Advanced Econometrics

Lecture 4: Conditional Expectation, Linear Projection and Linear Causal Model (Hansen Chapter 2)

Instructor: Ma, Jun

Renmin University of China

October 8, 2021

Conditional expectation function

- Given the joint density $f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})$, $\boldsymbol{X}$ has the marginal density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} f_{Y,\boldsymbol{X}}(y,\boldsymbol{x})\, dy.$$

- The conditional density of Y given $\boldsymbol{X}$:

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

- The CEF of Y given $\boldsymbol{X} = \boldsymbol{x}$ is the mean of the conditional density:

$$m(\boldsymbol{x}) = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, dy.$$

  $m(\boldsymbol{x})$ 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 $m(\boldsymbol{X})$.
- We call this the conditional expectation function (CEF). The CEF is a function of $\boldsymbol{X}$.

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

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

  and when $\boldsymbol{X}$ is continuous

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

- In this course, we ignore conditions such as $\mathbb{E}|Y| < \infty$.

Theorem

For any random vectors $\boldsymbol{X}_1$and $\boldsymbol{X}_2$,

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

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

Theorem (Conditioning Theorem)

$$\begin{aligned} \mathbb{E}(g(\boldsymbol{X})Y \mid \boldsymbol{X}) &= g(\boldsymbol{X})\mathbb{E}(Y \mid \boldsymbol{X}) \\ \mathbb{E}(g(\boldsymbol{X})Y) &= \mathbb{E}(g(\boldsymbol{X})\mathbb{E}(Y \mid \boldsymbol{X})). \end{aligned}$$

CEF error

- The CEF error is defined as
$$e = Y - m(\boldsymbol{X}).$$
By construction,
$$Y = m(\boldsymbol{X}) + e.$$
- By the linearity of expectations, the definition $m(\boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X})$ and the Conditioning Theorem
$$\begin{aligned} \mathbb{E}(e \mid \boldsymbol{X}) &= \mathbb{E}((Y - m(\boldsymbol{X})) \mid \boldsymbol{X}) \\ &= \mathbb{E}(Y \mid \boldsymbol{X}) - \mathbb{E}(m(\boldsymbol{X}) \mid \boldsymbol{X}) \\ &= m(\boldsymbol{X}) - m(\boldsymbol{X}) \\ &= 0. \end{aligned}$$
- The unconditional mean is also zero:
$$\mathbb{E}(e) = \mathbb{E}(\mathbb{E}(e \mid \boldsymbol{X})) = \mathbb{E}(0) = 0.$$

Theorem

Properties of the CEF error

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

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

3. For any function $h(\cdot)$, $\mathbb{E}(h(\boldsymbol{X})\, e) = 0.$

- 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.

- The equation $\mathbb{E}\left(e \mid \boldsymbol{X}\right) = 0$ is called a conditional mean restriction, since the conditional mean of the error is restricted to equal zero.
- The property is also 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 $\mathrm{N}\left(0, 1\right)$. Then conditional on X the error e has the distribution $\mathrm{N}\left(0, x^2\right)$. Thus $\mathbb{E}\left(e \mid X\right) = 0$ and e is mean independent of X, yet e is not fully independent of X. Mean independence does not imply full independence.

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

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

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

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

$$\begin{aligned} Y &= \mathbb{E}(Y \mid \boldsymbol{X}_1) + e_1 \\ Y &= \mathbb{E}(Y \mid \boldsymbol{X}_1, \boldsymbol{X}_2) + e_2. \end{aligned}$$

- 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.

Theorem

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

Best predictor

- Suppose that given a realized value of $\boldsymbol{X}$, we want to create a prediction or forecast of Y.
- 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 $\mathbb{E}\left((Y - g(\boldsymbol{X}))^2\right)$.
- What function is the best predictor? It turns out that the answer is the CEF:

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

Theorem

Conditional Mean as Best Predictor

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

Definition

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

The conditional variance of the regression error e is

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

- Generally, $\sigma^2(\boldsymbol{X}) > 0$ is a function of $\boldsymbol{X}$.
- Notice as well that $\sigma^2(\boldsymbol{X}) = \operatorname{Var}(Y \mid \boldsymbol{X})$ 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

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

- 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

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

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

Homoskedasticity and heteroskedasticity

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

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

- 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.
- We will still frequently impose homoskedasticity when making theoretical investigations. In many cases homoskedasticity greatly simplifies the theoretical calculations.

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 $\frac{\partial}{\partial X_1} m\left(X_1, \ldots, X_k\right)$.
- 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

$$m\left(1, X_2, ..., X_k\right) - m\left(0, X_2, ..., X_k\right).$$

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

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

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

$$\boldsymbol{\nabla} m(\boldsymbol{X}) = \left[\begin{array}{c} \nabla_1 m(X) \\ \nabla_2 m(\boldsymbol{X}) \\ \vdots \\ \nabla_k m(X) \end{array} \right].$$

Linear CEF

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

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

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

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{pmatrix}.$$

- With this redefinition, the CEF is

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

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

- 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.
- The CEF could take the quadratic form

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

  This is also a linear CEF in the sense of being linear in the coefficients.
- The regression derivatives:

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

  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

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

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

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

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

The minimizer

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

is called the Linear Projection Coefficient.

- By calculations,

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

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

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

  Solving for the first-order condition, $\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\boldsymbol{Q}_{\boldsymbol{X}Y}$ where $\boldsymbol{Q}_{\boldsymbol{X}Y} = \mathbb{E}(\boldsymbol{X}Y)$ is $k \times 1$ and $\boldsymbol{Q}_{\boldsymbol{XX}} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}')$ is $k \times k$.
- We now have an explicit expression for the best linear predictor:

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

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

- 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

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

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

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

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

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

Best linear approximation

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

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

- 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 $d(\boldsymbol{b})$:

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

- It turns out that the best linear predictor and the best linear approximation are identical:

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

Causal effects

- 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:

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

  where X_1 and $\boldsymbol{X}_2$ are observed variables, $\boldsymbol{U}$ is some unobserved random factor and h is a functional relationship.

Definition

The causal effect of X_1 on Y is

$$C(X_1, \boldsymbol{X}_2, \boldsymbol{U}) = \nabla_1 h(X_1, \boldsymbol{X}_2, \boldsymbol{U})$$

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.

Average causal effect

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 $X_1, \boldsymbol{X}_2$.

Linear causal model

- Suppose that $h(X_1, \boldsymbol{X}_2, U) = g(X_1, \boldsymbol{X}_2) + U$ (U is one-dimensional). Then, $C(X_1, \boldsymbol{X}_2, \boldsymbol{U}) = \nabla_1 g(X_1, \boldsymbol{X}_2)$ and $ACE(X_1, \boldsymbol{X}_2) = \nabla_1 g(X_1, \boldsymbol{X}_2)$.
- The linear causal model specifies that $g(X_1, \boldsymbol{X}_2)$ is linear in parameters. E.g., $g(X_1, \boldsymbol{X}_2) = \gamma X_1 + \boldsymbol{X}_2'\boldsymbol{\beta}$ ($\nabla_1 g(X_1, \boldsymbol{X}_2) = \gamma$) or $g(X_1, \boldsymbol{X}_2) = \gamma X_1 + \boldsymbol{X}_2'\boldsymbol{\beta} + X_1 \boldsymbol{X}_2'\boldsymbol{\delta}$ ($\nabla_1 g(X_1, \boldsymbol{X}_2) = \gamma + \boldsymbol{X}_2'\boldsymbol{\delta}$), where $\gamma, \boldsymbol{\beta}, \boldsymbol{\delta}$ are unknown parameters.
- Without additional assumptions, it is impossible to recover the parameters $\gamma, \boldsymbol{\beta}, \boldsymbol{\delta}$.
- If we assume the stronger condition $\mathbb{E}(U \mid X_1, \boldsymbol{X}_2) = 0$, then $g(X_1, \boldsymbol{X}_2) = \mathbb{E}(Y \mid X_1, \boldsymbol{X}_2)$.
- If $g(X_1, \boldsymbol{X}_2) = \gamma X_1 + \boldsymbol{X}_2'\boldsymbol{\beta}$ and a weaker condition $\mathbb{E}(X_1 U) = 0$ and $\mathbb{E}(\boldsymbol{X}_2 U) = 0$ holds, then $g(X_1, \boldsymbol{X}_2) = \mathcal{P}(Y \mid X_1, \boldsymbol{X}_2)$.
- Under either stronger or weaker condition, the unobservable factor U is uncorrelated with the observable factors.

Omitted variable bias

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

$$\begin{aligned} Y &= \boldsymbol{X}_1'\boldsymbol{\gamma}_1 + e \\ \mathbb{E}(\boldsymbol{X}_1 e) &= \mathbf{0}. \end{aligned}$$

- Suppose that the linear causal model holds: $Y = \boldsymbol{X}_1'\boldsymbol{\beta}_1 + \boldsymbol{X}_2'\boldsymbol{\beta}_2 + U$, with $\mathbb{E}(\boldsymbol{X}_1 U) = 0$ and $\mathbb{E}(\boldsymbol{X}_2 U) = 0$.
- Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$:

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

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 $\boldsymbol{\beta}_2 = \mathbf{0}$.