# Introduction to Econometrics

*Lecture 3 : OLS Regression*

**Zhaopeng Qu**

**Business School,Nanjing University**

Sep. 27, 2018

# Outlines


1. Review the Previous Lecture
2. Make Regression Make Sense
   - Economic Relationships and the CEF
   - The Properties of CEF
   - Linear Regression and the CEF: Why Regress?

# Review the Previous Lecture

# Causal Inference in Social Science

- Causality is our main goal in the studies of empirical social science.
- The existence of Selection Bias makes social science more difficult than science.
- Although experimental method is a powerful tool for economists, every project can not be carried on by it.
- It is the main reason *why modern econometrics exists and develops.*
- *"The modern menu of econometric methods can seem confusing, even to an experienced number cruncher. Luckily, not everything on the menu is equally valuable or important. Some of the more exotic items are needlessly complex and may even be harmful. On the plus side, the core methods of applied econometrics remain largely unchanged, while the interpretation of basic tools has become more nuanced and sophisticated."* *Angrist and Pischke(2009)*

## Furious Seven Weapons（七种武器）

- Build a reasonable counterfactual world or find a proper control group is the core of econometrical methods.
  1. **Random Trials(随机试验)**
  2. **Regression(Ordinary Least Squares)(OLS 回归)**
  3. **Matching and Propensity Score（匹配与倾向得分）**
  4. **Decomposition（分解）**
  5. **Instrumental Variable（工具变量）**
  6. **Regression Discontinuity（断点回归）**
  7. **Difference in Differences （双差分或倍差法)**
- The most basic of these tools is regression, which compares treatment and control subjects who have the *same observed* characteristics.
- Regression concepts are foundational, paving the way for the more elaborate tools used in the class that follow.
  - *So let's start our exciting journey from it.*

# Make Regression Make Sense

# Regression: What You Need to Know

- We spend our lives running regressions (I should say: "regressions run me"). And yet this basic empirical tool is often misunderstood. So I begin with a recap of key regression properties. (Angrist, 2014)
- Our Regression Agenda
  1. The CEF is all you need
  2. What is Regression and Why We Regress
  3. Regression and Causality

# Conditional Expectation Function(CEF): Education and Earnings

- Most of what we want to do in the social science is learn about how **two variables** are related, such as *Education and Earnings*.
- On average, people with more schooling earn more than people with less schooling.
  - The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it's clear that education predicts earnings in a narrow statistical sense.
- This predictive power is compellingly summarized by the **Conditional Expectation Function.**

# Probability Review: Conditional Expectation Function(CEF)

- Both X and Y are r.v., then conditional on X, Y's probability density function is

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

- Conditional on X, Y's expectation is

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

- So Conditional Expectation Function(CEF) is a function of x, since x is a random variable, so CEF is also a random variable
- **直观理解：期望就是求平均值，而条件期望就是"分组取平均"或"在... 条件下的均值"。**

# Conditional Expectation Function(CEF): Education and Earnings

- Most of what we want to do in the social science is learn about how two variables are related, such as Education and Earnings.
- On average, people with more schooling earn more than people with less schooling.
  - The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it's clear that education predicts earnings in a narrow statistical sense.
- This predictive power is compellingly summarized by the **Conditional Expectation Function.**
- The Law of Iterated Expectations(LIE)

$$E[Y] = E[E[Y \mid X]]$$

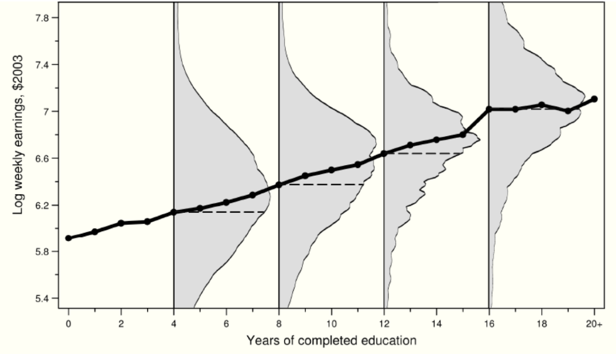
# The Law of Iterated Expectations(LIE)

## Proof.

$$
\begin{aligned}
E[E(Y|X)] &= \int E[Y \mid X = u] g_X(u)\, du \\
&= \int \left[ \int t f_Y(t \mid X = u)\, dt \right] g_X(u)\, du \\
&= \iint t f_Y(t \mid X = u) g_X(u)\, dt\, du \\
&= \int t \left[ \int f_Y(t \mid X = u) g_X(u)\, du \right] dt \\
&= \int t \left[ \int f_{XY}(t, u)\, du \right] dt \\
&= \int t f_Y(t)\, dt
\end{aligned}
$$

□

# Education and Wage: CEF View




- The figure plots the CEF of **log weekly wages** given **schooling** for a sample of middle-aged white men from the 1980 census.
- The CEF in the figure captures the fact that the enormous variation individual circumstances notwithstanding - people with more schooling generally earn more, on average.

# The CEF Decomposition Property

**Theorem**

*Every random variable such as $Y_i$ can be written as*

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i$$

*where $\varepsilon_i$ is mean-independent of $X_i$, i.e., $E[\varepsilon_i \mid X_i] = 0$. and therefore $\varepsilon_i$ is uncorrelated with any function of $X_i$.*

**Proof.**

Skipped □

- This theorem says that any random variable, $Y_i$, can be decomposed into two parts
  - a piece that's "explained by $X_i$", i.e. the CEF,
  - a piece left over which is orthogonal to (i.e. uncorrelated with) any function of $X_i$.

# The CEF-Prediction Property

## Theorem

*Let $m(X_i)$ be any function of $X_i$. The CEF is **the Minimum Mean Squared Error**(MMSE) predictor of $Y_i$ given $X_i$. Thus*

$$E[Y_i \mid X_i] = \underset{m(X_i)}{argmin E}\left[[Y_i - M(X_i)]^2\right]$$

## Proof.

$$\begin{aligned}
(Y - m(X_i))^2 &= [(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))]^2 \\
&= (Y_i - E[Y_i \mid X_i])^2 \\
&+ 2(Y_i - E[Y_i \mid X_i])(E[Y_i \mid X_i] - m(X_i)) \\
&+ (E[Y_i \mid X_i] - m(X_i))^2
\end{aligned}$$

The last term is minimized at zero when $m(X_i)$ is the CEF. □

# The CEF-Prediction Property

- Suppose we are interested in predicting Y using some function $m(X_i)$, the optimal predictor under the **MMSE** (Minimized Mean Squared Error) criterion is CEF.
- Therefore ,CEF is a natural summary of the relationship between Y and X under MMSE.
- *It means that if we can know CEF, then we can describe the relationship of Y and X.*

# CEF and Regression

- We have already learned CEF is a natural summary of the relationships which we would like to know it.
- But CEF is an unknown functional form, so the next question is How to model CEF, $E(Y \mid X)$?
- Answer: Two basic approaches
  - Nonparametric(Matching, Kernel Density etc.)
  - Parametric(Regression)
- Regression estimates provides a valuable baseline for almost all empirical research because Regression is tightly linked to CEF.

# Estimating the CEF: two discrete values

- Suppose a binary $X$ case： $X$ only take on two values, 0 and 1( like our formal example: here X is a treatment).
- We’ ve been writing and for the means in different groups。
  - Then the mean in each group is just a conditional expectation:
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it’s clear that education predicts earnings in a narrow statistical sense.
- How to estimate $\hat{E}[Y \mid X_i = x]$? it means that we have to use sample data to inference the population.
- we could use **sample means** within each group.

$$
\begin{aligned}
\hat{E}[Y \mid X_i = 1] &= \frac{1}{n_1} \sum_{i:X_i=1} Y_i \\
\hat{E}[Y \mid X_i = 0] &= \frac{1}{n_0} \sum_{i:X_i=0} Y_i
\end{aligned}
$$

here $n_0$and $n_1$ are numbers of men and women in the sample.

# Estimating the CEF: multiple discrete values

- What if $X$ takes on $> 2$ discrete values?
- we can still estimate $\hat{E}[Y \mid X_i = x]$ with the sample mean among those who have $X_i = x$, thus

$$\hat{E}[Y \mid X_i = x] \quad = \quad \frac{1}{n_x} \sum_{i:X_i = x} Y_i$$

where $n_x$and is the number of group $x$ in the sample.

# Estimating the CEF: continuous

- What if $X$ is continuous? Can we calculate a mean for every value of $X_i$.
- Because the probability could take values only in an interval for a continuous variable. So we could turn it into a discrete variable. This is called as **stratification**.
  - Once it' s discrete, we can just calculate the means within each strata.
- The stratification approach was fairly crude: it assumed that means were constant within strata, but that seems wrong.
- Now we will think about $E[Y \mid X_i = x]$ as a function. What does this function look like?
  - unknown functions in the population! make producing an estimator very difficult!

# Population Regression: What is a Regression?

**Definition**

population regression ("regression" for short) as the solution to the population least squares problem. Specifically, the K×1 regression coefficient vector $\beta$ is defined by solving

$$\beta = \underset{b}{arg\,min} E\left[(Y_i - X_i'b)^2\right]$$

- Using the first order condition

$$E[X_i(Y_i - X_i'b)] = 0$$

- The solution for $b$ can be written

$$\beta = E[X_iX_i']^{-1} E[X_iY_i']$$

# Three Reasons to Regress

- There are three reasons (three justifications) why the vector of population regression coefficient might be of interest.
    1. The Best Linear Predictor Theorem
    2. The Linear CEF Theorem
    3. The Regression-CEF Theorem

# Regression Justification I

**Theorem**

*The Best Linear Predictor Theorem*
*Regression solves the population least squares problem and is therefore the Best Linear Predictor(BLP) of $Y_i$ given $X_i$.*

**Proof.**

By definition of regression. $\square$

- In other words, just as CEF, which is the best predictor of $Y_i$ given $X_i$ in the class of all functions of $X_i$, the population regression function is the best we can do in the class of linear functions.

# Regression Justification II

**Theorem**

*The Linear CEF Theorem*

*Suppose the CEF is linear. Then the Regression function is it.*

**Proof.**

Suppose $E(Y_i|X_i) = X_i'\beta^*$ for a K×1 vector of coefficients. By the CEF decomposition property, we have

$$E[X_i(Y_i - E[Y_i \mid X_i])] = 0$$

Then substitute using $E(Y_i|X_i) = X_i'\beta^*$

$$E[X_i(Y_i - X_i'\beta^*)] = 0$$

At last find that

$$\beta^* = \beta = E[X_iX_i']^{-1} E[X_iY_i]$$

# Regression Justification III

## Theorem

*The Regression-CEF Theorem*

*The population regression function $X_i'\beta$ provides the* ***MMSE linear approximation*** *to $E(Y_i|X_i)$, thus*

$$\beta = \arg\min_b E\left[\left(E[Y_i|X_i] - X_i'b\right)^2\right]$$

## Proof.

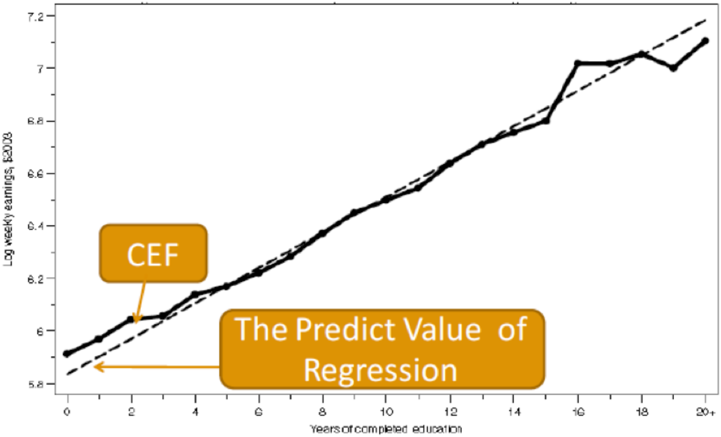
$$\begin{aligned}\left(Y_i - X_i'b\right)^2 &= \left[\left(Y_i - E[Y_i \mid X_i]\right) + \left(E[Y_i \mid X_i] - X_i b\right)\right]^2 \\ &= \left(Y_i - E[Y_i \mid X_i]\right)^2 + \left(E[Y_i \mid X_i] - X_i b\right)^2 \\ &+ 2\left(Y_i - E[Y_i \mid X_i]\right)\left(E[Y_i \mid X_i] - X_i b\right)\end{aligned}$$

The first term has no $b$ and the last term by the CEF-decomposition property. Therefore the minimized problem has the same solution as

# Regression Justification

- The Best Linear Predictor Theorem and The Regression-CEF Theorem show us two more ways to view regression:
  - Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable.
  - If we prefer to think about approximating $E(Y_i|X_i)$, as opposed to predicting $Y_i$, the Regression-CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
- Actually, The regression-CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships, without necessarily trying to pin them down exactly.
- We are not really interested in predicting individual $Y_i$; it's the distribution of $Y_i$ that we care about.

# The CEF and Regression




Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling