Chapter 1. Introduction*

The term "econometrics" was first used by Pawel Ciompa in 1910 in a somewhat obscure book published in Germany. To Ciompa, the goals of "oekonometrie" were to describe economic data series mathematically and to display them geometrically and graphically. According to the Nobel Laureate Ragnar Frisch (1936), however, Ciompa's view of econometrics was too narrow, since it emphasized only the descriptive side of econometrics. Writing as founding editor in the inaugural issue of *Econometrica* in 1933, Frisch defined econometrics in more general terms:

> 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 synonomous with the application of mathematics to economics. Experience has shown that each of these three view-points, 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.

To Frisch, econometrics embodies a creative tension between theory and observation:

> Theory, in formulating its abstract quantitative notions, must be inspired to a larger extent by the technique of observation. And fresh statistical and other factual studies must be the healthy element of disturbance that constantly threatens and disquiets the theorist and prevents him from coming to rest on some inherited, obsolete set of assumptions.

With the general notion of econometrics by Frisch in mind, we in this chapter first use a famous example in labor economics to put linear regression (the main topic of this course) in a general framework, then discuss the objective of econometrics and microeconometrics and the role of economic theory in econometrics, followed by main econometric approaches used in this course, and conclude with a summary of notations.

1 Linear Regression and Its Extensions

Suppose we observe $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where y_i is the response variable and $\mathbf{x}_i$ is the covariates. The objective is to study the relationship between y_i and $\mathbf{x}_i$. To be specific, we can think y_i is the wage rate, $\mathbf{x}_i$ includes education and experience, and the target is to study the return to schooling.

*Email: pingyu@hku.hk

The most general model is

$$y = m(\mathbf{x}, \mathbf{u}), \tag{1}$$

where $\mathbf{x} = (x_1, x_2)'$ with x_1 being education and x_2 being experience, $\mathbf{u}$ is a vector of unobservable errors (e.g., the innate ability, skill, quality of education, work ethic, interpersonal connection, preference, and family background), which may be correlated with $\mathbf{x}$ (why?), and $m(\cdot)$ can be any (nonlinear) function. To simplify our discussion, suppose u is one-dimensional and represents the ability of individuals. In this model, the return to schooling is

$$\frac{\partial m(x_1, x_2, u)}{\partial x_1},$$

which depends on the levels of x_1 and x_2 and also u. In other words, for different levels of education, the returns to schooling are different; furthermore, for different levels of experience (which is observable) and ability (which is unobservable), the returns to schooling are also different. This model is called the **nonadditively separable nonparametric model** (NSNM) since u is not additively separable. When u is additively separable, we get the **additively separable nonparametric model** (ASNM),

$$y = m(\mathbf{x}) + u.$$

In this model, the return to schooling is

$$\frac{\partial m(x_1, x_2)}{\partial x_1},$$

which depends only on observables. A special case of this model is the **additive separable model** (ASM) where $m(x) = m_1(x_1) + m_2(x_2)$. In this case, the return to schooling is $\frac{\partial m(x_1)}{\partial x_1}$, which depends only on x_1. There is also the case where the return to schooling depends on the unobservable but not other covariates. For example, suppose

$$y = \alpha(u) + m_1(x_1)\beta_1(u) + m_2(x_2)\beta_2(u),$$

and then the return to schooling is

$$\frac{\partial m_1(x_1)}{\partial x_1}\beta_1(u),$$

which does not depend on x_2 but depend on x_1 and u. A special case of this model is the **random coefficient model** (RCM) of Hildreth and Houck (1968) where $m_1(x_1) = x_1$ and $m_2(x_2) = x_2$.[1] In this case, the return to schooling is $\beta_1(u)$ which depends only on u. Of course, the return to schooling may depend only on x_2 and u. For example, if

$$y = \alpha(x_2, u) + x_1\beta_1(x_2, u),$$

then the return to schooling is $\beta_1(x_2, u)$ which does not depend on x_1. A special case is the **varying**

[1] The RCM dates back as early as Rubin (1950).

coefficient model (VCM) originated by Robinson (1989, 1991), Cleveland et al. (1992) and Hastie and Tibshirani (1993), where

$$y = \alpha(x_2) + x_1\beta_1(x_2) + u,$$

and the return to schooling is $\beta_1(x_2)$ depending only on x_2. When the return to schooling does not depend on either (x_1, x_2) or u, we get the linear regression model (LRM),

$$y = \alpha + x_1\beta_1 + x_2\beta_2 + u \equiv \mathbf{x}'\boldsymbol{\beta} + u,$$

where $\mathbf{x} \equiv (1, x_1, x_2)'$, $\boldsymbol{\beta} \equiv (\alpha, \beta_1, \beta_2)'$, and the return to schooling is β_1 which is constant.

Table 1 summarizes the models above.

x_1	✓	✓	✓		✓			
x_2	✓	✓		✓		✓		
u	✓		✓	✓			✓	
Model	NSNM	ASNM	?	?	ASM	VCM	RCM	LRM

Table 1: Models Based on What the Return to Schooling Depends on

The models in the table can be divided into two subclasses: $\mathbf{x}$ and u are uncorrelated (or even independent) and $\mathbf{x}$ and u are correlated. In the former case, $\mathbf{x}$ is called **exogenous**, and in the latter case, $\mathbf{x}$ is called **endogenous**. Further extensions include models with limited dependent variables (LDV) and multiple equations. We can also study different characteristics of the conditional distribution of y given $\mathbf{x}$. Two popular choices are **conditional mean**

$$m(\mathbf{x}) = E[y|\mathbf{x}] = \int y f(y|\mathbf{x})dy = \int m(\mathbf{x}, u) f(u|\mathbf{x})du$$

and **conditional quantile**

$$Q_\tau(\mathbf{x}) = \inf\{y|F(y|\mathbf{x}) \geq \tau\}, \tau \in (0,1),$$

where $m(\mathbf{x})$ is often called the **conditional expectation function** (CEF), and $Q_{.5}(\mathbf{x})$ is the conditional median. Other characteristics include **conditional variance**

$$\sigma^2(\mathbf{x}) = Var(y|\mathbf{x}) = E\left[(y - m(\mathbf{x}))^2 \middle| \mathbf{x}\right],$$

which measures the dispersion of $f(y|\mathbf{x})$,[2] conditional skewness $E\left[\left(\frac{y-m(\mathbf{x})}{\sigma(\mathbf{x})}\right)^3 \middle| \mathbf{x}\right]$ which measures the asymmetry of $f(y|\mathbf{x})$, and conditional kurtosis $E\left[\left(\frac{y-m(\mathbf{x})}{\sigma(\mathbf{x})}\right)^4 \middle| \mathbf{x}\right]$ which measures the heavy-tailedness of $f(y|\mathbf{x})$. Figure 1 displays the hourly wage densities for male and female workers from the 1985 Current Population Survey (CPS).[3] These are conditional densities - the density of hourly

[2] $\sigma(\mathbf{x}) = \sqrt{\sigma^2(\mathbf{x})}$ is called the **conditional standard deviation**.

[3] The sample has 528 individuals who were full-time employed (defined as those who had worked at least 36 hours

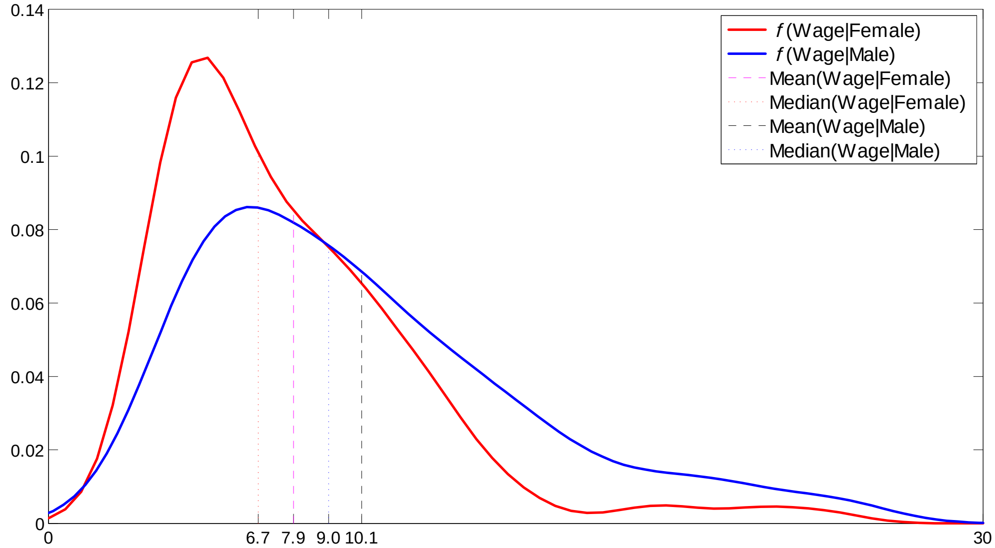

Figure 1: Wage Densities for Male and Female from the 1985 CPS

wages conditional on gender. From this figure, we can read out many interesting features of the female and male wage distributions. First, both mean and median of male wage are larger than those of female wage. Second, for both male and female wage, median is less than mean, which indicates that wage distributions are positively skewed. This is corroborated by the fact that the skewness of both male and female wage is greater than zero (1.0 and 2.9, respectively). Third, the variance of male wage (27.9) is greater than that of female wage (22.4). Fourth, the right tail of male wage is heavier than that of female wage.

Exercise 1 *Suppose that the random variables Y and X only take the values 0 and 1, and have the following joint probability distribution*

	$X = 0$	$X = 1$
$Y = 0$	.1	.2
$Y = 1$	.4	.3

Find $E[Y|X = x]$, $E[Y^2|X = x]$ and $Var(Y|X = x)$ for $x = 0$ and $x = 1$.

Exercise 2 *Suppose $Y|X = x$ follows the uniform distribution on $[0, x + 1]$, and X follows the Bernoulli distribution with the success probability $p = 2/3$. What is the conditional mean function $E[Y|X = x]$? What is the conditional variance function $Var(Y|X = x)$? What is the conditional median function $Med(Y|X = x)$? What is the conditional distribution $X|Y$? What is $E[X|Y = y]$, $Var(X|Y = y)$ and $Med(X|Y = y)$?*

per week for at least 48 weeks the past year), and are not in the military, 244 of which are female and the rest are male.

There are some other simplifications of the general model or combinations of simplified models. For example, the VCM can be simplified to the partially linear model (PLM) of Robinson (1988) where $y = \alpha(x_2) + x_1\beta_1 + u$, $E[u|x_1, x_2] = 0$.[4] Combining the LRM and the ASNM, we get the single index model (SIM) of Ichimura (1993) where $y = m(\mathbf{x}'\boldsymbol{\beta}) + u$, $E[u|x_1, x_2] = 0$.

This course will concentrate on the conditional mean estimation in the linear regression model (with one equation) with and without endogeneity. We will also discuss some LDV models without endogeneity.

2 Econometrics, Microeconometrics and Economic Theory

In modern econometrics, any economy is viewed as a stochastic process $\{W_{it} : t \in (-\infty, \infty), i = 1, \cdots, n_t\}$ which summarizes the economic behavior of all individuals at time t, and any economic phenomenon (i.e., a data set) is viewed as a (partial) *realization* of this stochastic process, where W_{it} can be infinite-dimensional, and n_t is the number of individuals at time t. Typically, three types of data are collected. (i) cross-sectional data. The observations are $\{w_i : i = 1, \cdots, n\}$ at a fixed time point t, where w is a subset of W (e.g., wage, consumption, education, etc) or a transformation of W (e.g., aggregations such as unemployment rates in different countries, consumption at the household level and investment of different coporations), and $n \leq n_t$. (ii) time series data. The observations are $\{w_t : t = 1, \cdots, T\}$ for the same target of interest (e.g., GDP, CPI, stock price, etc), where the time unit can be year, quarter, month, day, hour or even second. (iii) panel data or longitudinal data. The observations are $\{w_{it} : t = 1, \cdots, T; i = 1, \cdots, n\}$. If specify to the setup in the last section, we can think $w = (y, \mathbf{x}')'$. Since w_t in time series data can be a vector or for a group of individuals, the difference between time series data and panel data is blurred. Usually, the distinction is from a technical perspective: if T is much larger than n, the data is treated as time series; if n is much larger than T, the data is treated as a panel. Of course, there is literature considering the case where both n and T are large.

The objective of econometrics is to infer (characteristics of) the probability law of this economic stochastic process (i.e., the data generating process) using observed data, and then use the obtained knowledge to explain what has happened (i.e., *internal validity*), and predict what will happen (i.e., *external validity*). The internal validity concerns three problems: What is a plausible value for the parameter? (point estimation) What are a plausible set of values for the parameter? (set/interval estimation) Is some preconceived notion or economic theory on the parameter "consistent" with the data? (hypothesis testing). In other words, the objectives of econometrics are estimation, inferences (including hypothesis testing and confidence interval (CI) construction) and prediction.

This course will concentrate on microeconometrics, i.e., the main data types analyzed in this course are cross-sectional data and panel data.[5] Our discussion will be close to Hayashi (2000), Cameron and Trivedi (2005), Hansen (2007) and Wooldridge (2010). We also use part of materials from Ruud (2000). Other popular text books include Amemiya (1985), Goldberger (1991), Davidson

[4] Different from the models in Table 1, x_1 and x_2 are not symmetrically treated in the PLM.

[5] Maybe only cross-sectional data will be discussed due to time constraint.

and MacKinnon (1993, 2004), Davidson (1999), Angrist and Pischke (2009) and Greene (2012).

One main objective of microeconometrics is to explore causal relationships between a response variable y and some covariates $\mathbf{x}$.[6] For example, we may be interested in the effect of class sizes on test scores, police expenditures on crime rates, climate change on economic activity, years of schooling on wages, baby-bearing on the labor force participation of women, institutional structure on growth, the effectiveness of rewards on behavior, the consequences of medical procedures on health outcomes, or any variety of possible causal relationships. Sometimes, we estimate parameters that are inputs of the measurements of causal effects; sometimes, our targets are causal effects directly. One caveat is that causality is different from correlation. For example, using umbrellas can predict raining but we cannot claim umbrellas cause raining. Noncausal relationships describe only associations, so are of less economic interests.

(**)Two inherent barriers are that the causal effect is typically specific to an individual and that it is unobserved. Consider the effect of schooling on wages. The causal effect is the actual difference one would receive in wages if we could change his/her level of education *holding all else constant*. This is specific to each individual as their employment outcomes in these two distinct situations is individual. The causal effect is unobserved because we can only observe their actual level of education and actual wage, not the *counterfactual* wage if their education had been different. This is termed as the fundamental problem of causal inference in Holland (1986). Briefly stated, all causal inference involves comparison of a factual with a counterfactual outcome.

A variable x_1 can be said to have a causal effect on the response variable y if y changes with x_1 when all other inputs are held constant. In the formulation of (1), the **causal effect** of x_1 on y is

$$\Delta(x_1, x_2, u) = dm(x_1, x_2, u)/dx_1. \tag{2}$$

Sometimes it is useful to write this relationship as a potential outcome function

$$y(x_1) = m(x_1, x_2, u),$$

where the notation implies that $y(x_1)$ is holding x_2 and u constant. A popular example arises in the analysis of treatment effects with a binary regressor x_1. Often, x_1 is denoted as d. Let $d = 1$ indicate treatment (e.g. a medical procedure or a training program) and $d = 0$ indicate non-treatment (or control). In this case, $y(d)$ can be written as

$$\begin{aligned} y_1 &= m_1(x_2, u), \\ y_0 &= m_0(x_2, u). \end{aligned}$$

In the literature on treatment effects, it is common to refer to y_0 and y_1 as the latent (or potential) outcomes associated with non-treatment and treatment, respectively. This potential outcome

[6] In a letter to J.S. Switzer in 1953, Albert Einstein said, development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

approach goes back at least to J.P. Neyman (1923) and R.A. Fisher (1925, 1935)[7], but the modern version is usually attributed to Rubin (1973a, 1973b, 1974, 1977, 1978). In this case, the causal effect of treatment (or the **treatment effect**) is

$$\Delta(x_2, u) = y_1 - y_0,$$

which is the *ceteris paribus* change of outcomes for an agent across states 0 and 1. $\Delta(x_2, u)$ is random (a function of x_2 and u) as both potential outcomes y_0 and y_1 are different across individuals. Also, we cannot observe both outcomes from the same individual, we only observe the realized value $y = dy_1 + (1 - d)y_0$.

As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the **average treatment effect** (ATE).[8] The ATE of x_1 on y conditional on x_2 is

$$\Delta(x_1, x_2) = E\left[\Delta(x_1, x_2, u)|x_1, x_2\right] = \int \Delta(x_1, x_2, u) f(u|x_1, x_2)du,$$

or

$$\Delta(x_2) = E\left[\Delta(x_2, u)|x_1, x_2\right] = \int \Delta(x_2, u) f(u|x_2)du,$$

where $f(u|x_1, x_2)$ is the conditional density of u given x_1, x_2 and $f(u|x_2)$ is similarly defined. We can think of the ATE $\Delta(x_1, x_2)$ or $\Delta(x_2)$ as the average effect in the general population with a specific value of x_2 and/or x_1.

When conduct a regression analysis (that is, consider the regression of observed wages on educational attainment), we may hope that the regression reveals the average causal effect, that is, $dm(x_1, x_2)/dx_1 = \Delta(x_1, x_2)$ or $m(1, x_2) - m(0, x_2) = \Delta(x_2)$. But this is not generally true.

$$\begin{aligned}
\frac{dm(x_1, x_2)}{dx_1} &= \frac{d \int m(x_1, x_2, u) f(u|x_1, x_2)du}{dx_1} \\
&= \int \frac{dm(x_1, x_2, u)}{dx_1} f(u|x_1, x_2)du + \int m(x_1, x_2, u) \frac{df(u|x_1, x_2)}{dx_1} du \\
&= \Delta(x_1, x_2) + \int m(x_1, x_2, u) \frac{df(u|x_1, x_2)}{dx_1} du,
\end{aligned}$$

so unless $df(u|x_1, x_2)/dx_1 = 0$, $dm(x_1, x_2)/dx_1 \neq \Delta(x_1, x_2)$. In other words, only if $u \perp x_1|x_2$ or u is independent of x_1 conditional on x_2, regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.[9] This condition is not easy to hold. Consider the return to schooling example. This condition means that the education decision does not depend on idiosyncratic characteristics such as expectation of the future wage after controlling observables x_1

[7] Jerzy Neyman (1894-1981) and Ronald A. Fisher (1890-1962) are two iconic founders of modern statistical theory.

[8] The quantile treatment effect (QTE) is also popular nowadays.

[9] The conditional independence notation $u \perp x_1|x_2$ was introduced by Dawid (1979). Note that $u \perp x_1|x_2$ is weaker than $u \perp (x_1, x_2)$. Roughly speaking, u could have correlation with x_1 but only indirectly through x_2. Full independence implies the CIA and implies that each regression derivative equals that variable's average causal effect, but full independence is not necessary in order to causally interpret a subset of the regressors.

and x_2, or the decision of education choice can be fully explained by observables. In the treatment literature, this condition is termed as "*ignorable treatment assignment*", "*conditional independence assumption (CIA)*",[10] "*selection on observables*" or "*unconfoundedness*".

In the linear regression,

$$\Delta(x_1, x_2) = \int \Delta(x_1, x_2, u) f(u|x_1, x_2) du = \int \beta_1 f(u|x_1, x_2) du = \beta_1,$$

and

$$\begin{aligned}\frac{dm(x_1, x_2)}{dx_1} &= \frac{d(x_1\beta_1 + x_2\beta_2 + E[u|x_1, x_2])}{dx_1} \\ &= \beta_1 + \frac{dE[u|x_1, x_2]}{dx_1},\end{aligned}$$

so if $E[u|x_1, x_2] = 0$, the regression coefficients have causal interpretation.

The CIA assumption essentially assumes that we can impose an exogenous variation on x_1, holding other covariates at controlled settings. This can happen in a controlled **social experiment** (famous social experiments include the National Supported Work Demonstration (NSW) program and the National Job Training Partnership Act (JTPA) of 1982), but such experiments are generally expensive to organize and run. Therefore, it is more attractive to implement causal modeling using data generated by a **natural experiment** or **quasi-experiment**, where some causal variable changes exogenously and independently of other explanatory variables, so "naturally" provide treated and untreated subjects. For example, Card and Krueger (1994) estimate the minimum wage effects on employment by noticing that New Jersey increases minimum wage while neighboring Pennsylvania does not, creating a natural experiment in which observations from the "treated" state can be compared with those from the "control" state. More often, program evaluation or treatment evaluation is based on **observational data** (or survey or census data) where the causal variables themselves reflect individual decisions and hence are potentially endogenous. Understanding the individual choice process provides not only estimates of the "effect of cause" but additional insights on the "cause of effect", which are important to the external validity for a new policy in a common environment or existing policies in new environments. See Heckman and Vytlacil (2007a,b) and Imbens and Wooldridge (2009) for a comprehensive summary of the existing literature, but we will not discuss this topic in this course. (**)

Economic theory or model is not a general framework that embeds an econometric model. In contrast, economic theory is often formulated as a restriction on the probability law of the economic stochastic process or the data generating process (DGP). Such a restriction can be used to validate economic theory, and to improve forecasts if the restriction is valid or approximately valid. Usually, the economic theory play the following roles in econometric modeling: (i) indication of the nature (e.g., conditional mean, conditional variance, etc) of the relationship between y and $\mathbf{x}$: which moments are important and of interest? (ii) choice of economic variables $\mathbf{x}$ (e.g., theoretical

[10] CIA is also a short for Central Intelligence Agency which may be more famous.

considerations may suggest that certain variables have no direct effect on others because they do not enter into agents' utility function, nor do they affect the constraints these agents face); (iii) restriction on the functional form or parameters of the relationship; (iv) help judge causal relationship (e.g., whether women's fertility choice affects their employment statuses and hours worked). In summary, any economic theory can be formulated as a restriction on the probability distribution of the economic stochastic process. Economic theory plays an important role in simplifying statistical relationships so that a parsimonious econometric model can eventually capture essential economic relationships.

3 Econometric Approaches

There are two econometric traditions: the *frequentist* approach and the *Bayesian* approach. The former treats the parameter as fixed (i.e., there is only one true value) and the samples as random, while the latter treats the parameter as random and the samples as fixed. This course will concentrate on the frequentist approach. Two main methods in this tradition are the likelihood method and the method of moments (MoM).

The estimator in the likelihood method is called the **maximum likelihood estimator** (MLE). The MLE was recommended, analyzed (with flawed attempts at proofs) and vastly popularized by R.A. Fisher between 1912 and 1922 (although it had been used earlier by Gauss, Laplace, T.N. Thiele,[11] and F.Y. Edgeworth[12]). Much of the theory of maximum-likelihood estimation was first developed for Bayesian statistics, and then simplified by later authors. The basic idea of the MLE is to guess the truth which could generate the phenomenon we observed most likely (practical examples here). Mathematically,

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} E[\ln(f(X|\boldsymbol{\theta}))] = \arg\max_{\boldsymbol{\theta}} \int f(x) \ln f(x|\boldsymbol{\theta}) dx = \arg\max_{\boldsymbol{\theta}} \int \ln f(x|\boldsymbol{\theta}) dF(x), \quad (3)$$

where X is a random vector, $f(x)$ is the true probability density function (pdf) or the true probability mass function (pmf), $f(x|\boldsymbol{\theta})$ is the specified parametrized pdf or pmf, and $F(x)$ is the true cumulative probability function (cdf). Equivalently, $\boldsymbol{\theta}_{MLE}$ minimizes the **entropy** of $f(x|\boldsymbol{\theta})$, $-E[\ln(f(X|\boldsymbol{\theta}))]$. Another explanation of the MLE is to minimize the **Kullback-Leibler information distance** between $f(x)$ and $f(x|\boldsymbol{\theta})$.[13] This distance is defined as

$$KLIC = \int f(x) \ln \left(\frac{f(x)}{f(x|\boldsymbol{\theta})} \right) dx.$$

Exercise 3 *Why are the two definitions of $\boldsymbol{\theta}_{MLE}$ equivalent?*

A useful property of MLEs is what has come to be known as the *invariance property* of MLEs.

[11] Thorvald Nicolai Thiele (1838-1910) was a Danish astronomer, actuary and mathematician. He wa the first to propose a mathematical theory of Brownian motion; he also introduced the cumulants in statistics.

[12] Francis Ysidro Edgeworth (1845-1926) was an Anglo-Irish philosopher and political economistis. He is most famous for Edgeworth box in microeconomics and Edgeworth expansion in econometrics.

[13] Equivalently,

Informally speaking, the invariance property of MLEs says that if $\widehat{\boldsymbol{\theta}}_{MLE}$ is the MLE of $\boldsymbol{\theta}$, then $\tau(\widehat{\boldsymbol{\theta}}_{MLE})$ is the MLE of $\tau(\boldsymbol{\theta})$. Another key advantage of the MLE is that it reaches the so-called **Cramér (1946)-Rao (1945) Lower Bound** (CRLB) asymptotically.[14] The CRLB is the asymptotic variance bound that a "regular"[15] estimator of $\boldsymbol{\theta}$ can reach. Informally speaking, for any regular estimator of $\boldsymbol{\theta}$, say, $\widehat{\boldsymbol{\theta}}$,

$$AVar\left(\widehat{\boldsymbol{\theta}}\right) \geq \text{CRLB},$$

while the MLE reaches this bound, where $AVar(\cdot)$ denotes the asymptotic variance of an estimator.

The MoM estimator was first introduced by Karl Pearson in 1894.[16] The original problem is to estimate k unknown parameters, say $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$, in $f(x)$. However, we are not fully sure about the functional form of $f(x)$. Nevertheless, we know the functional form of the moments of $X \in \mathbb{R}$ as a function of $\boldsymbol{\theta}$:

$$\begin{aligned} E[X] &= g_1(\boldsymbol{\theta}), \\ E[X^2] &= g_2(\boldsymbol{\theta}), \\ &\vdots \\ E[X^k] &= g_k(\boldsymbol{\theta}). \end{aligned} \tag{4}$$

There are k functions with k unknowns, so we can solve out $\boldsymbol{\theta}$ uniquely in principle. The MoM estimator uses only the moment information in X, while the MLE uses "all" information in X, so the MLE is more efficient than the MoM estimator. However, the MoM estimator is more robust than the MLE since it does not rely on the correctness of the full distribution but relies only on the correctness of the moment functions. *Efficiency* and *robustness* are a common trade-off among econometric methods.

The model based on the likelihood method uses all information so usually needs to specify the underlying economic behavior in details. Such kind of model is called the **structural model**. On the other side, the moment equations extract only partial "reliable" information from the full model. Such kind of model is called the **reduced form model**. In econometrics, structural models begin from deductive theories of the economy, while reduced form models begin by identifying *particular* relationships between variables. See Chapter 2 of Cameron and Trivedi (2005) for more discussions on these concepts.

In econometrics, moment conditions often originate from the first order conditions (FOCs) in an optimization problem. Consider the following microeconometric example. Suppose the firms are maximizing their profits conditional on the information in hand; then the problem for firm i is

$$\max_{d_i} E_{\nu|z}\left[\pi(d_i, z_i, \nu_i; \theta)\right]. \tag{5}$$

[14] In some cases, the MLE reaches this bound in finite samples.

[15] In finite samples, change "regular" to "unbiased" and "asymptotic variance" to "variance".

[16] Karl Pearson (1857-1936) is the father of Egon Pearson (1895-1980). The former is also famous for the Pearson correlation coefficient, and the latter is famous for the Neyman-Pearson (1933) Lemma in hypothesis testing.

Here, π is the profit function, e.g.,

$$\pi(d_i, z_i, \nu_i, \theta) = p_i f(L_i, \nu_i; \theta) - w_i L_i,$$

where $z_i = (p_i, w_i)'$ is all information used in decision and can be observed by both the firm and the econometrician, p_i is the output price and w_i is the wage rate, ν_i is the exogenous random error (e.g., weather, financial crisis, etc) and cannot be observed or controlled by either the firm or the econometrician, and $d_i = L_i$ is the decision of labor input. θ is the technology parameter, e.g., if $f(L_i, \nu_i; \theta) = L_i^{\phi} \cdot \exp(\nu_i)$, then $\theta = \phi$, which is known to the firm but unknown to the econometrician. Our goal is to estimate θ, which is relevant to measure the causal effect - the effect of labor input on profit. The FOCs of (5) are

$$E_{\nu|z}\left[\frac{\partial \pi(d_i, z_i, \nu_i, \theta)}{\partial d_i}\right] = m(d_i, z_i|\theta) = 0.$$

If there is randomness even in z_i,[17] then the objective function changes to $\max_{d_i} E_{v,z}\left[\pi(d_i, z_i, \nu_i; \theta)\right]$, and the FOCs change to

$$E\left[m(d_i, z_i|\theta)\right] = 0, \tag{6}$$

which are a special set of moment conditions. In macroeconomics, a model as follows is very standard.

$$\max_{\{c_t\}_{t=1}^{\infty}} \sum_{t=1}^{\infty} \rho^t E_0\left[u\left(c_t\right)\right]$$
$$\text{s.t. } c_{t+1} + k_{t+1} = k_t R_{t+1},\ k_0 \text{ is known},$$

where ρ is the discount factor, $E_0[u(\cdot)]$ is the conditional expected utility based on the information at $t = 0$, k_t is the capital accumulation at time period t, c_t is the consumption at t, and R_t is the gross return rate at t. From dynamic programming, we have the Euler equation

$$E_0\left[\rho \frac{u'\left(c_{t+1}\right)}{u'\left(c_t\right)} R_{t+1}\right] = 1.$$

If $u(c) = \frac{c^{1-\alpha}-1}{1-\alpha}$, $\alpha > 0$, then we get

$$E_0\left[\rho\left(\frac{c_t}{c_{t+1}}\right)^{\alpha} R_{t+1}\right] = 1. \tag{7}$$

Suppose ρ is known while α is unknown; then (7) is a moment condition for α.

(4), (6) and (7) are the population version of moment conditions. Although some econometricians treat "population" as a physical population (e.g., all individuals in the US census) in the real world, the term "population" is often treated *abstractly*, and is potentially infinitely large. Since the population distribution is unknown, we cannot solve the population moment conditions to estimate the parameters. In practice, we often have a set of data points from the population, so we

[17] The difference between z_i and ν_i is that z_i can be observed *ex post* while ν_i cannot. That z_i is random means that the decision is made before z_i is revealed, or the decision is made *ex ante*.

can substitute the population distribution in the moment conditions by the empirical distribution of the data. This is the **analog method** advocated in Manski (1988, 1994). Specifically, suppose the true distribution of a random vector X satisfies the following moment conditions,

$$E\left[m(X|\boldsymbol{\theta}_0)\right] = \mathbf{0},$$

i.e.,

$$\int m(x|\boldsymbol{\theta}_0)dF(x) = \mathbf{0},$$

where $m : \Theta \subset \mathbb{R}^k \to \mathbb{R}^k$, and $F(\cdot)$ is the true cumulative probability function (cdf) of X. Here, you need to pay attention to our notations. In elementary econometrics,

$$E\left[m(X|\boldsymbol{\theta}_0)\right] = \begin{cases} \int m(x|\boldsymbol{\theta}_0)f(x)dx, & \text{if } X \text{ is continuous,} \\ \sum_{j=1}^{J} m(x_j|\boldsymbol{\theta}_0)p_j, & \text{if } X \text{ is discrete,} \end{cases}$$

where $f(x)$ is the probability density function (pdf) of X, and $\{p_j = P(X = x_j)|j = 1, \cdots, J\}$ is the probability mass function (pmf) of X. We write $E\left[m(X|\boldsymbol{\theta}_0)\right]$ as a Riemann–Stieltjes integral $\int m(x|\boldsymbol{\theta}_0)dF(x)$ to cover both cases (and even more general cases). Substituting $F(\cdot)$ by $\widehat{F}_n$, the empirical distribution[18], we have

$$\int m(x|\boldsymbol{\theta})d\widehat{F}_n(x) = \mathbf{0},$$

which is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i|\boldsymbol{\theta}) = \mathbf{0}. \tag{8}$$

The MoM estimator $\widehat{\boldsymbol{\theta}}\left(X_1, \cdots, X_n\right)$ is the solution to (8). Similarly, the MLE can be constructed as the maximizer of the average log-likelihood

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln f\left(X_i|\boldsymbol{\theta}\right),$$

which is equivalent to the maximizer of the log-likelihood function

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f\left(X_i|\boldsymbol{\theta}\right)$$

or the likelihood function

$$L_n(\boldsymbol{\theta}) = \exp\left\{\mathcal{L}_n(\boldsymbol{\theta})\right\} = \prod_{i=1}^{n} f\left(X_i|\boldsymbol{\theta}\right).$$

[18] Recall that $\widehat{F}_n(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$.

Note that if $f(x|\boldsymbol{\theta})$ is smooth in $\boldsymbol{\theta}$, the FOCs for the MLE are

$$\frac{1}{n}\sum_{i=1}^{n} s(X_i|\boldsymbol{\theta}) = \mathbf{0},$$

where $s(\cdot|\boldsymbol{\theta}) = \partial \ln f(\cdot|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is called the **score function** in the likelihood literature. So the MLE is a special MoM estimator in this case. However, $f(x|\boldsymbol{\theta})$ can be discontinuous as a function of $\boldsymbol{\theta}$, so the objective function of the MLE is more general than that of the MoM estimator. In statistics, the former is termed as the **M-estimator** ("M" for maximization), and the later is termed as the **Z-estimator** ("Z" for zero).

The modern version of the likelihood method in the semiparametric setup is the empirical likelihood method introduced by Owen (1988, 2001); see also Gallant and Nychka (1987) for the semi-nonparametric maximum likelihood estimation and Geman and Hwang (1982) for the non-parametric maximum likelihood estimation. The modern version of the method of moments is the generalized method of moments (GMM) introduced by Hansen (1982). We will cover only the GMM method in this course.

Another principle, which is useful especially in linear models, is projection. We will dicuss this principle in the next chapter. This principle provides more straightforward interpretations of the above-mentioned estimators by geometric intuitions.

4 Notations

Real numbers are written using lower case italics. Vectors are defined as column vectors and represented using lowercase bold. For example, in linear regression the regressor vector $\mathbf{x}$ is a $k \times 1$ column vector with jth entry x_j and the parameter vector $\boldsymbol{\beta}$ is a $k \times 1$ column vector with jth entry β_j, i.e.,

$$\underset{(k\times 1)}{\mathbf{x}} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \text{ and } \underset{(k\times 1)}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Then the linear regression model $y = \beta_1 x_1 + \cdots + \beta_k x_k + u$ is expressed as $y = \mathbf{x}'\boldsymbol{\beta} + u$. Without further specification, $x_1 = 1$. So we can express $\mathbf{x} = (1, \underline{\mathbf{x}}')'$ and $\boldsymbol{\beta} = (\beta_1, \underline{\boldsymbol{\beta}}')'$, where $\underline{\mathbf{x}} = (x_2, \cdots, x_k)'$ and $\underline{\boldsymbol{\beta}} = (\beta_2, \cdots, \beta_k)'$. At times a subscript i is added to denote the typical ith observation. The linear regression equation for the ith **observation** is then

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i.$$

In this course observations are assumed to be independent over i (not between y_i and $\mathbf{x}_i$!) since we consider only cross-sectional data. Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case we say that the data are **independent and identically distributed**, or iid. We call this a **random**

sample.

Matrices are represented using uppercase bold. In matrix notation the **sample** (**data**, or **dataset**) is $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is an $n \times 1$ vector with ith entry y_i and $\mathbf{X}$ is a matrix with ith row $\mathbf{x}_i'$, i.e.,

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ and } \underset{(n\times \dim(\mathbf{x}))}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix},$$

where the first column of $\mathbf{X}$ is assumed to be ones if without further specification, i.e., the first column of $\mathbf{X}$ is

$$\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The bold zero, $\mathbf{0}$, denotes a vector or matrix of zeros. Usually the dimensions will be clear from the context. Sometimes, we need to express $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_k \end{pmatrix},$$

where different from $\mathbf{x}_i$, $\mathbf{X}_j$, $j = 1, \cdots, k$, represents the jth column of $\mathbf{X}$ and is all the observations for jth variable. The linear regression model upon stacking all n observations is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{u}$ is an $n \times 1$ column vector with ith entry u_i.

Sometimes, the true value of a parameter is indexed by a subscript 0. When the subscript 0 is absent, it means that the result holds for a *generic* true value of the parameter. Sometimes (especially in nonlinear models), the capital letters such as X denote random variables or random vectors and the corresponding lower case letters such as x denote the potential values they may take. Generic notation for a parameter in nonlinear environments (e.g., nonlinear models or nonlinear constraints) is $\boldsymbol{\theta}$, while in linear environments is $\boldsymbol{\beta}$. All notations should be clear from the context.

Usually, $\|\cdot\|$ means the Euclidean norm. For a vector $\mathbf{a} \in \mathbb{R}^n$, $\|\mathbf{a}\| = \left(\sum_{i=1}^n a_i^2\right)^{1/2}$. For a $m \times n$ matrix $\mathbf{A}$, $\|\mathbf{A}\| = \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2\right)^{1/2} = [\text{trace}(\mathbf{A}'\mathbf{A})]^{1/2}$,[19] where for an $n \times n$ matrix $\mathbf{A}$, $\text{trace}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$. For an $n \times n$ symmetric matrix $\mathbf{A}$, $\mathbf{A} \geq 0$ means $\mathbf{A}$ is positive semi-definite; $\mathbf{A} > 0$ means $\mathbf{A}$ is positive definite; $\mathbf{A} \leq 0$ and $\mathbf{A} < 0$ are similarly defined. For a vector $\mathbf{a} \in \mathbb{R}^n$, $\text{diag}(\mathbf{a})$ means a diagonal matrix with diagonal elements $a_1, \cdots, a_n$. $\mathbf{I}_n$ is the $n \times n$ identity matrix. All convergences are as $n \to \infty$, so we do not write out "$n \to \infty$" explicitly throughout the course.

$\square$ is used to signal the end of an example, and $\blacksquare$ the end of a proof. $\equiv$ means "defined as".

Analytical exercises are given in the relevant context. Empirical exercises are given at the end of each chapter.

[19] This matrix norm is also called the **Frobenius norm** or the **Hilbert–Schmidt norm**.

This course will concentrate on linear models. Nonlinear models will be briefly discussed with emphasis on intuition rather than rigorous proof. Sections, proofs, exercises or footnotes indexed by * are optional. Paragraphs started and ended with (**) are also optional.