微观计量经济学

阮 睿

中央财经大学
中国财政发展协同创新中心

February 27, 2025

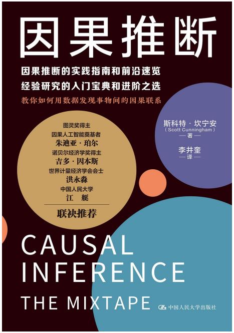
1. Introduction
    - What is Econometrics
    - The Probability Approach to Econometrics
    - Econometric Terms and Notation
    - Standard Data Structures
2. Conditional Expectation and Projection
    - The Distribution of Wages
    - Conditional Expectation
    - Logs and Percentages
    - Conditional Expectation Function
    - Continuous Variables
    - Law of Iterated Expectations
    - CEF Error
    - Regression Variance
    - Best Predictor
    - Conditional Variance
    - Homoskedasticity and Heteroskedasticity
    - Regression Derivative
    - Linear CEF
    - Linear CEF with Nonlinear Effects
    - Linear CEF with Dummy Variables
    - Best Linear Predictor
    - Linear Predictor Error Variance
    - Regression Coefficients
    - Regression Sub-Vectors
    - Coefficient Decomposition
    - Omitted Variable Bias
    - Limitations of the Best Linear Projection
    - Causal Effects

推荐的先修课教材

- 推荐的先修课教材
  - 《计量经济学导论：现代观点》(Introductory Econometrics: A Modern Approach), 伍德里奇 (Jeffrey M Wooldridge) 著。
  - 《计量经济学》(Introduction to Econometrics), 斯托克、沃森 (James H Stock and Mark W Watson) 著。
- 推荐阅读教材
  - 《因果推断》，斯科特 · 坎宁安，中国人民大学出版社，2023 年。
  - 《因果推断初步》，姚东旻，清华大学出版社，2022 年。
  - 《因果推断实用计量方法》，邱嘉平，上海财经大学出版社，2020 年。
  - 《基本无害的计量经济学》(Mostly Harmless Econometrics: An Empiricist’s Companion), 安格里斯特 (Joshua D Angrist)、皮施克 (Jorn-Steffen Pischke) 著。格致出版社，2012 年。
  - Applied Causal inference powered by ML and AI
  - Mastering ’ Metrics: The Path from Cause to Effect, by Joshua D Angrist and Jorn-Steffen Pischke, 2015.

推荐阅读教材

- Microeconometrics Using Stata （用 STATA 学微观计量经济学）, Revised Edition, by A Colin Cameron and Pravin K Trivedi, 2010.
- «横截面与面板数据的计量经济分析» (Econometric Analysis of Cross Section and Panel Data), 伍德里奇 (Jeffrey M Wooldridge) 著。中文第二版，中国人民大学出版社，2016 年。
- Econometrics, Bruce Hansen. 2022 年。
- «计量经济分析» (Econometric Analysis), 格林 (William H Greene) 著。中文第六版，中国人民大学出版社，2011 年。英文第六版，中国人民大学出版社，2009 年。英文最新版，7th edition, 2011.
- «计量经济学» (Econometrics), 林文夫 (Fumio Hayashi) 著。中文版，上海财经大学出版社，2005 年。



因果推断

因果推断的实践指南和前沿速览

经验研究的入门宝典和进阶之选

教你如何用数据发现事物间的因果联系

图灵奖得主
因果人工智能奠基者
朱迪亚 · 珀尔
诺贝尔经济学奖得主
吉多 · 因本斯
世界计量经济学会会士
洪永淼
中国人民大学
江　艇

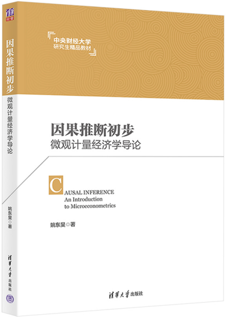
联袂推荐

斯科特 · 坎宁安
（Scott Cunningham）
— 著 —

李井奎
— 译 —

CAUSAL
INFERENCE
THE MIXTAPE

中国人民大学出版社



因果推断初步 微观计量经济学导论 姚东旻○著 清华大学出版社

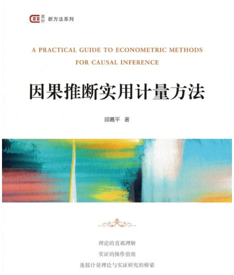
中央财经大学
研究生精品教材

因果推断初步
微观计量经济学导论

CAUSAL INFERENCE
An Introduction
to Microeconometrics

姚东旻○著

清华大学出版社



匡时 新方法系列

A PRACTICAL GUIDE TO ECONOMETRIC METHODS
FOR CAUSAL INFERENCE

因果推断实用计量方法

邱嘉平　著

理论的直观理解

实证的操作指南

连接计量理论与实证研究的桥梁

What is Econometrics

- ”Econometrics” 这个词是 Ragnar Frisch (1895-1973) 发明的.
- Ragnar Frisch
    - the three principal founders of the Econometric Society
    - first editor of the journal *Econometrica*
    - co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969

A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: “The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems....”

But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three view- points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern eco- nomic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics.
Ragnar Frisch, Econometrica, (1933), 1, pp. 1-2.

- 计量经济学是对经济模型、数理统计和经济数据的统一研究。
- 计量经济学理论涉及工具和方法的发展，以及计量经济学方法的特性研究。
- 应用计量经济学是描述定量经济模型的发展以及使用经济数据将计量经济学方法应用于这些模型。

The Probability Approach to Econometrics

Trygve Haavelmo (1911-1999, winner of the 1989 Nobel Memorial Prize in Economic Sciences) argued that quantitative economic models must necessarily be probability models (by which today we would mean stochastic). Deterministic models are blatently inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics.

- **structural approach**：A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as “taking their model seriously”. The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.
- **quasi-structural approach**: A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation.
  The quasi-structural approach to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

- **semiparametric approach**: A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least squares and the Generalized Method of Moments. The semiparametric approach dominates contemporary econometrics
- **calibration approach** Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics (deeming classical theory as inappropriate for approximate models) and instead selects parameters by matching model and data moments using non-statistical *ad hoc* methods.

Econometric Terms and Notation

- data, dataset, sample: 对一组变量进行一组重复测量，例如，对于一个关于劳动的研究中，变量可能包括每周收入、教育程度、年龄和其他描述性特征。
- observations：对变量的不同重复测量。一个单独的 observation 通常对应于一个特定的经济单位，或对应于某一时间点测量。“观测”。
- 变量：用 Y, X, Z 表示变量。惯例：Y 表示被解释变量，X, Z 表示解释变量。
- 用小写英文字母，如 x，表示一个 scalar。向量可以用 x 或 $\mathbf{x}$ 表示。

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$$

Econometric Terms and Notation

- $\mathbf{X}$ 表示矩阵。
- β, θ, σ^2 表示未知参数（estimand）。
- $\hat{\beta}, \hat{\theta}$ 表示 estimator。
- 对于一组数据，可以得到一组 estimate。

Standard Data Structures

- cross-sectional（横截面）, time series（时间序列）, panel（面板）, clustered, and spatial.
- clustered: 在聚类抽样中，观测值被分组为” 群组”，这些群组被视为相互独立的，但允许在群组内有依赖性。与面板数据的主要区别在于，聚类抽样通常并不明确地对误差的结构进行建模。
- spatial: 空间依赖是另一种相互依赖的模型。根据一个空间尺度（例如，地理上的接近性)，观察值被视为相互依赖。与聚类不同，空间模型允许所有的观测值都是相互依赖的，并且通常依赖于依赖关系的明确建模。
- 大部分关于相互独立的假设都是针对横截面数据。

Definition

The variables (Y_i, X_i) are a sample from the distribution F if they are identically distributed with distribution F.

Definition

The variables (Y_i, X_i) are a random sample if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.

Chapter 2 Conditional Expectation and Projection

- The most commonly applied econometric tool is least squares estimation, also known as **regression**.
- Least squares is a tool to estimate the conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).
- In this chapter we abstract from estimation and focus on the probabilistic foundation of the conditional expectation model and its projection approximation.

The Distribution of Wages

- We view the wage of an individual worker as a random variable wage with the **probability distribution**

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

- When a distribution function F is differentiable we define the probability density function

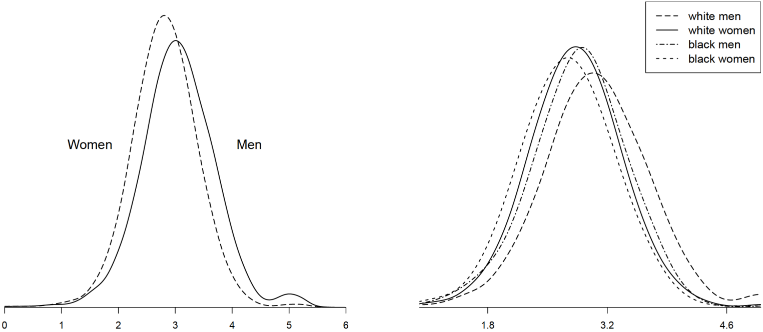
$$f(u) = \frac{d}{du}F(u)$$

- **mean** or **expectation**

$$\mu = \mathbb{E}[Y] = \sum_{j=1}^{\infty} \tau_j \mathbb{P}\left[Y = \tau_j\right]$$

$$\mu = \mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)dy$$

Conditional Expectation



(a) Women and Men

(b) By Gender and Race

Figure 2.2: Log Wage Density by Gender and Race

- **conditional expectations**

$$\begin{aligned}\mathbb{E}[\log(\text{ wage }) \mid \text{ gender } = \text{ man }] &= 3.05\\ \mathbb{E}[\log(\text{ wage }) \mid \text{ gender } = \text{ woman }] &= 2.81\\ \mathbb{E}[\log(\text{ wage }) \mid \text{ gender } = \text{ woman, race } = \text{ Black }] &= 2.73.\end{aligned}$$

Logs and Percentages

Take two positive numbers a and b. The percentage difference between a and b is

$$p = 100\frac{a-b}{b}$$

$$\frac{a}{b} = 1 + \frac{p}{100}$$

Taking natural logarithms:

$$\log a - \log b = \log\left(1 + \frac{p}{100}\right)$$

泰勒展开：$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = x + O\left(x^2\right)$，所以 $\log(1+x) \simeq x$.
得到：$p \simeq 100(\log a - \log b)$

expectation of log transformed random variables

Take two random variables $X_1, X_2 > 0$. It will be useful to define their geometric means $\theta_1 = \exp(\mathbb{E}[\log X_1])$ and $\theta_2 = \exp(\mathbb{E}[\log X_2])$. The difference in the expectation of the log transforms (multiplied by 100) is

$$100\left(\mathbb{E}\left[\log X_2\right] - \mathbb{E}\left[\log X_1\right]\right) = 100\left(\log\theta_2 - \log\theta_1\right) \simeq p$$

The difference between the average of the log transformed variables is (approximately) the percentage difference in the geometric means.

$$\begin{aligned} & \mathbb{E}[\log(X_1)] \\ = \quad & \frac{1}{n}\textstyle\sum_{i=1}^{n}\log(X_{1i}) \\ = \quad & \frac{1}{n}\log(\textstyle\prod_{i=1}^{n}X_{1i}) \\ = \quad & \log\left(\textstyle\prod_{i=1}^{n}X_{1i}\right)^{\frac{1}{n}} \end{aligned}$$

Conditional Expectation Function

- Conditional expectations can be written with the generic notation

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

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

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$$

- CEF can be written as :

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

- 当 X 取值为 x 时，Y 的均值为 $m(x)$。

Continuous Variables

- 对于密度函数为 $f(y,x)$ 的联合分布，变量 x 的边缘密度函数为

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

- 对于任意满足 $f_X(x) > 0$ 的 x，给定 X 的 Y 的条件密度函数为：

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

- The CEF of Y given X = x is the expectation of the conditional density

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

Law of Iterated Expectations

Theorem

Simple Law of Iterated Expectations

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

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$$

条件期望的期望等于无条件期望。

对于离散的 X,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_{j=1}^{\infty} \mathbb{E}\left[Y \mid X = x_j\right] \mathbb{P}\left[X = x_j\right]$$

对于连续的 X,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \int_{\mathbb{R}^k} \mathbb{E}[Y \mid X = x] f_X(x) dx$$

CEF Error

- The CEF error e is defined as the difference between Y and the CEF evaluated at X:

$$e = Y - m(X)$$

Theorem

Properties of the CEF error *If* $\mathbb{E}|Y| < \infty$ *then*

- $\mathbb{E}[e|X] = 0$
- $\mathbb{E}[e] = 0$
- *If* $\mathbb{E}|Y|^r < \infty$ *for* $r \geq 1$ *then* $\mathbb{E}|e|^r < \infty$

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

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e \mid X]] = \mathbb{E}[0] = 0$$

- The condition $\mathbb{E}[e \mid X] = 0$ is implied by the definition of e as the difference between Y and the CEF $m(X)$. The equation $\mathbb{E}[e \mid X] = 0$ is sometimes called a conditional mean restriction, since the conditional mean of the error e is restricted to equal zero. The property is also sometimes called mean independence, for the conditional mean of e is 0 and thus independent of X. However, it does not imply that the distribution of e is independent of X. Sometimes the assumption " e is independent of X " is added as a convenient simplification, but it is not generic feature of the conditional mean. Typically and generally, e and X are jointly dependent even though the conditional mean of e is zero.

Regression Variance

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

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

Theorem (2.5)

If $\mathbb{E}\left[Y^2\right] < \infty$ *then* $\sigma^2 < \infty$.

- 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 expectation $\mathbb{E}[Y \mid X]$.

Theorem

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

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

This Theorem says that the variance of the difference between Y and its conditional expectation (weakly) decreases whenever an additional variable is added to the conditioning information.

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 $\mathbb{E}\left[(Y - g(X))^2\right]$. What function is the best predictor? It turns out that the answer is the CEF $m(X)$. This holds regardless of the joint distribution of (Y, X).

To see this, note that the mean squared error of a predictor $g(X)$ is

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

Theorem

Conditional Expectation as Best Predictor *If* $\mathbb{E}\left[Y^2\right] < \infty$, *then for any predictor* $g(X)$,

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

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

Conditional Variance

- While the conditional mean is a good measure of the location of a conditional distribution it does not provide information about the spread of the distribution. A common measure of the dispersion is the conditional variance. We first give the general definition of the conditional variance of a random variable W.

Definition

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

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

The conditional variance treated as a random variable is $\operatorname{var}[W \mid X] = \sigma^2(X)$.

Definition

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

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

The conditional variance of e treated as a random variable is $\operatorname{var}[e \mid X] = \sigma^2(X)$.

条件方差和无条件方差之间的关系

Theorem

If $\mathbb{E}\left[X^2\right] < \infty$ *then*

$$\operatorname{var}[X] = \mathbb{E}[\operatorname{var}[X \mid W]] + \operatorname{var}[\mathbb{E}[X \mid W]].$$

Homoskedasticity and Heteroskedasticity

同方差和异方差

Definition

The error is homoskedastic if $\sigma^2(x) = \sigma^2$ does not depend on x.

Definition

The error is heteroskedastic if $\sigma^2(x)$ depends on x.

Regression Derivative

One way to interpret the CEF $m(x) = \mathbb{E}[Y \mid X = x]$ is in terms of how marginal changes in the regressors x imply changes in the conditional mean of the response variable Y. It is typical to consider marginal changes in a single regressor, say X_1, holding the remainder fixed. When a regressor X_1 is continuously distributed, we define the marginal effect of a change in X_1, holding the variables $X_2, \ldots, X_k$ 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

$$m\left(1, x_2, \ldots, x_k\right) - m\left(0, x_2, \ldots, x_k\right).$$

Regression Derivative

We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(x) = \begin{cases} \frac{\partial}{\partial x_1} m\left(x_1, \ldots, x_k\right), & \text{if } X_1 \text{ is continuous} \\ m\left(1, x_2, \ldots, x_k\right) - m\left(0, x_2, \ldots, x_k\right), & \text{if } X_1 \text{ is binary.} \end{cases}$$

Collecting the k effects into one $k \times 1$ vector, we define the regression derivative with respect to X :

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

When all elements of X are continuous, then we have the simplification $\nabla m(x) = \frac{\partial}{\partial x} m(x)$, the vector of partial derivatives.

Linear CEF

- An important special case is when the CEF $m(x) = \mathbb{E}[Y \mid X = x]$ is linear in x. In this case we can write the mean equation as

$$m(x) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \beta_{k+1}.$$

线性模型在这里是求条件期望的一种方式。

- Notationally it is convenient to write this as a simple function of the vector x. An easy way to do so is to augment the regressor vector X by listing the number " 1 " as an element. We call this the "constant" and the corresponding coefficient is called the "intercept". Equivalently, specify that the final element [9] of the vector x is $x_k = 1$.

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

- With this redefinition, the CEF is

  $$m(x) = x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k = x'\beta \tag{1}$$

  where

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

  is a $k \times 1$ coefficient vector. This is the linear CEF model. It is also often called the linear regression model, or the regression of Y on X.

Definition

Linear CEF Model

$$
\begin{aligned}
Y &= X'\beta + e \\
\mathbb{E}[e \mid X] &= 0
\end{aligned}
$$

Definition

Homoskedastic Linear CEF Model

$$
\begin{aligned}
Y &= X'\beta + e \\
\mathbb{E}[e \mid X] &= 0 \\
\mathbb{E}\left[e^2 \mid X\right] &= \sigma^2
\end{aligned}
$$

Linear CEF with Nonlinear Effects

$$m\left(x_1, x_2\right)=x_1 \beta_1+x_2 \beta_2+x_1^2 \beta_3+x_2^2 \beta_4+x_1 x_2 \beta_5+\beta_6 \tag{2}$$

To simplify the expression we define the transformations $x_3=x_1^2, x_4=x_2^2, x_5=x_1 x_2$, and $x_6=1$, and redefine the regressor vector as $x=\left(x_1, \ldots, x_6\right)^{\prime}$. With this redefinition, $m\left(x_1, x_2\right)=x^{\prime} \beta$ which is linear in β. For most econometric purposes (estimation and inference on β) the linearity in β is all that is important.

An exception is in the analysis of regression derivatives. In nonlinear equations such as 2 the regression derivative should be defined with respect to the original variables not with respect to the transformed variables. Thus

$$\frac{\partial}{\partial x_1} m\left(x_1, x_2\right)=\beta_1+2 x_1 \beta_3+x_2 \beta_5$$
$$\frac{\partial}{\partial x_2} m\left(x_1, x_2\right)=\beta_2+2 x_2 \beta_4+x_1 \beta_5 .$$

We see that in the model 2, the regression derivatives are not a simple coefficient, but are functions of several coefficients plus the levels of (x_1, x_2). Consequently it is difficult to interpret the coefficients individually. It is more useful to interpret them as a group.

We typically call β_5 the **interaction effect**. Notice that it appears in both regression derivative equations and has a symmetric interpretation in each. If $\beta_5>0$ then the regression derivative with respect to x_1 is increasing in the level of x_2 (and the regression derivative with respect to x_2 is increasing in the level of x_1), while if $\beta_5<0$ the reverse is true.

Linear CEF with Dummy Variables

- When all regressors take a finite set of values it turns out the CEF can be written as a linear function of regressors.
- This simplest example is a **binary** variable which takes only two distinct values. For example, in traditional data sets the variable gender takes only the values man and woman (or male and female). Binary variables are extremely common in econometric applications and are alternatively called **dummy variables** or **indicator variables**.
- the conditional mean can only take two distinct values. For example:

$$\mathbb{E}[Y \mid \text{ gender }] = \begin{cases} \mu_0 & \text{if} \quad \text{gender} = \text{man} \\ \mu_1 & \text{if} \quad \text{gender} = \text{woman} \,. \end{cases}$$

- To facilitate a mathematical treatment we record dummy variables with the values $0, 1$. For example

$$X_1 = \begin{cases} 0 & \text{if} \quad \text{gender} = \text{man} \\ 1 & \text{if} \quad \text{gender} = \text{woman} \,. \end{cases}$$

- Given this notation we write the conditional mean as a linear function of the dummy variable X_1. Thus $\mathbb{E}\left[Y \mid X_1\right] = \beta_1 X_1 + \beta_2$ where $\beta_1 = \mu_1 - \mu_0$ and $\beta_2 = \mu_0$. In this simple regression equation the intercept β_2 is equal to the conditional mean of Y for the $X_1 = 0$ subpopulation (men) and the slope β_1 is equal to the difference in the conditional means between the two subpopulations.
- Alternatively,

$$X_1 = \begin{cases} 1 & \text{if} \quad \text{gender} = \text{man} \\ 0 & \text{if} \quad \text{gender} = \text{woman} . \end{cases}$$

- 这里有一个给变量取名的问题。如果把 X_1 命名为“性别”，则无法区分上面两种情形。所以应该取名为“男性”或“女性”。

Now suppose we have two dummy variables X_1 and X_2. For example, $X_2 = 1$ if the person is married, else $X_2 = 0$. The conditional mean given X_1 and X_2 takes at most four possible values:

$$\mathbb{E}\left[Y \mid X_1, X_2\right] = \begin{cases} \mu_{00} & \text{if} \quad X_1 = 0 \text{ and } X_2 = 0 \qquad \text{(unmarried men)} \\ \mu_{01} & \text{if} \quad X_1 = 0 \text{ and } X_2 = 1 \qquad \text{(married men)} \\ \mu_{10} & \text{if} \quad X_1 = 1 \text{ and } X_2 = 0 \qquad \text{(unmarried women)} \\ \mu_{11} & \text{if} \quad X_1 = 1 \text{ and } X_2 = 1 \qquad \text{(married women).} \end{cases}$$

In this case we can write the conditional mean as a linear function of X, X_2 and their product $X_1 X_2$:

$$\mathbb{E}\left[Y \mid X_1, X_2\right] = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4$$

where $\beta_1 = \mu_{10} - \mu_{00}, \beta_2 = \mu_{01} - \mu_{00}, \beta_3 = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$, and $\beta_4 = \mu_{00}$. We can view the coefficient β_1 as the effect of gender on expected log wages for unmarried wage earners, the coefficient β_2 as the effect of marriage on expected log wages for men wage earners, and the coefficient β_3 as the difference between the effects of marriage on expected log wages among women and among men.

- Alternatively, it can also be interpreted as the difference between the effects of gender on expected log wages among married and non-married wage earners. Both interpretations are equally valid. We often describe β_3 as measuring the interaction between the two dummy variables, or the interaction effect, and describe $\beta_3 = 0$ as the case when the interaction effect is zero.
  In this setting we can see that the CEF is linear in the three variables $(X_1, X_2, X_1 X_2)$. To put the model in the framework of Section 2.13 we define the regressor $X_3 = X_1 X_2$ and the regressor vector as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ 1 \end{pmatrix}.$$

  So even though we started with only 2 dummy variables, the number of regressors (including the intercept) is 4 .

- 为什么说，只要所有的解释变量都只能取有限个值，那么 CEF 就一定是线性的？例如：

$$X_3 = \begin{cases} 1 & \text{if} \quad \text{white} \\ 2 & \text{if} \quad \text{Black} \\ 3 & \text{if} \quad \text{other} \end{cases} \qquad \mathbb{E}\left[Y \mid X_3\right] = \begin{cases} \mu_1 & \text{if} \quad X_3 = 1 \\ \mu_2 & \text{if} \quad X_3 = 2 \\ \mu_3 & \text{if} \quad X_3 = 3 \end{cases}$$

- 设：

$$X_4 = \begin{cases} 1 & \text{if} \quad \text{Black} \\ 0 & \text{if} \quad \text{not Black} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{if} \quad \text{other} \\ 0 & \text{if} \quad \text{notother.} \end{cases}$$

- 则可以得到：

$$X_3 = \begin{cases} 1 & \text{if} \quad X_4 = 0 \text{ and } X_5 = 0 \\ 2 & \text{if} \quad X_4 = 1 \text{ and } X_5 = 0 \\ 3 & \text{if} \quad X_4 = 0 \text{ and } X_5 = 1. \end{cases}$$

$$\mathbb{E}\left[Y \mid X_3\right] = \mathbb{E}\left[Y \mid X_4, X_5\right] = \beta_1 X_4 + \beta_2 X_5 + \beta_3.$$

- 这就是为什么在实操中，我们非常推荐大家把分类变量转换成 dummies 放入到回归当中。
- 这是“最好的”求条件期望的方式。

Best Linear Predictor

- While the conditional mean $m(X) = \mathbb{E}[Y \mid X]$ is the best predictor of Y among all functions of X, its functional form is typically unknown. In particular, the linear CEF model is empirically unlikely to be accurate unless X is discrete and low-dimensional so all interactions are included. Consequently, in most cases it is more realistic to view the linear specification (1) as an approximation. In this section we derive a specific approximation with a simple interpretation.
- Theorem **Conditional Expectation as Best Predictor** showed that the conditional mean $m(X)$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.

Assumption (2.1)

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.*

- $\|X\| = (x'x)^{1/2}$，表示向量 x 的 Euclidean length （欧几里得长度）.
- The first two parts of Assumption (3.1) imply that the variables Y and X have finite means, variances, and covariances.
- The third part of the assumption is more technical, and its role will become apparent shortly. It is equivalent to imposing that the columns of the matrix $\mathbf{Q}_{XX} = \mathbb{E}\left[XX'\right]$ are linearly independent, or that the matrix is invertible.
- A linear predictor for Y is a function $X'\beta$ for some $\beta \in \mathbb{R}^k$. The mean squared prediction error is

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

- The best linear predictor of Y given X, written $\mathscr{P}[Y \mid X]$, is found by selecting the β which minimizes $S(\beta)$.

Definition

The **Best Linear Predictor** of Y given X is

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

where β minimizes the mean squared prediction error

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

The minimizer

$$\beta = \underset{b \in \mathbb{R}^k}{\text{argmin}} S(b) \tag{2.18}$$

is called the **Linear Projection Coefficient**.

推导

- $S(\beta) = \mathbb{E}\left[(Y - X'\beta)^2\right]$ can be written out as a quadratic function of β:

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

- $S(\beta)$ 写成二次函数形式，是为了最小化问题有解。一阶条件为：

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

$$2\mathbb{E}[XY] = 2\mathbb{E}\left[XX'\right]\beta$$

- 令 $\boldsymbol{Q}_{XY} = \mathbb{E}[XY]$ is $k \times 1$，$\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is $k \times k$。则上面的解为：

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

或

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

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

- $\boldsymbol{Q}_{XX}$ 是一个 $k \times k$ 阶矩阵，$\boldsymbol{Q}_{XY}$ 是一个 $k \times 1$ 的列向量。
- 所以 $\frac{\mathbb{E}[XY]}{\mathbb{E}[XX']}$ 和 $\mathbb{E}[XY](\mathbb{E}[XX'])^{-1}$ 的写法都不对。
- 假设“$\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is positive definite”意味着 $\boldsymbol{Q}_{XX}$ 可以求逆，所以一阶条件一定是有解的。
- best linear predictor，**linear projection** of Y on X（Y 在 X 上的线性投影）:

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

- The **projection error** is
$$e = Y - X'\beta$$
- 把 Y 分解为 linear predictor 和 linear error:
$$Y = X'\beta + e \tag{3}$$
- $X'\beta$ is the best linear predictor of Y given X, or the linear projection of Y on X. 方程3也被称为 Y 对 X 的回归。
- “把 Y 跑在 X 上。”

- 如果只有一个 x，是求一条直线。
- 如果有两个 x, x_1, x_2，回归求的是什么?

- 如果只有一个 x，是求一条直线。
- 如果有两个 x, x_1, x_2，回归求的是什么?
    - 求的是一个平面。

- projection error e 的性质

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

- 推导

$$\begin{aligned}\mathbb{E}[Xe] &= \mathbb{E}\left[X(Y - X'\beta)\right] \\ &= \mathbb{E}[XY] - \mathbb{E}\left[XX'\right]\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] \\ &= 0\end{aligned}$$

- X 是矩阵，e 是向量，所以 $\mathbb{E}[Xe] = 0$ 是一个 0 向量。所以对于每一行 j 都有：

$$\mathbb{E}[X_j e] = 0$$

$X_k = 1$ 时，则有：

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

Theorem (2.9 Properties of Linear Projection Model)

Under Assumption 2.1,

1. *The moments* $\mathbb{E}[XX']$ *and* $\mathbb{E}[XY]$ *exist with finite elements.*
2. *The linear projection coefficient (2.18) exists, is unique, and equals*

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

3. *The best linear predictor of y given x is*

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

4. *The projection error* $e = Y - X'\beta$ *exists. It satisfies* $\mathbb{E}\left[e^2\right] < \infty$ *and* $\mathbb{E}[Xe] = 0$.
5. *If X contains an constant, then* $\mathbb{E}[e] = 0$.
6. *If* $\mathbb{E}|Y|^r < \infty$ *and* $\mathbb{E}\|X\|^r < \infty$ *for* $r \geq 2$ *then* $\mathbb{E}|e|^r < \infty$.

需要强调的事情

It is useful to reflect on the generality of Theorem 2.9. The only restriction is Assumption 2.1. Thus for any random variables (Y, X) with finite variances we can define a linear equation $Y = X'\beta + e$ with the properties listed in Theorem 2.9. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model $Y = X'\beta + e$ exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation $Y = X'\beta + e$ is defined as the **best linear predictor**. It is **not** necessarily a **conditional mean**, nor a **parameter of a structural or causal economic model**.

Linear CEF Model：

$$
\begin{aligned}
Y &= X'\beta + e \\
\mathbb{E}[e \mid X] &= 0
\end{aligned}
$$

Invertibility and Identification

- The linear projection coefficient $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is invertible. The matrix $\boldsymbol{Q}_{XX}$ is often called the design matrix as in experimental settings the researcher is able to control $\boldsymbol{Q}_{XX}$ by manipulating the distribution of the regressors X.
- Observe that for any non-zero $\alpha \in \mathbb{R}^k$,

$$\alpha'\boldsymbol{Q}_{XX}\alpha = \mathbb{E}\left[\alpha' XX'\alpha\right] = \mathbb{E}\left[\left(\alpha' X\right)^2\right] \geq 0$$

so $\boldsymbol{Q}_{XX}$ by construction is positive semi-definite, conventionally written as $\boldsymbol{Q}_{XX} \geq 0$. The assumption that it is positive definite means that this is a strict inequality, $\mathbb{E}\left[\left(\alpha' X\right)^2\right] > 0$. This is conventionally written as $\boldsymbol{Q}_{XX} > 0$. This condition means that there is no non-zero vector α such that $\alpha' X = 0$ identically. Positive definite matrices are invertible. Thus when $\boldsymbol{Q}_{XX} > 0$ then $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]$ exists and is uniquely defined. In other words, if we can exclude the possibility that a linear function of X is degenerate, then β is uniquely defined.

- Theorem 2.5 shows that the linear projection coefficient β is identified (uniquely determined) under Assumption 2.1. The key is invertibility of $\boldsymbol{Q}_{XX}$. Otherwise, there is no unique solution to the equation

$$\boldsymbol{Q}_{XX}\beta = \boldsymbol{Q}_{XY}.$$

  When $\boldsymbol{Q}_{XX}$ is not invertible there are multiple solutions to (2.29). In this case the coefficient β is not identified as it does not have a unique value.

Linear Predictor Error Variance (定义)

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 can write σ^2 as

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

Regression Coefficients

现在有一个线性回归方程：X 表示一个变量

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

其中 α 是截距项，X 不再包含常数项。两边取期望：

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

也可以写成：$\mu_Y = \mu_X'\beta + \alpha$，其中 $\mu_Y = \mathbb{E}[Y]$，$\mu_X = \mathbb{E}[X]$。
由于 $\mathbb{E}[e] = 0$ ，可以得到 $\alpha = \mu_Y - \mu_X'\beta$。进而得到线性方程：

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

根据线性回归模型的结果得到：

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

系数是 X 和 Y 的协方差。

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

$$\begin{aligned} Y &= X'\beta + e \\ &= X_1'\beta_1 + X_2'\beta_2 + e \qquad (2.42) \\ \mathbb{E}[Xe] &= 0. \end{aligned}$$

In this section we derive formulae for the sub-vectors β_1 and β_2. Partition $\boldsymbol{Q}_{XX}$ conformably with X

$$\boldsymbol{Q}_{XX} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[X_1X_1'\right] & \mathbb{E}\left[X_1X_2'\right] \\ \mathbb{E}\left[X_2X_1'\right] & \mathbb{E}\left[X_2X_2'\right] \end{bmatrix}$$

and similarly

$$\boldsymbol{Q}_{XY} = \begin{bmatrix} \boldsymbol{Q}_{1Y} \\ \boldsymbol{Q}_{2Y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[X_1Y\right] \\ \mathbb{E}\left[X_2Y\right] \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.3)

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21} & \\ \boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

where $\boldsymbol{Q}_{11\cdot 2} \overset{\text{def}}{=} \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot 1} \overset{\text{def}}{=} \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus

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

We have shown that $\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1Y\cdot 2}$ and $\beta_2 = \boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{2Y\cdot 1}$.

Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors β_1 and β_2. We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection. Take equation (2.42) for the case $\dim X_1 = 1$ so that $\beta_1 \in \mathbb{R}$.

$$Y = X_1 \beta_1 + X_2' \beta_2 + e \tag{2.44}$$

Now consider the projection of X_1 on X_2 :

$$\begin{aligned} X_1 &= X_2' \gamma_2 + u_1 \\ \mathbb{E}\left[X_2 u_1\right] &= 0. \end{aligned}$$

$\gamma_2 = \boldsymbol{Q}_{22}^{-1} \boldsymbol{Q}_{21}$ and $\mathbb{E}\left[u_1^2\right] = \boldsymbol{Q}_{11 \cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12} \boldsymbol{Q}_{22}^{-1} \boldsymbol{Q}_{21}$. We can also calculate that

$$\mathbb{E}\left[u_1 Y\right] = \mathbb{E}\left[\left(X_1 - \gamma_2' X_2\right) Y\right] = \mathbb{E}\left[X_1 Y\right] - \gamma_2' \mathbb{E}\left[X_2 Y\right] = \boldsymbol{Q}_{1Y} - \boldsymbol{Q}_{12} \boldsymbol{Q}_{22}^{-1} \boldsymbol{Q}_{2Y} = \boldsymbol{Q}_{1Y \cdot 2}.$$

We have found that

$$\beta_1 = \boldsymbol{Q}_{11 \cdot 2}^{-1} \boldsymbol{Q}_{1Y \cdot 2} = \frac{\mathbb{E}\left[u_1 Y\right]}{\mathbb{E}\left[u_1^2\right]}$$

the coefficient from the simple regression of Y on u_1.

What this means is that in the multivariate projection equation (2.44) , the coefficient β_1 equals the projection coefficient from a regression of Y on u_1, the error from a projection of X_1 on the other regressors X_2. The error u_1 can be thought of as the component of X_1 which is not linearly explained by the other regressors. Thus the coefficient β_1 equals the linear effect of X_1 on Y after stripping out the effects of the other variables.

There was nothing special in the choice of the variable X_1. This derivation applies symmetrically to all coefficients in a linear projection. Each coefficient equals the simple regression of Y on the error from a projection of that regressor on all the other regressors. Each coefficient equals the linear effect of that variable on Y after linearly controlling for all the other regressors.

Omitted Variable Bias

- If variables X2 are not observed:

$$\begin{aligned} Y &= X_1'\gamma_1 + u \\ \mathbb{E}\left[X_1 u\right] &= 0 \end{aligned} \tag{2.45}$$

Notice that we have written the coefficient as γ_1 rather than β_1 and the error as u rather than e. This is because (2.45) is different than (2.42). Goldberger (1991) introduced the catchy labels **long regression** for (2.42) and **short regression** for (2.45) to emphasize the distinction. Typically, $\beta_1 \neq \gamma_1$, except in special cases. To see this, we calculate

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

where $\Gamma_{12} = \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$ is the coefficient matrix from a projection of X_2 on X_1 where we use the notation from Section **Regression Sub-Vectors**.

title

- Observe that $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 \neq \beta_1$ unless $\Gamma_{12} = 0$ or $\beta_2 = 0$. Thus the short and long regressions have different coefficients. They are the same only under one of two conditions. First, if the projection of X_2 on X_1 yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on X_2 in (2.42) is zero. The difference $\Gamma_{12}\beta_2$ between γ_1 and β_1 is known as omitted variable bias. It is the consequence of omission of a relevant correlated variable.
- Γ_{12} 表示 X_1 和 X_2 的相关性，β_2 表示 X_2 和 Y 的相关性。
- 避免遗漏变量问题的标准建议是尽可能把可能的相关的变量都囊括到回归中。但是现实中无法做到。
    - 更现实的做法是，研究者要知晓自己的回归可能面临的遗漏变量问题，并尽可能地去讨论遗漏变量问题造成的后果，以及为自己辩护。

Unfortunately the above simple characterization of omitted variable bias does not immediately carry over to more complicated settings, as discovered by Luca, Magnus, and Peracchi (2018). For example, suppose we compare three nested projections

$$\begin{aligned} Y &= X_1'\gamma_1 + u_1 \\ Y &= X_1'\delta_1 + X_2'\delta_2 + u_2 \\ Y &= X_1'\beta_1 + X_2'\beta_2 + X_3'\beta_3 + e \end{aligned}$$

We can call them the short, medium, and long regressions. Suppose that the parameter of interest is β_1 in the long regression. We are interested in the consequences of omitting X_3 when estimating the medium regression, and of omitting both X_2 and X_3 when estimating the short regression. In particular we are interested in the question: Is it better to estimate the short or medium regression, given that both omit X_3 ? Intuition suggests that the medium regression should be ”less biased” but it is worth investigating in greater detail. By similar calculations to those above, we find that

$$\begin{aligned} \gamma_1 &= \beta_1 + \Gamma_{12}\beta_2 + \Gamma_{13}\beta_3 \\ \delta_1 &= \beta_1 + \Gamma_{13\cdot 2}\beta_3 \end{aligned}$$

where $\Gamma_{13\cdot 2} = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{13\cdot 2}$ using the notation from Section **Regression Sub-Vectors**.

We see that the bias in the short regression coefficient is $\Gamma_{12}\beta_2 + \Gamma_{13}\beta_3$ which depends on both β_2 and β_3, while that for the medium regression coefficient is $\Gamma_{13\cdot 2}\beta_3$ which only depends on β_3. So the bias for the medium regression is less complicated and intuitively seems more likely to be smaller than that of the short regression. However it is impossible to strictly rank the two. It is quite possible that γ_1 is less biased than δ_1. Thus as a general rule it is strictly impossible to state that estimation of the medium regression will be less biased than estimation of the short regression.

Best Linear Approximation

There are alternative ways we could construct a linear approximation $X'\beta$ to the conditional mean $m(X)$. In this section we show that one alternative approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of $X'\beta$ to $m(X)$ as the expected squared difference between $X'\beta$ and the conditional mean $m(X)$, 先定义 $X'\beta$ 和 $m(X)$ 之间偏差：

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

The function $d(\beta)$ is a measure of the deviation of $X'\beta$ from $m(X)$. If the two functions are identical then $d(\beta) = 0$, otherwise $d(\beta) > 0$. We can also view the mean-square difference $d(\beta)$ as a density-weighted average of the function $\left(m(X) - X'\beta\right)^2$ since

$$d(\beta) = \int_{\mathbb{R}^k} \left(m(x) - x'\beta\right)^2 f_X(x) dx$$

where $f_X(x)$ is the marginal density of X.

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

$$\beta = \underset{b\in\mathbb{R}^k}{\operatorname{argmin}} d(b) \tag{2.46}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.18) selects β to minimize the expected squared prediction error, while the best linear approximation (2.46) selects β to minimize the expected squared approximation error. Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.18) plus an application of conditional expectations we can find that

$$\begin{aligned} \beta &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[Xm(X)] \qquad (2.47)\\ &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] \qquad (2.48)\end{aligned}$$

(see Exercise 2.19). Thus (2.46) equals (2.18). We conclude that the definition (2.46) can be viewed as an alternative motivation for the linear projection coefficient.

Limitations of the Best Linear Projection

Let's compare the linear projection and linear CEF models.
From Theorem 2.4.4 we know that the CEF error has the property $\mathbb{E}[Xe] = 0$. Thus a linear CEF is the best linear projection. However, the converse is not true as the projection error does not necessarily satisfy $\mathbb{E}[e \mid X] = 0$. Furthermore, the linear projection may be a poor approximation to the CEF.
To see these points in a simple example, suppose that the true process is $Y = X + X^2$ with $X \sim \mathrm{N}(0, 1)$. In this case the true CEF is $m(x) = x + x^2$ and there is no error. Now consider the linear projection of Y on X and a constant, namely the model $Y = \beta X + \alpha + e$. Since $X \sim \mathrm{N}(0, 1)$ then X and X^2 are uncorrelated and the linear projection takes the form $\mathscr{P}[Y \mid X] = X + 1$. This is quite different from the true CEF $m(X) = X + X^2$. The projection error equals $e = X^2 - 1$ which is a deterministic function of X yet is uncorrelated with X. We see in this example that a projection error need not be a CEF error and a linear projection can be a poor approximation to the CEF.

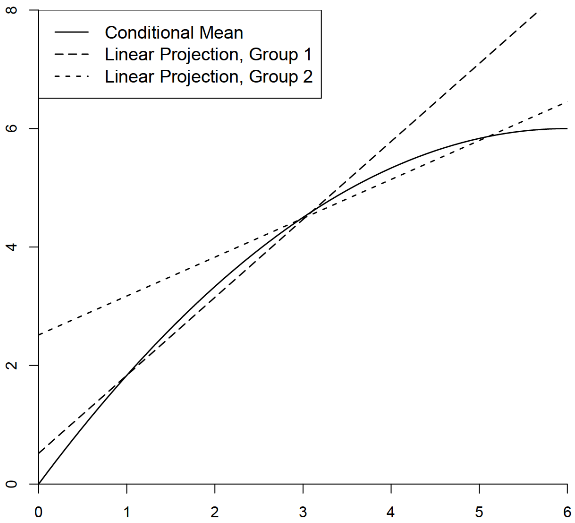


Figure: Conditional Mean and Two Linear Projections

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is nonlinear. We illustrate the issue in Figure 1 for a constructed joint distribution of Y and X. The solid line is the nonlinear CEF of Y given X. The data are divided in two groups - Group 1 and Group 2 - which have different marginal distributions for the regressor X, and Group 1 has a lower mean value of X than Group 2. The separate linear projections of Y on X for these two groups are displayed in the figure by the dashed lines. These two projections are distinct approximations to the CEF. A defect with linear projection is that it leads to the incorrect conclusion that the effect of X on Y is different for individuals in the two groups. This conclusion is incorrect because in fact there is no difference in the conditional mean function. The apparent difference is a by-product of linear approximations to a nonlinear mean combined with different marginal distributions for the conditioning variables.

Causal Effects

- 以上内容都不涉及因果性。但是我们的研究经常关心因果性。
- 研究因果性有两个障碍：
  - 因果效应对每个个体的作用不同；
  - 因果效应不可观测。

总结

- CEF 是对 Y 的最佳预测（best predictor）。
- 优化问题：

  $$\min_{\beta \in \mathbb{R}^p} E[(Y - X'\beta)^2]$$

  得到的解 $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$，$X'\beta$ 是对 Y 的最佳线性预测（best linear predictor, BLP），也是给定 X 的条件期望 $E(Y|X)$ 的 BLP.
- $X'\beta$ 是对 CEF 的最佳线性逼近（best linear approximation, BLA）
- $\mathbb{E}[(Y - X'\beta)X] = 0$，根据迭代期望定律可以得到：

  $$\mathbb{E}[(\mathbb{E}[Y|X] - X'\beta)X] = 0$$

- 对 X 进行线性变换，可以构造出 X 的多项式，可以十分接近 CEF。

Approximating a Smooth Function with a Poly-nomial Dictionary

Suppose $W \sim U(0,1)$ where U denotes the continuous uniform distribution, and

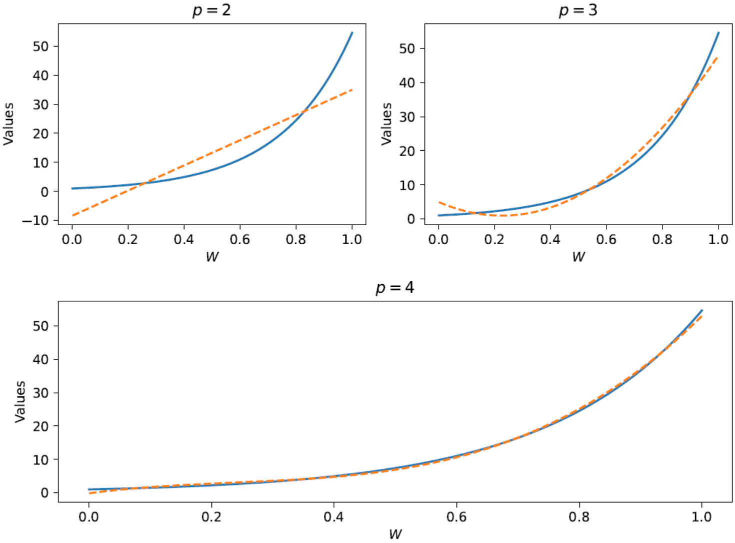
$$g(W) = \exp(4 \cdot W).$$

We use

$$T(W) = \underbrace{(1, W, W^2, \ldots, W^{p-1})'}_{p \text{ terms}}$$

to form the BLA/BLP, $\beta' T(W)$. Figure 1.2 provides a sequence of panels that illustrate the approximation properties of the BLA/BLP corresponding to $p = 2$, 3, and 4:

- With $p = 2$ we get a linear in W approximation to $g(W)$. As the figure shows, the quality of this approximation is poor.
- With $p = 3$ we get a quadratic-in-W approximation to $g(W)$. Here, the approximation quality is markedly improved relative to $p = 2$ though approximation errors are still clearly visible.
- With $p = 4$ we get a cubic-in-W approximation to $g(W)$, and the quality of approximation appears to be excellent.


$p = 2$
Values
50
40
30
20
10
0
−10
0.0
0.2
0.4
0.6
0.8
1.0
W
$p = 3$
Values
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
W
$p = 4$
Values
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
W

When we have multiple variables, we may generate transforma- tions of each of the variables and employ interactions – products involving these terms. As a simple concrete example, consider a case with two raw regressors, W_1 and W_2. We could build polynomials of second order in each of the raw regressors – $(1, W_1, W_1^2)$, $(1, W_2, W_2^2)$. We may then collect these variables along with the interaction in the raw regressors, $W_1 W_2$ in a vector

$$(1, W_1, W_2, W_1^2, W_2^2, W_1 W_2)$$

for use in a regression model. There are, of course, many other possibilities such as considering higher order polynomial terms, e.g. W_1^3; higher order interactions, e.g. $W_1^2 W_2$; and other nonlinear transformations, e.g. $\log(W_1)$.