

# Introduction

Ping Yu

School of Economics and Finance  
The University of Hong Kong

## Course Information

- Time: 9:30-10:45am and 11:00-12:15pm on Saturday.
- Location: KKL103
- Office Hour: 2:00-3:00pm on **Friday**, KKL1108 (extra?)
- Textbook: My lecture notes posted on Moodle.
  - Others: Hayashi (2000), Ruud (2000), Cameron and Trivedi (2005), Hansen (2007) and Wooldridge (2010).
  - References: no need to read unless necessary.
- Exercises: no need to turn in, but necessary to pass this course.
  - Solve the associated analytical exercises in the lecture notes covered during this week, and I will post answer keys to these exercises **just** before the next class (remind me if I did not).
  - At the end of each chapter, there are one or two empirical exercises. Do them only after the whole chapter is finished. Use matrix programming languages such as Matlab or Gauss; do not use Stata like softwares.
- Evaluation: Midterm Test (40%), Final Exam (60%)
- The exams are open-book and open-note.

## What is Econometrics? What will This Course Cover?

- Ragnar Frisch (1933): unification of statistics, economic theory, and mathematics.
- Linear regression and its extensions.
- The objective of econometrics and microeconometrics, and the role of economic theory in econometrics.
- Main econometric approaches.
  
- We will concentrate on **linear** models, i.e., linear regression and linear GMM, in this course. Nonlinear models are discussed only briefly.
- Sections, proofs, exercises, paragraphs or footnotes indexed by \* are optional and will not be covered in this course.
- I may neglect or add materials beyond my notes (depending on your backgrounds and time constraints). Just follow my slides and read the corresponding sections.

# Linear Regression and Its Extensions

## Return to Schooling: Our Starting Point

- Suppose we observe  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ , where  $y_i$  is the wage rate,  $\mathbf{x}_i$  includes education and experience, and the target is to study the return to schooling or the relationship between  $y_i$  and  $\mathbf{x}_i$ .
- The most general model is

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

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.

- **Notations:** real numbers are written using lower case italics. Vectors are defined as column vectors and represented using lowercase bold.

## Nonadditively Separable Nonparametric Model

- In (1), 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 **NSNM** since  $u$  is not additively separable.

# 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  $m(\mathbf{x}) = m_1(x_1) + m_2(x_2)$ .
- In this case, the return to schooling is  $\frac{\partial m(x_1)}{\partial x_1}$ , which depends only on  $x_1$ .

## Random Coefficient Model

- 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 is the **RCM** where  $m_1(x_1) = x_1$  and  $m_2(x_2) = x_2$ .
- In this case, the return to schooling is  $\beta_1(u)$  which depends only on  $u$ .



## Varying Coefficient Model

- 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 **VCM** 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  $x_2$ .

## Linear Regression Model

- When the return to schooling does not depend on either  $(x_1, x_2)$  or  $u$ , we get the **LRM**,

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

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

- Summary:

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

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

- Other popular models:
  - The VCM can be simplified to the **partially linear model** (PLM), where  $y = \alpha(x_2) + x_1\beta_1 + u$ .
  - Combining the LRM and the ASNM, we get the **single index model** (SIM) where  $y = m(\mathbf{x}'\beta) + u$ .

## Other Dimensions

- $\mathbf{x}$  and  $u$  are uncorrelated (or even independent) and  $\mathbf{x}$  and  $y$  are correlated. In the former case,  $\mathbf{x}$  is called **exogenous**, and in the latter case,  $\mathbf{x}$  is called **endogenous**.
- Limited dependent variables (LDV): part information about  $y$  is missing.
- Single equation versus Multiple equations.
- Different characteristics of the conditional distribution of  $y$  given  $\mathbf{x}$ .
  - **Conditional mean or conditional expectation function (CEF)**

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

where  $f(y|\mathbf{x})$  is the conditional probability density function (pdf) or the conditional probability mass function (pmf) of  $y$  given  $\mathbf{x}$ .

- **Conditional quantile**

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

where  $F(y|\mathbf{x})$  is the conditional cumulative probability function (cdf) of  $y$  given  $\mathbf{x}$ . Especially,  $Q_{.5}(\mathbf{x})$  is the conditional median.

## What We will Cover?

### - Conditional variance

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

which measures the dispersion of  $f(y|\mathbf{x})$ .

- **Conditional skewness**  $E \left[ \left( \frac{y - m(\mathbf{x})}{\sigma(\mathbf{x})} \right)^3 \middle| \mathbf{x} \right]$  which measures the asymmetry of  $f(y|\mathbf{x})$ .

- **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  $f(y|\mathbf{x})$ .

• LRM × Conditional Mean × Exogeneity × Single Equation  
 ∴ ∴ × Endogeneity × Multiple Equation

• LDV × Conditional Mean × Exogeneity × Single Equation  
 ∴ ∴ × Endogeneity × Multiple Equation

## A Real Example

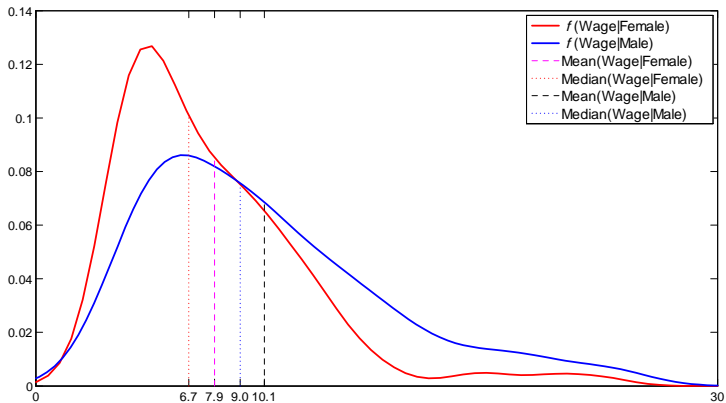


Figure: Wage Densities for Male and Female from the 1985 CPS

## What Can We Get From The Figure?

- These are conditional densities - the density of hourly wages conditional on gender.
- 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.

# Econometrics, Microeconometrics and Economic Theory

## Econometric Data Types

- In modern econometrics, any economy is viewed as a stochastic process  $\{W_{it} : t \in (-\infty, \infty), i = 1, \dots, n_t\}$  which summarizes the economic behavior of **all** individuals at time  $t$ , where  $W_{it}$  can be infinite-dimensional, and  $n_t$  is the number of individuals at time  $t$ .
- **Any** economic phenomenon (i.e., a data set) is viewed as a (partial) *realization* of this stochastic process.
- Typically, three types of data are collected.
  - **Cross-sectional** data. The observations are  $\{w_i : i = 1, \dots, n\}$  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 corporations), and  $n \leq n_t$ .
  - **Time series** data. The observations are  $\{w_t : t = 1, \dots, T\}$  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.
  - **Panel** data or **longitudinal** data. The observations are  $\{w_{it} : t = 1, \dots, T; i = 1, \dots, n\}$ .
- If specify to the setup in the return-to-schooling example, we can think  $w = (y, \mathbf{x}')'$ .



# The Objective of Econometrics

- The objective of econometrics is to infer (characteristics of) the probability law of this economic stochastic process (i.e., the **data generating process** or **DGP**) 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.

## The Objective of Microeconometrics

- This course will concentrate on microeconometrics, i.e., the main data types analyzed in this course are cross-sectional data and panel data.<sup>1</sup>
- One main objective of microeconometrics is to explore causal relationships between a response variable  $y$  and some covariates  $\mathbf{x}$ .
  - 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
- **Caveat:** causality is different from correlation.
  - using umbrellas can predict raining but we cannot claim umbrellas cause raining.
- Noncausal relationships describe only associations, so are of less economic interests.

---

<sup>1</sup>Maybe only cross-sectional data will be discussed due to time constraint.

## Roles of Economic Theory

- 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 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:
  - 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?
  - choice of economic variables  $\mathbf{x}$  (e.g., theoretical 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);
  - restriction on the functional form or parameters of the relationship (e.g., for Cobb-Douglas production function,  $Y = AL^{\beta_1} K^{\beta_2}$ , constant-return-to-scale implies that  $\beta_1 + \beta_2 = 1$ );
  - help judge causal relationship (e.g., whether women's fertility choice affects their employment statuses and hours worked).

# 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.
  - the latter treats the parameter as random and the samples as fixed.
- This course will concentrate on the frequentist approach.
- Two main methods in the frequentist approach are the **likelihood method** and the **method of moments** (MoM).
- We will concentrate on the MoM and only briefly discuss the likelihood method.
- We will use the estimation problem to illustrate these two methods.

## The Maximum Likelihood Estimator

- The MLE was popularized by R.A. Fisher (1890-1962), one iconic founder of modern statistical theory.
- The basic idea of the MLE is to guess the truth which could generate the phenomenon we observed most likely (practical examples here).
- Mathematically,

$$\theta_{MLE} = \arg \max_{\theta} E[\ln(f(X|\theta))] = \arg \max_{\theta} \int f(x) \ln f(x|\theta) dx = \arg \max_{\theta} \int \ln f(x|\theta) dF(x) \quad (2)$$

where  $X$  is a random vector,  $f(x)$  is the true pdf or the true pmf,  $f(x|\theta)$  is the specified parametrized pdf or pmf, and  $F(x)$  is the true cdf.

- Another explanation of the MLE is to minimize the **Kullback-Leibler information distance** between  $f(x)$  and  $f(x|\theta)$ ,

$$KLIC = \int f(x) \ln \left( \frac{f(x)}{f(x|\theta)} \right) dx = E[\ln(f(X)) - \ln(f(X|\theta))].$$

- Two Good Properties of the MLE: (i) *Invariant*: if  $\hat{\theta}_{MLE}$  is the MLE of  $\theta$ , then  $\tau(\hat{\theta}_{MLE})$  is the MLE of  $\tau(\theta)$ . (ii) *Efficient*: it reaches the **Cramér-Rao Lower Bound** (CRLB) asymptotically.

# The MoM Estimator

- The MoM estimator was invented by Karl Pearson (1857-1936).
- The original problem is to estimate  $k$  unknown parameters, say  $\theta = (\theta_1, \dots, \theta_k)$ , in  $f(x)$ . But 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  $\theta$ :

$$\begin{aligned} E[X] &= g_1(\theta), \\ E[X^2] &= g_2(\theta), \\ &\vdots \\ E[X^k] &= g_k(\theta). \end{aligned} \tag{3}$$

There are  $k$  functions with  $k$  unknowns, so we can solve out  $\theta$  uniquely in principle.

## Efficiency and Robustness

- 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.



## A Microeconomic Example

- Moment conditions often originate from the first order conditions (FOCs) in an optimization problem.
- Suppose the firms are maximizing their profits conditional on the information in hand; then the problem for the firm  $i$  is

$$\max_{d_i} E_{v|z} [\pi(d_i, z_i, v_i; \theta)]. \quad (4)$$

- $\pi$  is the profit function, e.g.,

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

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,  $v_i$  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.

- $\theta$  is the technology parameter, e.g., if  $f(L_i, v_i; \theta) = L_i^\phi \cdot \exp(v_i)$ , then  $\theta = \phi$ , and is known to the firm but unknown to the econometrician. Our goal is to estimate  $\theta$ , which is relevant to measure the causal effect - the effect of labor input on profit.

continue...

- The first-order conditions (FOCs) of (4) are

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

- When there is randomness even in  $z_i$ ,<sup>2</sup> then the objective function changes to

$$\max_{d_i} E[\pi(d_i, z_i, v_i; \theta)],$$

and the FOCs change to

$$E[m(d_i, z_i | \theta)] = 0, \tag{5}$$

which are a special set of moment conditions.

---

<sup>2</sup>The difference between  $z_i$  and  $v_i$  is that  $z_i$  can be observed *ex post* while  $v_i$  cannot. That  $z_i$  is random means that the decision is made before  $z_i$  is revealed, or the decision is made *ex ante*.

## A Macroeconomic Example

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

- $\rho$  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

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

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

$$E_0 \left[ \rho \left( \frac{c_t}{c_{t+1}} \right)^\alpha R_{t+1} \right] = 1. \quad (6)$$

- Suppose  $\rho$  is known while  $\alpha$  is unknown; then (6) is a moment condition for  $\alpha$ .

## Population Version vs Sample Version of Moment Conditions

- (3), (5) and (6) 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 can substitute the population distribution in the moment conditions by the empirical distribution of the data, which generates the **sample version** of the moment conditions.
- This is called the **analog method**.

## (The Sample Version of) the MoM Estimator

- Suppose the true distribution of  $X$  satisfies

$$E[m(X|\theta_0)] = \mathbf{0} \text{ or } \int m(x|\theta_0) dF(x) = \mathbf{0},$$

where  $m: \Theta \subset \mathbb{R}^k \rightarrow \mathbb{R}^k$ , and  $F(\cdot)$  is the true cdf of  $X$ .

- **Notations:**

$$E[m(X|\theta_0)] = \begin{cases} \int m(x|\theta_0) f(x) dx, & \text{if } X \text{ is continuous,} \\ \sum_{j=1}^J m(x_j|\theta_0) p_j, & \text{if } X \text{ is discrete,} \end{cases}$$


where  $f(x)$  is the pdf of  $X$ , and  $\{p_j = P(X = x_j) | j = 1, \dots, J\}$  is the pmf of  $X$ . We write  $E[m(X|\theta_0)]$  as  $\int m(x|\theta_0) dF(x)$  to cover both cases.

- The essence of the MoM estimator is to substitute the true distribution  $F(\cdot)$  by the empirical distribution  $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$ :

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

which is equivalent to

$$\frac{1}{n} \sum_{i=1}^n m(X_i|\theta) = \mathbf{0}. \quad (7)$$

- The MoM estimator  $\hat{\theta}(X_1, \dots, X_n)$  is the solution to (7). 

## (The Sample Version of) the MLE

- Similarly, the MLE can be constructed as the maximizer of the average log-likelihood function

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(X_i|\theta),$$

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

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

or the likelihood function

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

- If  $f(x|\theta)$  is smooth in  $\theta$ , the FOCs for the MLE are

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

where  $s(\cdot|\theta) = \partial \ln f(\cdot|\theta) / \partial \theta$  is called the **score function**.<sup>3</sup> So the MLE is a special MoM estimator in this case.

<sup>3</sup>More often,  $\sum_{i=1}^n s(X_i|\theta)$  is called the score function.

## Extensions of the Two Methods

- (Parametric) Likelihood

$$\rightarrow \left( \begin{array}{ll} \text{semi-parametric:} & \text{empirical likelihood} \\ \text{semi-nonparametric:} & \text{semi-nonparametric likelihood} \\ \text{nonparametric:} & \text{nonparametric likelihood} \end{array} \right)$$

- MoM  $\rightarrow$  GMM

- We will cover only the GMM method in this course.

- Another principle, which is useful especially in linear models, is **projection**, which is the topic of our next chapter.

- This principle provides more straightforward interpretations of the above-mentioned estimators by geometric intuitions.

- Keep the three principles in mind: likelihood, GMM and projection.