Advanced Econometrics

Lecture 2: Conditional Expectation and Projection (Hansen Chapter 2)

Instructor: Ma, Jun

Renmin University of China

Fall 2018

Conditional Expectation Function

- Conditional expectations can be written with the generic notation

$$\mathbb{E}\left(Y \mid X_1, X_2, \ldots, X_k\right) = m\left(X_1, X_2, \ldots, X_k\right).$$

  We call this the conditional expectation function (CEF). The CEF is a function of $(X_1, X_2, \ldots, X_k)$ as it varies with the variables.

- For greater compactness, we will typically write the conditioning variables as a vector in $\mathbb{R}^k$:

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

  Given this notation, the CEF can be compactly written as

$$\mathbb{E}\left(Y \mid \boldsymbol{X}\right) = m\left(\boldsymbol{X}\right).$$

Conditional Expectation Function

- Given the joint density $f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})$ the variable $\boldsymbol{X}$ has the marginal density
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})\,dy.$$
- For any $\boldsymbol{x}$ such that $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ the conditional density of Y given $\boldsymbol{X}$ is defined as
$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \frac{f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})}.$$
- The CEF of Y given $\boldsymbol{X} = \boldsymbol{x}$ is the mean of the conditional density
$$m(\boldsymbol{x}) = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\,dy.$$
Intuitively, $m(\boldsymbol{x})$ is the mean of for the idealized subpopulation where the conditioning variables are fixed at $\boldsymbol{x}$.
- $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ or $\mathbb{E}(Y \mid \boldsymbol{x})$ is interpreted as $m(\boldsymbol{x})$; $\mathbb{E}(Y \mid \boldsymbol{X})$ is interpreted as $m(\boldsymbol{X})$.

Law of Iterated Expectations

Theorem (Simple Law of Iterated Expectations)

If $\mathbb{E}|Y| < \infty$ then for any random vector $\boldsymbol{X}$,

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) = \mathbb{E}(Y).$$

When $\boldsymbol{X}$ is discrete

离散

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) = \sum_{j=1}^{\infty} \mathbb{E}\left(Y \mid \boldsymbol{x}_j\right) \Pr\left(\boldsymbol{X} = \boldsymbol{x}_j\right)$$

and when $\boldsymbol{X}$ is continuous

连续

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) = \int_{\mathbb{R}^k} \mathbb{E}(Y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}.$$

Law of Iterated Expectations

Theorem

If $\mathbb{E}\,|y| < \infty$ *then for any random vectors* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$*,*

$$\mathbb{E}\left(\mathbb{E}\left(Y \mid \boldsymbol{X}_1, \boldsymbol{X}_2\right) \mid \boldsymbol{X}_1\right) = \mathbb{E}\left(Y \mid \boldsymbol{X}_1\right)$$

Law of Iterated Expectations

A property of conditional expectations is that when you condition on a random vector x you can effectively treat it as if it is constant. For example, $\mathbb{E}(\boldsymbol{X}|\boldsymbol{X}) = \boldsymbol{X}$ and $\mathbb{E}(g(\boldsymbol{X})|\boldsymbol{X}) = g(\boldsymbol{X})$ for any function $g(\cdot)$. The general property is known as the Conditioning Theorem.

求关于X的条件期望，那么所有有关X的都看成一个常数.

Theorem (Conditioning Theorem)

If $\mathbb{E}|y| < \infty$ *then*

$$\mathbb{E}(g(\mathsf{x})\, y \mid \mathsf{x}) = g(\mathsf{x})\, \mathbb{E}(y \mid \mathsf{x})$$

If in addition $\mathbb{E}|g(\mathsf{x})\, y| < \infty$, *then*

$$\mathbb{E}(g(\mathsf{x})\, y) = \mathbb{E}(g(\mathsf{x})\, \mathbb{E}(y \mid \mathsf{x})).$$

In this course, we can safely ignore conditions such as $\mathbb{E}|Y| < \infty$ and $\mathbb{E}|g(\boldsymbol{X})Y| < \infty$.

CEF Error

- The CEF error is defined as the difference between and the CEF evaluated at the random vector X:

CEF error $\leftarrow$ $$e = Y - m(\boldsymbol{X}).$$ $\rightarrow$ 条件期望函数(CEF)

By construction, this yields the formula

$$Y = m(\boldsymbol{X}) + e.$$

- A key property of the CEF error is that it has a conditional mean of zero. To see this, by the linearity of expectations, the definition $m(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ and the Conditioning Theorem

$E(X+Y|Z)=E(X|Z)+E(Y|Z)$

$$\begin{aligned}\mathbb{E}(e \mid \boldsymbol{X}) &= \mathbb{E}((Y - m(\boldsymbol{X})) \mid \boldsymbol{X}) \\ &= \mathbb{E}(Y \mid \boldsymbol{X}) - \mathbb{E}(m(\boldsymbol{X}) \mid \boldsymbol{X}) \\ &= m(\boldsymbol{X}) - m(\boldsymbol{X}) \\ &= 0.\end{aligned}$$

- The unconditional mean is also zero:

$$\mathbb{E}(e) = \mathbb{E}(\mathbb{E}(e \mid \boldsymbol{X})) = \mathbb{E}(0) = 0.$$

$X \perp\!\!\!\perp Y \Rightarrow E(X|Y)=E(X)$

$\Rightarrow cov(X,Y)=0$

$\Updownarrow$

$E(XY)=E(X)E(Y)$

$=E\,E(XY|Y)$

$=E[Y\cdot E(X|Y)]$

$=E(Y)E(X)$

$E(X|Y)=E(X)$

$E(X^2|Y)=$ function of Y

CEF Error

Theorem

Properties of the CEF error

If $\mathbb{E}|y| < \infty$ *then*

1. $\mathbb{E}(e \mid \boldsymbol{X}) = 0.$
2. $\mathbb{E}(e) = 0.$
3. *For any function* $h(\boldsymbol{X})$ *such that* $\mathbb{E}|h(\boldsymbol{x})e| < \infty$ *then* $\mathbb{E}(h(\boldsymbol{X})e) = 0.$

$E(h(X)e) = E(E(h(X)e|X))$

$= E(h(X)E(e|X))$

$= E(0) = 0$

The equations

$$y = m(\boldsymbol{X}) + e$$

$$\mathbb{E}(e \mid \boldsymbol{X}) = 0$$

together imply that $m(\boldsymbol{X})$ is the CEF of Y given $\boldsymbol{X}$. It is important to understand that this is not a restriction. These equations hold true by definition.

CEF Error

- The equation $\mathbb{E}(e \mid \boldsymbol{X}) = 0$ is sometimes called a conditional mean restriction, since the conditional mean of the error is restricted to equal zero. The property is also sometimes called **mean independence**, for the conditional mean of is 0 and thus independent of $\boldsymbol{X}$. However, it does not imply that the distribution of e is independent of $\boldsymbol{X}$.
- As a simple example of a case where X and e are mean independent yet dependent, let $e = X\epsilon$ where X and e are independent $N(0, 1)$. Then conditional on X the error e has the distribution $N\left(0, x^2\right)$. Thus $\mathbb{E}(e|X) = 0$ and e is mean independent of X, yet e is not fully independent of X. Mean independence does not imply full independence.

$\varepsilon \perp\!\!\!\perp X \quad \varepsilon \sim N(0,1)$

$X \sim N(0,1)$

$e = X \cdot \varepsilon$

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

$e|X \sim N(0, X^2)$

（e关于X的条件分布，X可以看作一个常数．ε是正态的，e也是正态的）

Regression Variance

- An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error e. We write this as

$$\sigma^2 = \operatorname{Var}(e) = \mathbb{E}\left((e - \mathbb{E}e)^2\right) = \mathbb{E}\left(e^2\right).$$

$$= E(e^2 - 2eEe + (Ee)^2) = Ee^2 - \underbrace{(Ee)^2}_{=0} = Ee^2$$

- We can call σ^2 the regression variance or the variance of the regression error. The magnitude of σ^2 measures the amount of variation in Y which is not "explained" or accounted for in the conditional mean $\mathbb{E}(Y|\boldsymbol{X})$.

$(Y, X) \in \mathbb{R} \times \mathbb{R}^k$

$$\left.\begin{array}{l} m(X) = E(Y|X) \\ \underbrace{e = Y - m(X)}_{\text{CEF error}} \end{array}\right\} \Rightarrow E(e|X) = 0 \Rightarrow \operatorname{Cov}(e, X) = 0$$

Y中能被X解释的是m(X).

Regression Variance

- The regression variance depends on the regressors. Consider two regressions

$$
\left.\begin{aligned}
Y &= \mathbb{E}(Y \mid \boldsymbol{X}_1) + e_1 \\
Y &= \mathbb{E}(Y \mid \boldsymbol{X}_1, \boldsymbol{X}_2) + e_2.
\end{aligned}\right\} \rightarrow \text{一定有 } e_2 \le e_1
$$

- The simple relationship we now derive shows that the variance of this unexplained portion decreases when we condition on more variables. This relationship is monotonic in the sense that increasing the amount of information always decreases the variance of the unexplained portion.

Jensen不等式

$g(E(X)) \le E[g(X)]$

if $g(\cdot)$是凸的.

eg. $g(X) = X^2$

Theorem

If $\mathbb{E}\left(y^2\right) < \infty$ *then*

$$\operatorname{Var}(Y) \geq \operatorname{Var}(Y - \mathbb{E}(Y \mid \boldsymbol{X}_1)) \geq \operatorname{Var}(Y - \mathbb{E}(Y \mid \boldsymbol{X}_1, \boldsymbol{X}_2)).$$

Best Predictor

- Suppose that given a realized value of $\boldsymbol{X}$, we want to create a prediction or forecast of We can write any predictor as a function $g(\boldsymbol{X})$ of $\boldsymbol{X}$. A non-stochastic measure of the magnitude of the prediction error $Y - g(\boldsymbol{X})$ is the expectation of its square

衡量了 $g(X)$ 作为 Y 的 predictor 的误差的大小.

$$\mathbb{E}\left((Y - g(\boldsymbol{X}))^2\right). \tag{1}$$

- We can define the best predictor as the function $g(\boldsymbol{X})$ which minimizes (1). What function is the best predictor? It turns out that the answer is the CEF. This holds regardless of the joint distribution of $(Y, \boldsymbol{X})$:

$m(X)$ 是所有预测结果中误差最小的一项.

$$\begin{aligned}
\mathbb{E}\left((Y - g(\boldsymbol{X}))^2\right) &= \mathbb{E}\left((e + m(\boldsymbol{X}) - g(\boldsymbol{X}))^2\right) \\
&= \mathbb{E}\left(e^2\right) + 2\mathbb{E}\left(e\left(m(\boldsymbol{X}) - g(\boldsymbol{X})\right)\right) + \mathbb{E}\left((m(\boldsymbol{X}) - g(\boldsymbol{X}))^2\right) \\
&= \mathbb{E}\left(e^2\right) + \mathbb{E}\left((m(\boldsymbol{X}) - g(\boldsymbol{X}))^2\right) \\
&\geq \mathbb{E}\left(e^2\right) \\
&= \mathbb{E}\left((Y - m(\boldsymbol{X}))^2\right).
\end{aligned}$$

$$\begin{aligned}
&E(e(m(X) - g(X))) \\
&= E\left(E(e(m(X) - g(X)) \mid X\right) \\
&= E\left((m(X) - g(X))\underbrace{E(e \mid X)}_{=0}\right) \\
&= 0
\end{aligned}$$

Best Predictor

$m(\cdot) = \underset{g(\cdot)}{\arg\min}\, E(Y - g(X))^2$

Theorem

Conditional Mean as Best Predictor

If $\mathbb{E}\left(Y^2\right) < \infty$, *then for any predictor* $g(\boldsymbol{X})$,

$$\mathbb{E}\left((Y - g(\boldsymbol{X}))^2\right) \geq \mathbb{E}\left((Y - m(\boldsymbol{X}))^2\right)$$

where $m(\boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X})$

Conditional Variance

$Var(W) = E(W - E(W))^2$

把所有期望换成关于X的条件期望

$Var(W|X) = E((W - E(W|X))^2 | X)$

Definition

If $\mathbb{E}\left(W^2\right) < \infty$, the conditional variance of W given $\boldsymbol{X}$ is

$$\operatorname{Var}(W \mid \boldsymbol{X}) = \mathbb{E}\left((W - \mathbb{E}(W \mid \boldsymbol{X}))^2 \mid \boldsymbol{X}\right)$$

Definition 回归误差 e 的条件方差

If $\mathbb{E}\left(e^2\right) < \infty$, the conditional variance of the regression error e is

$$\sigma^2(\boldsymbol{X}) = \operatorname{Var}(e \mid \boldsymbol{X}) = \mathbb{E}\left(e^2 \mid \boldsymbol{X}\right)$$

Conditional Variance

- Generally, $\sigma^2(\boldsymbol{X})$ is a non-trivial function of $\boldsymbol{X}$ and can take any form subject to the restriction that it is non-negative.
- Notice as well that $\sigma^2(\boldsymbol{X}) = \operatorname{Var}(Y|\boldsymbol{X})$ so it is equivalently the conditional variance of the dependent variable.
- We define the **conditional standard deviation** as its square root $\sigma(\boldsymbol{X}) = \sqrt{\sigma^2(\boldsymbol{X})}$.
- The unconditional error variance and the conditional variance are related by the law of iterated expectations

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

Conditional Variance

$E(\,g(X)e|X) = g(X)\,E(e|X)$

Conditioning Theorem

- Given the conditional variance, we can define a rescaled error

$$\varepsilon = \frac{e}{\sigma(\boldsymbol{X})}.$$

- We can calculate that since $\sigma(\boldsymbol{X})$ is a function of $\boldsymbol{X}$

$$\mathbb{E}(\varepsilon \mid \boldsymbol{X}) = \mathbb{E}\left(\frac{e}{\sigma(\boldsymbol{X})} \mid \boldsymbol{X}\right) = \frac{1}{\sigma(\boldsymbol{X})}\mathbb{E}(e \mid \boldsymbol{X}) = 0$$

and

$$\operatorname{Var}(\varepsilon \mid \boldsymbol{X}) = \mathbb{E}\left(\varepsilon^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(\frac{e^2}{\sigma^2(\boldsymbol{X})} \mid \boldsymbol{X}\right) = \frac{1}{\sigma^2(\boldsymbol{X})}\mathbb{E}\left(e^2 \mid \boldsymbol{X}\right) = \frac{\sigma^2(\boldsymbol{X})}{\sigma^2(\boldsymbol{X})} = 1$$

Thus ε has a conditional mean of zero, and a conditional variance of 1.

Homoskedasticity and Heteroskedasticity

Definition 同方差 是一个常数. 和X无关.

The error is homoskedastic if $\underline{\mathbb{E}\left(e^2 \mid \boldsymbol{X}\right) = \sigma^2}$ does not depend on $\boldsymbol{X}$.

Definition 异方差

The error is heteroskedastic if $\mathbb{E}\left(e^2 \mid \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X})$ depends on $\boldsymbol{X}$.

Some older or introductory textbooks describe heteroskedasticity as the case where "the variance of e varies across observations". This is a poor and confusing definition. It is more constructive to understand that heteroskedasticity means that the conditional variance $\sigma^2(\boldsymbol{X})$ depends on observables.

Homoskedasticity and Heteroskedasticity

- Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification, and describe heteroskedasticity as an exception or deviance.
- The correct view is that heteroskedasticity is generic and “standard”, while homoskedasticity is unusual and exceptional. The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.
- We will still frequently impose the homoskedasticity assumption when making theoretical investigations into the properties of estimation and inference methods. The reason is that in many cases homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning.

老书认为同方差是正常的.异方差是特例.
但现代观点认为异方差才是一般的情况.同方差才是特例.
实证研究中都假设误差是异方差.
现在常用 White heteroskedasticity robust s.e. 怀特异方差稳健标准差

Regression Derivative

- When a regressor X_1 is continuously distributed, we define the **marginal effect** of a change in X_1, holding the variables $X_2, ..., X_k$ fixed, as the partial derivative of the CEF $\frac{\partial}{\partial X_1} m(X_1, \ldots, X_k)$.
- When X_1 is discrete we define the marginal effect as a discrete difference. For example, if X_1 is binary, then the marginal effect of X_1 on the CEF is

$$m(1, X_2, ..., X_k) - m(0, X_2, ..., X_k).$$

- We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(\boldsymbol{X}) = \begin{cases} \frac{\partial}{\partial X_1} m(X_1, \ldots, X_k), & \text{if } x_1 \text{ is continuous} \\ m(1, X_2, ..., X_k) - m(0, X_2, ..., X_k), & \text{if } x_1 \text{ is binary.} \end{cases}$$

- Collecting the k effects into one $k \times 1$ vector, we define we define the regression derivative to be

$$\boldsymbol{\nabla} m(\boldsymbol{X}) = \begin{bmatrix} \nabla_1 m(X) \\ \nabla_2 m(\boldsymbol{X}) \\ \vdots \\ \nabla_k m(X) \end{bmatrix}.$$

Linear CEF

- An important special case is when the CEF $m(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ is linear in $\boldsymbol{X}$:

$$m(\boldsymbol{X}) = X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \beta_{k+1}.$$

$= \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k$

- An easy way to do so is to augment the regressor vector $\boldsymbol{X}$ by listing the number "1" as an element. We call this the "constant" and the corresponding coefficient is called the "intercept":

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \\ 1 \end{pmatrix}.$$

截距项

With this redefinition, the CEF is

$$\begin{aligned} m(\boldsymbol{X}) &= X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \beta_{k+1} \\ &= \boldsymbol{X}'\boldsymbol{\beta} \end{aligned}$$

线性条件期望模型

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k+1})'$. This is the **linear CEF model**. It is also often called the **linear regression model**.

Linear CEF

- In the linear CEF model, the regression derivative is simply the coefficient vector: $\boldsymbol{\nabla} m(\boldsymbol{X}) = \boldsymbol{\beta}$. The coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{X}'\boldsymbol{\beta} + e \\ \mathbb{E}(e \mid \boldsymbol{X}) &= 0 \end{aligned}$$

Homoskedastic Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{X}'\boldsymbol{\beta} + e \\ \mathbb{E}(e \mid \boldsymbol{X}) &= 0 \\ \mathbb{E}\left(e^2 \mid \boldsymbol{X}\right) &= \sigma^2 \end{aligned}$$

Linear CEF with Nonlinear Effects

- We can include as regressors nonlinear transformations of the original variables. In this sense, the linear CEF framework is flexible and can capture many nonlinear effects.

线性指的是对参数$\vec{\beta}$是线性的.

- The CEF could take the quadratic form

交互项 计 $\beta_5 > 0$, X_1的边际效应对X_2是单增的.

$$m(X_1, X_2) = X_1\beta_1 + X_2\beta_2 + X_1^2\beta_2 + X_2^2\beta_4 + X_1X_2\beta_5 + \beta_6.$$

教育×性别 — 教育回报在性别间存在不平等.

  This is also a linear CEF in the sense of being linear in the coefficients.

- The regression derivatives:

$$\frac{\partial}{\partial X_1} m(X_1, X_2) = \beta_1 + 2X_1\beta_3 + X_2\beta_5$$
$$\frac{\partial}{\partial X_2} m(X_1, X_2) = \beta_2 + 2X_2\beta_4 + X_1\beta_5.$$

  We typically call β_5 the **interaction effect**. If $\beta_5 > 0$ then the regression derivative with respect to X_1 is increasing in the level of X_2.

Best Linear Predictor 最优线性预测

A linear predictor for is a function of the form $\boldsymbol{X}'\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. The mean squared prediction error is

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

Definition

The Best Linear Predictor of Y given $\boldsymbol{X}$ is

$$\mathcal{P}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ minimizes the mean squared prediction error

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

The minimizer

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

线性投影

is called the Linear Projection Coefficient.

Best Linear Predictor

$E(X'\beta)^2 = E(\beta' X X' \beta) = \beta' E(XX')\beta$

- By calculations,

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(Y^2\right) - 2\boldsymbol{\beta}'\mathbb{E}(\boldsymbol{X}Y) + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{\beta}.$$

- By matrix calculus, the first-order condition for minimization is

??

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\mathbb{E}(\boldsymbol{X}Y) + 2\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{\beta}.$$

Solving for the first-order condition, $\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1}\boldsymbol{Q}_{\boldsymbol{X}Y}$ where $\boldsymbol{Q}_{\boldsymbol{X}Y} = \mathbb{E}(\boldsymbol{X}Y)$ is $k \times 1$ and $\boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}')$ is $k \times k$.

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

$$\mathcal{P}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\left(\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{X}Y).$$

最优线性近似

This expression is also referred to as the **linear projection** of Y on $\boldsymbol{X}$.

Best Linear Predictor

- The **projection error** is

$$e = y - \boldsymbol{X}'\boldsymbol{\beta}.$$

Rewriting, we obtain a decomposition of Y into linear predictor and error

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + e.$$

An important property of the projection error is

$$\begin{aligned}
\mathbb{E}(\boldsymbol{X}e) &= \mathbb{E}\left(\boldsymbol{X}\left(Y - \boldsymbol{X}'\boldsymbol{\beta}\right)\right) \\
&= \mathbb{E}(\boldsymbol{X}Y) - \mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\left(\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{X}Y) \\
&= \mathbf{0}.
\end{aligned}$$

Best Linear Predictor

Theorem (Properties of Linear Projection Model)

1. *The Linear Projection Coefficient equals*

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{X}Y\right).$$

2. *The best linear predictor of Y given $\boldsymbol{X}$ is*

$$\mathcal{P}\left(Y \mid \boldsymbol{X}\right) = \boldsymbol{X}'\left(\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{X}Y\right).$$

3. *The projection error $e = Y - \boldsymbol{X}'\boldsymbol{\beta}$ satisfies*

投影误差

$$\mathbb{E}\left(\boldsymbol{X}e\right) = \mathbf{0}.$$

4. *If $\boldsymbol{X}$ contains an constant, then*

$$\mathbb{E}\left(e\right) = 0.$$

Best Linear Predictor

Linear Projection Model

$$
\begin{aligned}
y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\
\mathbb{E}(\boldsymbol{x}e) &= \mathbf{0} \\
\boldsymbol{\beta} &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y)
\end{aligned}
$$

不考

Regression Sub-Vectors

- Let the regressors be partitioned as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

We can write the projection of Y on X as

$$\begin{aligned} y &= X'\beta + e \\ &= X_1'\beta_1 + X_2'\beta_2 + e \\ \mathbb{E}(Xe) &= \mathbf{0}. \end{aligned}$$

$$\underset{k\times 1}{X} = \begin{pmatrix} \underset{k_1\times 1}{X_1} \\ \underset{k_2\times 1}{X_2} \end{pmatrix} \quad k_1+k_2=k$$

$$\underset{1\times k}{X'} = (\underset{1\times k_1}{X_1'}, \underset{1\times k_2}{X_2'})$$

- Partition:

$$Q_{XX} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(X_1X_1') & \mathbb{E}(X_1X_2') \\ \mathbb{E}(X_2X_1') & \mathbb{E}(X_2X_2') \end{bmatrix}$$

$$Q_{XX} = E(XX')_{k\times k}$$

$$Q_{XY} = E(XY)_{k\times 1}$$

and

$$Q_{XY} = \begin{bmatrix} Q_{1Y} \\ Q_{2Y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(X_1Y) \\ \mathbb{E}(X_2Y) \end{bmatrix}.$$

Regression Sub-Vectors

- By the partitioned matrix inversion formula,

$$Q_{XX}^{-1} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Q_{11\cdot 2}^{-1} & -Q_{11\cdot 2}^{-1} Q_{12} Q_{22}^{-1} \\ -Q_{22\cdot 1}^{-1} Q_{21} Q_{11}^{-1} & Q_{22\cdot 1}^{-1} \end{bmatrix},$$

  where $Q_{11.2} = Q_{11} - Q_{12} Q_{22}^{-1} Q_{21}$ and $Q_{22.1} = Q_{22} - Q_{21} Q_{11}^{-1} Q_{12}$.
- Thus,

$\beta = Q_{XX}^{-1} Q_{XY}$

$$\begin{aligned} \beta &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\ &= \begin{bmatrix} Q_{11\cdot 2}^{-1} & -Q_{11\cdot 2}^{-1} Q_{12} Q_{22}^{-1} \\ -Q_{22\cdot 1}^{-1} Q_{21} Q_{11}^{-1} & Q_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} Q_{1Y} \\ Q_{2Y} \end{bmatrix} \\ &= \begin{pmatrix} Q_{11\cdot 2}^{-1} \left(Q_{1Y} - Q_{12} Q_{22}^{-1} Q_{2Y} \right) \\ Q_{22\cdot 1}^{-1} \left(Q_{2Y} - Q_{22\cdot 1}^{-1} Q_{21} Q_{1Y} \right) \end{pmatrix} \end{aligned}$$

Coefficient Decomposition

分块回归

- $\beta_1 \in \mathbb{R}$ and

$$Y = X_1\beta_1 + \boldsymbol{X}_2'\boldsymbol{\beta}_2 + e.$$

X_1是第一个变量

$\boldsymbol{X}_2$是其他变量的向量

- Now consider the projection of X_1 on $\boldsymbol{X}_2$:

$$X_1 = \boldsymbol{X}_2'\boldsymbol{\gamma}_2 + U_1$$
$$\mathbb{E}(\boldsymbol{X}_2U_1) = \boldsymbol{0}.$$

γ_2是X_1在$\boldsymbol{X}_2$上的线性投影.

$$\gamma_2 = \underbrace{E(X_2X_2')^{-1}}_{Q_{22}} \underbrace{E(X_2X_1)}_{Q_{21}}$$

- $\boldsymbol{\gamma}_2 = \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and

$$\begin{aligned}
\mathbb{E}\left(U_1^2\right) &= \mathbb{E}\left(\left(X_1 - \boldsymbol{X}_2'\boldsymbol{\gamma}_2\right)^2\right) \\
&= \mathbb{E}\left(X_1^2\right) - 2\mathbb{E}\left(X_1\boldsymbol{X}_2'\right)\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2'\mathbb{E}\left(\boldsymbol{X}_2\boldsymbol{X}_2'\right)\boldsymbol{\gamma}_2 \\
&= Q_{11} - 2\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} \\
&= Q_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}.
\end{aligned}$$

Coefficient Decomposition

- Also calculate

$$\mathbb{E}(U_1 Y) = \mathbb{E}\left(\left(X_1 - \gamma_2' \boldsymbol{X}_2\right) Y\right) = \boldsymbol{Q}_{1Y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2Y}.$$

- We found

$$\boldsymbol{\beta}_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{Q}_{1Y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2Y}\right) = \frac{\mathbb{E}(u_1 y)}{\mathbb{E}\left(u_1^2\right)}$$

the coefficient from the simple regression of Y on U_1.

Omitted Variable Bias

- Consider the projection of Y on $\boldsymbol{X}_1$ only:

$$\begin{aligned} y &= \boldsymbol{X}_1'\boldsymbol{\gamma}_1 + U \\ \mathbb{E}\left(\boldsymbol{X}_1 U\right) &= \mathbf{0}. \end{aligned}$$

- Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$:

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \left(\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_1 Y\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{\beta}_1 + \boldsymbol{X}_2'\boldsymbol{\beta}_2 + e\right)\right) \\ &= \boldsymbol{\beta}_1 + \left(\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_2'\right)\boldsymbol{\beta}_2 \\ &\neq \boldsymbol{\beta}_1 \end{aligned}$$

$$\underbrace{E(eX)}_{k\times 1} = 0 \iff E\begin{pmatrix} eX_1 \\ eX_2 \end{pmatrix} = 0$$

$$\Rightarrow E(eX_1) = 0$$

unless $\left(\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_2'\right) = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$.

重要

Best Linear Approximation

- We start by defining the mean-square approximation error of $\boldsymbol{X}'\boldsymbol{\beta}$ to $m(\boldsymbol{X})$ as the expected squared difference between $\boldsymbol{X}'\boldsymbol{\beta}$ and the conditional mean $m(\boldsymbol{X})$:

$$d(\boldsymbol{\beta}) = \mathbb{E}\left((m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{\beta})^2\right) = \int_{\mathbb{R}^k} (m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta})^2 f_X(\boldsymbol{x})\, d\boldsymbol{x}.$$

$\min_{b \in \mathbb{R}^k} E((\underbrace{Y}_{m(X)} - X'b)^2)$

- We can then define the best linear approximation to the conditional mean $m(\boldsymbol{X})$ as the function $\boldsymbol{X}'\boldsymbol{\beta}$ obtained by selecting $\boldsymbol{\beta}$ to minimize $d(\boldsymbol{\beta})$:

$$\boldsymbol{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}}\, d(\boldsymbol{b}).$$

$E(XY) = E\,E(XY|X)$

$= E_X E(Y|X)$

$= E(X m(X))$

- It turns out that the best linear predictor and the best linear approximation are identical: by law of iterated expectation,

$$\begin{aligned}\boldsymbol{\beta} &= (\mathbb{E}(\boldsymbol{X}\boldsymbol{X}'))^{-1}\,\mathbb{E}(\boldsymbol{X}m(\boldsymbol{X})) \\ &= (\mathbb{E}(\boldsymbol{X}\boldsymbol{X}'))^{-1}\,\mathbb{E}(\boldsymbol{X}Y).\end{aligned}$$

$P(Y|X) = X'E(XX')^{-1}E(XY)$

① Y对X的投影

② 对E(Y|X)的最优线性近似

Random Coefficient Model

- A linear random coefficient model takes the form

$$Y = \boldsymbol{X}'\boldsymbol{\eta},$$

where the individual-specific coefficient $\boldsymbol{\eta}$ is random and independent of $\boldsymbol{X}$.

- It is interesting to discover that the linear random coefficient model implies a linear CEF. Let

$$\begin{aligned} \boldsymbol{\beta} &= \mathbb{E}(\boldsymbol{\eta}) \\ \boldsymbol{\Sigma} &= \operatorname{Var}(\boldsymbol{\eta}) \end{aligned}$$

and decompose $\boldsymbol{\eta} = \boldsymbol{\beta} + \boldsymbol{U}$. Now $\boldsymbol{U}$ is distributed independently of $\boldsymbol{X}$ with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

- Then

$$\mathbb{E}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\mathbb{E}(\boldsymbol{\eta} \mid \boldsymbol{X}) = \boldsymbol{X}'\mathbb{E}(\boldsymbol{\eta}) = \boldsymbol{X}'\boldsymbol{\beta}$$

so the CEF is linear in $\boldsymbol{X}$, and the coefficients $\boldsymbol{\beta}$ equal the mean of the random coefficient $\boldsymbol{\eta}$.

Random Coefficient Model

Theorem

In the linear random coefficient model $Y = \boldsymbol{X}'\boldsymbol{\eta}$ *with* $\boldsymbol{\eta}$ *independent of* $\boldsymbol{x}$, *then*

$$\mathbb{E}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta}$$
$$\operatorname{Var}(e \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}$$

where $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\eta})$, $\boldsymbol{\Sigma} = \operatorname{Var}(\boldsymbol{\eta})$ *and* $e = Y - \boldsymbol{X}'\boldsymbol{\beta}$. *So the error is conditionally heteroskedastic with its variance a quadratic function of* $\boldsymbol{X}$.

Causal Effects

因果效应

因果关系是单向的、

- Consider the effect of schooling on wages. The causal effect is the actual difference a person would receive in wages if we could change their level of education holding all else constant.
- The causal effect is unobserved because the most we can observe is their actual level of education and their actual wage, but not the counterfactual wage if their education had been different.
- A variable X_1 can be said to have a causal effect on the response variable if the latter changes when all other inputs are held constant.
- A full model:

$$Y = h(X_1, \boldsymbol{X}_2, \boldsymbol{U}),$$

Y的数据生成过程. X_1、X_2被解释. U观测不到的因素、

  where X_1 and $\boldsymbol{X}_2$ are observed variables, $\boldsymbol{U}$ is some unobserved random factor and h is a functional relationship.
- This framework is called the **potential outcomes** framework.

Causal Effects

Definition

In the model (2.52) the causal effect of x_1 on y is

因果效应 $\Rightarrow$ $$C(x_1, \boldsymbol{x}_2, \boldsymbol{u}) = \nabla_1 h(x_1, \boldsymbol{x}_2, \boldsymbol{u})$$

the change in y due to a change in x_1, holding $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.
We define the causal effect of X_1 within this model as the change in due to a change in X_1 holding the other variables $\boldsymbol{X}_2$ and $\boldsymbol{U}$ constant.

Causal Effects

- A popular example arises in the analysis of treatment effects with a binary regressor X_1. Let $X_1 = 1$ indicate treatment (e.g. a medical procedure) and $X_1 = 0$ indicate non-treatment.
- In this case: $Y(0) = h(0, \boldsymbol{X}_2, \boldsymbol{U})$ and $Y(1) = h(1, \boldsymbol{X}_2, \boldsymbol{U})$.
- $Y(0)$ and $Y(1)$ are the latent outcomes associated with non-treatment and treatment. That is, for a given individual, $Y(0)$ is the health outcome if there is no treatment, and $Y(1)$ is the health outcome if there is treatment.
- The causal effect of treatment for the individual is the change in their health outcome due to treatment — the change in as we hold both $\boldsymbol{X}_2$ and $\boldsymbol{U}$ constant:

$$C(\boldsymbol{X}_2, \boldsymbol{U}) = Y(1) - Y(0).$$

- In a sample, we cannot observe both outcomes from the same individual, we only observe the realized value

$$Y = \begin{cases} Y(0) & \text{if } X_1 = 0 \\ Y(1) & \text{if } X_1 = 1. \end{cases}$$

treatment effect 就是 Cause effect 中的一个特例，只不过 X_1 是离散的。

不能同时观测到 $Y(0)$, $Y(1)$，是同一个体的两种不同状态。

Average Causal Effect

$C(X_1, X_2, U) = \nabla_1 h(X_1, X_2, U)$

$E[C(X_1, X_2, U) \mid X_1, X_2]$

$= ACE(X_1, X_2)$ 平均因果效应

Definition

The average causal effect of X_1 on Y conditional on $\boldsymbol{X}_2$ is

$$
\begin{aligned}
ACE\left(X_1, \boldsymbol{X}_2\right) &= \mathbb{E}\left(C\left(X_1, \boldsymbol{X}_2, \boldsymbol{U}\right) \mid X_1, \boldsymbol{X}_2\right) \\
&= \int \nabla_1 h\left(X_1, \boldsymbol{X}_2, \boldsymbol{u}\right) f_{\boldsymbol{U} \mid X_1, \boldsymbol{X}_2}\left(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2\right) d\boldsymbol{u}
\end{aligned}
$$

where $f_{\boldsymbol{U} \mid X_1, \boldsymbol{X}_2}\left(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2\right)$ is the conditional density of $\boldsymbol{U}$ given $X_1, \boldsymbol{X}_2$.

What is the relationship between the average causal effect $ACE\left(X_1, \boldsymbol{X}_2\right)$ and the regression derivative $\nabla_1 m(X_1, \boldsymbol{X}_2)$?

Average Causal Effect

- Since $Y = h(X_1, \boldsymbol{X}_2, \boldsymbol{U})$, the CEF is

$$m(X_1, \boldsymbol{X}_2) = \mathbb{E}(h(X_1, \boldsymbol{X}_2, \boldsymbol{U}) \mid X_1, \boldsymbol{X}_2)$$
$$= \int h(X_1, \boldsymbol{X}_2, \boldsymbol{u}) f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2)\, d\boldsymbol{u}$$

  the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{U}$.
- The regression derivative is

$$\begin{aligned}\nabla_1 m(X_1, \boldsymbol{X}_2) &= \int \nabla_1 h(X_1, \boldsymbol{X}_2, \boldsymbol{u}) f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2)\, d\boldsymbol{u} \\ &\quad + \int h(X_1, \boldsymbol{X}_2, \boldsymbol{u}) \nabla_1 f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2)\, d\boldsymbol{u} \\ &= ACE(X_1, \boldsymbol{X}_2) \\ &\quad + \int h(X_1, \boldsymbol{X}_2, \boldsymbol{u}) \nabla_1 f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2)\, d\boldsymbol{u}.\end{aligned}$$

- The regression derivative and ACE equal in the special case when $\nabla_1 f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid X_1, \boldsymbol{X}_2) = 0$, that is, when the conditional density of $\boldsymbol{U}$ given $X_1, \boldsymbol{X}_2$: $f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2)$ does not depend on x_1.

$$\frac{\partial m(x_1, x_2)}{\partial x_1} = \frac{\partial}{\partial x_1} \int h(x_1, x_2, u) f(u|x_1 x_2)\, du$$
$$= \int \frac{\partial}{\partial x_1}\left(h(x_1, x_2, u)\, f(u|x_1, x_2)\right) du$$
$$= \int \left\{ \frac{\partial h(x_1, x_2, u)}{\partial x_1} f(u|x_1, x_2) + h(x_1, x_2, u) \frac{\partial f}{\partial x_1} \right\} du$$

Average Causal Effect

条件独立假设

$X \perp\!\!\!\perp W \iff f_{X|W} = f_X$

$X \perp\!\!\!\perp W \mid Z \iff \underbrace{f_{X|W,Z}}_{\text{与}w\text{无关}} = f_{X|Z}$

Definition (**Conditional Independence Assumption(CIA)**)

Conditional on $\boldsymbol{X}_2$, the random variables X_1 and $\boldsymbol{U}$ are statistically independent.

- Like (unconditional) independence ($f_{\boldsymbol{U}|X_1} = f_{\boldsymbol{U}}$), the conditional independence means $f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2} = f_{\boldsymbol{U}|\boldsymbol{X}_2}$ and thus $\nabla_1 f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2}(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = 0$.
- Thus CIA implies $\nabla_1 m(X_1, \boldsymbol{X}_2) = ACE(X_1, \boldsymbol{X}_2)$.
- CIA is weaker than full independence of $\boldsymbol{U}$ from the regressors $X_1, \boldsymbol{X}_2$: $f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2} = f_{\boldsymbol{U}} \implies f_{\boldsymbol{U}|X_1,\boldsymbol{X}_2} = f_{\boldsymbol{U}|\boldsymbol{X}_2}$.

$f_{u|X_1,X_2} = f_u \Rightarrow f_{u|X_2} = f_u \Rightarrow \underbrace{f_{u|X_2} = f_{u|X_1,X_2}}_{\text{CIA}}$

$$
\begin{aligned}
f_{u,x_2}(u,x_2) &= \int f_{u,x_1,x_2}(u,x_1,x_2)\,dx_1 \\
&= \int f_u(u) f_{x_1,x_2}(x_1,x_2)\,dx_1 \\
&= f_u(u) \underbrace{\int f_{x_1,x_2}(x_1,x_2)\,dx_1}_{f_{x_2}(x_2)}
\end{aligned}
$$