Econometric Analysis of Cross Section and Panel Data

Lecture 1: Introduction, CEF and Projection

Zhian Hu

Central University of Finance and Economics

Autumn 2024

This Lecture

Hansen (2022): Chapter 1 and 2

- Basic econometric concepts
- Conditional expectation function
- Best linear predictor

Econometrics

Econometrics typically studies the following model:

$$h(X, Y, U) = 0.$$

- We don't know the joint distribution of (X, Y, U).
- We have observed a **sample** of (X, Y). But we are not able to observe U.
- We want to learn something about $h(\cdot)$ according to the sample.

Econometrics

In this course we only study parametric models half-parametric and non-parametric

$$h(X, Y, U; \beta) = 0$$

norm distribution a typical parametric model

where the functional form of $h(\cdot)$ is known up to a parameter β. More specifically, we will be focusing on linear models with various assumptions of U:

linear is the linear parameter, and n dimension of x_n is allowed

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

Gaussion-Markov assumption:
1.random sample, or observation are iid(为满足大数定理和中心极限定理以便估计\beta)
2.linear 3.不完全共线(ols中求(伪)逆) 4.E(U|X)=0均值独立 无偏性,E(\beat|X)=\beta;大样本下无偏性对应一致性

- The estimation of β is aimed at finding a statistic $\hat{\beta}$ that is "reasonably" close to β 's true value.
- The statistical inference on β involves hypothesis testing and constructing confidence intervals for β.

假设E(U|X)强于E(UX)，若E(UX)=0不满足，OLS失效，而使用工具变量方法

Observational Data

- A common econometric question is to quantify the causal impact of one set of variables on another variable.
  - the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant
- Experimental Data & Observational Data
- Experimental Data: draw data from experiments (Randomized Control Trials are prevailing)
  - Banerjee and Duflo, *Poor Economics* (2019 Nobel Prize with Michael Kremer)
  - Money burning, immoral, general equilibrium effects, external validity concern
- Observational Data: survey data, census data （人口普查）
  - draw balls from a black box, call people to survey their characteristics

实验方法：不道德，贵，一般均衡效应（发钱扰动市场），不具有现实指导意义

Standard Data Structures

1. **Cross-sectional data:** one observation per individual
   - survey data (one period), census data
   - It is conventional to assume that cross-sectional observations are mutually independent.
2. **Times series data:** several periods for one variable
   - macroeconomic aggregates, prices, and interest rates
   - Serial dependence: $Y_t = \rho Y_{t-1} + e_t$
3. **Panel data:** a set of individuals measured repeatedly over time
   - independent across individuals, but dependent across periods for a given individual
4. **Clustered sample**
5. **Spatial dependent sample**

Standard Data Structures

- **Independently Distributed:** i^{th} observation (Y_i, X_i) is independent of j^{th} observation (Y_j, X_j)
- **Identically Distributed:** Observations are drawn from the same probability distribution.
  - The variables (Y_i, X_i) are a sample from the distribution $F(Y, X)$ if they are identically distributed with distribution $F(Y, X)$.
  - $F(Y, X)$ is called as **the population**.
- Think of drawing balls from one box or two box with different ball-color distribution
- **Independent and identically distributed, i.i.d., or a random sample:** The variables (Y_i, X_i) are a random sample if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.

Standard Data Structures

- **The distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample.**
- The assumption of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.
- Law of Large Number, Central Limit Theorem

The Distribution of Wages

- Remember, we treat variables as random!
- For example, we view the wage of an individual worker as a random variable *wage* with the probability distribution

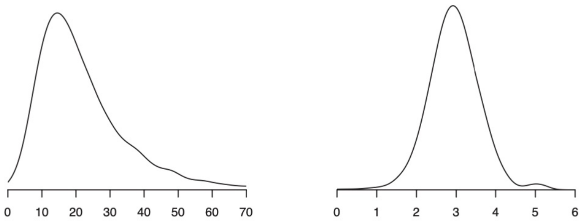
$$F(y) = \mathbb{P}[wage \leq y]$$

Condition expectation(CEF):E(e|x)=0
BLP: E(xe)=0

- Probability density function:

$$f(y) = \frac{d}{dy} F(y)$$

The Distribution of Wages



(a) Wage Density (b) Log Wage Density

Figure 2.1: Density of Wages and Log Wages

- Median m: $F(m) = \frac{1}{2}$
- Mean or expectation of a random variable Y:
  - Discrete: $\mu = \mathbb{E}[Y] = \sum_{j=1}^{\infty} \tau_j \mathbb{P}[Y = \tau_j]$
  - Continuous: $\mu = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y f(y) dy$
- The expectation is not robust, so usually we take log of wage.

Conditional Expectation

- Conditional expectation is defined as $\mathbb{E}[Y|X = x]$ or $\mathbb{E}[Y|X]$.
- Conditional expectation of log wages given gender:

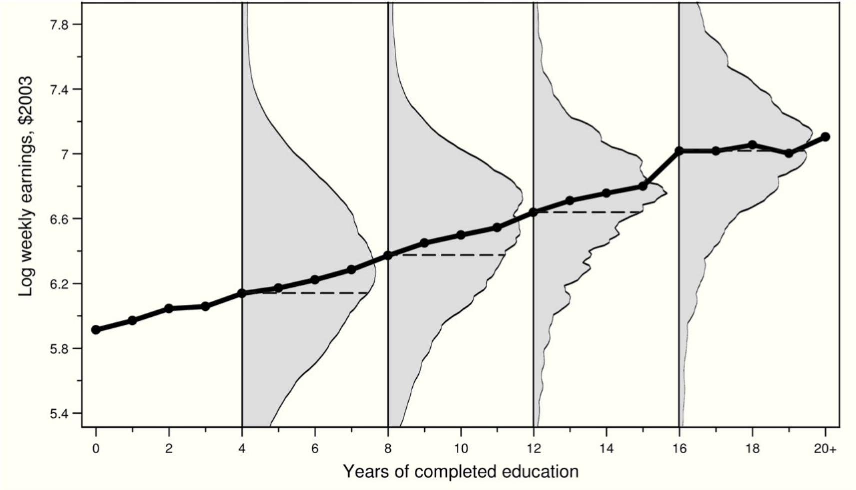
$$\mathbb{E}[\log(\mathit{wage}) \mid \mathit{gender} = \mathit{man}] = 3.05$$

$$\mathbb{E}[\log(\mathit{wage}) \mid \mathit{gender} = \mathit{woman}] = 2.81$$

E(Y|X)=\int Yf(Y|x)dY= \int \frac{f(x,y)}{f(x)}dY

Conditional Expectation

It reminds me of EM algorithm.



Conditional Expectation Function

- More generally, we can define the conditional expectation as a function of many random variables

$$\mathbb{E}\left[Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right] = m\left(x_1, x_2, \ldots, x_k\right)$$

- We call this the **conditional expectation function** (CEF).
- We typically write the conditioning variables as a vector in $\mathbb{R}^k$:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

- Given this notation, the CEF can be compactly written as

$$\mathbb{E}[Y \mid X = x] = m(x)$$

- When X takes the value x then the average value of Y is $m(x)$.

CEF with Continuous Variables

- Assume that the variables (Y, X) are continuously distributed with a joint density function $f(y, x)$.
- The marginal density of x is

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

- For any x such that $f_X(x) > 0$ the conditional density of Y given X is defined as

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

- The CEF of Y given $X = x$ is the expectation of the conditional density (Note: $\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy$)

$$m(x) = \mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy$$

Law of Iterated Expectations

Theorem 2.1 Simple Law of Iterated Expectations

- If $\mathbb{E}|Y| < \infty$ then for any random vector X, $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.

Theorem 2.2 Law of Iterated Expectations

- If $\mathbb{E}|Y| < \infty$ then for any random vector X_1 and X_2, $\mathbb{E}\left[\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1\right] = \mathbb{E}\left[Y \mid X_1\right]$.

Theorem 2.3 Conditioning Theorem

- If $\mathbb{E}|Y| < \infty$ then $\mathbb{E}[g(X)Y \mid X] = g(X)\mathbb{E}[Y \mid X]$
- If in addition $\mathbb{E}|g(X)| < \infty$ then $\mathbb{E}[g(X)Y] = \mathbb{E}[g(X)\mathbb{E}[Y \mid X]]$.

CEF Error

Define **CEF error**:

$$e = Y - m(X)$$ X`\beta=m(x)

Theorem 2.4 Properties of the CEF error

If $\mathbb{E}|Y| < \infty$ then

1. $\mathbb{E}[e \mid X] = 0$ (**mean independence**)
2. $\mathbb{E}[e] = 0$ 无条件期望，迭代法则求解
3. For any function $h(x)$ such that $\mathbb{E}|h(X)e| < \infty$ then $\mathbb{E}[h(X)e] = 0$.

- The equations that $Y = m(X) + e$ and $\mathbb{E}[e \mid X] = 0$ together imply that $m(X)$ is the CEF of Y given X.
  g(x)=E[Y|x]=m(x)

Regression Variance

Unconditional variance of the CEF error:

$$\sigma^2 = \text{var}[e] = \mathbb{E}\left[(e - \mathbb{E}[e])^2\right] = \mathbb{E}\left[e^2\right]$$

- The magnitude of σ^2 measures the amount of variation in Y which is not "explained" or accounted for in the conditional expectation $E[Y|X]$.
- The variance of this unexplained portion e decreases when we condition on more variables.

Theorem 2.6 If $\mathbb{E}\left[Y^2\right] < \infty$ then

$$\text{var}[Y] \geq \text{var}\left[Y - \mathbb{E}\left[Y \mid X_1\right]\right] \geq \text{var}\left[Y - \mathbb{E}\left[Y \mid X_1, X_2\right]\right]$$

- See the proof in Section 2.33 in Hansen (2022) using Jensen's inequality.

Andrew Ng used to teach me

Best Predictor

Suppose that given a random vector X, we want to predict or forecast Y.

- We can write any predictor as a function $g(X)$ of X. The (ex-post) prediction error is the realized difference $Y - g(X)$.
- A non-stochastic measure of the magnitude of the prediction error is the expectation of its square
$$\mathbb{E}\left[(Y - g(X))^2\right]$$
- We can define the **best predictor** as the function $g(X)$ which **minimizes** the above equation.

Best Predictor

The CEF is the best predictor.

Prove:

$$\begin{aligned}
\mathbb{E}\left[(Y-g(X))^2\right] &= \mathbb{E}\left[(e+m(X)-g(X))^2\right] \qquad \text{CEF 3:E[h(x)e]=0} \\
&= \mathbb{E}\left[e^2\right] + 2\mathbb{E}[e(m(X)-g(X))] + \mathbb{E}\left[(m(X)-g(X))^2\right] \\
&= \mathbb{E}\left[e^2\right] + \mathbb{E}\left[(m(X)-g(X))^2\right] \\
&\geq \mathbb{E}\left[e^2\right] \\
&= \mathbb{E}\left[(Y-m(X))^2\right]
\end{aligned}$$

- It means that the CEF minimizes the squared error among all the functions $g(X)$.

Conditional Variance

If $\mathbb{E}\left[Y^2\right] < \infty$, the **conditional variance** of Y given $X = x$ is

$$\sigma^2(x) = \text{var}[Y \mid X = x] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x\right]$$

var(Y)=E[Y-E(Y)]^2, and E(Y)=E[Y|X=x]

- The conditional variance is a function of the conditioning variables.
- It can be easily seen that the conditional variance of Y is exactly the conditional variance of the CEF error.

$$\sigma^2(x) = \mathbb{E}\left[e^2 \mid X = x\right] = \text{var}[e \mid X = x]$$

- Relate to unconditional error variance:

需自行推导

$$\sigma^2 = \mathbb{E}\left[e^2\right] = \mathbb{E}\left[\mathbb{E}\left[e^2 \mid X\right]\right] = \mathbb{E}\left[\sigma^2(X)\right]$$

Homoskedasticity and Heteroskedasticity

最佳线性无偏估计量：同方差、不完全共线、随机抽样等

- The error is homoskedastic if $\sigma^2(x) = \sigma^2$ does not depend on x.
- The error is heteroskedasticity if $\sigma^2(x)$ depend on x. 异方差
- **Heteroskedasticity is generic and "standard", while homoskedasticity is unusual and exceptional.**

Linear CEF

The CEF can be both linear or non-linear in x.

- An important special case is when the CEF is linear in x:

$$m(x) = x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k = x'\beta$$

  where

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

- It is called the **linear CEF model**, **linear regression model**, or the regression of Y on X .

Linear CEF with Nonlinear Effects

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$

- This equation is quadratic in the regressors (x_1, x_2), but we still call it a linear CEF because it is a linear function of the coefficients.
- Define $x_3 = x_1^2, x_4 = x_2^2, x_5 = x_1x_2$, and $x_6 = 1$, and transform the equation into a common and linear one.

$$\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + 2x_1\beta_3 + x_2\beta_5$$

- We typically call β_5 the interaction effect.

Best Linear Predictor

While the conditional expectation $m(X) = \mathbb{E}[Y \mid X]$ is the best predictor of Y among all functions of X, its functional form is typically unknown. (It is rarely linear.)

- Let's find a linear approximation.
- Assumptions:
  1. $\mathbb{E}\left[Y^2\right] < \infty$
  2. $\mathbb{E}\|X\|^2 < \infty$
  3. $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ **is positive definite**

Best Linear Predictor

Define the **Best Linear Predictor or Projection** of Y given X as

$$\mathscr{P}[Y \mid X] = X'\beta$$

where β minimizes the mean squared prediction error

$$S(\beta) = \mathbb{E}\left[\left(Y - X'\beta\right)^2\right]$$

not vector

The minimizer

$$\beta = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} S(b)$$

is called the **Linear Projection Coefficient**.

Best Linear Predictor

We now calculate an explicit expression for its value.

$$S(\beta) = \mathbb{E}\left[Y^2\right] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}\left[XX'\right]\beta$$

dim =1

The first-order condition for minimization is

$$0 = \frac{\partial}{\partial \beta} S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}\left[XX'\right]\beta$$

[\beatE(XX`)]^T, 不管对左右\beta

Then we have

$$\beta = \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}[XY]$$

Sometimes we define $\boldsymbol{Q}_{XY} = \mathbb{E}[XY]$ and $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$. So

$$\beta = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY}$$

Best Linear Predictor

We now have an explicit expression for the best linear predictor:

$$\mathscr{P}[Y \mid X] = X'\beta = X'\mathbb{E}\left[XX'\right]^{-1}\mathbb{E}[XY]$$

Define the **projection error** as

$$e = Y - X'\beta$$

An important property of the projection error e is

$$\mathbb{E}[Xe] = 0$$

proof:
=E[X(Y-X^{T}\beta)]
=E[XY]-E[XX^T]^{-1}E[XY]E[XY]
=0

- It is equivalent to $\mathbb{E}\left[X_j e\right] = 0$ for $j = 1, \ldots, k$.
- If the X contains a constant, then $\mathbb{E}[e] = 0$.

Y=\beat_0+\sum_i \beta_i x_i+e, where \beta_0 can be regarded as \beat_{0}x_0,x_0=1。If E[x_{0}e]=0.E[e]=0

不具有常数项的回归方程穿过0，且R^2可能为负？

Linear Predictor Error Variance

As in the CEF model, we define the error variance as

$$\sigma^2 = \mathbb{E}\left[e^2\right]$$

Setting $Q_{YY} = \mathbb{E}\left[Y^2\right]$ and $\boldsymbol{Q}_{YX} = \mathbb{E}\left[YX'\right]$ we have

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\left(Y - X'\beta\right)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - 2\mathbb{E}\left[YX'\right]\beta + \beta'\mathbb{E}\left[XX'\right]\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\overset{\text{def}}{=} Q_{YY\cdot X}.
\end{aligned}$$

Regression Coefficients with Intercept

Sometimes it is useful to separate the constant from the other regressors

$$Y = X'\beta + \alpha + e$$

where α is the intercept and X does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}[Y] = \mathbb{E}\left[X'\beta\right] + \mathbb{E}[\alpha] + \mathbb{E}[e]$$

E[e]=0

Let $\mu_Y = \mathbb{E}[Y]$ and $\mu_X = \mathbb{E}[X]$, then we have

同上所证

$$Y - \mu_Y = \left(X - \mu_X\right)' \beta + e$$

$$\begin{aligned}\beta &= \left(\mathbb{E}\left[\left(X - \mu_X\right)\left(X - \mu_X\right)'\right]\right)^{-1} \mathbb{E}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] \\ &= \operatorname{var}[X]^{-1} \operatorname{cov}(X, Y)\end{aligned}$$

$$\alpha = \mu_Y - \mu_X'\beta$$

Omitted Variable Bias

Consider two projections of Y:

1. $Y = X_1'\beta_1 + X_2'\beta_2 + e$, $\mathbb{E}[Xe] = 0$, called long regression
2. $Y = X_1'\gamma_1 + u$, $\mathbb{E}[X_1 u] = 0$, called short regression

The relation between γ_1 and β_1 is

$$\begin{aligned}
\gamma_1 &= \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}\left[X_1 Y\right] \\
&= \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}\left[X_1 \left(X_1'\beta_1 + X_2'\beta_2 + e\right)\right] \\
&= \beta_1 + \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}\left[X_1 X_2'\right] \beta_2 \\
&= \beta_1 + \Gamma_{12}\beta_2
\end{aligned}$$

where $\Gamma_{12} = \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$ is the coefficient matrix from a projection of X_2 on X_1.

Omitted Variable Bias

- Observe that $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 \neq \beta_1$ unless $\Gamma_{12} = 0$ (X_1 and X_2 are uncorrelated) or $\beta_2 = 0$.
- The difference $\Gamma_{12}\beta_2$ between γ_1 and β_1 is known as **omitted variable bias**.

Best Linear Approximation

前文被跳过了

The best linear predictor $X'\beta$ is also the best linear approximation of the CEF $m(X)$.

Define the mean-square approximation error

$$d(\beta) = \mathbb{E}\left[\left(m(X) - X'\beta\right)^2\right]$$

Select β to minimize $d(\beta)$. Then we have

$$\begin{aligned}
\beta &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[Xm(X)] \\
&= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[X(m(X)+e)] \\
&= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]
\end{aligned}$$

区别：
E[(Y-X^T\beat)^2]

Limitations of the Best Linear Projection

- The linear CEF means $\mathbb{E}[e|X] = 0$, so apparently $\mathbb{E}[Xe] = 0$.
- The best linear predictor means $\mathbb{E}[Xe] = 0$, but it may not be true that $\mathbb{E}[e|X] = 0$.

Furthermore, the linear projection may be a poor approximation to the CEF.

- Consider the true process is $Y = X + X^2$ with $X \sim \mathrm{N}(0, 1)$. Then the CEF is $m(X) = X + X^2$. when have a constant, \beta=var(X)^{-1}cov(X,Y)
- Now consider project Y on X and a constant: $Y = \beta X + \alpha + e$. Then we have $\beta = 1$ and $\alpha = 1$. So $\mathscr{P}[Y \mid X] = X + 1$.
- The projection error is $e = X^2 - 1$, which increases infinitely as X increases.

can CEF E[e|x]=0 demenstrate BLP E[xe]=0?
consider e=x^2,x\sim N(0,1),E[xe]=0 while E[e|x]=x^2