\beta^*)$ evaluated at β^* 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 $\{S_t(\beta^*)\}$ is MDS (we note that $S_t(\beta^*)$ 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

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

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_* as long as $\{S_t(\beta^*)\}$ is an MDS process.

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

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

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

where

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

In other words, we have to estimate the long-run variance-covariance matrix for V when $\{S_t(\beta^*)\}$ 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 $E\left[S_t\left(\beta^*\right)|I_{t-1}\right] = 0$ nor

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

where $E(\cdot|\Psi_t)$ is taken under the true conditional distribution which differs from the model $f(y_t|\Psi_t, \beta^*)$ 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 9.13 [Asymptotic Normality of QMLE]: *Suppose Assumptions 9.1–9.6 hold. Then*

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, H_*^{-1}V_*H_*^{-1}),$$

where $V_* = V(\beta^*) \equiv avar[\sqrt{n}\hat{S}(\beta^*)]$ and $H_* = H(\beta^*) \equiv E\left[\frac{\partial^2 \ln f(Y_t|\Psi_t,\beta^*)}{\partial\beta\partial\beta'}|\Psi_t\right]$.

Remarks:
Without the MDS property of the score function, we have to estimate $V_* \equiv \text{avar}[\sqrt{n}\hat{S}(\beta^*)]$ 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 $H_*^{-1}V_*H_*^{-1}$ to $-H_*^{-1}$ 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 $-H_*^{-1}$ of MLE, the asymptotic variance $-H_*^{-1}V_*H_*^{-1}$ 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_* has to be estimated using a kernel-based method when $\{S_t(\beta^*)\}$ is not an MDS.

In the literature, the variance $H_*^{-1}V_*H_*^{-1}$ 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, $H_*^{-1}V_*H_*^{-1}$ 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 $-H_*^{-1}$, 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 $H_*^{-1}V_*H_*^{-1}$ of the QMLE is not the same as the asymptotic variance $-H_*^{-1}$ 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.

9.4.1 Asymptotic Variance Estimation

Question: How to estimate the asymptotic variance $H_*^{-1}V_*H_*^{-1}$ 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 $\hat{H}(\hat{\beta}) \xrightarrow{p} H_*$.

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

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

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

In this case,

$$V_* = E[S_t(\beta^*)S_t(\beta^*)']$$

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 II: When $\{Z_t\}$ is not independent, $\{S_t(\beta^*)\}$ 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_*.

Assumption 9.7: $\hat{V} \xrightarrow{p} V_*$, where V_* is finite and nonsingular.

Lemma 9.14 [Asymptotic Variance Estimator for QMLE]: *Suppose Assumptions 9.1–9.7 hold. Then as* $n \to \infty$,

$$\hat{H}^{-1}(\hat{\beta})\hat{V}\hat{H}^{-1}(\hat{\beta}) \xrightarrow{p} H_*^{-1}V_*H_*^{-1}.$$

9.4.2 QMLE-based 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

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

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

Wald Test Under Model Misspecification

We first consider a Wald test.

Theorem 9.15 [QMLE-based Hypothesis Testing, Wald Test]: *Suppose Assumptions 9.1–9.7 hold. Then under* $\mathbf{H}_0 : R(\beta^*) = r$, *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

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

where we have made use of the fact that$\sqrt{n}(\hat{\beta}-\beta^*) \xrightarrow{d} N(0, H_*^{-1}V_*H_*^{-1})$, 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

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

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

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

where

$$\begin{aligned} \Omega &= [R'(\beta^*)H_*^{-1}R'(\beta^*)']^{-1} \\ &\quad \times R'(\beta^*)H_*^{-1}V_*H_*^{-1}R'(\beta^*)' \\ &\quad \times [R'(\beta^*)H_*^{-1}R'(\beta^*)']^{-1}. \end{aligned}$$

Then a robust LM test statistic

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

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: $\tilde{V} \xrightarrow{p} V_*$, 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 9.16 [QMLE-based LM Test]: *Suppose Assumptions 9.1–9.6 and 9.8 and* $\mathbf{H}_0$: $R(\beta^*) = r$ *holds. Then as* $n \to \infty$,

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

Remarks:

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 $\{S_t(\beta^*)\}$ and the information matrix equality ($V_* = -H_*$), 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 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 I: When $\{Z_t = (Y_t, X_t')'\} \sim$ i.i.d.

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

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

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

$$\begin{aligned} n^{1/2}\hat{m}(\hat{\beta}) &= n^{-1/2} \sum_{t=1}^{n} [m_t(\beta^o) - D_0 H_0^{-1} S_t(\beta^o)] \\ &\xrightarrow{d} N(0, W), \end{aligned}$$

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

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

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

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

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 II: $\{Z_t = (Y_t, X_t')'\}$ 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-square 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 9.17: *If $f(y|\Psi_t, \beta^o)$ coincides with the true conditional pdf of Y_t given Ψ_t, then*

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

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 modeling 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 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.

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) $\text{var}(Y_t|X_t) = \Phi(X_t'\beta^o)[1 - \Phi(X_t'\beta^o)]$.

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

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

(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

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

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.

CHAPTER 10 CONCLUSION

Abstract: In this chapter, we first review what we have covered in the previous chapters, and then discuss other econometric courses needed for various fields of economics and finance.

Key words: Microeconometrics, Financial econometrics, Nonparametric econometrics, Panel data econometrics, Time series econometrics.

10.1 Summary

Question: What have we learnt from this course?

In this chapter, we will first summarize what we have learnt in this book.

The modern econometric theory developed in this book is built upon the following fundamental axioms:

- Any economy can be viewed as a stochastic process governed by some probability law.
- Any economic phenomena can be viewed as a realization of the stochastic economic process.

The probability law of the data generating process can be called the "law of economic motions." The objective of econometrics is to infer the probability law of economic motions using observed data, and then use the obtained knowledge to explain what has happened, to predict what will happen, and to test economic theories and economic hypotheses.

Suppose the conditional pdf $f(y_t|\Psi_t)$ of Y_t given $\Psi_t = (X_t, Z^{t-1})$, is available. Then we can obtain various attributes of the conditional distribution of Y_t given Ψ_t, such as

- conditional mean;
- conditional variance;
- conditional skewness;
- conditional kurtosis;
- conditional quantile.

An important question in economic analysis is: what aspect of the conditional pdf will be important in economics and finance? Generally speaking, the answer is dictated by the nature of the economic problem one has at hand. For example, the efficient market hypothesis states that the conditional expected asset return given the past information is equal to the long-run market

average return; rational expectations theory suggests that conditional expectational errors given the past information should be zero. In unemployment duration analysis, one should model the entire conditional distribution of the unemployment duration given the economic characteristics of the unemployed workers.

It should be emphasized that the conditional pdf or its various aspects only indicate a predictive relationship between economic variables, that is, when one can use some economic variables to predict other variables. The predictive relationship may or may not be the causal relationship between or among economic variables, which is often of central interest to economists. Economic theory often hypothesizes a causal relationship and such economic theory is used to interpret the predictive relationship as a causal relationship.

Economic theory or economic model is not a general framework that embeds an econometric model. In contrast, economic theory is often formulated as a restriction on the conditional pdf or its certain aspect. Such a restriction can be used to validate economic theory, and to improve forecasts if the restriction is valid or approximately valid.
Question: What is the role that economic theory plays in economic modeling?

- Indication of the nature (e.g., conditional mean, conditional variance, etc) of the relationship between Y_t and X_t : Which moments are important and of interest?
- Choice of economic variables X_t.
- Restriction on the functional form or parameters of the relationship.
- Helping judge causal relationships.

In summary, any economic theory can be formulated as a restriction on the conditional probability distribution of the economic stochastic process. Economic theory plays an important role in simplifying statistical relationships so that a parsimoneus econometric model can eventually capture essential economic relationships.

Motivated by the fact that economic theory often has implication on and only on the conditional mean of economic variables of interest, we first develop a comprehensive econometric theory for linear regression models where by linearity we mean the conditional mean is linear in parameters and not necessarily linear in explanatory variables. We start in Chapter 3 with the classical linear regression model, for which we develop a finite sample statistical theory when the regression disturbance is i.i.d. normally distributed, and is independent of the regressor. The normality assumption is crucial for the finite sample statistical theory. The essence of the

classical theory for linear regression models is i.i.d., which implies conditional homoskedasticity and serial uncorrelatedness, which ensures the BLUE property for the OLS estimator. When conditional heteroskedasticity and autocorrelation exist, the GLS estimator illustrates how to restore the BLUE property by correcting conditional heteroskedasticity and differencing out serial correlation.

With the classical linear regression model as a benchmark, we have developed a modern econometric theory for linear regression models by relaxing the classical assumptions in subsequent chapters. First of all, we relax the normality assumption in Chapter 4. This calls for asymptotic analysis because finite sample theory is no longer possible. It is shown that when the sample size is large, the classical results are still approximately applicable for linear regression models with independent observations under conditional homoskedasticity. However, under conditional heteroskedasticity, the classical results, such as the popular t-test and F-test statistics, are no longer applicable, even if the sample size goes to infinity. This is due to the fact that the asymptotic variance of the OLS estimator has a different structure under conditional heteroskedasticity. We need to use White's (1980) heteroskedasticity-consistent variance-covariance estimator and use it to develop robust hypothesis tests. It is therefore important to test conditional homoskedasticity, and White (1980) develops a regression-based test procedure.

The asymptotic theory developed for linear regression models with independent observations in Chapter 4 is extended to linear regression models with time series observations. This covers two types of regression models: one is called a static regression model where the explanatory variables or regressors are exogenous variables. The other is called a dynamic regression model whose regressors include lagged dependent variables and exogenous variables. It is shown in Chapter 5 that when the asymptotic theory of Chapter 4 is applicable when the regression disturbance is a martingale difference sequence. Because of its importance, we introduce tests for martingale difference sequence of regression disturbances by checking serial correlation in the disturbance. The tests include the popular Lagrange multiplier test for serial correlation. We have also considered a Lagrange multiplier test for autoregressive conditional heteroskedasticity (ARCH) and discussed its implication on the inference of static and dynamic regression models respectively.

For many static regression models, it is evident that the regression disturbance displays serial correlation. This affects the asymptotic variance of the OLS estimator. When serial correlation is of a known structure up to a few unknown parameter, we can use the Ornut-Cochrance procedure to obtain asymptotically efficient estimator for regression parameters. When serial correlation is of unknown form, we have to use a long-run variance estimator to estimate the asymptotic variance of the OLS estimator. A leading example is the kernel-based estimator such as the Newey-West variance estimator. With such a variance estimator, robust test procedures for hypotheses of interest can be constructed. These are discussed in Chapter 6.

The estimation and inference of linear regression models are complicated when the condition of $E(\varepsilon_t|X_t) = 0$ does not hold, which can arise due to measurement errors, simultaneous equations bias, omitted variables, and so on. In Chapter 7 we discuss a popular method—the two-stage least squares—to estimate model parameters in such scenarios.

Chapter 8 introduces the GMM method, which is particularly suitable for estimating both linear and nonlinear econometric models that can be characterized by a set of moment conditions. A prime economic example is the rational expectations theory, which is often characterized by an Euler equation. In fact, the GMM method provides a convenient framework to view most econometric estimators, including the least squares, and instrumental variables regression.

Chapter 9 discusses conditional probability distribution models and other econometric models that can be estimated by using pseudo probability likelihood methods. Conditional distribution models have found wide applications in economics and finance, and MLE is the most popular and most efficient method to estimate parameters in conditional distribution models. On the other hand, many econometric models can be conveniently estimated by using a pseudo likelihood function. These include nonlinear least squares, ARMA, GARCH models, as well as limited dependent variables and discrete choice models. Such an estimation method is called the Quasi-MLE. There is an important difference between MLE and QMLE. The forms of their asymptotic variances are different. In certain sense, the asymptotic variance of MLE is similar in structure to the asymptotic variance of the OLS estimator under conditional homoskedasticity and serial uncorrelatedness, while the asymptotic variance of the QMLE is similar in structure to the asymptotic variance of the OLS estimator under conditional heteroskedasticity and autocorrelation.

Chapters 2 to 9 are treated in a unified and coherent manner. The theory is constructed progressively from the simplest classical linear regression models to nonlinear expectations models and then to conditional distributional models. The book has emphasized the important implication of conditional heteroskedasticity and autocorrelation as well as misspecification of conditional distributional models on the asymptotic variance of the related econometric estimators. With a good command of the econometric theory developed in Chapters 2 to 9, we can conduct a variety of empirical analysis in economics and finance, including all motivating examples introduced in Chapter 1. In addition to asymptotic theory, the book has also shown students how to do asymptotic analysis via the progressive development of the asymptotic theory in Chapters 2 to 9. Moreover, we have also introduced a variety of basic asymptotic analytic tools concepts, including various convergence concepts, limit theorems, and basic time series concepts and models.

10.2 Directions for Further Study in Econometrics

The econometric theory presented in this book has laid down a solid foundation in econometric study. However, it does not cover all econometric theory. For example, we only cover stationary time series models, nonstationary time series models, such as unit root models and cointegrated models, have not been covered, which call for a different asymptotic theory (see, e.g., Hamilton 1994). Panel data models also require a separate and independent treatment (see, e.g., Hsiao 2002). Due to the unique features of financial time series, particularly high-frequency financial time series, financial econometrics has emerged as a new field in econometrics that is not covered by standard time series econometrics. On the other hand, although our theory can be applied to models for limited dependent variables and discrete choice variables, more detailed treatment and comprehensive coverage are needed. Moreover, topics on asymptotic analytic tools may be covered to train students' asymptotic analysis ability in a more comprehensive manner.

References

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedastcity", *Journal of Econometrics* 31, 307-327.

Box, G.E.P. and D.A. Pierce (1970), "Distribution of Residual Autorrelations in Autoregressive Moving Average Time Series Models," *Journal of the American Statistical Association* 65, 1509-1526.

Campbell, J.Y. and J. Cochrance (1999), "“By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior” *Journal of Political Economy* 107, 205-251.

Chen, D. and Y. Hong (2003), "Has Chinese Stock Market Become Efficient? Evidence from a New Approach," *China Economic Quarterly* 1 (2), 249-268.

Chow, G. C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica* 28, 591-605.

Cournot, A. (1838), *Researches into the Mathematical Properties of the Theory of Wealth*, trans. Nathaniel T. Bacon, with an essay and an biography by Irving Fisher. McMillan: New York, 2nd edition, 1927.

Cox, D. R. (1972), "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society, Series B,* 34, 187-220,

Engle, R. (1982), "Autoregressive Conditional Hetersokedasticity with Estimates of the Variance of United Kingdom Inflation,” *Econometrica* 50, 987-2008.

Engle, R. and C.W.J. Granger (1987), "Cointegration and Error-Corretion Representation, Estimation and Testing,” *Econometrica* 55, 251-276.

Fisher, I. (1933), "Report of the Meeting," *Econometrica* 1, 92-93

Frisch, R. (1933), “Propagation Problems and Impulse Problems in Dynamic Economics.” In *Economic Essays in Honour of Gustav Cassel*. London: Allen and Unwin, 1933.

Granger, C.J.W. (2001), "Overview of Nonlinear Macroeconometric Empirical Models," *Journal of Macroeconomic Dynamics* 5, 466-481.

Granger, C.J.W. and T. Teräsvirta (1993), *modelling Nonlinear Economic Relationships*, Oxford University Press: Oxford.

Groves, T., Hong, Y., McMillan, J. and B. Naughton (1994), "Incentives in Chinese State-owned Enterprises," *Quarterly Journal of Economics* CIX, 183-209.

Gujarati, D.N. (2006), *Essentials of Econometrics*, 3rd Edition, McGraw-Hill: Boston.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica* 50, 1029-1054.

Hansen, L.P. and K. Singleton (1982), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50, 1269-1286.

Hardle, W. (1990), *Applied Nonparametric Regression*. Cambridge University Press: Cambridge.

Hong, Y. and Y.J. Lee (2005), "Generalized Spectral Testing for Conditional Mean Models in Time Series with Conditional Heteroskedasticity of Unknown Form," *Review of Economic Studies* 72, 499-451.

Hsiao, C. (2003), *Panel Data Analysis, 2nd Edition*, Cambridge University Press: Cambridge.

Keynes, J. M. (1936), *The General Theory of Employment, Interest and Money*, McMillan Cambridge University Press: Cambridge, U.K.

Kiefer, N. (1988), "Economic Duration Data and Hazard Functions," *Journal of Economic Literature* 26, 646-679.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge University Press: Cambridge, U.K.

Lucas, R. (1977), "Understanding Business Cycles," in *Stabilization of the Domestic and International Economy*, Karl Brunner and Allan Meltzer (eds.), Carnegie-Rochester Conference Series on Public Policy, Vol. 5. North-Holland: Amsterdam.

Mehra, R. and E. Prescott (1985), "The Equity Premium: A Puzzle," *Journal of Monetary Economics* 15, 145-161.

Nelson, C.R. and C. I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics* 10, 139-162.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press: Cambridge.

Phillips, P.C. (1987), "Time Series Regression with a Unit Root," *Econometrica* 55, 277-301.

Samuelson, L. (2005), "Economic Theory and Experimental Economics," *Journal of Economic Literature* XLIII, 65-107.

Samuelson, P. (1939), "Interactions Between the Multiplier Analysis and the Principle of Acceleration," *Review of Economic Studies* 21, 75-78.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, edited, with an Introduction, Notes, Marginal Summary and an Enlarged Index, by Edwin Cannan; with an Introduction by Max Lerner. New York :The Modern library, 1937.

Von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press: Princeton.

Walras, L. (1874), *Elements of Pure Economics, or, The Theory of Social Wealth,* translated by William Jaffe. Fairfield, PA; Kelley,1977.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heterokedasticity," *Econometrica* 48, 817-838.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 1-26.

White, H. (1994), *Estimation, Inference and Specification Analysis.* Cambridge University Press: Cambridge.

About the Author: Yongmiao Hong received his Bachelor Degree in Physics in 1985, and his MA degree in Economics in 1988, both from Xiamen University. He received his PHD in Economics from University of California, San Diego, in 1993. In the same year, he became a tenure track assistant professor in Department of Economics, Cornell University, where he became a tenured faculty in 1998, and a full professor in 2001. He has also been a special-term visiting professor in the School of Economics and Management, Tsinghua University since 2002, and a Cheung Kong Visiting Professor in the Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, since 2005. He is the President of the Chinese Economists Society in North America, 2009-2010. Yongmiao Hong's research interests have been econometric theory, time series analysis, financial econometrics, and empirical study on the Chinese economy and financial markets. He has published dozens of academic papers in a number of top academic journals in economics, finance and statistics, such as *Econometrica*, *Journal of Political Economy*, *Journal of Quarterly Economics*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Review of Financial Studies*, *Journal of Econometrics*, *Econometric Theory*, *Biometrika*, *Journal of Royal Statistical Society Series B*, and *Journal of American Statistical Association*.