计量模型及应用

李世纪

2025-11-26

目录

说明

econometric models（计量模型）

简介

写本书是因为计量经济学已经发生了巨大改变，因果效应成为了研究的主流目标，而传统的线性回归模型已经不能满足这种需求。近年来，随着计算能力的提升和数据获取的便利，越来越多的新方法被提出并应用于实际问题中。

I 数据与模型

1 线性回归基础

本章导读

本章旨在系统梳理线性回归模型作为计量经济学核心方法的理论基石。我们将在先修课程的基础上，深化对数据生成过程的理解，并严格审视支撑经典推断的基本假设体系。本章不仅阐述在理想条件下（假设成立）模型所具有的最优性质，更将重点剖析当关键假设被违反时，对参数估计、统计推断与模型预测所造成的具体影响及其根本原因。这种" 建构-解构" 的视角，旨在培养严谨的实证思维习惯，为后续学习处理复杂现实数据的计量方法奠定坚实基础。

1.1 计量视角下的回归分析：一个回顾与统一

1.1.1 条件期望函数与最优线性预测

在计量经济学中，我们关注的是因变量 Y 在给定解释变量 X 的条件下的行为。条件期望函数（Conditional Expectation Function, CEF）定义为：

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

其中 $f(X)$ 是 X 的任意函数。条件期望函数具有一个重要性质：它是给定 X 下对 Y 的最小均方误差预测。即对于任意函数 $g(X)$，有：

$$E[(Y - f(X))^2|X] \leq E[(Y - g(X))^2|X]$$

在实际应用中，我们通常不知道 $f(X)$ 的函数形式。线性回归模型提供了一个简洁的近似框架：

$$E(Y|X) \approx X\beta$$

这里，$X\beta$ 是对真实条件期望函数的最佳线性预测。

1.1.2 总体回归与样本回归：统计推断的桥梁

计量经济学区分了两个关键概念：

1. 总体回归函数（Population Regression Function, PRF）：

$$Y_i = X_i'\beta + \varepsilon_i, \quad i = 1, ..., N$$

   其中 β 是未知的总体参数，ε_i 是随机扰动项，满足 $E(\varepsilon_i|X_i) = 0$。

2. 样本回归函数（Sample Regression Function, SRF）：

$$Y_i = X_i'\hat{\beta} + \hat{\varepsilon}_i$$

   其中 $\hat{\beta}$ 是基于样本数据对 β 的估计，$\hat{\varepsilon}_i$ 是残差项。

统计推断的核心任务就是从样本回归中获取关于总体回归的可靠信息。普通最小二乘（OLS）估计量 $\hat{\beta}$ 通过最小化残差平方和得到：

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$$

在矩阵形式下，解为：

$$\hat{\beta} = (X'X)^{-1}X'Y$$

其中 X 为 $n \times k$ 的解释变量矩阵，Y 为 $n \times 1$ 的被解释变量向量。

1.1.3 多元线性回归的矩阵表述与几何解释

考虑包含 n 个观测值和 k 个解释变量（包含常数项）的多元线性回归模型：

$$Y = X\beta + \varepsilon$$

其中：- Y 为 $n \times 1$ 的被解释变量向量 - X 为 $n \times k$ 的解释变量矩阵（秩为 k）- β 为 $k \times 1$ 的未知参数向量 - ε 为 $n \times 1$ 的随机扰动向量

OLS 估计量 $\hat{\beta}$ 的几何解释为：通过将 Y 投影到 X 的列空间上，得到 Y 在该空间上的正交投影 $\hat{Y} = X\hat{\beta} = P_X Y$，其中 $P_X = X(X'X)^{-1}X'$ 为投影矩阵。残差向量 $\hat{\varepsilon} = Y - \hat{Y} = (I_n - P_X)Y$

垂直于 X 的列空间。

1.2 经典线性回归模型：假设体系、性质与关联

1.2.1 线性于参数：模型设定与误设偏误

经典线性回归模型假设因变量 Y 与参数 β 呈线性关系：

$$Y = X\beta + \varepsilon$$

这里的" 线性" 指的是对参数线性，而非对变量线性。模型可以包含变量的非线性变换（如对数、平方项等）。

违反后果：若真实关系非线性但误设为线性，则产生模型设定偏误，导致 OLS 估计量有偏且不一致。

1.2.2 严格外生性：无偏性与一致性的基石

严格外生性假设要求：

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

这意味着给定所有解释变量 X，扰动项的条件均值为零。一个较弱但足够的条件是均值独立：

$$E(\varepsilon_i|X_i) = 0, \quad i = 1, ..., n$$

违反后果：当 $E(\varepsilon_i|X_i) \neq 0$ 时（即 X 与 ε 相关），产生内生性问题，导致 OLS 估计量有偏且不一致。

常见原因包括：

1. 遗漏重要变量

2. 测量误差

3. 联立性（双向因果关系）

4. 样本选择偏误

1.2.3 无完全多重共线性：估计可行性与精度前提

无完全多重共线性要求解释变量矩阵 X 列满秩：

$$\text{rank}(X) = k$$

即不存在严格线性关系：$\sum_{j=1}^{k} a_j X_j = 0$（除非所有 $a_j = 0$）。

违反后果：1. 完全共线性：$(X'X)$ 不可逆，参数无法唯一识别 2. 近似共线性：$(X'X)$ 接近奇异，导致：- 估计方差增大，估计精度下降 - 估计值对样本微小变化敏感 - t 检验功效降低（难以拒绝原假设）

衡量共线性的常用指标是方差膨胀因子（VIF）：

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

其中 R_j^2 是 X_j 对其他解释变量回归的决定系数。通常认为 VIF > 10 表明严重共线性。

1.2.4 球形扰动项假设：有效性基础

球形扰动项假设包含两个部分：

(a) 同方差性：

$$\text{Var}(\varepsilon_i|X) = \sigma^2, \quad \forall i = 1, ..., n$$

即扰动项的条件方差为常数。

(b) 无自相关：

$$\text{Cov}(\varepsilon_i, \varepsilon_j|X) = 0, \quad \forall i \neq j$$

即不同观测的扰动项相互独立。

在矩阵形式下，球形扰动项假设等价于：

$$\text{Var}(\varepsilon|X) = \sigma^2 I_n$$

违反后果：当球形扰动项假设不成立时，OLS 估计量仍是无偏和一致的，但不再是有效的（即不再具有最小方差），且常规标准误估计是有偏的，导致假设检验失效。

1.2.5 正态性假设：精确推断的充分条件

正态性假设要求：

$$\varepsilon|X \sim N(0, \sigma^2 I_n)$$

这是对扰动项分布的强化假设。

影响：

1. 有限样本：正态性假设下，OLS 估计量服从精确的正态分布，t 和 F 统计量分别服从精确的 t 和 F 分布

2. 大样本：由中心极限定理，即使扰动项非正态，OLS 估计量也具有渐近正态性，大样本推断仍然有效

3. 小样本非正态：假设检验和置信区间的准确性可能受影响

1.2.6 关键性质与假设的关联总结

下表总结了经典线性回归模型主要性质所依赖的核心假设：

性质	核心依赖假设	违背后果
无偏性	严格外生性 $E(\varepsilon\|X) = 0$	估计量有偏
一致性	均值独立 $E(\varepsilon_i\|X_i) = 0$ + 随机抽样	估计量不一致
有效性（最小方差）	同方差性 + 无自相关	不再是最优线性无偏估计
精确正态推断	正态性假设	小样本下检验可能不准确
大样本渐近推断	较弱的矩条件	通常仍成立

高斯-马尔可夫定理：在假设 1.2.1-1.2.4 下（线性性、严格外生性、无完全共线性、球形扰动项），OLS 估计量是最佳线性无偏估计量（Best Linear Unbiased Estimator, BLUE），即在线性无偏估计量类中具有最小方差。

1.3 假设违反的后果：系统性影响分析

1.3.1 对参数估计量性质的影响

有偏性与非一致性：外生性违背的灾难

当严格外生性假设 $E(\varepsilon|X) = 0$ 被违反时，设真实模型为：

$$Y = X\beta + \varepsilon, \quad E(\varepsilon|X) \neq 0$$

则 OLS 估计量的概率极限为：

$$\text{plim}_{n\to\infty}\hat{\beta}_{OLS} = \beta + \text{plim}\left(\frac{X'X}{n}\right)^{-1} \text{plim}\left(\frac{X'\varepsilon}{n}\right)$$

由于 $E(X'\varepsilon) \neq 0$，第二项非零，导致估计量不一致。在有限样本下，估计量也有偏：

$$E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'E(\varepsilon|X) \neq \beta$$

有效性丧失：球形扰动项违背的影响

当同方差性或无自相关假设被违反时，设 $\text{Var}(\varepsilon|X) = \Omega \neq \sigma^2 I_n$，则 OLS 估计量的条件方差为：

$$\text{Var}(\hat{\beta}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

而如果使用正确的广义最小二乘（GLS）估计量 $\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$，其方差为 $(X'\Omega^{-1}X)^{-1}$。根据高斯-马尔可夫定理，对于任意 $\Omega \neq \sigma^2 I_n$，有：

$$\text{Var}(\hat{\beta}_{GLS}|X) \leq \text{Var}(\hat{\beta}_{OLS}|X)$$

即 OLS 估计量不是有效的（方差不是最小）。

估计不稳定与方差膨胀：多重共线性的影响

在近似多重共线性下，$(X'X)$ 接近奇异，其特征值中至少有一个非常小。OLS 估计量的方差可以表示为：

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}$$

其中 $(1-R_j^2)$ 项反映了 X_j 与其他解释变量的相关性。当 $R_j^2 \to 1$ 时，方差趋于无穷大。虽然估计量仍无偏，但估计精度严重下降，估计值对样本微小变化极为敏感。

1.3.2 对统计假设检验的影响

标准误估计偏误：异方差与自相关的后果

当存在异方差或自相关时，OLS 的常规标准误估计：

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}, \quad \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

是有偏的。实际上，$\hat{\sigma}^2(X'X)^{-1}$ 收敛于：

$$\text{plim}\hat{\sigma}^2(X'X)^{-1} = \text{plim}\left(\frac{X'X}{n}\right)^{-1} \text{plim}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) \text{plim}\left(\frac{X'X}{n}\right)^{-1}$$

而真实渐近方差为：

$$\text{Avar}(\hat{\beta}) = \text{plim}\left(\frac{X'X}{n}\right)^{-1} \text{plim}\left(\frac{X'\Omega X}{n}\right) \text{plim}\left(\frac{X'X}{n}\right)^{-1}$$

两者相等仅当 $\Omega = \sigma^2 I_n$。标准误的偏误方向取决于 Ω 的结构：- 异方差：常规标准误可能高估或低估真实标准误 - 正自相关：常规标准误通常低估真实标准误，导致过度拒绝原假设（第一类错误增加）

推断结论失真：检验水平与功效扭曲

错误的 t 统计量：

$$t_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

不再服从标准正态或 t 分布（即使在渐近意义上），除非使用正确的标准误。这导致：1. 检验水平扭曲：实际显著性水平偏离名义水平（如 5%）2. 功效变化：检验发现真实效应的能力改变 3. 置信区间失效：实际覆盖率偏离置信水平

例如，当存在正自相关时，使用常规标准误的 t 检验会过度拒绝原假设（第一类错误膨胀）；而使用常规标准误的 F 检验也会有类似问题。

1.3.3 对模型预测的影响

点预测偏误：参数估计偏误的传递

如果 $\hat{\beta}$ 是有偏或不一致的，则点预测 $\hat{Y}_0 = X_0'\hat{\beta}$ 也是有偏的：

$$E(\hat{Y}_0|X) = X_0'E(\hat{\beta}|X) \neq X_0'\beta$$

偏误大小为 $X_0'[E(\hat{\beta}|X) - \beta]$。

预测区间不准确：方差与分布误设

即使 $\hat{\beta}$ 无偏，若标准误估计错误或扰动项分布误设，预测区间也会不准确。对于新观测 (X_0, Y_0)，常规预测区间为：

$$\hat{Y}_0 \pm t_{\alpha/2,n-k} \times \hat{\sigma}\sqrt{1 + X_0'(X'X)^{-1}X_0}$$

当存在异方差时，$\hat{\sigma}^2$ 不能准确估计预测方差；当存在自相关时，预测误差项 ε_0 与样本扰动项可能相关，进一步复杂化预测方差计算。这些都会导致预测区间的实际覆盖率偏离名义水平。

1.4 诊断、应对与因果推断的桥梁

1.4.1 诊断的逻辑：从理论先验到统计检验

计量经济学诊断应遵循系统化流程：

1. 理论先验分析：基于经济学理论判断可能存在的问题
    - 变量间理论上是否存在双向因果关系？
    - 是否可能遗漏了重要变量？
    - 数据生成过程是否暗示异方差或自相关？
2. 数据可视化探索：
    - 残差图：残差 vs. 拟合值、残差 vs. 解释变量
    - 分量加残差图（Partial Residual Plot）
    - Q-Q 图检验正态性
3. 正式统计检验：
    - 异方差检验：Breusch-Pagan 检验、White 检验
    - 自相关检验：Durbin-Watson 检验、Breusch-Godfrey 检验
    - 函数形式检验：RESET 检验、Ramsey 检验
    - 正态性检验：Jarque-Bera 检验、Shapiro-Wilk 检验

1.4.2 常见违背情形的应对路径导引

下表总结了主要假设违背的诊断方法及后续章节将详细讨论的应对策略：

违背类型	诊断方法	应对策略（后续章节）	关键思想
内生性	经济理论判断、Hausman 检验	工具变量法（第 4 章）、面板数据模型（第 3 章）	寻找外生变异来源
异方差	残差图、BP 检验、White 检验	稳健标准误（第 2 章）、WLS/FGLS（第 2 章）	修正标准误或加权估计
自相关	残差图、DW 检验、BG 检验	聚类标准误（第 2 章）、时间序列模型（第 5 章）	修正标准误或模型设定
非线性	残差图、RESET 检验	函数形式变换（第 2 章）、非线性模型（第 6 章）	更灵活的函数形式
共线性	相关系数矩阵、VIF	变量筛选、主成分回归（第 2 章）	降维或获取更多数据

1.4.3 从预测到因果：外生性假设的核心地位

线性回归可用于两个不同目的：

1. 预测：关注 $E(Y|X)$ 的准确估计
    - 允许黑箱方法
    - 侧重样本内拟合与样本外预测精度
    - 不要求 X 外生
2. 因果推断：关注 β 的因果解释
    - 要求 $E(\varepsilon|X) = 0$（条件均值独立）
    - β_j 解释为：X_j 外生变化一单位引起 Y 的平均因果效应
    - 需要证明/论证 X 的外生性

因果识别的基本问题：即使观测到 X 与 Y 相关，也无法区分：1. X 引起 Y 变化（因果效应）2. Y 引起 X 变化（反向因果）3. 第三因素 Z 同时影响 X 和 Y（混杂偏误）

内生性问题的来源：1. 遗漏变量偏误：未观测到的重要变量与 X 相关 2. 测量误差：X 的测量误差导致其与 ε 相关 3. 联立性偏误：X 与 Y 相互决定 4. 样本选择偏误：样本非随机导致 X 与 ε 相关

这些内生性问题使得 $E(\varepsilon|X) \neq 0$，破坏了因果解释的基础。第二部分（因果推断）将系统介绍解决这些问题的现代方法。

1.5 案例分析与代码实现

案例主题：教育回报率的实证分析——假设的审视与警示

核心目标：通过实际数据操作与模拟对比，直观感受假设成立与违反的差异。

分析步骤：

基准回归：使用 OLS 估计标准的 Mincer 方程，解释系数。

诊断探索：

绘制残差-拟合值图与残差-QQ 图，进行直观诊断。

计算方差膨胀因子（VIF），诊断共线性。

后果演示（模拟辅助）：

内生性演示：模拟一个遗漏重要能力变量的场景，展示 OLS 估计量的偏误。

异方差后果：在一个存在已知异方差结构的数据中，对比普通标准误与异方差稳健标准误的差异，展示其对 t 检验结论的影响。

报告对比：整理并对比 “理想情况” 与 “问题情况” 下的回归结果表，强调正确诊断与报告的重要性。

代码实现要点 (R/Python 双版本)：

数据操作与估计：pandas / statsmodels (Python) 或 dplyr / lm() (R)。

图形诊断：seaborn / matplotlib (Python) 或 ggplot2 (R)。

共线性诊断：statsmodels.stats.outliers_influence (Python) 或 car::vif (R)。

稳健标准误：statsmodels 的 cov_type 选项 (Python) 或 sandwich::vcovHC (R)。

数据模拟：使用 numpy (Python) 或自定义函数 (R) 生成具有特定数据缺陷的数据。

本章总结本章构建了线性回归分析的完整逻辑框架：首先，确立了一个由线性性、严格外生性、无完全共线性、球形扰动项等构成的假设体系，并阐明了其与估计量无偏性、一致性、有效性等理论性质的严密关联。其次，系统分析了各类假设违反对估计、推断与预测三大环节的具体影响，揭示了传统 OLS 在现实应用中的局限性。最后，指明了诊断的初步思路与后续修正的基本路径，并凸显了外生性假设作为通往因果推断的关键桥梁地位。这一 “假设-性质-违反-后果-诊断” 的思维链条，是贯穿整个计量经济学学习的核心方法论。

2 横截面数据：假设违反的诊断与修正

本章导读

在第一章中，我们系统学习了经典线性回归模型的基本假设及其违反的后果。本章将转向现实世界中的横截面数据分析，聚焦于诊断、修正和处理各类违反经典假设的问题。横截面数据作为计量经济学中最常见的数据类型，其分析面临多重挑战：异方差、自相关、多重共线性、异常值、缺失数据以及空间相关性等。

本章遵循" 问题识别 → 理论修正 → 实践应用" 的逻辑框架，系统介绍针对各类问题的现代解决方法。我们不仅关注技术细节，更强调方法的选择逻辑与结果的合理解释。通过学习本章，您将掌握处理" 不完美" 横截面数据的完整工具箱，为进行严谨的实证研究奠定坚实基础。

2.1 异方差：诊断、修正与推断

2.1.1 异方差的来源与经济实例

异方差性（heteroskedasticity）指误差项的方差随解释变量变化而变化：

$$\text{Var}(\varepsilon_i|X_i) = \sigma_i^2 \neq \text{常数}$$

经济学中常见的异方差来源包括：

1. 规模效应：大企业、大城市的变量波动通常更大
2. 学习效应：经验积累减少行为不确定性
3. 数据聚集：按组平均导致方差系统性差异
4. 异质性反应：不同群体对相同政策反应不同

2.1.2 异方差的正式检验方法

图示法

- 残差-拟合值图：$\hat{\varepsilon}_i$ vs. $\hat{Y}_i$
- 残差-解释变量图：$\hat{\varepsilon}_i$ vs. X_{ij}

Breusch-Pagan 检验（**LM** 检验）

1. 估计原模型得残差 $\hat{\varepsilon}_i$
2. 辅助回归：$\hat{\varepsilon}_i^2 = \alpha_0 + Z_i'\gamma + v_i$
3. 检验统计量：$LM = nR^2 \sim \chi^2_{p-1}$，其中 R^2 为辅助回归决定系数

White 检验

1. 将 $\hat{\varepsilon}_i^2$ 对所有 X、X^2 及交叉项回归
2. 检验统计量：$W = nR^2 \sim \chi^2_q$，q 为辅助回归中解释变量个数

2.1.3 异方差稳健标准误

当存在异方差时，OLS 估计量的正确方差为：

$$\text{Var}(\hat{\beta}|X) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}, \quad \Omega = \text{diag}(\sigma_1^2, ..., \sigma_n^2)$$

White 异方差稳健估计量：

$$\widehat{\text{Var}}_{HC}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2 X_i X_i'\right)(X'X)^{-1}$$

小样本修正系列：- HC0：基本 White 估计 - HC1：$\frac{n}{n-k}\hat{\varepsilon}_i^2$ - HC2：$\frac{\hat{\varepsilon}_i^2}{1-h_{ii}}$ - HC3：$\frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^2}$

2.1.4 加权最小二乘法与可行 **GLS**

加权最小二乘法（**WLS**）

已知方差结构 $\text{Var}(\varepsilon_i|X_i) = \sigma^2 w_i$ 时：

$$\hat{\beta}_{WLS} = \left(\sum_{i=1}^{n} \frac{1}{w_i} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{w_i} X_i Y_i \right)$$

可行广义最小二乘法（**FGLS**）

两阶段估计：1. OLS 估计得 $\hat{\varepsilon}_i$ 2. 估计 $\log(\hat{\varepsilon}_i^2) = Z_i'\delta + v_i$ 3. 计算 $\hat{w}_i = \exp(Z_i'\hat{\delta})$ 4. 使用 WLS，权重为 $1/\hat{w}_i$

2.2 自相关与聚类标准误

2.2.1 自相关的来源与类型

横截面数据中的自相关可能源于：

1. 空间相依性：地理位置接近的观测相关
2. 聚类结构：同一组内观测相关（学校、企业、行业）
3. 网络效应：社会网络或经济网络中的相互影响
4. 时间维度：重复横截面中的时间相关性

2.2.2 自相关的检验方法

空间自相关检验

Moran's I 统计量：

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2}$$

其中 w_{ij} 为空间权重矩阵元素。

聚类内的自相关检验

对于聚类数据，可检验组内相关性：1. 计算组内相关系数（ICC）2. 进行 Breusch-Godfrey 类型检验的聚类版本

2.2.3 自相关的修正方法

空间计量模型

- 空间自回归模型（SAR）：$Y = \rho WY + X\beta + \varepsilon$
- 空间误差模型（SEM）：$Y = X\beta + u, \quad u = \lambda Wu + \varepsilon$
- 空间杜宾模型（SDM）：$Y = \rho WY + X\beta + WX\theta + \varepsilon$

聚类调整估计

对于已知聚类结构，可采用随机效应或固定效应模型。

2.2.4 聚类稳健标准误

基本聚类标准误

假设数据分为 G 个聚类，聚类内任意相关，聚类间独立：

$$\widehat{\text{Var}}_{\text{cluster}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g \right) (X'X)^{-1}$$

其中 X_g 和 $\hat{\varepsilon}_g$ 为第 g 个聚类的解释变量矩阵和残差向量。

多路聚类标准误

当存在多个聚类维度时（如企业-行业-年份）：

$$\widehat{\text{Var}}_{\text{multi}} = \widehat{\text{Var}}_1 + \widehat{\text{Var}}_2 - \widehat{\text{Var}}_{12}$$

少聚类问题的处理

当聚类数量 G 较小时：1. 使用 t_{G-1} 分布而非正态分布 2. Bell-McCaffrey 偏差修正 3. Wild bootstrap 方法

空间自相关稳健标准误

Conley（1999）空间 HAC 估计量：

$$\widehat{\text{Var}}_{\text{Conley}} = (X'X)^{-1} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} k(d_{ij}) X_i X_j' \hat{\varepsilon}_i \hat{\varepsilon}_j \right) (X'X)^{-1}$$

其中 $k(d_{ij})$ 为距离 d_{ij} 的核函数。

2.3 异常值、杠杆点与稳健估计

2.3.1 异常值的识别与诊断

杠杆值

杠杆值 $h_{ii} = X_i'(X'X)^{-1}X_i$，满足：- $0 \leq h_{ii} \leq 1$ - $\sum_{i=1}^{n} h_{ii} = k$ - 经验法则：$h_{ii} > 2k/n$ 为高杠杆点

影响度量

1. 学生化残差：$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \sim t_{n-k-1}$
2. **Cook** 距离：$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})' X'X (\hat{\beta}-\hat{\beta}_{(i)})}{k\hat{\sigma}^2}$
3. **DFFITS**：$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$

2.3.2 异常值的处理策略

数据核查

检查异常值是否数据错误，如是则修正或删除。

稳健回归方法

M 估计：最小化 $\sum_{i=1}^{n}\rho\left(\frac{Y_i-X_i'\beta}{\sigma}\right)$ 常见 ρ 函数：- Huber：$\rho(z)=\begin{cases}z^2/2 & |z|\le c\\ c|z|-c^2/2 & |z|>c\end{cases}$ - Tukey 双权：$\rho(z)=\begin{cases}\frac{c^2}{6}[1-(1-(z/c)^2)^3] & |z|\le c\\ c^2/6 & |z|>c\end{cases}$

S 估计与 **MM** 估计：- S 估计：最小化残差的尺度估计，高崩溃点 - MM 估计：结合高崩溃点 S 估计与高效率 M 估计

2.3.3 分位数回归

基本模型

τ 分位数回归估计量：

$$\hat{\beta}(\tau)=\arg\min_{\beta}\sum_{i=1}^{n}\rho_\tau(Y_i-X_i'\beta)$$

其中检验函数 $\rho_\tau(u)=u(\tau-I(u<0))$。

渐近性质

在独立同分布下：

$$\sqrt{n}(\hat{\beta}(\tau)-\beta(\tau))\xrightarrow{d}N(0,\tau(1-\tau)D_1^{-1}D_0D_1^{-1})$$

其中 $D_0=E[f_{Y|X}(0)X_iX_i']$，$D_1=E[X_iX_i']$。

优势与应用

1. 对异常值稳健
2. 描述条件分布全貌
3. 无需分布假设
4. 单调变换下性质良好

2.4 缺失数据处理

2.4.1 缺失数据机制

Rubin（1976）分类：1. 完全随机缺失（**MCAR**）：$P(M_i = 1|Y_i^{\text{obs}}, Y_i^{\text{mis}}, X_i) = P(M_i = 1)$ 2. 随机缺失（**MAR**）：$P(M_i = 1|Y_i^{\text{obs}}, Y_i^{\text{mis}}, X_i) = P(M_i = 1|Y_i^{\text{obs}}, X_i)$ 3. 非随机缺失（**MNAR**）：缺失依赖于未观测值

2.4.2 多重插补方法

Rubin 规则

设 m 个插补数据集，第 j 个数据集的估计为 $\hat{\theta}_j$，方差为 U_j：- 合并估计：$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$ - 合并方差：$T = \bar{U} + (1 + \frac{1}{m})B$ 其中 $\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$，$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \bar{\theta})^2$

MICE 算法

链式方程多重插补步骤：1. 为每个缺失变量指定条件分布 2. 通过迭代 Gibbs 抽样生成插补值 3. 重复生成 m 个完整数据集

2.4.3 选择模型与逆概率加权

Heckman 选择模型

两步估计法：1. 选择方程：$D_i^* = Z_i'\gamma + u_i$，$D_i = I(D_i^* > 0)$ 2. 结果方程：$Y_i = X_i'\beta + \varepsilon_i$，仅当 $D_i = 1$ 时观测

逆米尔斯比率调整：

$$E(Y_i|X_i, D_i = 1) = X_i'\beta + \rho\sigma_\varepsilon\lambda(Z_i'\gamma)$$

逆概率加权（**IPW**）

$$\hat{\beta}_{\text{IPW}} = \left(\sum_{i=1}^{n}\frac{D_i}{\hat{p}_i}X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n}\frac{D_i}{\hat{p}_i}X_iY_i\right)$$

其中 $\hat{p}_i = P(D_i = 1|Z_i)$ 为倾向得分。

双重稳健估计

结合回归调整与 IPW：

$$\hat{\beta}_{\text{DR}} = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} \left[\frac{D_i}{\hat{p}_i} (Y_i - X_i' \hat{\beta}_{\text{reg}}) + X_i' \hat{\beta}_{\text{reg}} \right] X_i$$

只要倾向得分模型或结果模型之一正确，估计即一致。

2.5 多重共线性：诊断、修正与推断

2.5.1 多重共线性的来源与识别

来源

1. 经济变量间的内在关联：如收入与消费、价格与需求量
2. 变量构造：多项式项、交互项与原始变量相关
3. 数据限制：样本变异不足，变量变化范围小
4. 过度参数化：模型包含过多解释变量

识别方法

1. 方差膨胀因子（**VIF**）：

   $$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

   其中 R_j^2 是 X_j 对其他解释变量回归的决定系数。通常 VIF > 10 表明严重共线性。

2. 条件数（**Condition Number**）：

   $$\kappa = \sqrt{\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}}$$

   其中 λ_{max} 和 λ_{min} 是 $X'X$ 的最大和最小特征值。$\kappa > 30$ 表明严重共线性。

3. 特征分析：小特征值对应的特征向量可识别近似线性关系。

2.5.2 多重共线性的后果

对估计量的影响

1. 无偏性：OLS 估计量仍无偏（在外生性成立下）
2. 方差增大：

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}$$

   共线性使 $R_j^2 \to 1$，方差 $\to \infty$
3. 估计不稳定性：微小数据变化导致估计值大幅变动
4. 系数符号反常：估计系数符号可能与理论预期相反

对推断的影响

1. **t** 检验失效：由于方差膨胀，t 统计量变小，难以拒绝 $\beta_j = 0$ 的原假设
2. 置信区间变宽：参数估计的不确定性增加
3. 模型预测能力下降：样本外预测方差增大

2.5.3 多重共线性的修正方法

理论与方法选择

1. 增加样本量：收集更多数据减少共线性
2. 变量变换：对变量进行中心化、标准化或差分处理
3. 删除冗余变量：基于理论或统计检验删除不必要变量
4. 降维
5. 正则化：

2.5.4 高维数据的正则化方法

岭回归

岭估计量：

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'Y$$

其中 $\lambda > 0$ 为调节参数。

偏差-方差权衡：- 偏差：$E(\hat{\beta}_{\text{ridge}}) - \beta = -\lambda(X'X + \lambda I)^{-1}\beta$ - 方差：$\text{Var}(\hat{\beta}_{\text{ridge}}) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$

Lasso 回归

Lasso 估计量：

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

性质：1. 产生稀疏解，实现变量选择 2. 当 $p > n$ 时仍可估计 3. 解路径为分段线性

弹性网

结合 L1 和 L2 惩罚：

$$\hat{\beta}_{\text{en}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n}(Y_i - X_i'\beta)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

优势：1. 处理高度相关变量时比 Lasso 更稳定 2. 当 $p > n$ 时最多可选择 n 个变量

高维因果推断

双重选择 **Lasso**（Belloni 等，2014）：1. 用 Lasso 选择与处理变量 D 相关的控制变量 2. 用 Lasso 选择与结果 Y 相关的控制变量 3. 合并两步选择的变量，用 OLS 估计处理效应

2.5.5 降维

主成分回归

设 X 的奇异值分解：$X = U\Lambda V'$ 取前 r 个主成分：$Z = XV_r$，其中 V_r 为前 r 个右奇异向量主成分回归估计：$\hat{\beta}_{PCR} = V_r(V_r'X'XV_r)^{-1}V_r'X'Y$

2.6 空间与网络数据建模引论

2.6.1 空间自相关的识别

空间权重矩阵

常见形式：1. 邻接矩阵：$w_{ij} = I(i$与j相邻) 2. 距离矩阵：$w_{ij} = 1/d_{ij}^{\alpha}$ 3. **k** 最近邻矩阵：每个单元与最近的 k 个单元连接

空间自相关检验

Moran's I 检验：原假设 H_0：无空间自相关标准化统计量：$Z = \frac{I - E(I)}{\sqrt{\text{Var}(I)}} \sim N(0,1)$

Geary's C 统计量：

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2\sum_i \sum_j w_{ij} \sum_i (Y_i - \bar{Y})^2}$$

2.6.2 空间计量模型简介

空间自回归模型（**SAR**）

$$Y = \rho W Y + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

其中 ρ 为空间自回归系数，W 为空间权重矩阵。

空间误差模型（**SEM**）

$$Y = X\beta + u, \quad u = \lambda W u + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

其中 λ 为空间误差系数。

空间杜宾模型（**SDM**）

$$Y = \rho W Y + X\beta + WX\theta + \varepsilon$$

包含因变量和解释变量的空间滞后。

2.6.3 网络自相关模型

网络权重矩阵

基于网络结构的连接矩阵 G：- 无向网络：$g_{ij} = g_{ji}$ - 有向网络：$g_{ij} \neq g_{ji}$（如引文网络）- 加权网络：g_{ij} 表示连接强度

网络自回归模型

$$Y = \alpha GY + X\beta + \varepsilon$$

其中 α 度量网络效应强度。

识别挑战

反射问题（reflection problem）：个体的行为影响邻居，邻居的行为又影响个体，导致双向因果。解决方法：1. 使用工具变量 2. 利用网络结构特征（如朋友的朋发特征）3. 实验或准实验设计

2.7 综合诊断框架与模型选择

2.7.1 系统化的诊断流程

诊断顺序建议

1. 理论先验分析：基于经济学理论判断可能问题
2. 描述性统计分析：数据可视化，发现异常模式
3. 基础假设检验：异方差、自相关、正态性检验
4. 模型设定检验：RESET 检验、非线性检验
5. 多重共线性诊断：VIF、条件数分析
6. 稳健性检查：不同方法、不同样本下的结果比较

诊断结果整合

建立诊断记录表，记录：- 检验方法 - 检验统计量与 p 值 - 问题严重程度评估 - 建议的修正措施

2.7.2 模型选择与模型平均

信息准则

1. **AIC**（Akaike Information Criterion）：

$$\text{AIC} = 2k - 2\ln(L)$$

   倾向于选择更复杂的模型。

2. **BIC**（Bayesian Information Criterion）：

$$\text{BIC} = k\ln(n) - 2\ln(L)$$

   对模型复杂度惩罚更重，倾向于更简洁的模型。

交叉验证

K 折交叉验证步骤：1. 随机将数据分为 K 份 2. 每次用 $K-1$ 份训练，1 份测试 3. 重复 K 次，计算平均预测误差

留一法交叉验证：$K = n$ 的特殊情况。

模型平均方法

1. 贝叶斯模型平均（**BMA**）：

$$\hat{\beta}_{\text{BMA}} = \sum_{m=1}^{M} w_m \hat{\beta}_m$$

   其中 $w_m = P(M_m|Y)$ 为后验模型概率。

2. 堆叠法（**Stacking**）：基于交叉验证表现确定权重，最小化预测误差。

2.7.3 实证研究中的报告规范

透明度原则

1. 数据描述：清晰说明数据来源、处理过程、样本选择
2. 方法报告：详细描述估计方法、检验方法、软件与版本
3. 结果呈现：报告所有相关结果，包括不显著的结果

4. 复制材料：提供代码、数据、详细结果供他人复制

敏感性分析

应报告以下敏感性分析：1. 模型设定敏感性：不同函数形式、不同控制变量集 2. 估计方法敏感性：不同估计方法（OLS、IV、GMM 等）3. 样本选择敏感性：不同子样本、不同时间区间 4. 异常值处理敏感性：包含/排除异常值的结果比较

谨慎解释

1. 区分统计显著性与经济显著性：不仅报告 p 值，还要讨论经济意义
2. 承认局限性：明确说明研究的假设、局限性和推广范围
3. 避免过度推断：基于证据的谨慎结论，不夸大发现

2.8 案例分析与代码实现

（本章案例将聚焦于一个综合性的实证研究，展示如何系统应用本章介绍的各种方法诊断和处理横截面数据中的多重问题。具体内容将在教材配套材料中详细展开。）

本章总结

本章系统介绍了处理横截面数据中违反经典回归假设的现代方法。我们从异方差这一最常见问题出发，逐步深入到更复杂的自相关、多重共线性、异常值、缺失数据以及空间网络相关问题。关键要点包括：

1. 诊断优先：任何修正方法的应用都应基于准确的诊断。异方差的图示法和统计检验、自相关的空间检验、多重共线性的 VIF 分析等，都是必要的前期工作。

2. 方法选择的权衡：效率与稳健性、偏差与方差、简洁性与准确性之间的权衡是方法选择的核心考量。例如：

   - 当异方差形式未知时，稳健标准误比 FGLS 更可靠
   - 当存在异常值时，分位数回归比 OLS 更稳健
   - 当存在多重共线性时，岭回归比删除变量更能保留信息

3. 标准误的稳健性至关重要：在实证研究中，正确的标准误是进行有效统计推断的基础。聚类标准误、异方差稳健标准误、空间 HAC 标准误等，都是应对不同相关结构的工具。

4. 处理复杂数据结构的现代方法：对于高维数据、缺失数据、空间网络数据，计量经济学发展了一系列现代方法：
    - 正则化方法（Lasso、弹性网）处理高维数据
    - 多重插补和双重稳健估计处理缺失数据
    - 空间计量模型处理空间相关性
5. 综合诊断与透明报告：良好的实证研究应遵循系统化的诊断流程，并透明报告所有步骤和结果。敏感性分析和稳健性检查是评估结论可靠性的关键。

通过本章的学习，您将具备处理现实世界横截面数据的全面能力，为进行严谨、可信的实证经济学研究打下坚实基础。在后续章节中，我们将把重点转向因果推断这一计量经济学的核心目标，这些处理数据问题的技术将成为我们进行可信因果推断的重要前提。

要点回顾：- 异方差使 OLS 无效但依然一致，可使用稳健标准误或 WLS/FGLS 修正 - 自相关（空间相关、聚类相关）需使用聚类标准误或空间计量模型 - 异常值可通过稳健回归或分位数回归处理 - 缺失数据机制决定处理方法，多重插补和双重稳健估计是常用方法 - 多重共线性增加估计方差但不影响无偏性，可通过正则化方法处理 - 空间网络数据需要专门的模型和推断方法

掌握这些方法的关键不仅是理解其数学原理，更是能够在具体研究问题中做出恰当的方法选择，并对结果进行合理解释。这正是计量经济学作为一门应用科学的核心要义。

3 面板数据模型

本章导读

面板数据（Panel Data）结合了横截面与时间序列的双重维度，为识别因果关系提供了更丰富的信息。本章在前两章基础上，系统介绍面板数据的基本模型（混合 OLS、固定效应、随机效应）、估计方法及模型选择策略。重点理解组内变异与组间变异的区别，掌握豪斯曼检验，并通过案例体会面板数据在政策评估、劳动经济学等领域的应用价值。本章是连接经典横截面分析与现代因果推断的关键桥梁。

3.1 面板数据概述

3.1.1 面板数据的定义、结构与符号

面板数据是指对同一组个体（如个人、企业、国家）在多个时间点上进行重复观测所得到的数据。记 $i = 1, 2, \ldots, N$ 表示个体，$t = 1, 2, \ldots, T$ 表示时间，则观测值 y_{it} 表示第 i 个个体在第 t 期的因变量取值，$\mathbf{x}_{it}$ 表示相应的 k 维解释变量向量。

面板数据的基本结构如下表所示：

个体\时间	$t = 1$	$t = 2$	$\cdots$	$t = T$
$i = 1$	$(y_{11}, \mathbf{x}_{11})$	$(y_{12}, \mathbf{x}_{12})$	$\cdots$	$(y_{1T}, \mathbf{x}_{1T})$
$i = 2$	$(y_{21}, \mathbf{x}_{21})$	$(y_{22}, \mathbf{x}_{22})$	$\cdots$	$(y_{2T}, \mathbf{x}_{2T})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i = N$	$(y_{N1}, \mathbf{x}_{N1})$	$(y_{N2}, \mathbf{x}_{N2})$	$\cdots$	$(y_{NT}, \mathbf{x}_{NT})$

3.1.2 面板数据的优势与挑战

优势

1. 控制不可观测异质性：能够控制不随时间变化的个体固定效应（如个人能力、企业文化、地理位置），从而缓解遗漏变量偏误。
2. 提供更丰富的变异信息：既包含个体间的差异（组间变异，Between Variation），也包含个体随时间的变化（组内变异，Within Variation），有助于识别因果关系。
3. 提高估计效率：更多的观测值通常带来更小的标准误。
4. 研究动态行为：可以分析个体行为的动态调整过程。

挑战

1. 测量误差：变量的测量误差在面板数据中可能导致动态面板偏误。
2. 样本损耗：长期追踪调查中，个体可能退出，导致非平衡面板（Unbalanced Panel）。
3. 模型设定复杂性：需要选择合适的模型（混合、固定、随机效应），并处理可能的序列相关、异方差和截面相关。

3.1.3 基本模型设定

面板数据模型的三种基本设定如下：

1. 混合回归模型（**Pooled Model**）忽略个体差异，假设所有个体遵循相同的数据生成过程。

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

2. 固定效应模型（**Fixed Effects Model, FE**）允许每个个体拥有自己的截距项 α_i，且 α_i 可能与解释变量相关。

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

3. 随机效应模型（**Random Effects Model, RE**）将个体截距 α_i 视为随机变量，且与解释变量不相关。

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

其中，α_i 称为个体效应（Individual Effect）或不可观测异质性（Unobserved Heterogeneity）。

3.2 混合最小二乘估计（Pooled OLS）

3.2.1 混合 OLS 的模型形式与经典假设

混合 OLS 将 $N \times T$ 个观测值视为一个大的独立横截面数据进行回归，完全忽略面板数据结构：

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

其经典假设与横截面 OLS 相同：

1. 线性关系与严格外生性：$E(\varepsilon_{it}|\mathbf{x}_{it}) = 0$。

2. 随机抽样：观测值 $\{(y_{it}, \mathbf{x}_{it})\}$ 独立同分布。

3. 无完全共线性：解释变量矩阵 $\mathbf{X}$ 列满秩。

4. 同方差性：$\text{Var}(\varepsilon_{it}|\mathbf{x}_{it}) = \sigma^2$。

5. 无自相关：$\text{Cov}(\varepsilon_{it}, \varepsilon_{js}|\mathbf{x}_{it}, \mathbf{x}_{js}) = 0$，除非 $i = j$ 且 $t = s$。

3.2.2 混合 OLS 的适用条件与局限性

适用条件：- 个体效应 α_i 与所有解释变量 $\mathbf{x}_{it}$ 均不相关，即 $\text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$。- 研究目的仅关注 $\mathbf{x}_{it}$ 的 “平均” 效应，而非个体异质性效应。

局限性：1. 遗漏变量偏误：若 α_i 与 $\mathbf{x}_{it}$ 相关，则混合 OLS 的估计量 $\hat{\boldsymbol{\beta}}_{\text{Pooled}}$ 是有偏且不一致的。2. 效率损失：即使假设成立，混合 OLS 未利用面板数据的结构信息，其标准误可能不是最有效的（除非使用聚类稳健标准误进行修正）。

3.2.3 与横截面回归的比较

混合 OLS 本质上是将面板数据 “堆叠” 后进行横截面回归。与第 2 章横截面分析相比：- 数据层面：混合 OLS 虽然利用了重复观测，但假设不同期的观测相互独立，忽略了个体内的序列相关性。- 假设层面：横截面分析无法检验个体效应的存在，而面板数据允许我们通过比较混合、固定、随机效应模型来做出选择。

3.3 固定效应模型（Fixed Effects, FE）

3.3.1 固定效应的动机：消除不随时间变化的遗漏变量

固定效应模型的核心动机是控制不随时间变化的不可观测异质性 α_i。模型设定为：

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

其中 α_i 是待估参数（或通过变换消除），代表个体 i 特有的、不随时间变化的特征。

3.3.2 组内估计量（Within Estimator）的推导与性质

为了消除 α_i，对每个个体计算时间均值：

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \quad \bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$$

将原模型减去其时间均值模型 $\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i$，得到组内变换（Within Transformation）后的模型：

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

记 $\ddot{y}_{it} = y_{it} - \bar{y}_i$，$\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$，则上式简化为：

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{\varepsilon}_{it}$$

对上述模型进行 OLS 回归，得到的估计量 $\hat{\boldsymbol{\beta}}_{\text{FE}}$ 称为组内估计量。

估计量性质：- 一致性：只要 $E(\ddot{\varepsilon}_{it}|\ddot{\mathbf{x}}_{it}) = 0$（即 $\mathbf{x}_{it}$ 与 ε_{it} 不相关），FE 估计量是一致的。- 无法估计不随时间变化的变量：对于常数变量（如性别、种族），$\ddot{\mathbf{x}}_{it} \equiv 0$，其系数无法被识别。

3.3.3 虚拟变量法与一阶差分法（FD）的等价性

虚拟变量法（**LSDV**）：在原始模型中为每个个体加入一个虚拟变量（除一个基准个体外）：

$$y_{it} = \beta_0 + \sum_{i=2}^{N} \alpha_i D_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

其中 D_i 为个体虚拟变量。对该模型进行 OLS 回归，得到的 $\hat{\boldsymbol{\beta}}$ 与组内估计量完全等价。

一阶差分法（**First Difference, FD**）：对相邻两期数据进行差分，以消除 α_i：

$$\Delta y_{it} = y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1})$$

- 当 $T = 2$ 时，FD 估计量与 FE 估计量完全等价。- 当 $T > 2$ 时，两者略有不同。FD 假设扰动项 ε_{it} 无序列相关，否则 FE 估计量更有效。

3.3.4 固定效应模型的假设、优点与缺点

核心假设：1. 严格外生性：$E(\varepsilon_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \alpha_i) = 0$。即任意时期的扰动项 ε_{it} 与所有时期（过去、现在、未来）的解释变量均不相关。这是一个较强的假设。2. α_i 与 $\mathbf{x}_{it}$ 可以存在任意形式的相关性（这正是使用 FE 的动机）。

优点：1. 能有效控制不随时间变化的遗漏变量，缓解由此导致的偏误。2. 无需对 α_i 的分布做任何假设。

缺点：1. 无法估计不随时间变化的变量的效应。2. 若关注变量 x_{it} 本身随时间变化很小（即组内变异小），则 FE 估计量的标准误会很大，估计不精确。3. 不能控制随时间变化的遗漏变量。

3.4 随机效应模型（**Random Effects, RE**）

3.4.1 随机效应的设定与假设

随机效应模型将个体效应 α_i 视为随机变量，并纳入复合扰动项中：

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it}$$

核心假设：1. $\alpha_i \sim$ i.i.d.$(0, \sigma_\alpha^2)$，且与 ε_{it} 相互独立。2. $\varepsilon_{it} \sim$ i.i.d.$(0, \sigma_\varepsilon^2)$。3. α_i 与所有解释变量 $\mathbf{x}_{it}$ 均不相关，即 $\text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$。

在此假设下，复合扰动项 u_{it} 的协方差结构为：

$$\text{Var}(u_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$$

$$\text{Cov}(u_{it}, u_{is}) = \sigma_\alpha^2, \quad \text{for } t \neq s$$

即同一个体不同期的扰动项之间存在相关性，相关系数为 $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$。

3.4.2 广义最小二乘估计（GLS）的原理与实现

由于扰动项存在组内自相关，OLS 不再有效。随机效应模型采用广义最小二乘法（**GLS**）进行估计。其核心是对数据进行准离差变换（Quasi-demeaning Transformation）：

$$y_{it} - \theta\bar{y}_i = (\mathbf{x}_{it} - \theta\bar{\mathbf{x}}_i)'\boldsymbol{\beta} + [(1-\theta)\beta_0 + (u_{it} - \theta\bar{u}_i)]$$

其中，

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}$$

- 当 $\theta = 0$ 时，等价于混合 OLS（$\sigma_\alpha^2 = 0$）。
- 当 $\theta = 1$ 时，等价于固定效应变换（$\sigma_\varepsilon^2 = 0$，或 $T \to \infty$）。
- 通常 $\theta \in (0,1)$，RE 估计量是混合 OLS 和 FE 估计量的加权平均。

实际操作中，需先估计 σ_ε^2 和 σ_α^2，然后进行 GLS 估计。现代软件可自动完成此过程。

3.4.3 随机效应模型的效率优势与一致性

效率优势：当 $\text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$ 的假设成立时，RE 估计量比 FE 估计量更有效（方差更小），因为它同时利用了数据的组内变异和组间变异。

一致性条件：RE 估计量的一致性完全依赖于 α_i 与 $\mathbf{x}_{it}$ 不相关的假设。若该假设不成立，则 RE 估计量是不一致的。因此，RE 模型的使用必须依赖于理论支持或统计检验。

3.5 模型选择：豪斯曼检验（Hausman Test）

3.5.1 豪斯曼检验的核心思想

豪斯曼检验用于判断个体效应 α_i 是否与解释变量 $\mathbf{x}_{it}$ 相关，从而在 FE 和 RE 模型之间做出统计选择。

逻辑基础：- 如果 $H_0 : \text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$ 成立，则 **FE** 和 **RE** 估计量都是一致的，但 RE 估计量更有效。- 如果 H_0 不成立，则 **FE** 估计量仍然一致，但 **RE** 估计量不一致。

因此，可以检验两种估计量之间的差异是否显著。若无显著差异，则选择更有效的 RE 模型；若有显著差异，则说明 RE 不一致，应选择 FE 模型。

3.5.2 检验统计量的构造、原假设与备择假设

定义两个估计量的差：$\mathbf{d} = \hat{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}$。在 H_0 下，$\text{plim}\,\mathbf{d} = 0$。豪斯曼证明，其渐近协方差矩阵为：

$$\text{Var}(\mathbf{d}) = \text{Var}(\hat{\boldsymbol{\beta}}_{\text{FE}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\text{RE}})$$

构造的检验统计量为：

$$H = \mathbf{d}'[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{FE}}) - \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{RE}})]^{-1}\mathbf{d} \overset{a}{\sim} \chi^2(k)$$

其中 k 为 $\boldsymbol{\beta}$ 的维数（不包括常数项）。

- 原假设 **(**H_0**)**：$\text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$，应选择随机效应模型。
- 备择假设 **(**H_1**)**：$\text{Cov}(\alpha_i, \mathbf{x}_{it}) \neq 0$，应选择固定效应模型。

决策规则：给定显著性水平 α（如 0.05），若 $H > \chi^2_\alpha(k)$，则拒绝 H_0，选择固定效应模型；否则，选择随机效应模型。

3.5.3 实践中的决策流程

1. 理论先导：首先根据经济理论或研究背景判断。如果理论上认为个体效应极有可能与解释变量相关（如企业不可观测的管理能力影响其研发投入），则应直接选择 FE 模型。
2. 统计检验：运行 FE 和 RE 模型，进行豪斯曼检验。注意，检验要求 RE 估计量是有效的（在 H_0 下），因此需要使用基于 GLS 的 RE 估计结果。
3. 稳健性报告：在学术论文中，通常同时报告 FE 和 RE 的估计结果，并注明豪斯曼检验的结论，以体现结论的稳健性。

3.6 面板数据模型的扩展与高级应用

3.6.1 动态面板数据模型：**Arellano-Bond GMM** 估计

当模型包含因变量的滞后项以刻画动态调整过程时，即为动态面板模型：

$$y_{it} = \rho y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

此时，即使使用 FE 变换，变换后的扰动项 $\ddot{\varepsilon}_{it}$ 与变换后的滞后因变量 $\ddot{y}_{i,t-1}$ 仍然相关，导致 FE 估计量有偏且不一致（尼克尔偏误，Nickell Bias）。

Arellano-Bond 估计法的解决思路：1. 先对原模型进行一阶差分以消除 α_i：$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta\mathbf{x}'_{it}\boldsymbol{\beta} + \Delta\varepsilon_{it}$。2. 差分后，$\Delta y_{i,t-1}$ 与 $\Delta\varepsilon_{it}$ 仍相关。但更早的滞后水平 $y_{i,t-2}, y_{i,t-3}, \ldots$ 与 $\Delta\varepsilon_{it}$ 不

相关，却与 $\Delta y_{i,t-1}$ 相关，因此可以作为有效的工具变量。3. 利用这些滞后变量作为工具变量，进行广义矩估计（**GMM**），即差分 **GMM**。

3.6.2 非平衡面板数据的处理方法

非平衡面板指不同个体拥有的时间期数 T_i 不同。对于 FE 和 RE 模型：- 固定效应模型：组内变换依然适用，只需将均值计算改为 $\bar{y}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} y_{it}$，其中 $\mathcal{T}_i$ 是个体 i 的观测时期集合。- 随机效应模型：GLS 估计中的变换参数 θ_i 变为个体特定：$\theta_i = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T_i\sigma_\alpha^2}}$。

处理时需关注样本损耗（Attrition）是否是随机的，若非随机可能导致样本选择偏误。

3.6.3 双向固定效应（**Two-Way Fixed Effects**）

在基础 FE 模型上，进一步加入时间固定效应 λ_t，以控制所有个体共同面临的时间趋势或宏观冲击：

$$y_{it} = \alpha_i + \lambda_t + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

估计方法：可在组内变换的基础上，再对时间均值进行离差变换（即对 $\ddot{y}_{it}$ 和 $\ddot{\mathbf{x}}_{it}$ 做 “时间离差”），或直接加入时间虚拟变量进行 LSDV 回归。双向固定效应模型是政策评估中双重差分法（**DID**）的标准设定框架。

3.7 面板数据在政策评估中的应用

面板数据是实施双重差分法（**Difference-in-Differences, DID**）的理想载体。基本 DID 模型设定如下：

$$y_{it} = \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{Treat}_i + \beta_3 (\text{Post}_t \times \text{Treat}_i) + \alpha_i + \lambda_t + \varepsilon_{it}$$

其中：- Treat_i：处理组虚拟变量（个体层面不随时间变化）。- Post_t：政策后时期虚拟变量（时间层面不随个体变化）。- $\text{Post}_t \times \text{Treat}_i$：交互项，其系数 β_3 是核心估计量，反映了政策处理的平均处理效应（**ATE**）。

在这个设定中，个体固定效应 α_i 吸收了 Treat_i，时间固定效应 λ_t 吸收了 Post_t，因此模型常简写为：

$$y_{it} = \beta_3 D_{it} + \alpha_i + \lambda_t + \varepsilon_{it}$$

其中 $D_{it} = \text{Post}_t \times \text{Treat}_i$ 表示个体 i 在时期 t 是否受到政策处理。这清晰地体现了面板数据通过控制不可观测的个体和时点特征来识别因果效应的优势。

本章总结

本章系统介绍了面板数据模型的核心内容，旨在利用数据在时间和个体两个维度的变异来更准确地识别经济关系与因果效应。

1. 面板数据的本质与优势：面板数据通过追踪同一组个体在不同时间点的信息，使研究者能够控制不随时间变化的个体异质性，这是其相对于横截面数据最根本的优势。

2. 三大基本模型：

    - 混合 **OLS**：忽略面板结构，假设所有观测独立。适用于个体效应与解释变量绝对不相关的理想情况，但通常风险较大。
    - 固定效应模型：通过组内变换消除个体效应 α_i，允许 α_i 与解释变量任意相关。是解决遗漏变量偏误的强有力工具，但无法估计不随时间变化变量的系数。
    - 随机效应模型：将个体效应视为随机扰动的一部分，要求 α_i 与解释变量不相关。若假设满足，则估计效率高于 FE 模型。

3. 模型选择的关键：豪斯曼检验提供了在 FE 与 RE 之间选择的统计依据。其核心是比较两个估计量的一致性。实践中应结合经济理论和统计检验综合判断。

4. 扩展与应用：

    - 动态面板：当包含滞后因变量时，需使用 **Arellano-Bond GMM** 等工具变量方法。
    - 双向固定效应：同时控制个体和时间效应，是更为稳健的设定。
    - 政策评估：面板数据为双重差分法提供了天然的实施框架，通过比较处理组和对照组在政策前后的变化来识别因果效应。

5. 核心思想贯穿始终：理解组内变异与组间变异的区别是掌握面板数据模型的钥匙。固定效应模型仅利用组内变异，随机效应模型则同时利用两种变异。选择何种模型，取决于研究者相信个体不可观测特征 α_i 与解释变量 $\mathbf{x}_{it}$ 是否相关，而这最终关系到估计结果的一致性与可靠性。

面板数据模型是现代计量经济学实证分析的基石。掌握本章内容，意味着具备了利用更丰富的数据结构去检验经济理论、评估政策效果的基本能力，为进一步学习更高级的微观计量与因果推断方法奠定了坚实基础。

4 时间序列分析

本章导读

时间序列数据是按照时间顺序收集的一系列观测值，例如国内生产总值（GDP）、消费者物价指数（CPI）、股票价格等。与前三章学习的横截面数据、面板数据不同，时间序列数据最大的特点是观测值之间存在时间依赖性或序列相关性，这违背了经典线性回归模型中 “观测值独立” 的核心假设。因此，直接对时间序列数据应用普通最小二乘法（OLS）可能导致 “伪回归” 等问题，即统计上显著的关系可能仅仅源于变量共同的时间趋势，而非真实的经济联系 [citation:3]。

本章旨在系统介绍时间序列计量经济学的核心理论与方法。我们将首先建立平稳性这一基石性概念，并学习单位根检验以诊断数据的平稳性。在此基础上，掌握对单变量序列进行建模和预测的 **ARIMA** 模型及其建模方法论。随后，我们将视野扩展至多变量系统，深入学习分析变量间动态交互作用的向量自回归（**VAR**）模型，及其相关的格兰杰因果检验、脉冲响应分析等工具。最后，为解决非平稳变量间的长期均衡问题，我们将引入协整理论与误差修正模型（**ECM**）。通过本章学习，你将能够恰当地处理、建模并科学解释经济与管理领域中常见的时间序列数据。

4.1 基本概念与平稳性

4.1.1 时间序列数据的特性

时间序列分析需特别关注其由时间维度引致的特殊性：1. 趋势性：数据在长期内呈现出持续向上或向下的系统性运动。2. 季节性/周期性：数据在固定时间间隔（如季度、月份）内呈现出规律的波动。3. 序列相关性：当期观测值 Y_t 通常与其自身的历史值 $Y_{t-1}, Y_{t-2}, \ldots$ 相关。4. 平稳性要求：许多经典时间序列模型要求数据是平稳的，即其统计特性不随时间变化。

4.1.2 平稳性

平稳性是时间序列分析的核心基础。一个平稳的时间序列其概率规律不随时间推移而改变。- 严格平稳：序列的联合概率分布在任何时间区间上都是相同的。定义严格但实践中难以验证。- 弱

平稳（协方差平稳）：更常用且实用的概念，要求满足：1. 均值恒定：$E(Y_t) = \mu$，对所有 t 成立。2. 方差恒定：$Var(Y_t) = E[(Y_t - \mu)^2] = \sigma^2$，对所有 t 成立。3. 协方差仅依赖于时间间隔：$Cov(Y_t, Y_{t+k}) = \gamma_k$，仅与滞后阶数 k 有关，与具体时点 t 无关。

4.1.3 平稳性检验：单位根检验

最常用的正式检验方法是单位根检验。其原假设 H_0 为：序列存在单位根（即非平稳）；备择假设 H_1 为：序列平稳。最经典的是 **Augmented Dickey-Fuller (ADF)** 检验。它通过估计以下回归式实现：

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \varepsilon_t$$

其中，Δ 为一阶差分算子，βt 为时间趋势项。检验的关键是判断系数 γ 是否显著小于 0。若不能拒绝 $\gamma = 0$ 的原假设，则认为序列存在单位根，是非平稳的。检验统计量需与 ADF 专用临界值比较 [citation:2]。

4.1.4 自相关与偏自相关函数

- 自相关函数（**ACF**）：描述序列 Y_t 与其自身滞后 k 期值 Y_{t-k} 之间的线性相关性，定义为 $\rho_k = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)Var(Y_{t-k})}}$。
- 偏自相关函数（**PACF**）：描述在控制了中间滞后项 $Y_{t-1}, ..., Y_{t-k+1}$ 的影响后，Y_t 与 Y_{t-k} 之间的条件相关性。ACF 和 PAC 图是识别时间序列模型类型和阶数的关键可视化工具。

4.2 单变量时间序列建模：**ARIMA** 模型

4.2.1 自回归模型（**AR(p)**）

p 阶自回归模型认为当前值 Y_t 是其过去 p 期值的线性组合加随机扰动：

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t$$

其中 ε_t 是白噪声。AR 模型的平稳性要求其特征方程 $1 - \phi_1 L - ... - \phi_p L^p = 0$ 的所有根都在单位圆外（L 为滞后算子）[citation:4]。

4.2.2 移动平均模型（MA(q)）

q 阶移动平均模型认为当前值由过去 q 期随机冲击的线性组合决定：

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

其中 ε_t 是白噪声。MA 模型总是平稳的。

4.2.3 ARMA 与 ARIMA 模型

- **ARMA(p, q)** 模型：结合 AR 和 MA 模型，用于平稳序列：$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$。
- **ARIMA(p, d, q)** 模型：针对非平稳序列，先通过 d 阶差分将其变为平稳序列，再对差分后的序列建立 ARMA(p, q) 模型。记 $\Delta^d Y_t$ 为 d 阶差分后的序列，则 ARIMA 模型为：$\Delta^d Y_t = c + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$[citation:8]。

4.2.4 Box-Jenkins 方法论

建立 ARIMA 模型通常遵循以下迭代步骤：1. 识别：通过观察序列图、ACF 和 PACF 图，初步判断差分阶数 d 以及 AR、MA 的阶数 p 和 q。2. 估计：使用最大似然估计法（MLE）等方法估计模型参数。3. 诊断检验：检验模型残差是否为白噪声（如使用 Ljung-Box Q 检验）。若拒绝白噪声假设，则返回第一步重新识别。4. 预测：利用估计好的模型进行未来值的点预测和区间预测。

4.3 多变量动态分析：向量自回归（VAR）模型

当需要分析多个时间序列变量之间的动态互动关系时，需要使用多变量模型。向量自回归（**VAR**）模型将系统中的每个内生变量表示为所有内生变量滞后值的函数，是分析联合内生变量动态性的标准工具 [citation:5]。

4.3.1 VAR 模型的基本形式

一个包含 m 个变量、滞后 p 阶的 **VAR(p)** 模型定义如下：

$$\mathbf{y}_t = \mathbf{c} + \square_1 \mathbf{y}_{t-1} + \square_2 \mathbf{y}_{t-2} + ... + \square_p \mathbf{y}_{t-p} + \square_t$$

其中：- $\mathbf{y}_t = (y_{1t}, y_{2t}, ..., y_{mt})'$ 是 $m \times 1$ 维内生变量向量。- $\mathbf{c}$ 是 $m \times 1$ 维常数项向量。- $\square_i$ 是 $m \times m$ 维系数矩阵。- $\square_t$ 是 $m \times 1$ 维白噪声扰动向量，满足 $E(\square_t) = \mathbf{0}$, $E(\square_t \square_t') = \square$（正定协

方差矩阵），且无序列相关。VAR 模型通常不施加基于经济理论的先验约束，故常被称为 “让数据自己说话” 的无约束模型 [citation:5]。

4.3.2 VAR 模型的建立、估计与诊断

1. 滞后阶数选择：使用信息准则确定最优滞后阶数 p，常用准则包括赤池信息准则（AIC）和贝叶斯信息准则（BIC/SBIC），通常选择使准则值最小的 p[citation:5][citation:10]。
2. 估计：由于每个方程的解释变量相同（均为所有变量的滞后项），对整个系统使用普通最小二乘法（**OLS**）进行方程-by-方程估计是一致的且有效的 [citation:10]。
3. 模型诊断：需进行稳定性检验（所有特征根的模长小于 1）、残差自相关检验（如 Portmanteau 检验）和异方差检验等，以确保模型设定合理 [citation:10]。

4.3.3 格兰杰因果关系检验

格兰杰因果关系检验旨在判断一个变量的过去值是否对预测另一个变量的当前值有统计上的显著贡献，它检验的是时间先后上的 “预测能力”。对于 VAR 中的变量 x 和 y，检验 “ x 不是 y 的格兰杰原因” 的原假设，即检验 y 的方程中所有 x 的滞后项系数是否联合为零。通常通过沃尔德检验（Wald test）实现 [citation:2]。

4.3.4 脉冲响应分析与方差分解

- 脉冲响应函数（**IRF**）：描绘系统中一个变量受到一单位标准冲击后，对所有变量（包括其自身）产生的动态影响路径。它直观展示了冲击在系统内的传导机制 [citation:5]。
- 方差分解：将每个变量的预测误差方差，按成因分解为来自系统中各变量冲击的贡献比例。它回答了 “某个变量的波动，有多大比例是由其他变量（或自身）的冲击造成的？” 这一问题 [citation:5]。

4.3.5 结构 VAR（SVAR）与扩展模型简介

- 结构 **VAR**（**SVAR**）：无约束 VAR 是简化式，其扰动项可能相关。SVAR 通过施加基于经济理论的识别约束（如短期零约束、符号约束等），试图估计出反映变量间同期结构性关系的模型，从而得到结构冲击 [citation:1][citation:2]。
- 扩展模型：为处理更复杂问题，发展出诸多扩展模型，如贝叶斯 **VAR**（**BVAR**）（用于解决参数过多问题）、时变参数 **VAR**（**TVP-VAR**）（捕捉参数随时间的变化）以及因子增广 **VAR**（**FAVAR**）（纳入大量信息）等 [citation:1]。

4.4 非平稳序列与协整分析

4.4.1 虚假回归问题

如果对两个或多个非平稳时间序列直接进行 OLS 回归，即使它们之间没有真实经济联系，也常会得到统计上显著的回归结果和高 R^2，即 “虚假回归”。其根本原因在于非平稳序列的共同趋势偶然匹配 [citation:3]。

4.4.2 协整的概念

协整为解决上述问题提供了钥匙：如果两个或多个非平稳序列（通常要求同阶单整，如 $I(1)$）的某个线性组合是平稳的（$I(0)$），则称这些变量之间存在协整关系 [citation:6]。- 经济含义：协整关系意味着变量之间存在长期均衡关系。虽然每个变量独自可能随机游走，但它们的某种组合却长期稳定在均衡水平附近。- 例子：居民消费 C_t 和可支配收入 Y_t 可能都是 $I(1)$，但差额 $(C_t - \beta Y_t)$ 是 $I(0)$，反映了消费与收入之间长期稳定的比例关系。

4.4.3 协整检验

1. Engle-Granger 两步法（针对双变量）

第一步：用 OLS 估计长期静态回归：$y_t = \alpha + \beta x_t + u_t$。第二步：对回归残差 $\hat{u}_t$ 进行单位根检验（ADF 检验，但需使用专门的 EG 临界值表）。若残差平稳，则 y_t 与 x_t 协整 [citation:6]。

2. Johansen 检验（针对多变量）

这是基于 VAR 模型的更一般方法。将 VAR 模型改写为向量误差修正模型（VECM）形式：

$$\Delta \mathbf{y}_t = \square \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \square_i \Delta \mathbf{y}_{t-i} + \square_t$$

其中，$\square = \square\square'$ 是关键。$\square$ 的秩 r 就是协整关系的个数。$\square$ 的每一行是一个协整向量，$\square$ 是调整系数矩阵。Johansen 方法通过迹检验或最大特征值检验来确定协整秩 r[citation:2]。

4.4.4 误差修正模型（**ECM**）

根据格兰杰表述定理，如果一组变量是协整的，则它们之间的短期动态关系必然可以由一个误差修正模型来描述 [citation:6]。ECM 将变量的短期变化 Δy_t 与两个因素联系起来：1. 其他变量的短

期变化（Δx_t）。2. 上一期对长期均衡的偏离（即误差修正项 $ECM_{t-1} = (y_{t-1} - \beta x_{t-1})$）。一个简单的双变量 ECM 形式为：

$$\Delta y_t = \gamma \Delta x_t + \lambda ECM_{t-1} + \varepsilon_t$$

其中，系数 λ 称为调整速度，理论上应小于 0。它衡量了系统从短期偏离向长期均衡回调的速度和力度 [citation:6]。

4.4.5 向量误差修正模型（**VECM**）

对于多变量协整系统，相应的模型是向量误差修正模型（**VECM**），它是包含协整约束的 VAR 模型，其一般形式如上文 Johansen 检验部分所示。VECM 同时刻画了变量间的长期均衡关系和短期动态调整机制 [citation:2]。

4.5 面板数据时间序列模型简介

当数据同时具有时间维度（T 期）和截面维度（N 个个体）时，即为面板数据。面板数据时间序列模型关注 “大 N，大 T” 情形下的动态建模 [citation:4]。- 面板单位根检验：检验面板数据的平稳性，方法包括 LLC 检验、IPS 检验等，分别适用于同质单位根和异质单位根情形 [citation:2]。- 面板协整检验：检验非平稳面板数据变量间是否存在长期均衡关系，常用方法如 Pedroni 检验、Kao 检验 [citation:2]。- 面板 **VAR** 模型：将 VAR 模型扩展到面板数据框架，能够分析多个变量在多个个体间的动态互动，并通常需要考虑个体异质性（固定效应或随机效应）[citation:1]。

本章总结

本章系统阐述了时间序列计量经济学的核心内容。我们从理解平稳性这一基本要求出发，掌握了单位根检验这一关键诊断工具，并区分了趋势平稳与差分平稳过程。

对于平稳单变量序列，我们学习了通过 **Box-Jenkins** 方法论建立 **ARIMA** 模型进行拟合与预测。当分析多个相互影响的时间序列变量时，我们引入了向量自回归（**VAR**）模型框架，并在此框架下学习了格兰杰因果检验、脉冲响应分析和方差分解等一系列实用的动态分析工具。

面对普遍存在的非平稳经济变量，我们揭示了虚假回归的风险，并引入了协整理论作为解决方案。协整关系刻画了变量间存在的长期均衡关系，而误差修正模型（**ECM**）和向量误差修正模型（**VECM**）则在此基础上，描述了系统从短期偏离向长期均衡调整的动态过程。

最后，我们简要介绍了将时间序列方法应用于面板数据的扩展模型。时间序列分析是理解宏观经济波动、金融市场动态等复杂经济现象不可或缺的计量工具。掌握本章内容是进一步学习结构 VAR、ARCH/GARCH 族波动率模型、状态空间模型等更高级专题的重要基础。

学习提示：时间序列计量经济学强调“干中学”。建议使用 EViews、Stata、R 或 Python 等软件，结合中国宏观或金融市场的实际数据（如国家统计局、中国人民银行网站数据），完整复现从数据导入、平稳性检验、模型估计到结果解读的全过程 [citation:4][citation:10]。这种实践对深刻理解理论方法至关重要。

5 离散数据与受限因变量模型

本章导读

在经典计量经济学模型中，因变量通常被假定为连续变量。然而，在经济决策与实证研究中，大量核心变量本质上是离散的（如是否购买、选择何种出行方式）或数值上受到限制的（如工作时间非负、消费数据在某一区间内聚集）。以这类变量作为被解释变量建立的模型，分别称为离散选择模型和受限因变量模型。

离散选择模型旨在分析决策者在有限个备选方案中作出选择的概率，其因变量取值为离散的类别。受限因变量模型则处理因数据搜集机制导致观测值不能完全反映总体分布的情形，例如样本截断或数据归并。这两类模型极大地扩展了计量经济学的应用范围，使其能够更贴切地分析微观个体行为，广泛应用于劳动经济学、金融学、卫生经济学及消费行为研究等领域。

本章将系统阐述这些模型的理论基础、估计方法及解释。首先从最简单的二元选择模型出发，逐步扩展到多元选择、排序选择及计数数据模型，最后讨论处理样本选择或数据截断等问题的受限因变量模型。

5.1 二元选择模型

5.1.1 线性概率模型及其局限

对于因变量 y_i 取值为 0 或 1 的二元选择问题，一个直观的起点是采用线性概率模型：

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

其中 $E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$。

该模型尽管估计简便，但存在两个主要缺陷：1. 预测值可能超出概率区间：线性组合 $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ 的取值可能不在 `[0, 1]` 之间，这与概率的定义相矛盾。2. 异方差性：误差项 ε_i 的方差为 $\text{Var}(\varepsilon_i|\mathbf{x}_i) = (\mathbf{x}_i^\top \boldsymbol{\beta})(1 - \mathbf{x}_i^\top \boldsymbol{\beta})$，必然存在异方差，导致普通最小二乘估计非有效。

这些缺陷促使我们发展出形式更为灵活的非线性概率模型。

5.1.2 潜变量框架与非线性设定

二元选择模型通常基于一个连续的潜变量 y_i^* 来构建：

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \quad y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0 \end{cases}$$

其中 u_i 是独立于 $\mathbf{x}_i$ 的随机扰动项，均值为 0，方差固定。潜变量 y_i^* 可理解为决策者的 “净收益” 或 “效用差”。

在此框架下，选择概率为：

$$P(y_i = 1|\mathbf{x}_i) = P(y_i^* > 0|\mathbf{x}_i) = P(u_i > -\mathbf{x}_i^\top \boldsymbol{\beta}|\mathbf{x}_i) = 1 - F(-\mathbf{x}_i^\top \boldsymbol{\beta})$$

其中 $F(\cdot)$ 是 u_i 的累积分布函数。为保证概率值在 `[0, 1]` 之间且模型设定可识别，需对 $F(\cdot)$ 的形式进行假设。

5.1.3 Probit 与 Logit 模型

两种最常用的分布假设对应了核心的二元选择模型：* **Probit** 模型：假设 u_i 服从标准正态分布，即 $u_i \sim N(0,1)$。此时，$P(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$，其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数。* **Logit** 模型：假设 u_i 服从逻辑分布。此时，$P(y_i = 1|\mathbf{x}_i) = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$。

两种模型的概率函数都是 $\mathbf{x}_i^\top \boldsymbol{\beta}$ 的单调递增函数，其曲线呈 S 形，能自动保证预测概率落在 0 到 1 之间。Probit 模型和 Logit 模型的估计结果通常非常接近。它们的差异主要在于逻辑分布的尾部略厚于正态分布，但这在实证中很少导致实质性区别。模型选择常取决于学术传统或软件实现的便利性。

5.1.4 估计与解释

由于潜变量 y_i^* 不可观测，二元选择模型通常采用极大似然法进行估计。对于样本量为 N 的数据，似然函数为：

$$L(\boldsymbol{\beta}) = \prod_{i:y_i=1} P(y_i = 1|\mathbf{x}_i) \cdot \prod_{i:y_i=0} [1 - P(y_i = 1|\mathbf{x}_i)]$$

对似然函数取对数后，通过数值优化方法求解使对数似然函数最大的参数估计值 $\widehat{\boldsymbol{\beta}}$。

参数 $\boldsymbol{\beta}$ 的估计值不能直接解释为边际效应。以 Probit 模型为例，第 j 个解释变量 x_{ij} 对选择概率的边际效应为：

$$\frac{\partial P(y_i = 1|\mathbf{x}_i)}{\partial x_{ij}} = \phi(\mathbf{x}_i^\top \boldsymbol{\beta})\beta_j$$

其中 $\phi(\cdot)$ 是标准正态概率密度函数。该边际效应不是常数，它依赖于所有解释变量在 $\mathbf{x}_i$ 处的取值。因此，报告边际效应时，通常需要计算在解释变量样本均值处的值，或计算每个观测个体的边际效应后再求样本平均。

5.2 多元选择模型

当决策者面临两个以上的离散选项时，需要使用多元选择模型。根据选项之间是否具有自然的排序，可以分为无序选择模型和有序选择模型。本节讨论无序选择模型。

5.2.1 多项 Logit 模型

多项 **Logit** 模型是最常用的无序选择模型。设个体 i 在 $J+1$ 个选项（编号为 `0, 1, ..., J`）中选择，以选项 `0` 为参照基准。个体选择选项 j 的概率为：

$$P(y_i = j|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{J} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}, \quad j = 1, \ldots, J$$

以及 $P(y_i = 0|\mathbf{x}_i) = \frac{1}{1+\sum_{k=1}^{J} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}$。

该模型的一个关键性质是无关选项的独立性：任意两个选项的选择概率之比仅与这两个选项的特性有关，与其他选项无关。这一性质在某些情境下可能构成限制。

5.2.2 多项 Probit 模型及其他模型

为克服 IIA 性质的限制，可采用多项 **Probit** 模型。该模型假设与各选项相关的随机误差项服从多元正态分布。由于其似然函数涉及高维数值积分，计算较为复杂，但在计算能力提升和模拟方法发展的背景下，其应用正逐渐增加。

此外，还有嵌套 **Logit** 模型和混合 **Logit** 模型等，它们通过引入更灵活的误差相关结构，来建模选项之间可能存在的相关性。

5.3 排序选择模型

当因变量的离散类别之间存在内在的顺序时，例如调查问卷中的满意度（非常不满意、不满意、一般、满意、非常满意）、信用评级或疾病严重程度分级，则应使用排序选择模型。

5.4.1 模型设定

排序选择模型同样基于潜变量框架：

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

所观测到的有序结果 y_i 由潜变量 y_i^* 穿越一系列递增的门槛值 $\mu_1 < \mu_2 < \cdots < \mu_{J-1}$ 决定：

$$y_i = \begin{cases} 0, & \text{if } y_i^* \le \mu_1 \\ 1, & \text{if } \mu_1 < y_i^* \le \mu_2 \\ \vdots & \\ J-1, & \text{if } y_i^* > \mu_{J-1} \end{cases}$$

其中，为了模型识别，通常将常数项设为零，并设定 μ_1（或 μ_0 为 $-\infty$，μ_J 为 $+\infty$）。

5.4.2 有序 Probit 与有序 Logit

根据扰动项 ε_i 的分布假设，可得到两种主要模型：* 有序 **Probit** 模型：假设 $\varepsilon_i \sim N(0,1)$。* 有序 **Logit** 模型：假设 ε_i 服从逻辑分布。

个体 i 选择类别 j 的概率为：

$$\begin{aligned} P(y_i = j|\mathbf{x}_i) &= P(\mu_j < y_i^* \le \mu_{j+1}) \\ &= F(\mu_{j+1} - \mathbf{x}_i^\top \boldsymbol{\beta}) - F(\mu_j - \mathbf{x}_i^\top \boldsymbol{\beta}) \end{aligned}$$

其中 $F(\cdot)$ 是标准正态或逻辑分布的累积分布函数。

解释变量 $\mathbf{x}_i$ 对处于特定类别 j 的概率的边际效应，其符号并不确定，需要进行计算。$\boldsymbol{\beta}$ 系数的符号可以解释为对潜变量 y_i^* 的影响方向。

5.4 计数数据模型

当因变量是非负整数计数时，例如一个家庭在一年内就医的次数、一家企业申请的专利数量，或一个区域发生的交通事故数，需要使用计数数据模型。

5.4.1 泊松回归模型

泊松回归是计数模型的基础。设 y_i 给定 $\mathbf{x}_i$ 的条件分布为泊松分布：

$$P(y_i|\mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ...$$

其中，条件均值（即发生率的期望）被设定为指数形式，以确保其为正：

$$E(y_i|\mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$$

泊松分布的一个重要性质是等离散性，即条件均值等于条件方差：$E(y_i|\mathbf{x}_i) = \text{Var}(y_i|\mathbf{x}_i) = \lambda_i$。

5.4.1 负二项回归模型

在实际数据中，方差常常大于均值，这种现象称为过度离散。忽视过度离散会导致标准误被低估。负二项回归模型通过引入一个额外的随机成分来放松等离散假设，其条件方差被设定为

$$\text{Var}(y_i|\mathbf{x}_i) = \lambda_i + \alpha\lambda_i^2$$

，其中 $\alpha > 0$ 是衡量过度离散程度的参数。当 $\alpha = 0$ 时，负二项回归即退化为泊松回归。

此外，对于数据中零值过多的情形，还有零膨胀泊松模型和零膨胀负二项模型等专门应对零值堆积问题的扩展模型。

5.5 受限因变量模型

当因变量的观测值由于数据收集过程而受到限制时，就需要用到受限因变量模型。主要分为两类：截取和断尾。

5.5.1 截取回归模型

在截取情况下，部分因变量的真实值无法被观测到，但我们知道它们是否被截取以及截取的界限。最经典的模型是 **Tobit** 模型，由 James Tobin 提出，常用于分析诸如家庭耐用消费品支出（有大量零值）等数据。

标准 Tobit 模型（Type I）设定如下：

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \quad u_i | \mathbf{x}_i \sim N(0, \sigma^2) \\
y_i &= \max(0, y_i^*)
\end{aligned}
$$

我们观测到的是 y_i，而非 y_i^*。对于 $y_i > 0$ 的观测，其条件分布是断尾正态分布；对于 $y_i = 0$ 的观测，我们仅知道 $y_i^* \le 0$。

该模型可用极大似然法估计，其对数似然函数由两部分构成：对应于 $y_i = 0$ 观测的概率部分，和对应于 $y_i > 0$ 观测的密度部分。Tobit 模型的一个关键特征是，解释变量 $\mathbf{x}_i$ 同时对决定 y_i^* 是否大于 0 的概率（即 “选择方程”）和给定 $y_i^* > 0$ 时 y_i^* 的水平（即 “水平方程”）产生影响，且两个效应的系数比例是固定的，这有时可能构成限制。

两阶段估计和工具变量 **Tobit** 模型也被发展出来，以处理内生解释变量等问题。

5.5.2 断尾回归模型

在断尾情况下，部分观测值被完全排除在样本之外，既不知道其存在，也不知道其解释变量的值。例如，只对收入高于某个门槛的家庭进行调查。

若断尾规则为 $y_i^* > c$，则我们观测到的 y_i 的条件分布为：

$$
f(y_i | y_i > c, \mathbf{x}_i) = \frac{f(y_i^* | \mathbf{x}_i)}{P(y_i^* > c | \mathbf{x}_i)}, \quad y_i > c
$$

其中 $f(y_i^* | \mathbf{x}_i)$ 是 y_i^* 的原始条件密度函数。忽略断尾而直接使用 OLS 估计，会导致参数估计有偏。

5.5.3 样本选择模型

样本选择模型（Heckman 模型）处理的是更一般的选择性问题：是否进入样本（选择方程）与关心的结果变量（结果方程）由两个虽有联系但不同的机制决定。

其经典设定如下：

$$
\begin{aligned}
&\text{选择方程:} \quad s_i^* = \mathbf{z}_i^\top \boldsymbol{\gamma} + v_i, \quad s_i = \mathbf{1}[s_i^* > 0] \\
&\text{结果方程:} \quad y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i \\
&\text{观测规则:} \quad y_i = y_i^* \text{ 当且仅当} s_i = 1\text{; 否则} y_i \text{ 缺失}
\end{aligned}
$$

其中 (u_i, v_i) 假设服从二元正态分布。如果误差 u_i 和 v_i 相关，那么仅对可观测样本进行 OLS 回归就会导致样本选择偏差。

Heckman 提出了一个广为使用的两阶段纠正方法：1. 利用全部样本，用 Probit 模型估计选择方程，得到逆米尔斯比 $\hat{\lambda}_i = \phi(\mathbf{z}_i^\top \hat{\boldsymbol{\gamma}}) / \Phi(\mathbf{z}_i^\top \hat{\boldsymbol{\gamma}})$。2. 在可观测子样本中，将 $\hat{\lambda}_i$ 作为额外控制变量加入结果

方程进行 OLS 回归，即估计 $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \rho\sigma_u \hat{\lambda}_i + \eta_i$。其中 $\rho\sigma_u$ 的系数显著性检验了选择偏差的存在性。

5.6 模型设定检验与前沿议题

5.6.1 对模型假设的检验

离散与受限因变量模型大多依赖于较强的分布假设（如正态性、逻辑分布）和函数形式假设（如线性指数）。近年来，针对这些假设的检验方法不断发展。

例如，对于 Tobit 模型，有文献系统性地发展了对其核心识别假设（线性指数、潜在误差的正态性、外生性）的可检验等式，并构建了相应的检验程序。还有研究专注于检验 Tobit 模型中的正态性假设，或将其与更灵活的两部分模型进行对比的设定检验。

当关键假设被拒绝时，研究者可以考虑使用对分布假设更稳健的半参数或非参数估计方法，或者转向基于较弱假设的部分识别分析框架。

5.6.2 动态模型与面板数据扩展

在面板数据背景下，离散和受限因变量模型可以纳入个体效应（固定效应或随机效应）以控制不随时间变化的异质性。更进一步，动态模型（包含因变量滞后项）被用来研究状态依赖性和调整成本，例如上一期的就业状态如何影响当期的就业概率。

这些扩展模型在估计上更具挑战，但为分析经济行为的持续性和动态变化提供了有力工具。

5.7 案例分析

本章节旨在提供一个或多个综合性的实证研究框架示例，展示如何根据研究问题与数据类型，从本章介绍的模型库中选择适当的模型，并完成从模型设定、估计、假设检验到结果解释的全过程。例如，可以分析影响个人高等教育选择（二元/多项选择）的因素，或研究家庭慈善捐款数额（截取数据）的决定因素。具体案例内容此处从略。

本章总结

本章系统介绍了当因变量为离散或受限变量时的一整套计量经济学建模方法。我们从二元选择的基础模型（Probit/Logit）出发，逐步扩展到多元无序选择、有序选择以及计数数据模型。最后，探

讨了处理数据截取、断尾和样本选择问题的受限因变量模型。

这些模型的共同核心是基于潜变量或直接设定非线性的条件期望函数，并通常依赖最大似然估计。在解释估计结果时，必须谨慎，因为系数通常不代表简单的边际效应，边际效应本身也常常依赖于其他变量的取值。

在选择模型时，首要准则是因变量的数据类型和经济问题的实质（有无序、是否有序、是否计数、是否受限）。同时，需要意识到各种经典模型背后的假设，并利用不断发展成熟的检验方法对模型设定进行诊断。当数据和方法允许时，考虑使用更稳健的估计量或探索面板数据、动态模型等扩展形式，能使实证分析更为深入和可靠。

掌握本章内容，将使研究者能够更恰当地处理微观实证研究中广泛存在的离散和受限因变量问题，从而得出更有效的经验证据和经济解释。

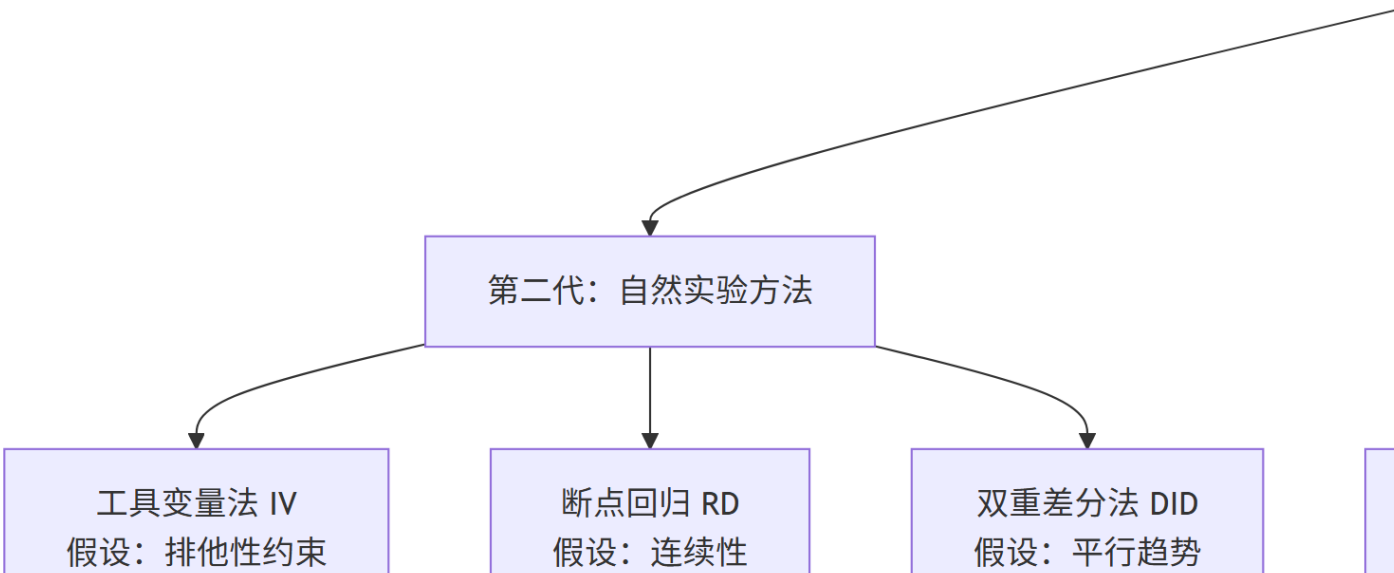
II 因果推断方法

6 因果推断框架

本章导读

“相关性不是因果性”——这一原则构成了现代计量经济学的基石。在前几章中，我们学习了如何使用回归模型描述变量间的相关关系。从本章开始，我们将回答一个更根本的问题：如何从观测数据中识别因果关系？本章将建立因果推断的基本理论框架，为后续章节的具体方法奠定基础。

在经济学研究中，我们经常关心诸如” 教育对收入的影响”、“最低工资对就业的影响”、“货币政策对经济增长的影响” 等问题。这些问题的本质都是因果问题——我们想知道如果改变某个变量（处理变量），结果变量会发生怎样的变化。然而，从观测数据中识别因果关系面临着选择偏差、内生性等根本挑战。本章将系统介绍因果推断的理论框架，为理解后续章节的具体方法提供坚实基础。


第二代：自然实验方法
工具变量法 IV
假设：排他性约束
断点回归 RD
假设：连续性
双重差分法 DID
假设：平行趋势

6.1 从相关到因果：问题的根本转变

6.1.1 相关性分析的局限

考虑以下三个经典例子：

1. 冰淇淋销量与溺水人数：两者呈现正相关，但这是因果关系吗？实际上，两者都受到季节（夏季）的影响。
2. 教育年限与收入：受教育程度高的人通常收入更高，但这是因为教育本身提高了生产率，还是因为能力高的人既倾向于接受更多教育又容易获得高收入？
3. 班级规模与学生成绩：小班教学的学生成绩更好，但这是因为班级规模的影响，还是因为资源丰富的学校既倾向于小班化又提供更好的教学条件？

这些例子揭示了相关性与因果性的根本区别。相关性描述的是变量间的统计关联，而因果性描述的是一个变量的变化如何导致另一个变量的变化。

6.1.2 经济研究中的因果问题类型

经济学中的因果问题主要分为三类：

1. 政策干预效果评估：评估某项政策（如税收改革、教育补贴）对经济结果的影响。
2. 行为反应机制分析：分析个体或企业对激励变化的反应。
3. 市场均衡效应识别：识别市场结构变化对均衡价格和数量的影响。

6.2 潜在结果框架：因果推断的统一语言

6.2.1 基本设定与符号体系

潜在结果框架（Potential Outcomes Framework），又称 Rubin 因果模型，由 Donald Rubin 等人发展，已成为现代因果推断的标准语言。

对于每个个体 i，我们定义：

- 处理状态：$T_i \in \{0, 1\}$，其中 $T_i = 1$ 表示接受处理（如参加培训项目），$T_i = 0$ 表示未接受处理（控制组）
- 潜在结果：$Y_i(1)$ 表示如果个体 i 接受处理时的结果，$Y_i(0)$ 表示如果个体 i 未接受处理时的结果

观测到的结果可以表示为：

$$Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

6.2.2 因果推断的**"** 根本问题**"**

因果推断面临的根本问题是：对于同一个体，我们只能观测到一种潜在结果。如果个体接受了处理（$T_i = 1$），我们观测到 $Y_i(1)$ 但无法观测 $Y_i(0)$；如果个体未接受处理（$T_i = 0$），我们观测到 $Y_i(0)$ 但无法观测 $Y_i(1)$。这个反事实结果（Counterfactual Outcome）的不可观测性被称为因果推断的根本问题。

6.2.3 主要因果参数

由于个体处理效应 $\tau_i = Y_i(1) - Y_i(0)$ 不可观测，我们转向估计群体层面的平均效应：

1. 平均处理效应（Average Treatment Effect, ATE）：

$$\tau_{\text{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

2. 处理组的平均处理效应（Average Treatment Effect on the Treated, ATT）：

$$\tau_{\text{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$$

3. 控制组的平均处理效应（Average Treatment Effect on the Controls, ATC）：

$$\tau_{\text{ATC}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 0]$$

6.3 稳定性假设与选择偏差

6.3.1 SUTVA 假设及其经济含义

稳定单位处理值假设（Stable Unit Treatment Value Assumption, SUTVA）包含两个部分：

1. 无干扰性：个体 i 的结果不受其他个体处理状态的影响。

$$Y_i(T_1, T_2, ..., T_n) = Y_i(T_i)$$

2. 处理一致性：同一处理对所有个体具有相同的含义和效果。

在经济学中，SUTVA 的违背常见于：- 溢出效应：一个地区的基础设施投资可能影响邻近地区 - 一般均衡效应：大规模政策可能改变市场价格和资源配置 - 网络效应：个体的行为可能受到社会网络的影响

6.3.2 选择偏差：因果推断的核心障碍

选择偏差源于处理组和对照组在潜在结果上的系统性差异。观测到的均值差异可以分解为：

$$\mathbb{E}[Y_i^{\text{obs}}|T_i=1]-\mathbb{E}[Y_i^{\text{obs}}|T_i=0]=\tau_{\text{ATT}}+\text{选择偏差}$$

其中选择偏差为：

$$\text{选择偏差}=\mathbb{E}[Y_i(0)|T_i=1]-\mathbb{E}[Y_i(0)|T_i=0]$$

当处理组个体即使不接受处理，其潜在结果 $Y_i(0)$ 也不同于对照组时，就产生了选择偏差。

6.3.3 选择偏差的类型与来源

偏差类型	来源	经济实例
可观测特征偏差	处理组和对照组在可观测特征上的差异	高收入者更可能参加培训项目
不可观测特征偏差	处理组和对照组在不可观测特征上的差异	能力高的人既倾向于接受教育又容易获得高收入
自选择偏差	个体基于预期结果选择是否接受处理	预期培训效果好的个体更可能参加培训
制度性选择偏差	制度规则导致的选择	贫困线以下的家庭自动获得福利

6.4 随机化实验：选择偏差的”黄金标准解”

6.4.1 随机化的理论保障

随机化实验通过随机分配处理状态，确保处理组和对照组在所有特征（包括可观测和不可观测特征）上具有可比性：

$$T_i\perp\{Y_i(1),Y_i(0),X_i,U_i\}$$

其中 X_i 表示可观测特征，U_i 表示不可观测特征。这一独立性意味着：

$$\mathbb{E}[Y_i(0)|T_i=1]=\mathbb{E}[Y_i(0)|T_i=0]$$

因此，选择偏差为零，ATE 的简单均值差估计量是一致的：

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i$$

6.4.2 经济学实验的设计类型

1. 实验室实验：在控制环境下进行，适用于检验理论机制。
2. 田野实验：在自然环境中进行，具有更高的外部有效性。
3. 自然实验：利用外生政策变化或自然事件作为处理分配。

6.4.3 随机化实验的局限与挑战

1. 外部有效性：实验环境可能无法反映真实世界
2. 伦理约束：某些处理（如有害物质）不能随机分配
3. 成本高昂：大规模实验需要大量资源
4. 依从性问题：实验对象可能不遵守分配
5. 处理效应异质性：简单均值差可能掩盖效应异质性

6.5 非混杂性：观测研究的识别基石

6.5.1 条件独立性的形式化表达

当随机化不可行时，我们需要依赖观测数据。强可忽略性假设（Strong Ignorability）是观测研究中因果识别的基础：

$$(Y_i(1), Y_i(0)) \perp T_i | X_i$$

这一假设要求：在给定协变量 X_i 的条件下，处理分配 T_i 与潜在结果独立。此外，还需要共同支持条件：

$$0 < \Pr(T_i = 1 | X_i = x) < 1 \quad \text{对于所有 } x$$

6.5.2 非混杂性的经济学解释

非混杂性假设意味着：所有同时影响处理选择和结果的变量都已包含在 X_i 中。在给定 X_i 的层内，处理分配如同随机。

考虑教育对收入的影响例子：- 如果能力既影响教育选择又影响收入，且能力可观测，则控制能力后，非混杂性可能成立。- 如果能力不可观测，则非混杂性被违背，估计将有偏。

6.5.3 假设的实践评估

在实践中，我们需要：

1. 基于经济理论选择控制变量
2. 检验平衡性：处理组和对照组在 X_i 上是否平衡
3. 进行敏感性分析：评估结论对未观测混杂的稳健性

6.6 内生性：计量经济学的经典难题

6.6.1 内生性的三个来源

内生性指解释变量与误差项相关，是因果推断的主要障碍：

1. 遗漏变量：未观测的混杂变量 U_i 既影响 T_i 又影响 Y_i
2. 反向因果：Y_i 影响 T_i，同时 T_i 影响 Y_i
3. 测量误差：T_i 的测量存在误差

6.6.2 内生性的统计后果

考虑线性模型：

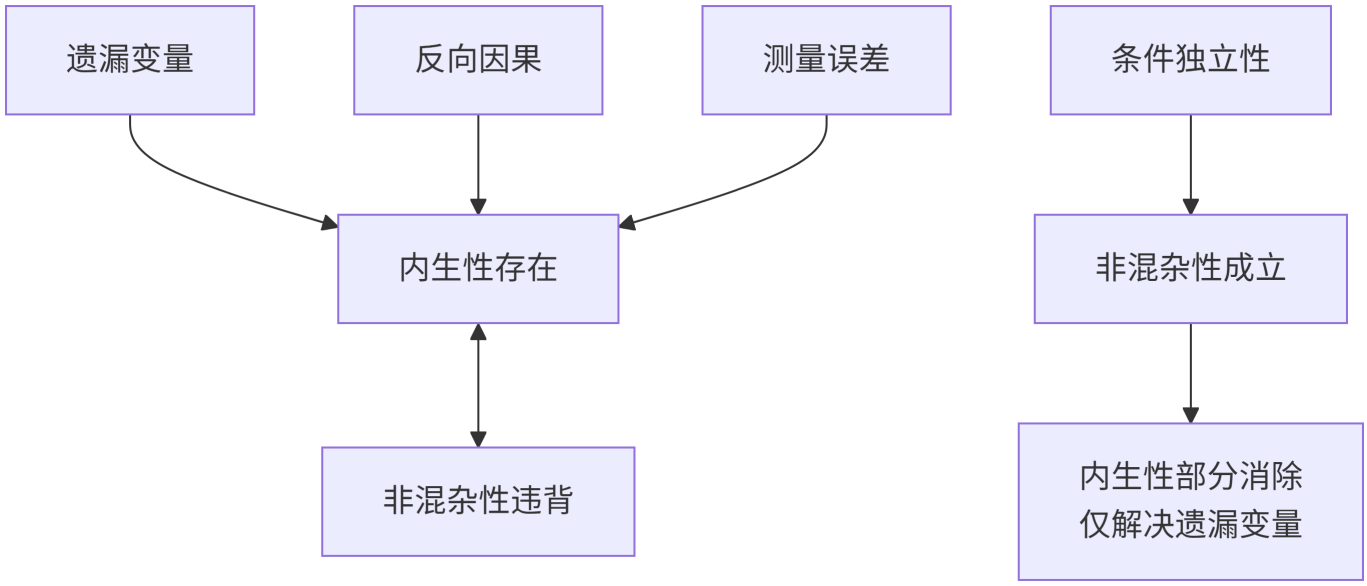
$$Y_i = \alpha + \beta T_i + \gamma' X_i + \epsilon_i$$

如果 $\mathbb{E}[\epsilon_i | T_i, X_i] \neq 0$，则：

1. OLS 估计量 $\hat{\beta}$ 有偏且不一致
2. 标准误差估计有偏
3. 假设检验失效
4. 预测和政策建议不可靠

6.6.3 内生性与非混杂性的关系

内生性和非混杂性是同一问题的两种表述。非混杂性成立意味着无遗漏变量问题，从而消除了一种内生性来源。具体关系如下：



6.7 线性回归的因果解释条件

6.7.1 随机化实验下的线性回归

在随机化实验中，即使是最简单的线性回归也能提供无偏的因果估计：

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

随机化保证 $\mathbb{E}[\epsilon_i | T_i] = 0$，因此 $\hat{\beta} = \widehat{\text{ATE}}$。

6.7.2 满足非混杂性条件的线性回归

当非混杂性成立时，包含所有混杂变量的线性回归可以提供无偏估计：

$$Y_i = \alpha + \beta T_i + \gamma' X_i + \epsilon_i$$

条件独立性 $(Y_i(1), Y_i(0)) \perp T_i | X_i$ 保证了 $\mathbb{E}[\epsilon_i | T_i, X_i] = 0$，因此 $\hat{\beta} = \widehat{\text{ATE}}$。

6.7.3 回归控制法：扩展与应用

回归控制法（Regression Adjustment）是线性回归在因果推断中的直接应用。当非混杂性成立时，通过控制所有混杂变量 X_i，我们可以获得处理效应的无偏估计。

在实践中，我们需要：

1. 正确设定函数形式（考虑非线性、交互项）
2. 检查共同支持条件
3. 进行模型诊断和稳健性检验

6.7.4 遗漏变量偏差的定量分析

考虑真实数据生成过程为：

$$Y_i = \alpha + \beta T_i + \gamma X_i + \delta U_i + \eta_i$$

但如果我们只控制 X_i，估计模型为：

$$Y_i = \tilde{\alpha} + \tilde{\beta} T_i + \tilde{\gamma} X_i + \tilde{\eta}_i$$

那么 $\tilde{\beta}$ 的概率极限为：

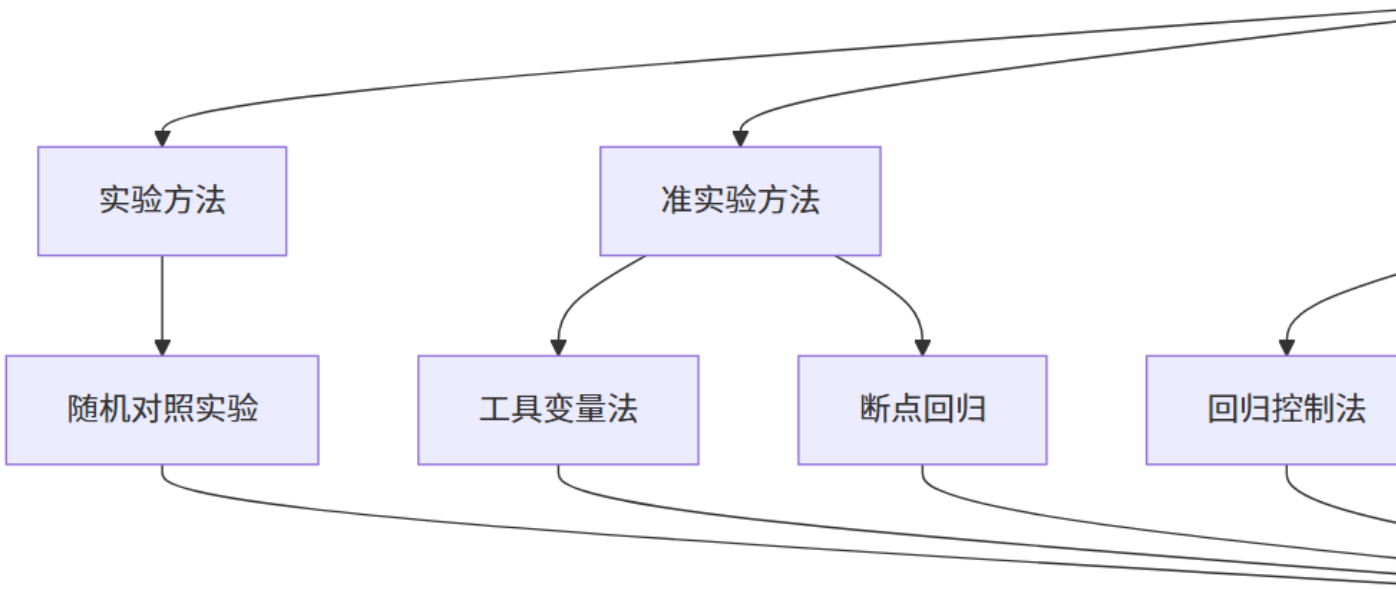
$$\text{plim}\, \tilde{\beta} = \beta + \delta \frac{\text{Cov}(T_i, U_i | X_i)}{\text{Var}(T_i | X_i)}$$

偏差的大小取决于：1. U_i 对 Y_i 的影响强度（δ）2. T_i 和 U_i 在给定 X_i 下的相关性

6.8 因果推断方法分类框架

6.8.1 基于识别策略的分类体系

现代因果推断方法可以根据其核心识别策略分为五大类：



6.8.2 各类方法的比较分析

方法类别	代表方法	关键假设	估计参数	数据要求	适用场景
实验方法	随机对照实验	完美随机化	ATE	实验数据	可实施随机化
准实验方法	工具变量法	外生工具变量	LATE	有效工具	有自然实验
准实验方法	断点回归	连续性假设	局部 ATE	运行变量	有清晰断点
可忽略性方法	回归控制法	条件独立性	ATE	丰富协变量	可测所有混杂
可忽略性方法	倾向得分匹配	强可忽略性	ATT	平衡协变量	对照组丰富
面板数据方法	固定效应模型	时不变混杂	ATE	面板数据	个体异质性
面板数据方法	双重差分法	平行趋势	ATT	面板数据	政策评估
合成控制方法	合成控制法	可合成性	ATT	时间序列	小样本政策评估

6.8.3 固定效应模型：处理时不变混杂

6.8.3.1 面板数据与因果识别的优势

固定效应模型是处理观测数据中未观测混杂的重要方法，特别适用于面板数据（Panel Data）或追踪数据（Longitudinal Data）。面板数据的核心特征是每个个体（如个人、企业、地区）在多个时间

点上被观测，这为我们处理因果推断问题提供了独特优势。

面板数据结构：

$$\{Y_{it}, T_{it}, X_{it}\}, \quad i = 1, ..., N; \quad t = 1, ..., T$$

其中：- i：个体标识（如个人、企业、城市）- t：时间标识（如年份、季度）- Y_{it}：个体 i 在时间 t 的结果变量 - T_{it}：个体 i 在时间 t 的处理状态（0 或 1）- X_{it}：个体 i 在时间 t 的可观测协变量

6.8.3.2 固定效应模型的因果识别机制

考虑以下固定效应模型：

$$Y_{it} = \alpha_i + \lambda_t + \beta T_{it} + \gamma' X_{it} + \epsilon_{it}$$

其中：- α_i：个体固定效应，捕捉所有不随时间变化的个体特征 - λ_t：时间固定效应，捕捉所有个体共同经历的时间趋势 - β：处理效应，是我们关心的因果参数 - ϵ_{it}：时变冲击，满足 $\mathbb{E}[\epsilon_{it}|T_{it}, X_{it}, \alpha_i, \lambda_t] = 0$

** 固定效应如何解决未观测混杂 **

固定效应模型的威力在于 α_i 可以吸收所有不随时间变化的未观测混杂。考虑一个具体例子：

研究问题：分析员工培训（T_{it}）对工资（Y_{it}）的因果效应。

未观测混杂：员工的能力（U_i）既影响是否参加培训，又影响工资水平。能力通常不可直接观测或难以准确测量。

传统截面数据分析的问题：如果使用截面数据，我们需要控制能力 U_i。但由于 U_i 不可观测，导致遗漏变量偏差：

$$\text{plim}\, \hat{\beta}_{OLS} = \beta + \frac{\text{Cov}(T_i, U_i)}{\text{Var}(T_i)}$$

固定效应模型的解决方案：在面板数据中，我们可以将能力分解为：

$$\text{能力}_i = \underbrace{\alpha_i}_{\text{时不变部分}} + \underbrace{v_{it}}_{\text{时变部分}}$$

固定效应模型通过 α_i 吸收了时不变的能力部分。只要能力的大部分变异是时不变的，固定效应模型就能有效消除能力混杂带来的偏差。

固定效应模型的有效性依赖于以下关键假设：

1. 严格外生性假设：

$$\mathbb{E}[\epsilon_{it}|T_{i1}, ..., T_{iT}, X_{i1}, ..., X_{iT}, \alpha_i, \lambda_t] = 0$$

这意味着给定个体固定效应和时间固定效应后，处理变量 T_{it} 和协变量 X_{it} 与误差项 ϵ_{it} 不相关。

2. 未观测混杂的时不变性：所有未观测的混杂变量 U_i 必须满足：

$$U_i = \alpha_i + v_{it}, \quad \text{其中 } \alpha_i \text{ 为时不变部分}$$

固定效应只能消除 α_i 部分，无法处理时变部分 v_{it}。

3. 处理效应同质性（或已知的异质性模式）：

$$Y_{it}(1) - Y_{it}(0) = \beta \quad \text{对所有 } i, t$$

或至少处理效应的异质性模式是已知且可建模的。

6.8.3.3 估计方法与实现

** 组内估计量（Within Estimator）**

固定效应模型最常用的估计方法是组内估计量，通过消除个体固定效应进行估计：

第一步：计算个体均值

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \quad \bar{T}_i = \frac{1}{T}\sum_{t=1}^{T} T_{it}, \quad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

第二步：进行组内变换

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i, \quad \tilde{T}_{it} = T_{it} - \bar{T}_i, \quad \tilde{X}_{it} = X_{it} - \bar{X}_i$$

第三步：估计变换后的模型

$$\tilde{Y}_{it} = \lambda_t + \beta \tilde{T}_{it} + \gamma' \tilde{X}_{it} + \tilde{\epsilon}_{it}$$

其中时间固定效应 λ_t 可以通过加入时间虚拟变量或进行时间均值差分来消除。

一阶差分估计量（**First-Difference Estimator**）

另一种常用方法是一阶差分法，特别适用于 $T = 2$ 的情况：

差分变换:

$$\Delta Y_i = Y_{i2} - Y_{i1}, \quad \Delta T_i = T_{i2} - T_{i1}, \quad \Delta X_i = X_{i2} - X_{i1}$$

估计模型：

$$\Delta Y_i = \beta \Delta T_i + \gamma' \Delta X_i + \Delta \epsilon_i$$

一阶差分法同样消除了个体固定效应 α_i。

两种方法的比较

方法	优点	缺点	适用场景
组内估计量	效率高（使用所有变异）	需要严格外生性	平衡面板，$T \geq 2$
一阶差分	对序列相关更稳健	损失信息，效率较低	$T = 2$ 或担心严格外生性

6.8.3.4 固定效应模型的因果解释

处理效应的识别来源

在固定效应模型中，处理效应 β 的识别来源于个体内部处理状态的变化。具体来说：

1. 处理组个体：通过比较同一个体在处理前后的变化
2. 处理状态变化个体：通过比较个体从未处理到处理（或相反）的变化
3. 始终处理/未处理个体：不贡献于 β 的识别（除非有处理效应异质性）

这种识别策略被称为" 利用个体内部变异"，其优势在于可以有效控制所有时不变的个体异质性。

图示说明：固定效应模型的识别机制

数学证明：固定效应模型的无偏性

假设真实数据生成过程为：

$$Y_{it} = \alpha_i + \lambda_t + \beta T_{it} + \theta' U_i + \gamma' X_{it} + \epsilon_{it}$$

其中 U_i 为未观测的时不变混杂变量。

在组内变换后：

$$\tilde{Y}_{it} = \lambda_t + \beta \tilde{T}_{it} + \gamma' \tilde{X}_{it} + \tilde{\epsilon}_{it}$$

因为 $\tilde{U}_i = U_i - \bar{U}_i = 0$，未观测混杂 U_i 被完全消除。只要 $\mathbb{E}[\tilde{\epsilon}_{it} | \tilde{T}_{it}, \tilde{X}_{it}] = 0$，$\hat{\beta}$ 就是 β 的无偏估计量。

** 主要局限性 **

1. 时变混杂问题：固定效应只能消除时不变混杂，无法处理时变未观测混杂

2. 动态选择问题：如果处理决策基于过去的冲击（$\epsilon_{i,t-1}$），严格外生性假设被违背
3. 处理效应异质性：如果处理效应因人而异，固定效应估计量可能不是有意义的平均
4. 测量误差偏误：组内变换可能放大测量误差的影响

解决方案

1. 滞后因变量模型：控制滞后结果变量

$$Y_{it} = \alpha_i + \lambda_t + \rho Y_{i,t-1} + \beta T_{it} + \gamma' X_{it} + \epsilon_{it}$$

2. 动态面板模型：使用 GMM 方法估计

$$Y_{it} = \alpha_i + \lambda_t + \rho Y_{i,t-1} + \beta T_{it} + \gamma' X_{it} + \epsilon_{it}$$

3. 事件研究法：检验处理前后的动态效应

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k=-K}^{-1} \beta_k \cdot D_{i,t+k} + \sum_{k=0}^{L} \beta_k \cdot D_{i,t+k} + \gamma' X_{it} + \epsilon_{it}$$

   其中 $D_{i,t+k}$ 是个体 i 在 $t+k$ 期是否处于处理期的虚拟变量。

固定效应模型在以下情况下特别适用：

1. 面板数据可得：每个个体有多个时间点的观测
2. 主要混杂时不变：理论判断主要混杂变量不随时间变化
3. 处理状态变化：个体处理状态随时间发生变化
4. 平行趋势假设：在没有处理的情况下，处理组和对照组的趋势相同 ### 6.8.4 合成控制法：小样本政策评估

合成控制法适用于处理单元较少（如一个州、一个国家）的政策评估问题：

基本思想：从未受处理的供体单元中构造一个” 合成控制组”，使其在处理前的特征和结果轨迹与处理单元尽可能相似。

优化问题：

$$\min_{w} \left\| X_1 - \sum_{j=2}^{J+1} w_j X_j \right\|_V \quad \text{s.t.} \quad w_j \geq 0, \sum_{j=2}^{J+1} w_j = 1$$

其中 X_1 为处理单元的预处理特征向量，X_j 为供体单元的特征向量，V 为权重矩阵，w_j 为供体单元的权重。

处理效应估计：

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}, \quad t > T_0$$

其中 T_0 为处理发生的时间。

关键假设：1. 供体池足够丰富，能够较好地拟合处理单元 2. 处理前拟合期足够长，能够捕捉趋势 3. 处理单元与合成控制组在处理前具有相似的特征和趋势

推断方法：排列检验（Placebo Test）

6.9 实证案例分析：最低工资的就业效应

6.9.1 研究背景与经典争议

最低工资政策对就业的影响是劳动经济学中的经典问题。传统理论预测，提高最低工资会减少就业，但实证证据并不一致。

6.9.2 Card 和 Krueger（1994）的自然实验

Card 和 Krueger 利用新泽西州提高最低工资而相邻的宾夕法尼亚州未提高的自然实验，采用双重差分法估计最低工资对快餐业就业的影响。

6.9.3 研究设计与识别策略

1. 处理组：新泽西州的快餐店
2. 对照组：宾夕法尼亚州的快餐店
3. 处理前后：1992 年 2 月（政策前）和 1992 年 11 月（政策后）
4. 识别假设：平行趋势假设——如果没有最低工资提高，两州的就业趋势相同

6.9.4 Stata 操作演示（目录）

- 6.9.4.1 数据导入与清理
- 6.9.4.2 描述性统计分析
- 6.9.4.3 平行趋势检验
- 6.9.4.4 双重差分估计
- 6.9.4.5 稳健性检验

- 6.9.4.6 结果可视化

6.9.5 R 操作演示（目录）

- 6.9.5.1 数据准备与探索
- 6.9.5.2 使用 `did` 包进行 DID 估计
- 6.9.5.3 事件研究法实现
- 6.9.5.4 敏感性分析
- 6.9.5.5 结果报告与可视化

本章总结

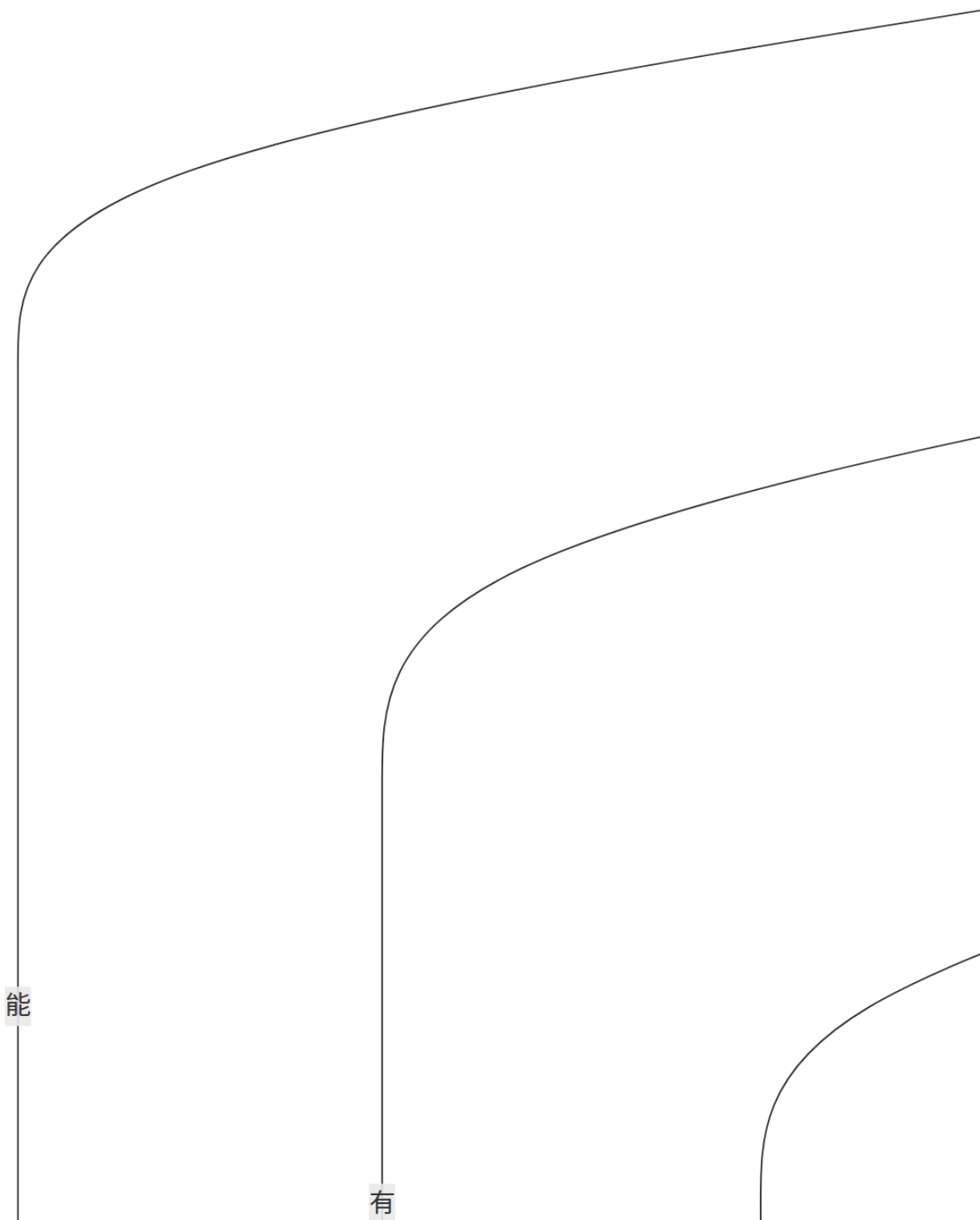
核心概念体系回顾

1. 潜在结果框架：定义了因果推断的基本语言和核心参数（ATE、ATT、ATC）
2. 根本问题：反事实结果的不可观测性是因果推断的根本障碍
3. 选择偏差：源于处理组和对照组在潜在结果上的系统性差异
4. **SUTVA** 假设：保证了处理效应的稳定性和可定义性
5. 非混杂性：观测研究中因果识别的核心假设
6. 内生性：计量经济学中的经典难题，有三个主要来源

方法体系梳理

因果推断方法可以根据其识别策略分为五大类：

1. 实验方法：随机化实验是因果识别的黄金标准
2. 准实验方法：利用自然实验或制度设计模拟随机化（IV、RD）
3. 基于可忽略性方法：通过控制所有混杂变量识别因果效应（回归控制、匹配）
4. 面板数据方法：利用时间维度消除时不变混杂（FE、DID）
5. 合成控制方法：为小样本政策评估提供解决方案


能
有

关键理论关系

1. 随机化的作用：
$$T_i \perp (Y_i(1), Y_i(0), X_i, U_i) \Rightarrow \text{无条件满足非混杂性}$$

2. 线性回归的因果解释条件：
   - 随机化条件下：$\mathbb{E}[\epsilon_i|T_i] = 0$
   - 非混杂性条件下：$\mathbb{E}[\epsilon_i|T_i, X_i] = 0$

3. 固定效应模型的识别力量：
$$Y_{it} = \alpha_i + \beta T_{it} + \epsilon_{it} \Rightarrow \text{消除所有时不变混杂}$$

从理论到实践的桥梁

本章建立的框架为后续章节的具体方法提供了理论基础：

1. 工具变量法（第 **7** 章）：通过寻找外生工具解决内生性问题
2. 断点回归（第 **8** 章）：利用制度断点创造局部随机化
3. 匹配方法（第 **9** 章）：基于可忽略性假设构造可比样本
4. 双重差分法（第 **10** 章）：结合面板数据和平行趋势假设
5. 合成控制法（第 **11** 章）：小样本政策评估的专门方法

实践指导原则

在进行因果推断研究时，应遵循以下原则：

1. 透明性：明确陈述识别假设和可能违背
2. 稳健性：使用多种方法和设定检验结论的稳健性
3. 诚实性：承认研究的局限性，避免过度解读
4. 理论指导：基于经济理论选择变量和设定模型
5. 敏感性分析：评估结论对关键假设的敏感性

扩展思考

1. 机器学习与因果推断：如何将机器学习方法用于协变量选择和模型设定？
2. 异质性处理效应：如何识别和处理效应的异质性？
3. 动态处理效应：如何处理处理效应的动态变化？

4. 溢出效应和一般均衡：如何放宽 SUTVA 假设？

因果推断不仅是统计学和计量经济学的方法论，更是一种科学的思维方式。它要求我们从” 是什么”（描述）转向” 如果…会怎样”（因果），从被动观察转向主动思考。掌握这一框架，将使你能够更严谨地评估经济理论和政策效果，成为更优秀的经济学家。

7 工具变量法

本章导读

在观察性研究中，当解释变量因测量误差、双向因果关系或不可观测的遗漏变量而与误差项相关时，常规的回归估计将失效。本章系统介绍解决此类内生性问题的核心方法之一——工具变量法。其基本思想是寻找一个满足特定条件的外部变量（工具变量），该变量与内生解释变量高度相关，但仅通过该内生变量影响结果变量，从而提供了一个可用于识别因果关系的 “外生变异” 来源。本章将从 IV 的直观逻辑入手，详细阐述其识别条件、核心估计方法（两阶段最小二乘法）、严格的统计检验以及在实际应用中的关键问题。

7.1 工具变量法的引入：动机与基本思想

7.1.1 OLS 的内生性问题回顾与 IV 的针对性

考虑线性回归模型 $y_i = \beta_0 + \beta_1 x_i + u_i$。当 x_i 与误差项 u_i 相关（即存在内生性）时，OLS 估计量 $\hat{\beta}_1^{OLS}$ 不一致：

$$\text{plim}\ \hat{\beta}_1^{OLS} = \beta_1 + \frac{Cov(x_i, u_i)}{Var(x_i)} \neq \beta_1.$$

内生性可能源于测量误差、双向因果或遗漏不可观测变量。工具变量法为此提供了一种解决方案：找到一个工具变量 z_i，它仅通过影响 x_i 来间接影响 y_i，从而利用 z_i 带来的外生变异识别 β_1。

7.1.2 工具变量法的直观类比与核心思想

工具变量的核心思想可通过 “分而治之” 来理解：1. 第一阶段：工具变量 z_i 与内生变量 x_i 相关（$Cov(z_i, x_i) \neq 0$），因此 z_i 的变化能预测 x_i 的变化。2. 排他性：工具变量 z_i 与误差项 u_i 不相关（$Cov(z_i, u_i) = 0$），因此 z_i 的变化不会直接干扰 y_i。3. 识别：因此，y_i 中与 z_i 相关的变动，只能是通过 x_i 传导的变动，从而可用于估计 x_i 对 y_i 的因果效应 β_1。

7.1.3 一个简单的工具变量模型设定

考虑模型 $y_i = \beta_0 + \beta_1 x_i + u_i$，其中 x_i 内生。假设存在有效的工具变量 z_i，其满足两个核心条件：1. 相关性：$Cov(z_i, x_i) \neq 0$。2. 外生性：$Cov(z_i, u_i) = 0$。

基于矩条件 $E[u_i] = 0$ 和 $E[z_i u_i] = 0$，可推导出工具变量估计量（简单情形下）：

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, x_i)}.$$

这直观地表示为 "z_i 对 y_i 的效应" 与 "z_i 对 x_i 的效应" 之比。

7.2 工具变量的定义与识别条件

7.2.1 工具变量的两个关键假设：相关性与外生性

对于模型 $y_i = \beta_0 + \beta_1 x_i + u_i$，变量 z_i 是 x_i 的一个有效工具变量，若满足：- 相关性条件：$Cov(z_i, x_i) \neq 0$。该条件可直接用数据检验。- 外生性条件：$Cov(z_i, u_i) = 0$。该条件涉及不可观测的 u_i，通常无法直接检验，必须基于经济理论与研究设计进行论证。它等价于要求 z_i 只能通过 x_i 影响 y_i，而不能存在其他直接或间接的路径。

7.2.2 排他性约束及其经济含义

外生性条件通常被称为 **"排他性约束"**。它要求工具变量 z_i 对结果变量 y_i 的所有影响都必须通过内生解释变量 x_i 这一唯一渠道。用因果图表示，有效的关系应为 $z_i \to x_i \to y_i$，而不能存在 $z_i \to y_i$ 的直接路径，或 $z_i \to W \to y_i$ 的间接路径（其中 W 为未包含在模型中的其他变量）。违反排他性约束将导致 IV 估计量不一致。

7.2.3 识别条件：恰好识别与过度识别

- 恰好识别：工具变量的数量（L）等于内生解释变量的数量（K），即 $L = K$。此时参数有唯一解。
- 过度识别：工具变量的数量多于内生解释变量的数量，即 $L > K$。过度识别提供了检验工具变量外生性的可能性（如过度识别检验），且通常可以提高估计效率。
- 不足识别：$L < K$。模型无法识别所有参数。

7.3 两阶段最小二乘法

7.3.1 2SLS 的估计步骤与几何解释

两阶段最小二乘法是估计工具变量模型最常用的方法，尤其适用于过度识别情形。设有单一内生变量 x_i 和多个工具变量 $z_{1i}, ..., z_{Li}$。

第一阶段：将内生变量 x_i 对所有工具变量 z_{li} 及模型中的外生变量 w_i（若有）进行回归：

$$x_i = \pi_0 + \pi_1 z_{1i} + ... + \pi_L z_{Li} + \delta w_i + v_i.$$

得到 x_i 的预测值 $\hat{x}_i$。$\hat{x}_i$ 是 z_i 的线性组合，因此与 u_i 不相关。

第二阶段：将结果变量 y_i 对第一阶段预测值 $\hat{x}_i$ 及外生变量 w_i 进行回归：

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \gamma w_i + \epsilon_i.$$

所得 $\hat{\beta}_1^{2SLS}$ 即为一致估计量。几何上，2SLS 先将 x_i 投影到工具变量张成的空间，再将 y_i 投影到该预测值上。

7.3.2 2SLS 估计量的统计性质（一致性、渐近正态性）

在工具变量相关性及外生性条件下，2SLS 估计量具有以下性质：- 一致性：$\text{plim}\, \hat{\beta}_1^{2SLS} = \beta_1$。- 渐近正态性：$\sqrt{n}(\hat{\beta}_1^{2SLS} - \beta_1) \xrightarrow{d} N(0, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1})$，其中 $Q_{ZX} = \text{plim}(Z'X/n)$ 等。这使得我们可以进行标准的假设检验。

7.3.3 2SLS 估计的标准误计算

2SLS 估计量的方差-协方差矩阵估计为 $\widehat{Var}(\hat{\beta}^{2SLS}) = \hat{\sigma}^2 (X' P_Z X)^{-1}$，其中 $P_Z = Z(Z'Z)^{-1}Z'$，$\hat{\sigma}^2$ 为第二阶段回归残差的方差估计。重要提示：直接使用第二阶段 OLS 回归的标准误是错误的，必须使用上述考虑了 $\hat{x}_i$ 是估计得来的公式。现代计量软件（如 Stata 的 `ivregress`，R 的 `AER` 包中的 `ivreg()`）会自动计算正确的标准误。

7.4 工具变量的检验

7.4.1 相关性检验：弱工具变量问题

当工具变量与内生变量的相关性较弱时，会引发严重的 弱工具变量问题，导致：1. 2SLS 估计量的小样本偏差可能很大。2. 即使在大样本下，估计量的分布也可能严重偏离正态分布，导致推断错误。

检验方法：对于单个内生变量，通常使用第一阶段回归的 **F** 统计量来检验联合显著性 $H_0 : \pi_1 = ... = \pi_L = 0$。- 经验准则（**Staiger & Stock, 1997**）：若第一阶段 F 统计量小于 10，则认为存在弱工具变量问题。- 对于多个内生变量，可使用 Cragg-Donald Wald F 统计量或 Kleibergen-Paap rk Wald F 统计量（适用于异方差或自相关情形）。

7.4.2 外生性检验：过度识别检验

当模型为过度识别（$L > K$）时，可以进行 过度识别检验（如 Sargan 检验或 Hansen J 检验），其原假设为 “所有工具变量均为外生”。- 检验统计量基于 2SLS 残差与工具变量的相关性构造，在原假设下服从 $\chi^2(L-K)$ 分布。- 重要提示：该检验只能检验 “过度识别” 的工具变量是否整体外生。若模型恰好识别（$L = K$），则无法进行此检验。拒绝原假设意味着至少有一个工具变量不满足外生性，但无法指出是哪一个。

7.4.3 内生性检验：是否需要使用 IV？——豪斯曼检验

有时我们不确定解释变量 x_i 是否真的内生。豪斯曼检验可用于比较 OLS 与 IV 估计量，检验 $H_0 : x_i$ 是外生的。- 基本思想：若 x_i 外生，则 OLS 与 IV 都是一致的，但 OLS 更有效；若 x_i 内生，则只有 IV 一致。因此，两者差异过大时拒绝原假设。- 实施方法：一种简便做法是在原模型中加入第一阶段回归的残差 $\hat{v}_i$ 作为额外控制变量，然后检验其系数是否显著。若显著，则拒绝外生性假设，支持使用 IV。

7.5 工具变量法的应用实例与分析

7.5.1 经典案例：教育回报率估计（使用邻近大学作为工具变量）

7.5.2 政策评估案例：工会身份对工资的影响（使用法律环境变化作为工具变量）

7.5.3 实例解读：如何论证工具变量的合理性

7.6 工具变量法的深入议题与扩展

7.6.1 局部平均处理效应理论

IV 估计量并不总是识别总体平均处理效应。在存在异质性处理效应且个体对工具变量的反应不同时，IV 估计的是局部平均处理效应（LATE），即“依从者”（其处理状态会因工具变量而改变的子群体）的平均处理效应。理解 LATE 对于正确解释 IV 估计结果至关重要。

7.6.2 多个内生变量与多个工具变量的情形

模型可以扩展至包含多个内生变量 X 和多个工具变量 Z。识别要求 $L \geq K$（阶条件），且工具变量与内生变量的协方差矩阵满秩（秩条件）。2SLS 估计步骤类似，第一阶段对每个内生变量进行回归，第二阶段将所有内生变量的预测值纳入回归。

7.6.3 控制函数法与 2SLS 的关系

控制函数法是另一种与 2SLS 等价的估计框架。其思路是：将内生变量 x_i 对工具变量回归得到残差 $\hat{v}_i$，然后将 $\hat{v}_i$ 作为控制变量加入原方程进行 OLS 回归。x_i 的系数即为处理效应的一致估计。该方法在非线性模型中尤其有用。

7.6.4 工具变量法的局限性与常见误区

1. 寻找有效工具变量极其困难：外生性条件在现实中难以满足。
2. 弱工具变量危害巨大：可导致比 OLS 更严重的偏差。
3. **LATE** 的解释局限：IV 估计的是特定群体的效应，不一定能推广到总体。
4. 忽视检验：不进行弱工具变量检验和过度识别检验（如果可能）。
5. 错误解释排他性约束：忽视工具变量可能通过其他渠道影响结果。

本章总结

工具变量法是解决内生性问题、识别因果效应的强大但要求严苛的方法。其有效性完全依赖于工具变量必须同时满足的相关性与外生性两个核心假设。本章系统介绍了从模型设定、工具变量选择、两阶段最小二乘估计到一系列诊断检验（弱工具变量检验、过度识别检验、内生性检验）的完整流程。

必须清醒认识到，工具变量的“外生性”是一个基于理论与研究设计的逻辑假设，无法被数据完全证实。因此，方法的成功应用不仅依赖于统计检验，更取决于工具变量选择的合理性与说服力。研究者必须深入理解其估计的局部平均处理效应内涵，并警惕弱工具变量可能带来的严重偏差。

工具变量法在应用计量经济学中占据中心地位，但只是因果推断工具箱中的一种。在实践中，应结合研究问题的具体背景，审慎评估其适用性，并考虑与其他方法（如 DID、RDD 等）相互印证，以得到更可靠的因果结论。

8 倾向得分匹配

本章导读

倾向得分匹配是处理观察性数据中由可观测混杂因素导致的选择偏差的核心方法。当个体是否接受某项处理并非随机分配，而是依赖于可观测的特征变量时，处理组和对照组在这些特征上的不平衡分布会导致简单的均值比较产生偏误。本章将系统介绍倾向得分匹配的基本原理，该方法通过为每个处理组个体寻找特征相似的对照组个体，构造" 近似可比" 的比较组，从而在观察性研究中模拟随机实验的平衡性特征。我们将详细讲解倾向得分的估计、匹配方法的实施、匹配质量的诊断以及前沿的扩展方法。

8.1 选择偏差问题与匹配方法的引入

8.1.1 观察性研究中的选择偏差：从反事实框架看比较组的不可比性

在观察性研究中，个体是否接受处理通常不是随机的，而是基于可观测（有时是不可观测）的特征进行自我选择或被选择。这导致处理组和对照组在潜在结果分布上存在系统性差异，即选择偏差。

反事实框架下的选择偏差：考虑处理效应 $ATT = E[Y_i(1) - Y_i(0)|D_i = 1]$，我们能够观测到的是：

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = ATT + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{选择偏差}}$$

如果选择偏差不为零，简单的组间均值比较不能无偏地估计处理效应。选择偏差的根源在于处理组和对照组在未处理状态下的潜在结果均值不同。

可观测选择偏差：当选择机制完全由可观测特征 X_i 决定时，即满足条件独立性假设：

$$(Y_i(1), Y_i(0)) \perp D_i|X_i$$

在这种情况下，选择偏差完全由可观测特征 X_i 的分布差异导致，可以通过统计方法进行调整。

8.1.2 匹配方法的基本逻辑：构造平衡可观测特征的比较组

匹配方法的核心思想是：对于每个处理组个体，从对照组中寻找一个或多个具有相似可观测特征的个体，用这些匹配个体的结果作为该处理组个体的反事实结果的近似。

精确匹配：最理想的情况是精确匹配，即对于每个处理组个体，找到在 X_i 上完全相同的对照组个体。但这种方法在实际中往往不可行，因为：1. X_i 通常是连续变量或多维变量，精确匹配很难实现 2. 即使找到精确匹配，样本量会急剧减少

近似匹配：在实践中，我们进行近似匹配，即寻找 X_i 上" 相似" 的个体。衡量相似性的方法包括：1. 马氏距离：$(X_i - X_j)'\Sigma^{-1}(X_i - X_j)$ 2. 欧氏距离：$\|X_i - X_j\|$ 3. 倾向得分距离：$|p(X_i) - p(X_j)|$

8.1.3 维度诅咒与倾向得分的提出

当可观测特征 X_i 的维度较高时，直接基于 X_i 进行匹配会遇到维度诅咒问题：随着维度增加，找到足够相似的匹配对象变得越来越困难。

Rosenbaum 和 **Rubin**（**1983**）的突破：Rosenbaum 和 Rubin 证明，如果条件独立性假设成立，那么匹配可以基于一维的倾向得分$p(X_i) = P(D_i = 1|X_i)$ 进行，而无需基于高维的 X_i。这是因为倾向得分具有以下重要性质：1. 平衡性：在给定 $p(X_i)$ 的条件下，D_i 与 X_i 独立 2. 可忽略性：在给定 $p(X_i)$ 的条件下，$(Y_i(1), Y_i(0))$ 与 D_i 独立

因此，基于倾向得分的匹配可以达到与基于 X_i 的匹配相同的平衡效果，同时避免了维度诅咒问题。

8.2 倾向得分的定义、性质与估计

8.2.1 倾向得分的定义：条件处理概率

倾向得分定义为给定可观测特征 X_i 的条件下，个体接受处理的概率：

$$p(X_i) = P(D_i = 1|X_i) = E[D_i|X_i]$$

倾向得分是一个介于 0 和 1 之间的数值，反映了在观察到 X_i 的情况下，个体接受处理的可能性。

倾向得分的解释：- 倾向得分接近 1：具有特征 X_i 的个体几乎肯定接受处理 - 倾向得分接近 0：具有特征 X_i 的个体几乎肯定不接受处理 - 倾向得分在中间范围：接受处理与否有一定不确定性

8.2.2 平衡得分性质与可忽略性假设

平衡得分：平衡得分 $b(X)$ 是 X 的任意函数，使得在给定 $b(X)$ 的条件下，处理分配 D 与特征 X 独立：

$$D \perp X|b(X)$$

定理（**Rosenbaum** 和 **Rubin**，**1983**）：倾向得分 $p(X)$ 是一个平衡得分。事实上，它是最粗糙的平衡得分（即包含信息最少但足以达到平衡）。

可忽略性假设：如果条件独立性假设在 X 条件下成立，那么在任意平衡得分 $b(X)$ 条件下也成立：

$$(Y(1), Y(0)) \perp D|X \Rightarrow (Y(1), Y(0)) \perp D|b(X)$$

特别地，在倾向得分 $p(X)$ 条件下：

$$(Y(1), Y(0)) \perp D|p(X)$$

这意味着，如果两组个体具有相同的倾向得分，那么他们的处理分配可以视为近似随机的，他们的潜在结果分布应该相似。

8.2.3 共同支持域条件

共同支持域是指处理组和对照组的倾向得分分布有重叠的区域。形式化地，共同支持域定义为：

$$S = \{p : 0 < f(p|D = 1) \text{ 且} 0 < f(p|D = 0)\}$$

其中 $f(p|D = d)$ 是倾向得分在组 d 中的密度函数。

实际操作中的共同支持域：我们通常要求：

$$0 < p(X_i) < 1 \quad \text{对于所有} X_i$$

在实践中，我们检查并确保：1. 处理组和对照组的倾向得分分布有显著重叠 2. 没有个体具有极端倾向得分（如接近 0 或 1）

样本修剪：如果存在极端倾向得分的个体，通常的做法是进行样本修剪，即删除倾向得分超出共同范围的个体。虽然这减少了样本量，但可以提高匹配质量，减少外推偏差。

8.2.4 倾向得分的估计：**Logit** 与 **Probit** 模型

倾向得分通常通过参数模型估计，最常用的是 Logit 和 Probit 模型。

Logit 模型：

$$p(X_i) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}$$

其中 β 是系数向量，通过最大似然估计得到。

Probit 模型：

$$p(X_i) = \Phi(X_i'\beta)$$

其中 $\Phi(\cdot)$ 是标准正态累积分布函数。

模型选择与设定：1. 应该包含所有同时影响处理状态和结果变量的协变量 2. 可以考虑高阶项（平方项、交互项）以改善平衡性 3. 不应包含仅影响结果但不影响处理状态的变量（这不会改善平衡性，但可能增加方差）4. 不应包含仅影响处理状态但不影响结果的变量（这可能导致” 坏控制” 问题）

模型诊断：1. 伪 R^2：衡量模型拟合优度，但高伪 R^2 不一定表示好的平衡性 2. 预测概率的分布：检查是否有很多接近 0 或 1 的预测值 3. Hosmer-Lemeshow 检验：检验预测概率与实际处理比例的一致性

8.3 倾向得分匹配的实施步骤

8.3.1 第一步：估计倾向得分与检验重叠性

估计倾向得分：使用 Logit 或 Probit 模型估计每个个体的倾向得分 $\hat{p}(X_i)$。

检验重叠性：1. 绘制处理组和对照组的倾向得分分布图（直方图或核密度图）2. 计算倾向得分的描述性统计量（最小值、最大值、分位数）3. 确定共同支持域：通常删除倾向得分小于对照组最大值且大于处理组最小值的个体，或删除分布两端一定比例（如 1% 或 5%）的个体

重叠性图形示例：

```
# R 代码示例：绘制倾向得分分布图
library(ggplot2)
ggplot(data, aes(x=pscore, fill=treat)) +
  geom_density(alpha=0.5) +
  labs(x="Propensity Score", y="Density", fill="Treatment")
```

8.3.2 第二步：选择匹配方法

常用的匹配方法包括：

最近邻匹配：为每个处理组个体寻找倾向得分最接近的一个或多个对照组个体。- 一对一匹配：每个处理组个体匹配一个最接近的对照组个体 - 一对多匹配：每个处理组个体匹配 k 个最接近的对照组个体 - 有放回 vs. 无放回：有放回允许对照组个体被多次匹配，通常能提高匹配质量

卡尺匹配：要求匹配个体的倾向得分差异不超过预设的阈值（卡尺）。卡尺通常设定为倾向得分标准差的 0.2-0.25 倍。

半径匹配：为每个处理组个体匹配所有在卡尺内的对照组个体，这些对照组个体获得相等的权重。

核匹配：使用所有对照组个体进行匹配，但根据倾向得分差异给予不同权重。权重由核函数 $K(\cdot)$ 决定：

$$w_{ij} = \frac{K\left(\frac{\hat{p}(X_j)-\hat{p}(X_i)}{h}\right)}{\sum_{k:D_k=0} K\left(\frac{\hat{p}(X_k)-\hat{p}(X_i)}{h}\right)}$$

其中 h 是带宽参数。

匹配方法的选择考量：1. 偏差-方差权衡：更严格的匹配（如一对一）可能减少偏差但增加方差 2. 计算复杂性：核匹配通常计算量更大 3. 样本利用率：半径匹配和核匹配利用了更多对照组信息

8.3.3 第三步：匹配后样本平衡性诊断

匹配后需要检验处理组和匹配后的对照组在可观测特征上是否平衡。

标准化差异：对于每个协变量 X_k，计算标准化差异：

$$SD_k = \frac{\bar{X}_{k,treated} - \bar{X}_{k,matched\_control}}{\sqrt{(s^2_{k,treated} + s^2_{k,matched\_control})/2}}$$

其中 $\bar{X}$ 是均值，s^2 是方差。

经验上，匹配后所有协变量的标准化差异应小于 0.1（理想情况下小于 0.05）。

方差比：对于每个协变量 X_k，计算处理组和匹配对照组的方差比：

$$VR_k = \frac{s^2_{k,treated}}{s^2_{k,matched\_control}}$$

匹配后，方差比应接近 1（如 0.8-1.25 之间）。

t 检验：对每个协变量进行两组均值差异的 t 检验。匹配后，这些检验应不再显著（p 值 >0.05）。

8.3.4 第四步：估计处理效应与统计推断

处理效应估计：匹配后，处理效应可以通过比较处理组和匹配对照组的结果均值来估计。对于一对一匹配：

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} \left[Y_i - \frac{1}{M} \sum_{j \in J_M(i)} Y_j \right]$$

其中 $J_M(i)$ 是处理组个体 i 的 M 个匹配的对照组个体集合。

对于核匹配：

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} \left[Y_i - \frac{\sum_{j:D_j=0} w_{ij} Y_j}{\sum_{j:D_j=0} w_{ij}} \right]$$

统计推断：由于倾向得分是估计得到的，且匹配过程引入了相关性，传统的标准误计算可能不正确。常用的方法包括：1. 自助法：对原始样本进行重复抽样，每次重新估计倾向得分并进行匹配 2. **Abadie-Imbens** 标准误：考虑了匹配不确定性的解析标准误 3. 稳健标准误公式：基于匹配后样本计算的稳健标准误

自助法步骤：1. 从原始样本中有放回地抽取一个自助样本 2. 在自助样本中重新估计倾向得分 3. 基于新的倾向得分重新进行匹配 4. 计算处理效应估计值 5. 重复 B 次（如 500 次），得到处理效应的自助分布 6. 基于自助分布计算标准误和置信区间

8.4 匹配质量的诊断与敏感性分析

8.4.1 平衡性检验：标准化差异、**t** 检验与方差比

平衡性检验表：研究报告应包含匹配前后的平衡性检验表，展示：1. 每个协变量的处理组和对照组均值 2. 标准化差异（匹配前后）3. 方差比（匹配前后）4. t 检验的 p 值（匹配前后）

可视化平衡性改进：可以绘制匹配前后标准化差异的图形，直观展示平衡性的改善。

经验准则：- 所有协变量的匹配后标准化差异应 <0.1 - 至少 90% 的协变量的标准化差异应 <0.05 - 方差比应在 0.8-1.25 之间 - t 检验的 p 值应 >0.05

8.4.2 倾向得分分布重叠图

分布重叠图：绘制处理组和对照组（匹配前后）的倾向得分分布图，直观展示：1. 匹配前两组分布的差异 2. 匹配后两组分布的相似性 3. 共同支持域的范围

分位数-分位数图：绘制处理组和对照组倾向得分分位数的 Q-Q 图，如果点在 45 度线附近，说明两组分布相似。

8.4.3 敏感性分析：评估未观测混杂因素的影响

倾向得分匹配只能控制可观测的混杂因素。如果存在未观测的混杂因素，估计结果可能仍有偏误。敏感性分析用于评估这种可能性。

Rosenbaum 界限方法：该方法评估需要多大的未观测混杂因素才能推翻研究结论。设 Γ 表示未观测混杂因素的最大影响，定义为两个具有相同可观测特征的个体接受处理概率的最大比值。

敏感性分析步骤：1. 对于不同的 Γ 值（如 1.5, 2.0, 2.5），计算处理效应的置信区间 2. 确定使结论变得不显著的 Γ 值 3. 评估这样的 Γ 值是否合理

经验解释：如果较小的 Γ（如 $\Gamma = 1.5$）就能使结论变得不显著，说明结果对未观测混杂因素敏感。如果较大的 Γ（如 $\Gamma = 3.0$）才能使结论不显著，说明结果相对稳健。

8.5 倾向得分方法的扩展

8.5.1 逆概率加权法

逆概率加权法不进行匹配，而是通过加权使处理组和对照组在特征分布上平衡。

ATE 的 **IPW** 估计量：

$$\hat{\tau}_{ATE}^{IPW} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right]$$

ATT 的 **IPW** 估计量：

$$\hat{\tau}_{ATT}^{IPW} = \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i Y_i - \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N}\frac{D_i \hat{p}(X_i)(1-D_i)Y_i}{1-\hat{p}(X_i)}$$

优点与缺点：- 优点：使用了所有样本，比匹配更有效率 - 缺点：对倾向得分模型设定敏感，特别是当倾向得分接近 0 或 1 时，权重会变得很大，导致估计不稳定

8.5.2 双重稳健估计

双重稳健估计结合了倾向得分加权和结果回归，只要其中一个模型正确设定，就能得到一致估计。

DR 估计量：

$$\hat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i(Y_i - \hat{m}_1(X_i))}{\hat{p}(X_i)} + \hat{m}_1(X_i)\right] - \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-D_i)(Y_i - \hat{m}_0(X_i))}{1-\hat{p}(X_i)} + \hat{m}_0(X_i)\right]$$

其中 $\hat{m}_1(X_i) = E[Y_i|D_i = 1, X_i]$ 和 $\hat{m}_0(X_i) = E[Y_i|D_i = 0, X_i]$ 是通过回归模型估计的。

优点：对模型误设更稳健，且通常比单独的匹配或 IPW 更有效。

8.5.3 广义倾向得分（处理变量连续或多值）

当处理变量是连续或多值时，可以使用广义倾向得分方法。

连续处理情况：对于连续处理 T_i，广义倾向得分定义为处理变量的条件密度：

$$r(t, X_i) = f_{T|X}(t|X_i)$$

剂量反应函数：剂量反应函数 $\mu(t) = E[Y_i(t)]$ 可以通过逆概率加权估计：

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{K_h(T_i - t)Y_i}{r(t, X_i)}$$

其中 $K_h(\cdot)$ 是核函数，h 是带宽。

8.5.4 边际结构模型简介

边际结构模型通过逆概率加权估计边际处理效应，特别适用于处理随时间变化的情况。

时变处理：设 T 个时期的处理序列为 $\bar{D} = (D_1, ..., D_T)$，协变量序列为 $\bar{X} = (X_1, ..., X_T)$。每个时期的权重为：

$$w_i = \prod_{t=1}^{T}\frac{1}{P(D_t = d_t|\bar{D}_{t-1} = \bar{d}_{t-1}, \bar{X}_t = \bar{x}_t)}$$

MSM 模型：估计边际结构模型：

$$E[Y(\bar{d})] = \beta_0 + \beta_1 \text{cum}(\bar{d})$$

其中 $\text{cum}(\bar{d})$ 是累计处理量。

优点：能处理时变混杂因素，但需要正确设定每个时期的倾向得分模型。

8.6 应用实例与操作实践

8.6.1 经典案例：劳动力市场培训项目评估

8.6.2 软件操作：Stata (`psmatch2`, `teffects`) 与 R (`MatchIt`, `WeightIt`)

8.6.3 研究报告规范与常见误区

本章总结

倾向得分匹配通过模拟随机实验的逻辑，在观察性研究中为处理组个体寻找特征相似的控制组个体，从而减少可观测混杂因素的影响。本章系统介绍了从倾向得分估计、匹配方法选择、平衡性检验到因果效应估计的完整流程。

需要强调的是，PSM 只能控制可观测的混杂变量，其有效性依赖于强可忽略性假设。对于不可观测的混杂因素，PSM 无法解决，需要借助其他方法（如工具变量、固定效应模型等）或进行敏感性分析。

成功的 PSM 应用不仅依赖于恰当的统计方法，更取决于：1. 对研究问题的深入理解 2. 对相关混杂因素的全面测量 3. 对匹配结果的严谨诊断检验 4. 对模型假设和局限性的清晰认识

倾向得分方法已发展出多种扩展形式，包括逆概率加权、双重稳健估计、广义倾向得分和边际结构模型，这些方法丰富了观察性研究中因果推断的工具箱。在实际应用中，研究者应根据具体研究问题和数据特征选择合适的方法，并通过敏感性分析评估估计结果的稳健性。

最后，无论使用哪种基于倾向得分的方法，都必须牢记：这些方法只能解决由可观测变量导致的选择偏差。对于因果推断，没有任何统计方法可以完全替代良好的研究设计和严谨的理论思考。

9 双重差分法

第 9 章双重差分法

本章导读

双重差分法是现代计量经济学中进行政策评估和因果推断的核心方法之一。当研究者无法进行随机实验，但可以观察到政策实施前后的变化时，DID 通过比较处理组和对照组在政策实施前后的差异变化，来识别政策的因果效应。本章将系统介绍双重差分法的基本思想、核心假设、模型设定、统计推断方法以及前沿扩展。通过学习，读者将掌握使用 DID 进行政策评估的完整框架，理解其应用前提与常见陷阱，为实际研究提供方法基础。

9.1 双重差分法的基本思想与模型设定

9.1.1 双重差分法的直观逻辑与基本框架

双重差分法（Difference-in-Differences, DID）是一种准实验设计方法，用于评估政策或处理的效果。其基本思想是通过比较处理组和对照组在政策实施前后结果变量的变化差异，来识别政策的净效应。

设 Y_{it} 为个体 i 在时期 t 的结果变量，D_{it} 为处理状态变量（$D_{it}=1$ 表示接受处理，$D_{it}=0$ 表示未接受处理）。假设政策在时期 t^* 实施，且实施后处理组持续接受处理。

双重差分估计量的直观计算为：

$$\hat{\tau}_{DID} = (\bar{Y}_{1,post} - \bar{Y}_{1,pre}) - (\bar{Y}_{0,post} - \bar{Y}_{0,pre})$$

其中：- $\bar{Y}_{1,post}$：处理组在政策后的平均结果 - $\bar{Y}_{1,pre}$：处理组在政策前的平均结果 - $\bar{Y}_{0,post}$：对照组在政策后的平均结果 - $\bar{Y}_{0,pre}$：对照组在政策前的平均结果

9.1.2 经典双重差分模型（**Two-way Fixed Effects Model**）

经典 DID 模型通常表示为双向固定效应模型：

$$Y_{it} = \alpha + \beta D_{it} + \gamma_i + \lambda_t + \epsilon_{it}$$

其中：- α：截距项 - D_{it}：处理状态变量（处理组在政策实施后取 1，否则取 0）- β：核心参数，表示处理效应 - γ_i：个体固定效应，控制个体不随时间变化的特征 - λ_t：时间固定效应，控制时间趋势和共同冲击 - ϵ_{it}：随机误差项

更一般地，对于平衡面板数据，DID 模型可以写为：

$$Y_{it} = \beta_0 + \beta_1 Treat_i \times Post_t + \beta_2 Treat_i + \beta_3 Post_t + \epsilon_{it}$$

其中：- $Treat_i$：个体是否属于处理组的虚拟变量（$Treat_i = 1$ 表示处理组，$Treat_i = 0$ 表示对照组）- $Post_t$：时间是否在政策后的虚拟变量（$Post_t = 1$ 表示政策后，$Post_t = 0$ 表示政策前）

此时，处理效应 β_1 即为双重差分估计量。

9.1.3 处理效应的图形化展示与直观理解

DID 的直观性部分来自于其图形化展示。一个典型的 DID 图形包括：1. 横轴：时间（政策实施时点标记）2. 纵轴：结果变量的均值 3. 两条线：分别代表处理组和对照组的时间趋势

在平行趋势假设下，政策实施前两条线应基本平行，政策实施后处理组的线应出现” 跳跃” 或趋势变化，而对照组的线应继续原有趋势。

图形化展示不仅有助于直观理解 DID 的逻辑，也是检验平行趋势假设的重要工具。

9.1.4 DID 与简单前后比较、截面比较的差异

DID 方法相对于简单方法的优势在于：

相对于简单前后比较：- 简单前后比较：$\hat{\tau}_{pre-post} = \bar{Y}_{1,post} - \bar{Y}_{1,pre}$ - 问题：无法区分政策效应与其他时间趋势（如宏观经济变化、技术进步等）

相对于截面比较：- 简单截面比较：$\hat{\tau}_{cross-section} = \bar{Y}_{1,post} - \bar{Y}_{0,post}$ - 问题：无法控制处理组和对照组的事前差异

DID 通过双重差分，既控制了时间趋势，也控制了两组的事前差异，从而更准确地识别政策效应。

9.2 平行趋势假设：DID 的识别基石

9.2.1 平行趋势假设的核心内涵与经济学解释

平行趋势假设是 DID 方法成立的核心识别假设。其正式表述为：

在缺乏政策干预的情况下，处理组和对照组的潜在结果随时间变化的趋势相同。即：

$$E[Y_{it}(0)|Treat_i = 1, t] - E[Y_{it}(0)|Treat_i = 1, t-1] = E[Y_{it}(0)|Treat_i = 0, t] - E[Y_{it}(0)|Treat_i = 0, t-1]$$

其中 $Y_{it}(0)$ 表示未接受处理时的潜在结果。

经济学解释：平行趋势假设要求，如果没有政策干预，处理组和对照组的结果变量会按照相同的趋势演变。这意味着两组除了是否接受政策干预外，在其他所有影响结果变量的因素上具有相似的时间趋势。

平行趋势假设的合理性取决于：1. 对照组的选取是否恰当 2. 是否有未观测到的时变混杂因素同时影响处理状态和结果变量

9.2.2 平行趋势的检验方法

事件研究图（**Event Study Plot**）

事件研究图是最直观的平行趋势检验方法。通过估计以下模型：

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k=-K}^{-1} \beta_k \cdot Treat_i \times 1(t = \text{政策前}k\ ) + \sum_{k=0}^{L} \beta_k \cdot Treat_i \times 1(t = \text{政策后}k\ ) + \epsilon_{it}$$

其中，将政策实施前的 K 期和政策实施后的 L 期分别与处理组虚拟变量交互。

检验方法：1. 政策前的交互项系数 β_k（$k < 0$）应统计上不显著（与 0 无差异）2. 政策后的交互项系数 β_k（$k \geq 0$）应显示政策效应

图形上，政策前的点应在 0 附近随机波动，政策后的点应显示政策效应。

动态效应模型

动态效应模型是事件研究图的参数化形式：

$$Y_{it} = \alpha_i + \lambda_t + \sum_{m=-M}^{M} \beta_m \cdot Treat_i \times Post_{t,m} + \epsilon_{it}$$

其中 $Post_{t,m}$ 是表示与政策时点相对距离的虚拟变量。

9.2.3 平行趋势不满足时的后果与应对策略

平行趋势不满足的后果：如果平行趋势假设不成立，DID 估计量将存在偏差：

$$\hat{\tau}_{DID} \xrightarrow{p} \tau + \text{趋势差异}$$

其中” 趋势差异” 是处理组和对照组在无政策情况下的趋势差异。

应对策略：1. 调整对照组：寻找更合适的对照组 2. 合成控制法：构造合成的对照组 3. 匹配 **DID**：先进行匹配，再进行 DID 分析 4. 三重差分法：引入第三个差分维度 5. 控制时间趋势：加入组别特定的时间趋势

9.2.4 替代性假设与敏感性分析

当平行趋势假设难以验证时，可以考虑以下替代方法：

共同趋势假设：要求处理组和对照组的趋势差异是固定的，不随时间变化。这比平行趋势假设稍弱。

敏感性分析：通过检验 DID 估计结果对平行趋势假设的敏感性，评估结论的稳健性。常见方法包括：1. 安慰剂检验：将政策时点提前，检验” 伪政策” 效应 2. 置换检验：随机分配处理状态，检验估计效应分布 3. 控制更多时变协变量

9.3 双重差分法的估计、推断与稳健标准误

9.3.1 DID 模型的参数估计与处理效应解释

DID 模型通常通过最小二乘法（OLS）进行估计。对于双向固定效应模型：

$$Y_{it} = \beta D_{it} + \gamma_i + \lambda_t + \epsilon_{it}$$

可以通过以下方法估计：1. **LSDV** 法：加入个体和时间虚拟变量 2. 组内估计法：先对个体和时间去均值，再回归 3. **Stata** 命令：`xtreg y d, fe` 或 `reghdfe y d, absorb(i t)`

处理效应的解释：- $\hat{\beta}$ 表示平均处理效应（ATE）- 如果处理效应异质，$\hat{\beta}$ 表示处理组平均处理效应（ATT）- 动态模型中，$\hat{\beta}_m$ 表示政策实施后第 m 期的处理效应

9.3.2 聚类稳健标准误：为何在 DID 中至关重要

在 DID 分析中，误差项通常存在：1. 组内自相关：同一组别内不同时期的误差相关 2. 组间异方差：不同组别的误差方差不同

这些问题会导致传统标准误低估真实不确定性，使得统计推断失效。

聚类稳健标准误：将标准误聚类在组别层面（如个体层面或更高层级），可以解决组内自相关问题。聚类稳健方差估计量为：

$$\widehat{Var}_{cluster}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \hat{\epsilon}_g \hat{\epsilon}_g' X_g \right) (X'X)^{-1}$$

其中 g 表示聚类，G 为聚类总数。

聚类层级选择：1. 个体层面聚类：处理组内自相关问题 2. 更高层级聚类：如县级、省级，处理更广泛的相关性 3. 双重聚类：同时考虑时间和个体维度

经验法则：- 至少聚类在处理组层面 - 当聚类数量较少时（如 <50），使用小样本调整（如 Wild bootstrap）

9.3.3 Conley 估计、bootstrap 等替代推断方法

当聚类数量较少或存在更复杂的误差结构时，可以考虑以下方法：

Conley 标准误：处理空间自相关的标准误估计方法，适用于地理相邻单元可能存在相关性的情况。

Bootstrap 方法：通过重复抽样计算标准误，特别适用于小样本情况：1. 对残差进行重抽样（残差 bootstrap）2. 对聚类进行重抽样（聚类 bootstrap）3. Wild bootstrap：特别适用于小样本和异方差情况

排列检验：通过随机置换处理状态，构建处理效应的经验分布，进行非参数推断。

9.3.4 处理组样本量较小时的统计推断问题

当处理组数量较少时（如政策只影响少数几个地区），传统推断方法可能失效：

问题：1. 聚类稳健标准误可能严重低估 2. 中心极限定理可能不适用 3. 统计检验势较低

解决方案：1. **Wild cluster bootstrap**：特别适用于少量聚类的情况 2. 随机化推断：基于随机分配处理的假设进行推断 3. 贝叶斯方法：引入先验信息 4. 合成控制法：特别适用于单一处理组的情况

9.4 双重差分法的扩展模型

9.4.1 多期 **DID** 与异质性处理效应（交错 **DID**）问题

在实际应用中，政策往往在不同时间点对不同个体实施，形成交错 DID 设计。

交错 **DID** 模型：设个体 i 在时期 G_i 首次接受处理，之后持续处于处理状态。传统双向固定效应模型为：

$$Y_{it} = \beta D_{it} + \gamma_i + \lambda_t + \epsilon_{it}$$

其中 $D_{it} = 1(t \geq G_i)$。

异质性处理效应问题：近期研究（如 Goodman-Bacon，2021；Sun & Abraham，2021）发现，当处理效应异质时（不同队列、不同时期效应不同），传统双向固定效应估计量可能是有偏的，它是不同”2×2 DID” 比较的加权平均。

9.4.2 异质性处理效应的识别

Sun & Abraham（2021）方法

通过估计事件研究模型，使用从未接受处理的组作为对照组：

$$Y_{it} = \gamma_i + \lambda_t + \sum_g \sum_{e \neq -1} \beta_{ge} \cdot 1(G_i = g) \cdot 1(E_{it} = e) + \epsilon_{it}$$

其中 $E_{it} = t - G_i$ 表示相对事件时间，以政策前一期（$e = -1$）为基准。

Callaway & Sant'Anna（2021）方法

基于反事实框架，使用从未接受处理的组或尚未接受处理的组作为对照组，估计组别-时期平均处理效应：

$$ATT(g,t) = E[Y_t - Y_{g-1}|G = g] - E[Y_t - Y_{g-1}|G > t]$$

Borusyak et al.（2021）方法

使用插补法：先用未处理样本估计反事实结果，再计算处理效应。

9.4.3 连续型处理与强度 DID

当处理强度在不同个体或不同时间不同时，可以使用强度 DID：

模型设定：

$$Y_{it} = \beta_0 + \beta_1 Intensity_i \times Post_t + \beta_2 Intensity_i + \beta_3 Post_t + \epsilon_{it}$$

其中 $Intensity_i$ 表示处理强度。

识别假设：除了平行趋势，还需要满足处理强度的外生性假设：处理强度与潜在结果变化无关。

9.4.4 三重差分法：排除混杂政策的影响

当存在同时影响处理组和对照组的混杂政策时，可以使用三重差分法：

模型设定：

$$Y_{ijt} = \beta_0 + \beta_1 Treat_i \times Post_t \times Group_j + \text{所有低阶交互项和主效应} + \epsilon_{ijt}$$

其中 $Group_j$ 表示第二个分组维度（如行业、地区类型等）。

识别假设：要求混杂政策对第二个分组维度的影响相同，即" 共同趋势中的共同趋势"。

应用场景：1. 全国性政策但部分群体豁免 2. 多个政策同时实施 3. 存在同时影响处理组和对照组的外部冲击

9.5 双重差分法的常见陷阱、批评与最新进展

9.5.1 "坏对照组**"** 问题与处理组选择偏误

坏对照组问题：对照组可能受到政策间接影响，或与处理组在政策实施后的趋势不再可比。

处理组选择偏误：处理组的选择可能非随机，与结果变量的趋势相关。解决方法：1. 使用匹配方法选择对照组 2. 使用合成控制法 3. 检验和处理选择方程

9.5.2 政策内生性与预期效应

政策内生性：政策实施可能基于前期趋势（如"Ashenfelter's dip"），导致估计偏误。

预期效应：政策实施前，个体可能基于预期调整行为，导致政策前效应。

解决方案：1. 检验政策前趋势 2. 使用更早的基线期 3. 考虑动态模型

9.5.3 最近的方法论争论与反思

"双向固定效应**"** 在交错 **DID** 中的偏误

Goodman-Bacon（2021）证明，当处理效应异质时，传统双向固定效应估计量是不同"2×2 DID" 比较的加权平均，可能产生偏误。

偏误来源：1. 早期处理组 vs 晚期处理组比较 2. 已处理组 vs 尚未处理组比较 3. 处理组 vs 从未处理组比较

动态偏误与静态偏误

de Chaisemartin & d'Haultfoeuille（2020）区分了：1. 动态偏误：来自处理组内部的比较 2. 静态偏误：来自处理组与对照组的比较

解决方案

1. 使用 Sun & Abraham（2021）、Callaway & Sant'Anna（2021）等方法
2. 使用 Borusyak et al.（2021）的插补法
3. 使用 de Chaisemartin & d'Haultfoeuille（2020）的估计量

9.5.4 合成控制法、匹配 DID 等混合方法

合成控制法：适用于处理组数量极少的情况，通过加权组合对照组单元构造” 合成对照组”。

匹配 **DID**：先进行倾向得分匹配或协变量匹配，再进行 DID 分析，以改善平行趋势假设。

双重稳健 **DID**：结合匹配和回归调整，提高估计的稳健性。

9.6 DID 的应用实例与 Stata/R 操作

9.6.1 经典案例研究：最低工资对就业的影响（Card and Krueger，1994）

9.6.2 中国语境下的典型案例解读（如” 四万亿” 投资、房产限购政策评估）

9.6.3 DID 在 Stata/R 中的实现步骤与代码示例（`reghdfe`、`did`、`did2s` 等命令）

本章总结

双重差分法以其清晰的逻辑和易于实现的特点，成为应用微观计量中最受欢迎的政策评估工具之一。本章系统介绍了 DID 的基本思想、核心假设、模型设定、统计推断方法以及前沿扩展。

DID 的有效性完全依赖于平行趋势假设的成立，而这一假设本质上不可直接验证，只能通过事前趋势检验和丰富的稳健性分析来提供支持证据。近年来，针对异质性处理效应和交错 DID 的讨论推动了该方法论的快速发展，研究者现在有更多工具来处理传统双向固定效应模型可能存在的偏误。

在使用 DID 时，研究者必须注意：1. 谨慎选择对照组，避免” 坏对照组” 问题 2. 正确计算聚类稳健标准误，考虑误差结构 3. 进行充分的平行趋势检验和稳健性分析 4. 对于交错 DID 设计，考虑使用最新的异质性处理效应估计方法 5. 结合其他方法（如合成控制法、匹配方法）提高估计可信度

通过深入理解 DID 的假设前提和最新方法论进展，研究者可以更准确地评估政策效应，为实证研究提供可靠的方法基础。

10 断点回归

本章导读

断点回归是一种利用制度或规则中存在的”断点”来识别因果效应的准实验方法。当个体是否接受处理（或处理强度）取决于某个连续变量是否超过某个确定的阈值时，在阈值附近的小邻域内，个体的分配可被视为近似随机。本章将系统介绍断点回归的设计思想、核心假设、模型设定、有效性检验方法以及实际应用中的关键问题。通过学习，读者将理解 RDD”局部随机实验”的独特逻辑，掌握执行与评估一项断点回归分析的完整流程。

10.1 断点回归的基本思想与设计逻辑

10.1.1 从”局部随机实验”的直观理解到正式定义

断点回归设计是一种准实验设计，它利用处理分配规则中的不连续性来识别因果效应。其核心思想是：当处理分配基于一个连续变量是否超过某个阈值时，在阈值附近的个体可以视为近似随机分配到处理组或对照组。

形式化定义：设 X_i 为驱动变量（running variable 或 forcing variable），c 为阈值。处理状态 D_i 由以下规则决定：

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

在精确断点回归中，该分配规则是确定性的；在模糊断点回归中，该规则是概率性的。

10.1.2 精确断点回归与模糊断点回归的区分

精确断点回归：处理状态 D_i 是驱动变量 X_i 的确定性函数：

$$D_i = \mathbb{1}(X_i \geq c)$$

其中 $\mathbb{1}(\cdot)$ 是指示函数。在精确 RDD 中，所有个体在阈值两侧的处理状态完全由驱动变量决定。

模糊断点回归：处理状态 D_i 不是完全确定的，而是概率性的：

$$P(D_i = 1|X_i) = \begin{cases} p_1(X_i) & \text{if } X_i \geq c \\ p_0(X_i) & \text{if } X_i < c \end{cases}$$

其中 $p_1(c) \neq p_0(c)$，即在阈值处处理概率存在跳跃。

10.1.3 驱动变量、阈值与处理变量的关系

在断点回归设计中，三个核心要素的关系如下：

1. 驱动变量X_i：连续变量，决定个体是否接受处理
2. 阈值c：驱动变量的临界值，决定处理分配的断点
3. 处理变量D_i：二值或多值变量，表示个体是否接受处理

在精确 RDD 中，当 $X_i \geq c$ 时，$D_i = 1$；当 $X_i < c$ 时，$D_i = 0$。结果变量 Y_i 可以表示为：

$$Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]D_i$$

其中 $Y_i(1)$ 和 $Y_i(0)$ 是潜在结果。

10.1.4 RDD 与其他因果推断方法（**IV**，**DID**）的比较

方法	识别假设	适用场景	估计效应
断点回归	连续性假设	存在明确分配规则和阈值	局部平均处理效应
工具变量	外生性和相关性	存在有效工具变量	依从者平均处理效应
双重差分	平行趋势假设	政策实施前后有面板数据	平均处理效应

RDD 的优势：1. 识别假设相对直观和可检验 2. 在阈值附近近似随机实验 3. 图形展示直观明了

RDD 的局限：1. 估计的是局部平均处理效应（LATE）2. 需要足够样本量在阈值附近 3. 对外部有效性有限制

10.2 精确断点回归：识别与估计

10.2.1 潜在结果框架下的识别假设：连续性假设

精确断点回归的识别依赖于连续性假设：

连续性假设：潜在结果函数 $E[Y_i(0)|X_i = x]$ 和 $E[Y_i(1)|X_i = x]$ 在 $x = c$ 处连续。即：

$$\lim_{x\uparrow c} E[Y_i(0)|X_i = x] = \lim_{x\downarrow c} E[Y_i(0)|X_i = x]$$

$$\lim_{x\uparrow c} E[Y_i(1)|X_i = x] = \lim_{x\downarrow c} E[Y_i(1)|X_i = x]$$

在连续性假设下，阈值处结果变量的跳跃只能归因于处理效应：

$$\tau_{SRD} = \lim_{x\downarrow c} E[Y_i|X_i = x] - \lim_{x\uparrow c} E[Y_i|X_i = x]$$

10.2.2 局部多项式回归：带宽选择与核函数

断点回归通常使用局部多项式回归进行估计。考虑以下回归模型：

$$Y_i = \alpha + \tau D_i + f(X_i - c) + \epsilon_i$$

其中 $f(\cdot)$ 是驱动变量的连续函数。

常用的估计方法是局部线性回归：

$$\min_{\alpha,\tau,\beta_l,\beta_r} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) [Y_i - \alpha - \tau D_i - \beta_l(X_i - c) \cdot \mathbb{1}(X_i < c) - \beta_r(X_i - c) \cdot \mathbb{1}(X_i \geq c)]^2$$

其中 $K(\cdot)$ 是核函数，h 是带宽。

常用核函数：1. 三角核：$K(u) = (1-|u|)\mathbb{1}(|u| \leq 1)$ 2. 均匀核：$K(u) = \frac{1}{2}\mathbb{1}(|u| \leq 1)$ 3. Epanechnikov 核：$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}(|u| \leq 1)$

10.2.3 参数化与非参数化估计策略

参数化方法：使用全样本，在回归中包含驱动变量的高阶多项式：

$$Y_i = \alpha + \tau D_i + \sum_{j=1}^{p} \beta_j (X_i - c)^j + \sum_{j=1}^{p} \gamma_j (X_i - c)^j \cdot D_i + \epsilon_i$$

其中 p 是多项式阶数。

非参数化方法：只使用带宽 h 内的样本，进行局部多项式回归。这是目前更常用的方法，因为它避免了多项式回归可能带来的外推问题。

10.2.4 最优带宽的选择方法

带宽选择是断点回归中的关键问题。带宽太小会导致估计方差大，带宽太大会导致估计偏差大。

IMSE 最优带宽：Imbens 和 Kalyanaraman（2012）提出了基于均方误差最小化的带宽选择方法：

$$h_{IK} = C_{IK} \cdot n^{-1/5}$$

其中 C_{IK} 是一个依赖于数据特征的常数。

交叉验证带宽：通过交叉验证选择使预测误差最小的带宽：

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \hat{Y}_{(-i)}(X_i) \right]^2$$

其中 $\hat{Y}_{(-i)}(X_i)$ 是剔除第 i 个观测后得到的预测值。

经验法则：在实际应用中，通常报告多个带宽下的估计结果，以检验估计的稳健性。

10.2.5 断点处的平均处理效应估计与统计推断

断点处的平均处理效应估计量为：

$$\hat{\tau}_{SRD} = \hat{\mu}_+ - \hat{\mu}_-$$

其中 $\hat{\mu}_+ = \lim_{x \downarrow c} \hat{E}[Y_i | X_i = x]$，$\hat{\mu}_- = \lim_{x \uparrow c} \hat{E}[Y_i | X_i = x]$。

统计推断：使用局部线性回归时，可以构造 t 统计量：

$$t = \frac{\hat{\tau}_{SRD}}{SE(\hat{\tau}_{SRD})}$$

其中标准误 $SE(\hat{\tau}_{SRD})$ 通常通过稳健标准误或自助法计算。

置信区间：可以使用常规方法构建置信区间：

$$CI_{1-\alpha} = \hat{\tau}_{SRD} \pm z_{1-\alpha/2} \cdot SE(\hat{\tau}_{SRD})$$

或使用自助法构建置信区间。

10.3 模糊断点回归：工具变量视角

10.3.1 作为工具变量法的 RDD：处理依从的非完全性

在模糊断点回归中，驱动变量超过阈值并不保证个体一定接受处理，而只是改变了接受处理的概率。因此，模糊 RDD 可以视为一种工具变量方法。

设定：设 $Z_i = \mathbb{1}(X_i \geq c)$ 为工具变量，D_i 为内生处理变量，Y_i 为结果变量。

第一阶段回归：

$$D_i = \pi_0 + \pi_1 Z_i + g(X_i - c) + v_i$$

其中 $g(\cdot)$ 是驱动变量的连续函数。

第二阶段回归：

$$Y_i = \alpha + \tau D_i + f(X_i - c) + \epsilon_i$$

10.3.2 两阶段最小二乘估计框架

模糊断点回归可以通过两阶段最小二乘法估计：

第一阶段：估计处理概率

$$\hat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i + \hat{g}(X_i - c)$$

第二阶段：使用 $\hat{D}_i$ 作为工具进行回归

$$Y_i = \alpha + \tau \hat{D}_i + f(X_i - c) + \epsilon_i$$

估计量 $\hat{\tau}_{FRD}$ 可以表示为：

$$\hat{\tau}_{FRD} = \frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x]}$$

10.3.3 识别假设：驱动变量的排他性与相关性

模糊断点回归的识别需要以下假设：

相关性假设：处理概率在阈值处存在跳跃：

$$\lim_{x \downarrow c} P(D_i = 1 | X_i = x) \neq \lim_{x \uparrow c} P(D_i = 1 | X_i = x)$$

排他性约束：驱动变量 X_i 只能通过处理状态 D_i 影响结果变量 Y_i。形式化地：

$$(Y_i(1), Y_i(0)) \perp Z_i | X_i$$

在给定驱动变量 X_i 的条件下，工具变量 Z_i 与潜在结果独立。

单调性假设：对于所有个体，有：

$$D_i(1) \geq D_i(0)$$

其中 $D_i(z)$ 表示当 $Z_i = z$ 时的潜在处理状态。

10.3.4 一阶段关系与工具变量有效性检验

在模糊断点回归中，需要检验工具变量的有效性：

一阶段关系检验：检验处理概率在阈值处是否存在显著跳跃。可以估计：

$$D_i = \pi_0 + \pi_1 Z_i + \sum_{j=1}^{p} \beta_j (X_i - c)^j + \sum_{j=1}^{p} \gamma_j (X_i - c)^j \cdot Z_i + v_i$$

并检验 $H_0 : \pi_1 = 0$。

F 统计量检验：计算第一阶段的 F 统计量，检验工具变量与内生变量的相关性。经验上，F 统计量应大于 10。

10.4 有效性检验与稳健性分析

10.4.1 连续性假设的间接检验：协变量在断点处的平衡性检验

由于潜在结果无法直接观测，我们通过检验可观测协变量在断点处的连续性来间接检验连续性假设。

方法：对于每个协变量 W_i，检验：

$$\lim_{x \downarrow c} E[W_i | X_i = x] = \lim_{x \uparrow c} E[W_i | X_i = x]$$

检验步骤：1. 将每个协变量作为结果变量进行断点回归 2. 检验断点处协变量是否存在显著跳跃 3. 如果多个协变量都不存在显著跳跃，则增强了对连续性假设的信心

图形展示：绘制协变量在驱动变量上的散点图和局部平滑线，观察在阈值处是否连续。

10.4.2 驱动变量的操纵检验：**McCrary** 密度检验

如果个体能够精确操纵驱动变量，可能会导致断点两侧的个体不可比。McCrary（2008）提出了检验驱动变量密度连续性的方法。

检验原理：如果个体不能操纵驱动变量，则驱动变量的密度函数在阈值处应该连续。如果存在操纵，则密度函数在阈值处会出现跳跃。

检验步骤：1. 将驱动变量的支持划分为若干区间 2. 计算每个区间的观测频数 3. 在阈值两侧分别拟合密度函数 4. 检验阈值处密度是否连续

原假设：驱动变量的密度函数在阈值处连续。

10.4.3 结果变量与协变量的预趋势检验

除了检验协变量在阈值处的平衡性，还可以检验协变量与驱动变量关系在阈值处的连续性。

方法：对于协变量 W_i，检验模型：

$$W_i = \alpha + \tau \mathbb{1}(X_i \geq c) + f(X_i - c) + \epsilon_i$$

中的 τ 是否显著不为零。

如果 τ 不显著，说明该协变量在阈值处没有跳跃，支持连续性假设。

10.4.4 带宽与函数形式的敏感性分析

断点回归的结果可能对带宽选择和函数形式设定敏感，因此需要进行敏感性分析。

带宽敏感性分析：1. 使用不同带宽（如 $0.5h_{opt}$、h_{opt}、$1.5h_{opt}$）进行估计 2. 比较不同带宽下的估计结果和标准误 3. 如果估计结果在不同带宽下稳定，则增强结果的可靠性

函数形式敏感性分析：1. 使用不同阶数的多项式（如线性、二次、三次）进行估计 2. 比较不同函数形式下的估计结果 3. 使用非参数方法作为基准

10.4.5 安慰剂检验：伪断点与伪结果检验

安慰剂检验是检验断点回归结果稳健性的重要方法。

伪断点检验：在真实的阈值之外选择其他点作为伪断点，检验在这些点上是否存在显著效应。如果在伪断点处也发现了显著效应，则可能表明观测到的效应不是由处理引起的。

伪结果检验：使用理论上不应受处理影响的变量作为结果变量进行断点回归。如果发现了显著效应，则可能表明存在混杂因素或模型设定有问题。

10.5 扩展议题与前沿讨论

10.5.1 多变量与多维断点回归

当处理分配基于多个驱动变量时，需要使用多维断点回归。

设定：设 X_{1i} 和 X_{2i} 为两个驱动变量，阈值为 (c_1, c_2)。处理状态为：

$$D_i = \mathbb{1}(X_{1i} \geq c_1, X_{2i} \geq c_2)$$

识别挑战：在多维情况下，需要定义" 接近" 阈值的邻域，并估计该邻域内的处理效应。

估计方法：1. 马氏距离法：使用 $d_i = \sqrt{(X_{1i} - c_1)^2 + (X_{2i} - c_2)^2}$ 作为单维驱动变量 2. 非参数方法：在高维空间中进行局部回归

10.5.2 分位数处理效应断点回归

标准断点回归估计的是平均处理效应，但有时我们关心处理效应在整个分布上的异质性。

分位数处理效应：设 $Q_Y(\tau|X)$ 为结果变量 Y 在给定 X 下的 τ 分位数。分位数处理效应定义为：

$$QTE(\tau) = Q_{Y(1)}(\tau|X = c) - Q_{Y(0)}(\tau|X = c)$$

估计方法：使用分位数回归框架，在阈值两侧分别拟合分位数回归模型，然后计算分位数处理效应。

10.5.3 非标准误差项的推断问题

在断点回归中，由于使用局部回归，误差项可能不满足经典假设，需要特殊处理。

误差结构：1. 异方差性：误差方差可能随驱动变量变化 2. 自相关性：空间或时间上的相关性

稳健推断方法：1. 自助法：特别是 wild bootstrap，适用于异方差情况 2. 聚类标准误：当观测值存在聚类结构时 3. 基于设计的方法：利用断点回归的随机化性质

10.5.4 机器学习方法在 **RDD** 中的应用

近年来，机器学习方法被引入断点回归，以解决高维协变量和模型选择问题。

应用场景：1. 最优带宽选择：使用机器学习方法选择最优带宽 2. 协变量调整：使用机器学习方法控制协变量 3. 异质性处理效应：使用机器学习方法识别处理效应的异质性

常用方法：1. 岭回归和 LASSO：用于高维协变量控制 2. 随机森林和梯度提升：用于灵活的函数形式 3. 神经网络：用于复杂的非线性关系

10.5.5 断点回归的内生阈值问题

在某些情况下，阈值本身可能是内生的，这会影响断点回归的识别。

内生阈值来源：1. 阈值基于处理前的结果变量设定 2. 阈值随时间或空间变化 3. 个体知道阈值并可能操纵驱动变量

解决方法：1. 使用工具变量法处理内生阈值 2. 使用双重差分法结合断点回归 3. 使用固定效应控制不可观测的异质性

10.6 应用案例与操作实践

10.6.1 经典案例回顾：班级规模对成绩的影响（**Angrist & Lavy**）

10.6.2 中国场景下的典型案例分析（如高考分数线、贫困线政策）

10.6.3 RDD 在 **Stata/R** 中的实现命令与步骤（**rdrobust**，**rdplot** 等）

10.6.4 研究设计中的常见陷阱与报告要点

本章总结

断点回归通过巧妙利用现实规则中存在的” 断点”，在断点附近构造了一个近似随机的实验环境，从而为因果识别提供了可信的框架。其核心力量源于一个可检验的连续性假设，即个体在驱动变量阈值两侧是可比的。本章详细阐述了精确与模糊断点回归的识别策略、非参数估计方法以及一套系统的有效性检验工具。

断点回归的主要优势在于其清晰的识别假设和直观的图形展示，但同时也存在一些局限：首先，它估计的是局部平均处理效应，其外部有效性有限；其次，估计结果对带宽选择、函数形式等建模选择较为敏感；最后，需要足够大的样本量在阈值附近才能进行可靠的统计推断。

研究者在应用断点回归时，必须如同进行一项严格的实验般，透明地报告所有检验结果与稳健性分析，包括但不限于：协变量平衡性检验、McCrary 密度检验、带宽敏感性分析、安慰剂检验等。同时，需要审慎地解释估计结果的局部性含义，避免过度外推。

随着计量经济学的发展，断点回归方法也在不断扩展，如多维断点回归、分位数处理效应断点回归、以及机器学习方法在 RDD 中的应用等，这些扩展丰富了断点回归的工具箱，使其能够应用于更复杂的研究场景。然而，无论方法如何扩展，对识别假设的深入理解和严格检验始终是应用断点回归的基石。

11 合成控制法

本章导读

合成控制法是一种专门用于政策评估和因果推断的计量经济学方法，特别适用于评估某一政策或事件对单一处理单元（如一个国家、一个省份、一个城市）的影响。其核心思想是为处理单元构造一个” 合成对照组”——即由未受处理的多个控制单元的加权组合，使其在处理前的结果变量路径与真实处理单元尽可能相似。本章将系统阐述 SCM 的基本原理、权重构造、有效性检验、统计推断方法及前沿进展，帮助读者掌握这一在小样本、少处理组情境下的重要因果推断工具。

11.1 为何需要合成控制法？单一个案评估的挑战

11.1.1 传统方法的局限：DID 与匹配方法在少处理组时的困境

在政策评估中，当处理组数量极少（甚至只有一个）时，传统方法面临严峻挑战：

双重差分法的局限：DID 要求处理组和对照组满足平行趋势假设，且通常需要较多的处理组和对照组单元进行统计推断。当只有一个处理单元时：1. 难以找到合适的对照组 2. 无法进行有效的统计推断（标准误计算困难）3. 平行趋势假设难以检验

倾向得分匹配的局限：PSM 需要足够多的控制单元进行匹配，当处理单元极少时：1. 匹配质量难以保证 2. 共同支持域可能很小 3. 统计推断不可靠

简单比较的局限：直接比较处理单元与某个特定控制单元，容易受到特殊因素的影响，缺乏反事实构造的科学性。

11.1.2 Abadie 与 Gardeazabal（2003）的奠基性研究

合成控制法的开创性工作来自 Abadie 和 Gardeazabal（2003），他们研究了西班牙巴斯克地区的恐怖主义活动对经济增长的影响。

研究背景：- 处理单元：巴斯克地区（受恐怖主义影响）- 控制单元：西班牙其他地区 - 政策/事件：持续的恐怖主义活动 - 时期：1955-1997 年

方法创新：1. 使用加权组合构造” 合成巴斯克” 2. 权重选择使合成巴斯克在处理前（恐怖主义活动开始前）的经济特征与真实巴斯克尽可能相似 3. 比较真实巴斯克与合成巴斯克在处理后的经济增长路径

研究结论：恐怖主义活动导致巴斯克地区人均 GDP 下降了约 10 个百分点。

11.1.3 SCM 的核心优势：透明性、数据驱动与避免外推

合成控制法相对于传统方法的优势：

透明性：权重向量明确显示每个控制单元的贡献，使得反事实的构造过程完全透明。

数据驱动：权重通过优化算法确定，最小化处理前的预测误差，减少了主观选择偏差。

避免外推：合成控制法通常要求权重非负且和为 1，这意味着合成控制单元是控制单元的凸组合，从而避免外推到数据范围之外。

适合小样本：特别适用于处理单元极少的情况，甚至只有一个处理单元。

11.2 合成控制法的基本原理与模型设定

11.2.1 潜在结果框架与反事实构造问题

设我们有 $J+1$ 个地区（单元），其中第一个地区（$j=1$）在时期 T_0 受到政策干预，其余 J 个地区（$j=2,\ldots,J+1$）未受干预。观测时期为 $t=1,2,\ldots,T$，其中 T_0 是政策实施时点。

潜在结果框架：令 Y_{jt}^N 表示地区 j 在时期 t 未受干预的潜在结果，Y_{jt}^I 表示受干预的潜在结果。实际观测结果为：

$$Y_{jt} = \begin{cases} Y_{jt}^N & \text{if } t \le T_0 \text{ 或} j \ne 1 \\ Y_{jt}^I & \text{if } t > T_0 \text{ 且} j = 1 \end{cases}$$

因果效应：处理地区 $j=1$ 在时期 $t>T_0$ 的处理效应为：

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N$$

问题在于 Y_{1t}^N（反事实结果）无法观测。合成控制法的目标是构造一个反事实估计 $\hat{Y}_{1t}^N$。

11.2.2 数据结构：处理前多期结果变量与协变量

合成控制法利用两类信息：

结果变量：Y_{jt}：地区 j 在时期 t 的实际观测结果。我们特别关注处理前时期（$t = 1, \dots, T_0$）的结果变量。

协变量：$X_j = (X_{j1}, X_{j2}, \dots, X_{jk})'$：地区 j 的 k 个预处理特征，这些特征应该与结果变量相关且不受政策影响。

数据结构要求：1. 处理前时期足够长（T_0 足够大）2. 控制单元足够多且与处理单元相关 3. 协变量包含影响结果的重要变量

11.2.3 权重向量的构造：最小化处理前预测误差

合成控制法的核心是找到权重向量 $W = (w_2, w_3, \dots, w_{J+1})'$，其中 $w_j \geq 0$ 且 $\sum_{j=2}^{J+1} w_j = 1$，使得合成控制单元在处理前的特征与处理单元尽可能相似。

目标：1. 协变量平衡：合成控制单元的协变量近似等于处理单元的协变量 2. 处理前结果路径相似：合成控制单元在处理前的结果变量路径近似于处理单元

优化问题：选择权重 W 最小化以下距离度量：

$$\|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

其中：- X_1 是处理单元的协变量向量（$k \times 1$）- X_0 是控制单元的协变量矩阵（$k \times J$）- V 是一个 $k \times k$ 的正定对角线矩阵，表示不同协变量的相对重要性

11.3 权重的估计与” 合成控制单元” 的构造

11.3.1 目标函数与约束条件

合成控制法的权重通过求解以下约束优化问题得到：

$$\min_W \|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

满足：1. $w_j \geq 0$ 对于所有 $j = 2, \ldots, J+1$ 2. $\sum_{j=2}^{J+1} w_j = 1$

其中 V 矩阵的选择至关重要。通常 V 是对角矩阵，对角线元素表示每个协变量在距离度量中的相对重要性。

11.3.2 数值求解方法

权重估计是一个带约束的二次规划问题，可以使用标准优化算法求解：

算法步骤：1. 初始化权重 W 2. 选择权重矩阵 V 3. 求解优化问题得到最优权重 W^* 4. 计算合成控制结果：$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j^* Y_{jt}$

V 矩阵的选择：V 通常通过以下方法选择：1. 简单平均：所有协变量同等重要 2. 基于预测能力：选择 V 使合成控制单元在处理前的结果变量路径与处理单元最接近 3. 交叉验证：将处理前时期分为训练期和验证期

在实践中，常用方法是选择 V 最小化处理前时期的预测误差：

$$\min_V \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V) Y_{jt} \right)^2$$

这是一个嵌套优化问题：内层优化求解 $W^*(V)$，外层优化求解 V。

11.3.3 协变量的作用：平衡处理前特征与趋势

协变量在合成控制法中扮演两个重要角色：

平衡处理前特征：确保合成控制单元在处理前的可观测特征与处理单元相似。

捕捉处理前趋势：通过包含处理前的结果变量（或它们的函数）作为协变量，可以确保合成控制单元在处理前的结果变量路径与处理单元相似。

常见的协变量处理方式：1. 处理前各期的结果变量均值 2. 处理前结果变量的线性趋势 3. 其他与结果变量相关的经济、社会、人口特征 4. 处理前关键时点的结果变量值

建议：包含处理前多期结果变量作为协变量，可以更好地捕捉处理前趋势，提高合成控制的质量。

11.4 效应评估、图形展示与安慰剂检验

11.4.1 处理效应的计算与图形化（路径对比图）

处理效应估计：在得到最优权重 W^* 后，合成控制单元在处理后时期 $t > T_0$ 的反事实结果为：

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

处理效应估计为：

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N, \quad t > T_0$$

平均处理效应：

$$\hat{\tau} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \hat{\tau}_{1t}$$

图形展示：1. 路径对比图：绘制处理单元和合成控制单元在全部时期的结果变量路径 2. 效应图：绘制处理效应 $\hat{\tau}_{1t}$ 随时间的变化 3. 权重图：展示各控制单元的权重，了解合成控制单元的构成

路径对比图是最重要的诊断工具，可以直观展示：- 处理前拟合质量 - 处理后的差异 - 处理效应的动态变化

11.4.2 "安慰剂检验**"**（或**"** 排列检验**"**）的原理与实施

由于只有一个处理单元，传统的统计推断方法不适用。合成控制法使用安慰剂检验进行推断。

基本思想：如果政策效应是真实的，那么：1. 处理单元的处理效应应该显著大于控制单元 2. 将处理" 虚假地" 分配给控制单元时，不应观察到显著效应

实施步骤：1. 将每个控制单元依次视为" 伪处理单元" 2. 用其他控制单元为其构造合成控制 3. 计算每个伪处理单元的" 伪处理效应" 4. 比较真实处理效应与伪处理效应的分布

安慰剂效应：对于每个控制单元 $j = 2, \ldots, J+1$，假设它在时期 T_0 受到处理，用其他控制单元为其构造合成控制，计算伪处理效应：

$$\hat{\tau}_{jt}^{placebo} = Y_{jt} - \hat{Y}_{jt}^{N,placebo}, \quad t > T_0$$

其中 $\hat{Y}_{jt}^{N,placebo}$ 是单元 j 的合成控制结果。

11.4.3 显著性推断：p 值的计算与解读

通过安慰剂检验可以计算非参数的 p 值：

p 值计算：比较真实处理效应与安慰剂效应的分布。对于每个时期 $t > T_0$，计算：

$$p_t = \frac{\text{安慰剂效应} \geq \text{真实效应}}{\text{安慰剂检验次数} + 1}$$

更常见的是计算平均处理效应的 p 值：

$$p = \frac{\text{平均安慰剂效应} \geq \text{平均真实效应}}{\text{安慰剂检验次数} + 1}$$

经验 **p** 值：如果进行 J 次安慰剂检验（对每个控制单元），则最小可能的 p 值为 $1/(J+1)$。因此，安慰剂检验需要足够多的控制单元才能提供有意义的推断。

显著性判断：如果真实处理效应位于安慰剂效应的极端位置（如最大的 5%），则认为处理效应统计显著。

图形展示：1. 安慰剂效应分布图：绘制所有安慰剂效应和真实效应 2. **p** 值序列图：绘制各时期处理效应的 p 值

11.5 合成控制法的扩展与稳健性讨论

11.5.1 多个处理单元与推广的合成控制法

当有多个处理单元时，可以推广合成控制法：

多处理单元 **SCM**：1. 为每个处理单元单独构造合成控制 2. 计算每个处理单元的处理效应 3. 汇总处理效应（简单平均或加权平均）

交互固定效应模型：将 SCM 视为一种特殊的因子模型：

$$Y_{jt}^N = \delta_t + \theta_t Z_j + \lambda_t \mu_j + \epsilon_{jt}$$

其中：- δ_t：时间固定效应 - Z_j：可观测协变量 - λ_t：共同因子 - μ_j：因子载荷 - ϵ_{jt}：idiosyncratic 冲击

合成控制法相当于用控制单元的加权组合估计 $\lambda_t \mu_1$。

11.5.2 安慰剂检验的变体与敏感性分析

安慰剂检验的变体：1. 时间安慰剂检验：将政策时点提前，检验” 伪政策” 效应 2. 空间安慰剂检验：使用地理上不相邻的地区作为安慰剂 3. 协变量安慰剂检验：使用理论上不应受影响的变量作为结果

敏感性分析：1. 控制池变化：使用不同的控制单元集合 2. 协变量变化：使用不同的协变量组合 3. 时期变化：使用不同的处理前时期 4. 权重约束变化：放松权重非负或和为 1 的约束

敏感性分析用于检验估计结果对模型设定的稳健性。

11.5.3 交叉验证与正则化方法

交叉验证：将处理前时期分为训练期和验证期：1. 使用训练期估计权重 2. 使用验证期评估预测效果 3. 选择预测误差最小的模型设定

正则化方法：当控制单元较多时，可以引入正则化防止过拟合：1. 岭回归型正则化：在目标函数中加入权重平方和惩罚项 2. **LASSO** 型正则化：加入权重绝对值和惩罚项，促进稀疏解 3. 弹性网络：结合岭回归和 LASSO 的优点

正则化可以帮助提高合成控制法的稳定性和泛化能力。

11.5.4 合成控制法与矩阵补全、因子模型的联系

矩阵补全视角：SCM 可以视为矩阵补全问题：我们有一个结果变量矩阵 Y，其中某些元素（处理单元在处理后的结果）缺失，目标是根据观测值补全缺失值。

因子模型视角：SCM 与因子模型密切相关。假设结果数据由因子模型生成：

$$Y_{jt} = \delta_t + \lambda_t \mu_j + \epsilon_{jt}$$

其中 λ_t 是共同因子，μ_j 是因子载荷。合成控制法用控制单元的加权组合估计处理单元的因子载荷组合。

广义合成控制法：将 SCM 推广到更一般的因子模型设定，允许更灵活的估计和推断。

11.6 应用实例与操作指南

11.6.1 经典案例回顾：加州烟草控制法案（**AB 08**）的效果评估

11.6.2 Stata (`synth`)、R (`Synth`) 中的实现步骤

11.6.3 研究报告的规范：如何展示结果与进行稳健性检验

本章总结

合成控制法为评估针对特定地区或个体的政策干预提供了强大的工具，特别适用于处理单元极少的情况。其核心优势在于透明性、数据驱动和避免外推。本章系统介绍了 SCM 的基本原理、权重估计、效应评估、统计推断方法以及各种扩展和稳健性分析。

然而，合成控制法的有效性依赖于几个关键条件：处理前时期足够长、控制单元池足够大且与处理单元相关、协变量选择恰当。统计推断严重依赖安慰剂检验，因此需要足够多的控制单元才能进行有意义的显著性检验。

在实践中，研究者应：1. 提供清晰的路径对比图，展示处理前拟合质量 2. 进行充分的安慰剂检验和敏感性分析 3. 透明报告权重向量，说明合成控制单元的构成 4. 谨慎解释结果，考虑可能的替代解释

近年来，合成控制法与计量经济学和机器学习方法的结合进一步拓展了其应用边界，如正则化 SCM、广义 SCM、矩阵补全方法等。这些发展为小样本政策评估提供了更丰富、更稳健的工具箱。然而，无论方法如何扩展，对数据要求的理解和严格的稳健性检验始终是应用合成控制法的基石。

12 回归控制法

本章导读

回归控制法是一种基于回归模型的政策评估方法，特别适用于处理单元数量较少的情境。与合成控制法类似，RC 方法也旨在为处理单元构造一个反事实的 “合成对照组”，但其核心思想是通过回归模型利用控制单元的面板数据来预测处理单元的反事实结果。本章将系统介绍回归控制法的两种主流框架——Hsiao 等人的方法与基于正则化回归的 ATC 方法，阐述其原理、估计、推断以及与合成控制法的比较，帮助读者掌握这一在小样本政策评估中的重要工具。

12.1 回归控制法的两种框架：Hsiao 等人的方法与 ATC 方法

12.1.1 核心问题：如何利用控制单元面板数据预测处理单元的反事实？

回归控制法的核心问题与合成控制法相同：当处理单元数量很少（甚至只有一个）时，如何利用未受处理的控制单元数据来预测处理单元如果没有接受处理时的结果（即反事实结果）?

基本设定：假设我们有 $N+1$ 个单元（如地区、企业等），其中第一个单元在时间 T_0 后接受处理，其余 N 个单元始终未接受处理。观测时间跨度为 $t = 1, 2, \dots, T$，其中 T_0 为政策干预时点。令 Y_{it} 表示单元 i 在时间 t 的结果变量。对于处理单元（$i = 1$），我们观测到的是处理后的结果 Y_{1t}^I（当 $t > T_0$），但我们想要估计的是其反事实结果 Y_{1t}^N（即如果没有接受处理的结果）。

回归控制法的基本思想是：处理单元的反事实结果可以通过控制单元结果的线性组合来预测，即

$$Y_{1t}^N = \sum_{j=2}^{N+1} \beta_j Y_{jt} + \varepsilon_t, \quad t = 1, \dots, T_0$$

然后利用估计出的关系来预测 $t > T_0$ 时的反事实结果。

12.1.2 Hsiao 等人的方法：将处理单元视为控制单元的线性组合加上误差项

Hsiao 等人（2012）提出了一种基于线性回归的预测方法。他们认为，处理单元的结果变量可以由控制单元的结果变量线性表示，再加上一个随机误差项。

模型设定：假设在处理前时期（$t \leq T_0$），处理单元的结果变量满足以下关系：

$$Y_{1t} = \alpha + \sum_{j=2}^{N+1} \beta_j Y_{jt} + \varepsilon_t, \quad t = 1, \dots, T_0$$

其中 α 是截距项，β_j 是系数，ε_t 是随机误差项，满足 $E(\varepsilon_t) = 0$。

关键假设：1. 线性关系：处理单元与控制单元之间的关系是线性的。2. 参数稳定性：这种线性关系在处理前和处理后保持不变（即系数 β_j 不随时间变化）。3. 控制单元不受政策影响：控制单元的结果变量 Y_{jt}（$j = 2, \dots, N+1$）在政策干预后不受影响，即 Y_{jt} 在 $t > T_0$ 时仍为 Y_{jt}^N。

在 Hsiao 等人的框架中，通常使用处理前的数据（$t = 1, \dots, T_0$）来估计模型参数，然后用估计的模型预测处理后的反事实结果：

$$\hat{Y}_{1t}^N = \hat{\alpha} + \sum_{j=2}^{N+1} \hat{\beta}_j Y_{jt}, \quad t > T_0$$

处理效应则通过比较实际观测值与预测值得到：$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N$。

12.1.3 ATC 方法：使用弹性网络等正则化回归直接预测结果

ATC（Augmented Synthetic Control）方法是一种结合了正则化回归的回归控制法，由 Ben-Michael、Feller 和 Rothstein（2021）提出。它通过正则化回归（如岭回归、LASSO 或弹性网络）来估计控制单元的权重，从而预测处理单元的反事实结果。

基本思想：与合成控制法类似，ATC 方法也试图用控制单元的加权组合来预测处理单元的结果，但它允许权重为负，并且通过正则化处理来解决控制单元数量较多时的过拟合问题。

模型设定：ATC 方法通常使用以下形式的回归模型：

$$Y_{1t} = \sum_{j=2}^{N+1} w_j Y_{jt} + \varepsilon_t, \quad t = 1, \dots, T_0$$

其中权重 w_j 通过正则化回归估计得到。例如，使用弹性网络时，我们求解以下优化问题：

$$\min_{w} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{N+1} w_j Y_{jt} \right)^2 + \lambda \left(\alpha \sum_{j=2}^{N+1} |w_j| + (1-\alpha) \sum_{j=2}^{N+1} w_j^2 \right)$$

其中 λ 是正则化参数，α 控制 LASSO 惩罚（L1）和岭回归惩罚（L2）的相对比例。

优势：1. 可以处理控制单元数量较多的情况，通过正则化避免过拟合。2. 权重可以为负，提供了更大的灵活性。3. 可以通过交叉验证选择正则化参数，提高预测精度。

12.2 Hsiao 等人的回归控制法：原理与估计

12.2.1 模型设定：因子模型视角

Hsiao 等人的方法可以从因子模型的角度来理解。假设每个单元的结果变量由以下因子模型生成：

$$Y_{it}^N = \mu_i + \lambda_t + \sum_{k=1}^{K} \theta_{ik} f_{kt} + \varepsilon_{it}$$

其中 μ_i 是单元固定效应，λ_t 是时间固定效应，f_{kt} 是 K 个不可观测的共同因子，θ_{ik} 是单元 i 在因子 k 上的载荷，ε_{it} 是 idiosyncratic 冲击。

对于处理单元 $i = 1$，假设其潜在结果可以表示为控制单元结果的线性组合，这是因为控制单元的结果变量中包含了共同因子 f_{kt} 的信息。如果控制单元足够多，且它们与处理单元受到相同的共同因子影响，那么处理单元的反事实结果就可以通过控制单元的线性组合来预测。

Hsiao 等人的模型：在实际应用中，Hsiao 等人通常使用以下形式的回归模型：

$$Y_{1t} = \alpha + \sum_{j=2}^{N+1} \beta_j Y_{jt} + \varepsilon_t, \quad t = 1, \dots, T_0$$

其中，他们假设 ε_t 为独立同分布的误差项。为了获得更好的预测，有时也会在回归中加入一些协变量。

12.2.2 最佳线性预测与系数估计

在 Hsiao 等人的框架中，系数 β_j 的估计是通过最小化处理前时期的预测误差来实现的。即，我们使用处理前数据 $(Y_{1t}, Y_{2t}, \dots, Y_{N+1,t})$ 对 $t = 1, \dots, T_0$ 来估计回归系数。

估计方法：通常使用普通最小二乘法（OLS）来估计参数 α 和 β_j。设 $Y_1 = (Y_{11}, \dots, Y_{1T_0})'$ 为处理单元在处理前时期的结果变量向量，X 为一个 $T_0 \times (N+1)$ 的矩阵，其中第一列为全 1 向量（对应截距），其余列为控制单元的结果变量（$Y_{jt}, j = 2, \dots, N+1$）。则回归模型可写为：

$$Y_1 = X\beta + \varepsilon$$

其中 $\beta = (\alpha, \beta_2, \dots, \beta_{N+1})'$。OLS 估计量为：

$$\hat{\beta} = (X'X)^{-1}X'Y_1$$

注意事项：当控制单元数量 N 较大，而处理前时期 T_0 相对较小时，OLS 估计可能会出现过拟合问题（即样本内拟合很好，但样本外预测很差）。因此，Hsiao 等人建议只选择一部分控制单元进入回归模型，或者使用主成分回归等降维技术。

12.2.3 处理效应的点估计与置信区间构造

点估计：在估计出回归系数后，处理单元在政策实施后的反事实结果预测为：

$$\hat{Y}_{1t}^N = \hat{\alpha} + \sum_{j=2}^{N+1} \hat{\beta}_j Y_{jt}, \quad t = T_0 + 1, \dots, T$$

处理效应的点估计为实际观测值与预测值之差：

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N, \quad t > T_0$$

平均处理效应（ATE）为：

$$\hat{\tau} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \hat{\tau}_{1t}$$

置信区间构造：由于只有一个处理单元，传统的标准误计算方法不适用。Hsiao 等人建议使用自助法（bootstrap）来构造置信区间。

自助法步骤：1. 从处理前时期的残差 $\{\hat{\varepsilon}_t\}_{t=1}^{T_0}$ 中有放回地抽取 T_0 个残差，得到自助样本残差 $\{\hat{\varepsilon}_t^*\}_{t=1}^{T_0}$。2. 生成自助样本的处理前结果：$Y_{1t}^* = \hat{\alpha} + \sum_{j=2}^{N+1} \hat{\beta}_j Y_{jt} + \hat{\varepsilon}_t^*$。3. 使用自助样本 $(Y_{1t}^*, Y_{2t}, \dots, Y_{N+1,t})$ 重新估计回归系数，得到 $\hat{\beta}^*$。4. 利用 $\hat{\beta}^*$ 预测处理后的反事实结果 $\hat{Y}_{1t}^{N*}$，并计算自助样本的处理效应 $\hat{\tau}_{1t}^* = Y_{1t} - \hat{Y}_{1t}^{N*}$（注意：这里 Y_{1t} 是实际观测值，因为政策实施后的结果没有重抽样）。5. 重复以上步骤多次（如 1000 次），得到处理效应的自助分布，然后利用该分布构造置信区间（如百分位数区间）。

注意：这种自助法假设误差项 ε_t 是独立同分布的，且模型设定正确。如果这些假设不成立，自助法可能无效。

12.3 基于正则化回归的回归控制法

12.3.1 高维控制问题与正则化（岭回归、**LASSO**、弹性网络）

当控制单元数量 N 较大，甚至超过处理前时期 T_0 时，传统的 OLS 估计会面临高维问题（即自变量数量多于观测值数量），此时 OLS 无法求解，或者即使可求也会导致严重的过拟合。

正则化方法：为了解决高维问题，我们可以使用正则化回归，通过在损失函数中加入惩罚项来约束系数的大小，从而获得更稳定的估计。

岭回归：岭回归在 OLS 损失函数中加入系数的 L2 惩罚项：

$$\min_{\beta} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{N+1} \beta_j Y_{jt} \right)^2 + \lambda \sum_{j=2}^{N+1} \beta_j^2$$

其中 $\lambda \geq 0$ 是正则化参数。岭回归的估计结果通常会使系数向零收缩，但不会将系数 exactly 设为零。

LASSO：LASSO 在 OLS 损失函数中加入系数的 L1 惩罚项：

$$\min_{\beta} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{N+1} \beta_j Y_{jt} \right)^2 + \lambda \sum_{j=2}^{N+1} |\beta_j|$$

LASSO 倾向于产生稀疏解，即将一些系数 exactly 设为零，从而实现了变量选择。

弹性网络：弹性网络结合了 L1 和 L2 惩罚项：

$$\min_{\beta} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{N+1} \beta_j Y_{jt} \right)^2 + \lambda \left(\alpha \sum_{j=2}^{N+1} |\beta_j| + (1-\alpha) \sum_{j=2}^{N+1} \beta_j^2 \right)$$

其中 $\alpha \in [0, 1]$ 控制两种惩罚的比例。弹性网络综合了岭回归和 LASSO 的优点，特别适用于高维且变量间存在相关性的情况。

12.3.2 **ATC** 估计量：双重稳健与渐进性质

ATC 估计量：Ben-Michael 等人（2021）提出的 ATC 方法实际上是一种双重稳健的估计量。它结合了回归调整和逆概率加权（IPW）的思想，但在这里我们主要关注其回归调整的部分。

模型设定：假设我们有一个面板数据集，其中 $i = 1, \ldots, N+1$ 个单元，$t = 1, \ldots, T$ 个时间点。处理发生在 T_0 之后，且只有第一个单元接受处理。我们想要估计处理单元的平均处理效应。

ATC 方法首先通过正则化回归（如弹性网络）来估计一个预测模型，用于预测处理单元的反事实结果。具体而言，我们使用处理前数据来估计以下模型：

$$Y_{1t} = \sum_{j=2}^{N+1} w_j Y_{jt} + \varepsilon_t, \quad t = 1, \dots, T_0$$

其中权重 w_j 通过弹性网络等正则化回归估计得到。

双重稳健性：ATC 估计量具有双重稳健性：只要预测模型（回归部分）或倾向得分模型（加权部分）其中之一设定正确，估计量就是一致的。但在回归控制法的语境下，通常我们只使用回归部分，因此双重稳健性并不直接体现。不过，ATC 方法通过正则化回归提高了预测的稳健性。

渐进性质：在一定的正则条件下，ATC 估计量是渐近正态的，并且可以通过自助法进行推断。当单元数量和时间维度都增加时，ATC 估计量收敛于真实处理效应。

12.3.3 交叉验证选择调优参数

在正则化回归中，正则化参数 λ（以及弹性网络中的 α）的选择至关重要。通常，我们使用交叉验证来选择这些调优参数。

交叉验证步骤：1. 将处理前时期的数据随机分成 K 折（通常 $K = 5$ 或 10）。2. 对于每一组候选参数 (λ, α)，进行以下操作：a. 对于 $k = 1, \dots, K$，使用除第 k 折外的所有数据拟合模型，得到权重估计 $\widehat{w}^{-k}$。b. 使用 $\widehat{w}^{-k}$ 预测第 k 折中处理单元的结果，计算预测误差。c. 将 K 折的预测误差平均，得到交叉验证误差。3. 选择使交叉验证误差最小的参数组合 (λ^*, α^*)。

注意事项：由于时间序列数据可能存在自相关，因此简单的随机分割可能会破坏时间结构。一种替代方法是使用滚动时间窗口交叉验证：用前 M 个时期的数据训练模型，预测下一个时期，然后移动窗口，重复此过程。

12.4 回归控制法的统计推断

12.4.1 基于残差自举法的推断

由于回归控制法通常用于小样本（尤其是处理单元很少），传统的渐进推断可能不适用。因此，自助法成为主要的推断工具。

残差自助法：假设我们已通过回归控制法得到处理效应的估计 $\widehat{\tau}_{1t}$。为了构造置信区间，我们可以对残差进行自助抽样。具体步骤如下（以 Hsiao 等人的方法为例）：

1. 估计处理前时期的模型：$Y_{1t} = \alpha + \sum_{j=2}^{N+1} \beta_j Y_{jt} + \varepsilon_t$，得到残差 $\widehat{\varepsilon}_t, t = 1, \dots, T_0$。

2. 对残差进行中心化处理：$\tilde{\varepsilon}_t = \hat{\varepsilon}_t - \frac{1}{T_0}\sum_{s=1}^{T_0}\hat{\varepsilon}_s$。
3. 从中心化后的残差 $\{\tilde{\varepsilon}_t\}$ 中有放回地抽取 T_0 个残差，得到自助残差 ε_t^*。
4. 生成自助样本的处理前结果：$Y_{1t}^* = \hat{\alpha} + \sum_{j=2}^{N+1}\hat{\beta}_j Y_{jt} + \varepsilon_t^*$。
5. 使用自助样本 $(Y_{1t}^*, Y_{2t}, \ldots, Y_{N+1,t})$ 重新估计回归系数 β^*。
6. 利用 β^* 预测处理后的反事实结果 $\hat{Y}_{1t}^{N*}$，并计算自助样本的处理效应 $\hat{\tau}_{1t}^* = Y_{1t} - \hat{Y}_{1t}^{N*}$（注意：政策实施后的 Y_{1t} 仍使用原始观测值）。
7. 重复步骤 3-6 多次（如 1000 次），得到处理效应的自助分布。
8. 基于自助分布构造置信区间，例如，取 2.5% 和 97.5% 分位数作为 95% 置信区间。

适用性：残差自助法假设模型设定正确且误差项独立同分布。如果这些假设不成立，可以考虑使用块自助法（block bootstrap）来捕捉时间序列相关性。

12.4.2 基于预测区间的推断

另一种推断方法是构造反事实结果的预测区间。预测区间反映了反事实预测的不确定性，从而可以判断处理效应是否显著不为零。

预测区间构造：假设反事实预测模型为 $\hat{Y}_{1t}^N = f(\{Y_{jt}\}_{j=2}^{N+1}; \hat{\beta})$，预测误差主要来自两个方面：参数估计的不确定性和模型误差。我们可以通过模拟来构造预测区间。

步骤：1. 估计模型参数 $\hat{\beta}$ 及其方差-协方差矩阵 $\hat{\Sigma}$（如果可用）。2. 从参数分布中抽取 $\beta^* \sim N(\hat{\beta}, \hat{\Sigma})$（或者使用自助法得到参数分布）。3. 对于每个 β^*，计算反事实预测 $\hat{Y}_{1t}^{N*} = f(\{Y_{jt}\}_{j=2}^{N+1}; \beta^*)$。4. 重复步骤 2-3 多次，得到反事实预测的分布。5. 取该分布的 $\alpha/2$ 和 $1-\alpha/2$ 分位数作为 $1-\alpha$ 预测区间。

如果实际观测值 Y_{1t} 落在预测区间之外，则表明处理效应显著。

12.4.3 与 **SCM** 安慰剂检验的对比

合成控制法通常使用安慰剂检验（permutation test）进行推断，即将处理状态随机分配给控制单元，观察 “伪处理效应” 的分布。回归控制法也可以采用类似的安慰剂检验。

安慰剂检验步骤：1. 依次将每个控制单元视为 “伪处理单元”，假设它在 T_0 后接受处理。2. 对每个伪处理单元，使用回归控制法估计其 “伪处理效应”。3. 将所有伪处理效应与真实的处理效应进行比较。4. 计算真实处理效应在伪处理效应分布中的位置，得到 p 值：$p = \frac{\text{伪处理效应} \geq \text{真实处理效应}}{\text{伪处理单元数量}+1}$。

与 **SCM** 安慰剂检验的异同：- 相似点：都是通过置换处理状态来构建经验分布。- 不同点：SCM 的安慰剂检验中，每个伪处理单元都需要重新计算权重（因为权重是非负且和为 1 的凸组合），而回归控制法中，伪处理单元的回归模型可能允许负权重，且不一定有凸组合约束。

注意事项：安慰剂检验要求控制单元之间相互独立且与处理单元可比。如果控制单元数量很少，安慰剂检验的功效可能很低。

12.5 回归控制法与合成控制法的比较与选择

12.5.1 方法逻辑对比：参数化 **vs.** 非参数化，外推 **vs.** 内插

回归控制法：- 参数化：RC 通常假设处理单元的反事实结果可以通过控制单元的线性组合来预测，且这种关系在处理前后保持不变。这是一种参数化假设。- 外推：RC 允许权重为负，因此合成控制单元可以是控制单元的外推（即超出控制单元观测值的范围），这可能导致不合理的预测（例如，预测值远超出控制单元的实际范围）。

合成控制法：- 非参数化：SCM 不预设具体的函数形式，而是通过数据驱动的方式寻找权重，使处理前拟合最优。它更接近于非参数方法。- 内插：SCM 要求权重非负且和为 1，这意味着合成控制单元是控制单元的凸组合，即内插。这通常被认为更安全，因为内插通常比外推更稳健。

可视化对比：在二维空间中，假设有两个控制单元 A 和 B，它们的结果变量构成一个平面。处理单元的反事实预测可以看作是这个平面上的一个点。SCM 要求这个点必须在 A 和 B 的连线上（凸组合），而 RC 允许这个点在线段的两侧延长线上（外推）。

12.5.2 适用场景与假设差异

适用场景：- 回归控制法：适用于控制单元数量较多，且处理单元与控制单元之间的关系可能是线性的情况。特别是当处理前时期 T_0 较大时，RC 可以通过回归估计更多参数。- 合成控制法：适用于控制单元数量适中，且我们希望合成控制单元是实际存在的控制单元的加权平均（即内插）的情况。SCM 对处理前时期长度的要求相对较低，因为它只需要拟合一条时间路径。

假设差异：- **RC** 的关键假设：线性关系稳定（即处理前后的系数不变），误差项满足一定条件（如独立同分布）。- **SCM** 的关键假设：存在一组权重，使得合成控制单元在处理前的结果变量路径与处理单元非常接近，且这种相似性在处理后仍保持（即无政策干预时，平行趋势成立）。

模型灵活性：RC 通常更灵活，因为允许负权重，并且可以通过加入更多控制变量（如协变量）来改进预测。SCM 则通过权重约束（凸组合）来避免外推，但可能因此损失一些拟合精度。

12.5.3 实证应用中的选择指南

在实际应用中，选择回归控制法还是合成控制法，可以考虑以下因素：

1. 数据规模：

- 如果控制单元数量很多（远大于处理前时期长度），考虑使用正则化的回归控制法（如 LASSO、弹性网络）来防止过拟合。
- 如果控制单元数量适中，且处理前时期较长，两种方法都可以尝试。

2. 理论基础：
   - 如果理论或先验知识表明处理单元可能是控制单元的凸组合（例如，处理单元是从控制单元所在群体中选取的），则 SCM 更合适。
   - 如果理论允许处理单元与控制单元之间存在更复杂的关系（可能涉及负权重），则 RC 可能更合适。
3. 处理前拟合：
   - 比较两种方法在处理前时期的拟合效果（如均方预测误差）。拟合更好的方法可能更可靠，但要注意过拟合问题。
4. 稳健性检验：
   - 尝试多种方法（包括不同设定下的 RC 和 SCM），观察处理效应估计是否稳健。如果不同方法给出的结论一致，则结果更可信。
5. 结果解释：
   - SCM 的权重通常更容易解释，因为它们是非负的且和为 1，可以看作是控制单元的 “贡献度”。RC 的权重可能为负，解释起来更复杂。

建议做法：在实证研究中，可以同时报告 RC 和 SCM 的结果，并进行比较。如果两者结果相似，则增强了结论的可信度。如果结果差异很大，则需要深入分析原因，并检查模型假设是否成立。

12.6 应用实例与操作指南

12.6.1 案例：欧元区对成员国经济增长的影响

12.6.2 R (`gsynth`, `fect`) 和 **Stata** 中的实现

12.6.3 结果报告与诊断图

本章总结

回归控制法为小样本政策评估提供了另一种强有力的、基于回归模型的工具。它通过利用控制单元的面板数据来预测处理单元的反事实结果，从而估计处理效应。本章介绍了两种主要的回归控制法框架：Hsiao 等人的方法和基于正则化回归的 ATC 方法。

Hsiao 等人的方法简单直观，通过线性回归建立处理单元与控制单元之间的关系，并假设这种关系在处理前后保持不变。但当控制单元数量较多时，容易产生过拟合问题。正则化回归方法（如岭回归、LASSO、弹性网络）通过引入惩罚项来解决高维问题，提高了预测的稳定性和准确性。

在统计推断方面，由于处理单元数量少，传统渐进推断不适用，因此我们介绍了基于自助法和安慰剂检验的推断方法。这些方法可以帮助我们评估处理效应的显著性。

与合成控制法相比，回归控制法更加灵活（允许负权重），但可能因此进行外推，导致不合理的预测。合成控制法通过权重约束（凸组合）确保内插，通常更稳健。在实际应用中，研究者应根据数据特征和理论背景选择合适的方法，或者同时使用两种方法以验证结果的稳健性。

回归控制法的有效性依赖于模型假设的正确性，包括线性关系稳定、误差项独立同分布等。因此，在使用回归控制法时，必须进行充分的稳健性检验，包括模型设定检验、残差诊断、安慰剂检验等。

随着计量经济学和机器学习的发展，回归控制法也在不断演进，例如与因子模型、矩阵补全等方法的结合。这些发展为小样本政策评估提供了更丰富、更稳健的工具箱。然而，无论方法如何扩展，对数据生成过程的理解和严格的模型检验始终是获得可靠因果推断的基石。

13 中介效应与调节效应

机制与异质性：中介效应与调节效应分析

本章导读

在运用第 7-12 章的方法可信地识别出”X 是否导致 Y” 之后，科学研究必然走向更深入的追问：这一影响通过何种机制传导？又在何种情境下更强或更弱？本章将系统介绍用于回答这两个问题的核心工具——中介效应与调节效应分析。需要预先强烈警示：本章方法并非用于解决核心因果关系的识别问题，其有效性严重依赖于”X→Y” 关系本身已得到无偏估计，且对机制变量（M）的测量与模型设定有极其严苛的要求。本章是因果推断链条的深化与拓展，而非起点。学习本章前，必须确保已掌握第 6 章的基本框架和前几章至少一种因果识别方法。

13.1 从因果识别到因果解释：机制与异质性分析的角色定位

13.1.1 中介效应：拆解” 黑箱”，揭示因果路径

在确认了”X 导致 Y” 的基本因果关系后，研究者自然希望了解这一影响是如何发生的。中介效应分析正是为了揭示这一” 黑箱” 机制而设计。其核心思想是：自变量 X 对因变量 Y 的影响并非全部直接发生，而是部分或全部通过一个或多个中间变量 M（称为中介变量）传递。

形式化表述：设 X 为自变量，Y 为因变量，M 为中介变量。完整的中介过程包含三条路径：1. X 对 Y 的总效应：$X \to Y$ 2. X 对 M 的影响：$X \to M$ 3. M 对 Y 的影响（控制 X 后）：$M \to Y$

中介效应分析旨在量化通过 M 传递的间接效应，并将其与 X 对 Y 的直接效应区分开来。

13.1.2 调节效应：界定边界，理解效应异质性

与中介效应关注” 如何发生” 不同，调节效应关注” 何时发生” 或” 对谁发生”。调节效应分析检验第三个变量 W（称为调节变量）如何改变 X 与 Y 之间关系的强度或方向。

形式化表述：调节效应表现为 X 与 W 的交互项对 Y 的影响。如果 X 与 Y 的关系随 W 的变化而变化，则 W 起到了调节作用。

调节的类型：1. 增强型调节：W 增强了 X 对 Y 的影响 2. 削弱型调节：W 削弱了 X 对 Y 的影响 3. 反转型调节：W 改变了 X 对 Y 影响的方向

13.1.3 两者核心区别与联系：过程机制 **vs.** 情境条件

核心区别：| 维度 | 中介效应 | 调节效应 | |——|————-|————-| | 研究问题 | X 如何影响 Y？ | X 何时/对谁影响 Y？ | | 理论角色 | M 是机制变量 | W 是边界条件变量 | | 统计模型 | 路径分析，效应分解 | 交互项分析 | | 变量时序 | 通常 $X \rightarrow M \rightarrow Y$ | X 和 W 通常同时影响 Y | | 关注焦点 | 解释过程 | 界定条件 |

联系：1. 中介和调节可以结合在同一个分析框架中 2. 两者都涉及第三个变量的作用 3. 在复杂的理论模型中，一个变量可能同时起到中介和调节作用

13.1.4 一个前置警告：忽略核心因果识别将使机制分析失去根基

重要警示：1. 中介效应的基础是已识别的 **X→Y** 关系：如果 X 与 Y 之间的因果关系因内生性问题而未能得到准确估计，那么在此基础上进行的中介分析将是无效的。2. 中介变量 **M** 的内生性：即使 X→Y 的关系得到准确识别，中介变量 M 本身可能也存在内生性问题（如遗漏变量、测量误差等），这会威胁中介效应估计的准确性。3. 调节变量的选择：调节变量 W 的选择应有充分的理论依据，而非数据挖掘的结果。

建议的研究流程：1. 首先使用第 7-12 章的方法准确识别 X→Y 的因果关系 2. 在确认 X→Y 关系的基础上，进行机制（中介）分析 3. 同时或在独立分析中，进行异质性（调节）分析 4. 对中介变量 M 可能存在的内生性进行充分讨论和检验

13.2 中介效应分析的传统方法与模型

13.2.1 中介变量的定义、选择与理论依据

中介变量的定义：中介变量 M 是自变量 X 影响因变量 Y 的中间机制或传递路径。它同时受到 X 的影响并影响 Y。

中介变量的选择标准：1. 理论依据：M 的选择应有充分的理论支持 2. 时间顺序：理想情况下应有 $X \rightarrow M \rightarrow Y$ 的时间顺序 3. 测量质量：M 应能得到准确、可靠的测量 4. 概念区分：M 应与 X 和 Y 在概念上明确区分

常见的中介变量类型：1. 心理机制：态度、信念、情绪等 2. 行为机制：具体的行为表现 3. 生理机制：生理指标、神经活动等 4. 社会机制：社会互动、网络关系等

13.2.2 经典三步回归法（**Baron & Kenny**）流程与局限

Baron 和 Kenny（1986）提出的三步回归法是中介效应分析最经典的方法。

三步回归法的步骤：

第一步：检验总效应

$$Y = i_1 + cX + e_1$$

检验系数 c 是否显著。如果 c 不显著，通常认为不存在中介效应（但有例外情况）。

第二步：检验 X 对 M 的影响

$$M = i_2 + aX + e_2$$

检验系数 a 是否显著。如果 a 不显著，说明 X 对 M 没有影响，中介效应不存在。

第三步：检验 M 对 Y 的影响（控制 X）

$$Y = i_3 + c'X + bM + e_3$$

检验系数 b 是否显著。如果 b 显著，且第一步中的 c 也显著，则可能存在中介效应。

判断标准：1. 如果 a 和 b 都显著，且 c' 变得不显著或显著减小，则为完全中介 2. 如果 a 和 b 都显著，且 c' 仍然显著但减小，则为部分中介

效应分解：- 总效应：c - 直接效应：c' - 间接效应（中介效应）：ab

三步回归法的局限：1. 低统计功效：需要三个独立的显著性检验，增加了 II 类错误的风险 2. **Sobel** 检验的局限：检验间接效应 ab 的 Sobel 检验要求 ab 服从正态分布，但通常不满足 3. 无法处理复杂模型：难以处理多重中介、链式中介等复杂情况 4. 忽略内生性：未考虑中介变量 M 可能存在的内生性问题

13.2.3 结构方程模型下的中介分析

结构方程模型为中介分析提供了更灵活的框架。

SEM 中的中介模型：在 SEM 框架下，中介模型可以表示为：

$$M = \alpha_1 + aX + \epsilon_1$$
$$Y = \alpha_2 + c'X + bM + \epsilon_2$$

SEM 的优势：1. 可以同时估计所有路径系数 2. 可以方便地处理测量误差 3. 可以估计模型的整体拟合度 4. 可以处理更复杂的模型（如多重中介）

SEM 的局限：1. 对大样本量的要求较高 2. 模型设定需要较强的理论指导 3. 同样面临中介变量内生性的问题

13.3 因果中介分析框架：迈向更严谨的机制检验

13.3.1 传统方法的困境：混淆与内生性威胁

传统中介分析方法面临的主要挑战是混淆偏误，特别是：1. **X-M** 关系的混淆：可能存在未观测变量同时影响 X 和 M 2. **M-Y** 关系的混淆（控制 X 后）：可能存在未观测变量同时影响 M 和 Y

这些混淆会导致中介效应估计的偏误。

以教育回报率为例：假设我们研究教育（X）通过认知能力（M）影响工资（Y）的中介机制。问题在于：1. 家庭背景可能同时影响教育选择和认知能力 2. 动机可能同时影响认知能力和工资

如果不控制这些混淆因素，中介效应估计将是有偏的。

13.3.2 基于潜在结果模型的因果中介效应定义

Imai 等人（2010）将中介分析置于反事实框架下，提出了因果中介分析。

定义：设 $Y_i(x, m)$ 表示当个体 i 的 X 取值为 x、M 取值为 m 时的潜在结果。$M_i(x)$ 表示当 X 取值为 x 时，个体 i 的 M 的潜在值。

因果中介效应：对于个体 i，在控制 X 从 x 变为 x^* 时，通过 M 传递的间接效应为：

$$\delta_i(x) = Y_i(x, M_i(x^*)) - Y_i(x, M_i(x))$$

直接效应：

$$\zeta_i(x) = Y_i(x^*, M_i(x)) - Y_i(x, M_i(x))$$

总效应：

$$\tau_i = Y_i(x^*, M_i(x^*)) - Y_i(x, M_i(x)) = \delta_i(x) + \zeta_i(x^*)$$

在个体层面，$\delta_i(x)$ 和 $\zeta_i(x)$ 通常不可识别，但我们关心的是平均因果中介效应和平均直接效应。

13.3.3 识别假设：序贯可忽略性及其不可检验性

因果中介效应的识别依赖于序贯可忽略性假设：

假设 **1**（可忽略处理分配）：

$$\{Y_i(x', m), M_i(x)\} \perp X_i | W_i = w$$

给定预处理协变量 W，处理分配 X 与潜在结果和潜在中介变量独立。

假设 **2**（可忽略中介变量分配）：

$$Y_i(x', m) \perp M_i | X_i = x, W_i = w$$

给定处理状态 X 和协变量 W，中介变量 M 与潜在结果独立。

假设 **3**（无交互作用）：对于所有的 $x \neq x'$ 和 m，

$$Y_i(x, m) - Y_i(x', m) = Y_i(x, m') - Y_i(x', m')$$

即直接效应不依赖于中介变量的取值。

这些假设的挑战：1. 特别是假设 2，要求在控制 X 和 W 后，M 与 Y 之间没有未观测的混淆 2. 这些假设无法直接检验 3. 在实践中很难完全满足

13.3.4 估计方法：参数化模型与半参数 **Bootstrap**

参数化方法：在序贯可忽略性假设下，可以使用参数模型估计因果中介效应。常见的方法是：

1. 用参数模型（如线性回归）估计中介方程：$M_i = \alpha_2 + \beta_2 X_i + \epsilon_{2i}$
2. 用参数模型估计结果方程：$Y_i = \alpha_3 + \beta_3 X_i + \gamma M_i + \epsilon_{3i}$
3. 计算平均因果中介效应：$\hat{\delta} = \hat{\beta}_2 \hat{\gamma}$

半参数 **Bootstrap** 方法：由于中介效应的抽样分布通常不是正态的，推荐使用 Bootstrap 方法进行推断。

Bootstrap 步骤：1. 从原始样本中有放回地抽取 B 个 Bootstrap 样本（通常 B=1000-5000）2. 在每个 Bootstrap 样本中估计中介效应 3. 基于 B 个估计值构建置信区间（如百分位数区间、偏差校正区间）

敏感性分析：由于序贯可忽略性假设无法检验，需要进行敏感性分析，评估结论对未观测混淆的稳健性。

13.4 中介效应的估计、检验与解读

13.4.1 效应分解：总效应、直接效应与间接（中介）效应

在中介分析中，总效应被分解为直接效应和间接效应。

效应分解公式：对于线性模型：

$$\text{总效应}：c = a \times b + c'$$
$$\text{间接效应}：a \times b$$
$$\text{直接效应}：c'$$

其中：- a: X 对 M 的效应 - b: M 对 Y 的效应（控制 X）- c': X 对 Y 的直接效应（控制 M）

效应量指标：1. 中介比例：$\frac{ab}{c} = \frac{ab}{ab+c'}$ 2. 效应大小：标准化间接效应（如完全标准化、部分标准化）

13.4.2 Bootstrap 法：检验间接效应的推荐方法

由于间接效应 ab 的乘积通常不服从正态分布，Bootstrap 法成为检验间接效应的首选方法。

百分位 **Bootstrap**：1. 从原始样本中有放回地抽取 B 个 Bootstrap 样本 2. 在每个样本中计算间接效应估计值 $\widehat{ab}^{(b)}$ 3. 将 B 个估计值从小到大排序 4. 取第 2.5 百分位数和第 97.5 百分位数作为 95% 置信区间

偏差校正 **Bootstrap**：对百分位 Bootstrap 进行偏差校正，可以提高置信区间的准确性，特别是在小样本或非对称分布的情况下。

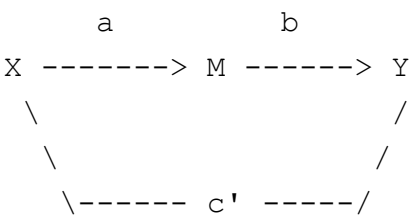
Bootstrap 样本量建议：- 至少 1000 次，推荐 5000 次 - 对于偏差校正 Bootstrap，可能需要更多次数

13.4.3 结果报告规范与图示

中介分析报告应包括：1. 描述性统计和相关矩阵 2. 各回归方程的结果（系数、标准误、显著性）3. 间接效应的点估计和置信区间 4. 直接效应和总效应的估计 5. 效应量指标（如中介比例）6. 模型检验信息（如 SEM 的拟合指数）

中介分析图示：标准的路径图应包括：1. 所有变量（X, M, Y）2. 路径系数（a, b, c’）3. 误差项 4. 必要时标注协变量

示例路径图：



13.4.4 多重中介与链式中介模型简介

多重中介模型：当有多个并行中介变量时，可以使用多重中介模型：

$$
\begin{aligned}
M_1 &= a_1 X + e_1 \\
M_2 &= a_2 X + e_2 \\
Y &= c'X + b_1 M_1 + b_2 M_2 + e_3
\end{aligned}
$$

总间接效应为：$a_1 b_1 + a_2 b_2$

链式中介模型：当中介变量之间存在序列关系时，可以使用链式中介模型：

$$
\begin{aligned}
M_1 &= a_1 X + e_1 \\
M_2 &= a_2 X + d_{21} M_1 + e_2 \\
Y &= c'X + b_1 M_1 + b_2 M_2 + e_3
\end{aligned}
$$

间接效应包括：1. 通过 M1：$a_1 b_1$ 2. 通过 M2：$a_2 b_2$ 3. 通过 M1 和 M2：$a_1 d_{21} b_2$

总间接效应为：$a_1 b_1 + a_2 b_2 + a_1 d_{21} b_2$

13.5 调节效应分析：模型、估计与展示

13.5.1 调节变量的定义与类型（分类/连续）

调节变量的定义：调节变量 W 影响自变量 X 与因变量 Y 之间关系的强度或方向。

调节变量的类型：1. 分类调节变量：如性别、种族、实验条件等 2. 连续调节变量：如年龄、收入、态度分数等 3. 类别与连续的交互：分类变量与连续变量的交互

13.5.2 含交互项的调节效应模型设定

基本调节模型：

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 X \times W + \epsilon$$

其中：- β_1: X 的主效应（当 W=0 时）- β_2: W 的主效应（当 X=0 时）- β_3: 交互效应，表示调节效应

连续调节变量的中心化：为了避免多重共线性并提高系数的可解释性，通常对连续变量进行中心化：

$$X_c = X - \bar{X}, \quad W_c = W - \bar{W}$$

然后估计模型：

$$Y = \beta_0 + \beta_1 X_c + \beta_2 W_c + \beta_3 X_c \times W_c + \epsilon$$

此时，β_1 表示在 W 取均值时，X 对 Y 的效应。

13.5.3 调节效应的图形化呈现：简单斜率分析

简单斜率分析是理解和呈现调节效应的关键工具。

简单斜率的计算：对于模型 $Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 XW + \epsilon$，X 对 Y 的简单斜率为：

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 W$$

这意味着 X 对 Y 的影响随 W 的值而变化。

简单斜率检验：检验在 W 的特定取值下，简单斜率是否显著不为零。

Johnson-Neyman 技术：确定 W 的哪些取值范围内，简单斜率是统计显著的。

13.5.4 Johnson-Neyman 区间与调节效应区域

Johnson-Neyman 技术：该方法确定调节变量 W 的" 显著性区域"，即在该区域内，X 对 Y 的简单斜率显著不为零。

计算步骤：1. 计算简单斜率的方差：$Var(\beta_1 + \beta_3 W) = Var(\beta_1) + W^2 Var(\beta_3) + 2WCov(\beta_1, \beta_3)$ 2. 构建 t 统计量：$t = \frac{\beta_1 + \beta_3 W}{\sqrt{Var(\beta_1 + \beta_3 W)}}$ 3. 解方程 $t^2 = t^2_{critical}$，得到 W 的临界值 4. 确定 W 的取值范围，使得 $|t| > t_{critical}$

结果解释：Johnson-Neyman 技术提供了调节变量 W 的取值区间，在该区间内 X 对 Y 有显著影响。这比选择几个特定点进行简单斜率检验更全面。

13.6 调节效应的估计、检验与结果解读

13.6.1 交互项系数的估计与假设检验

交互项系数的估计：使用 OLS 估计包含交互项的模型，重点关注交互项系数 β_3 的估计值和标准误。

假设检验：检验 $H_0 : \beta_3 = 0$，即不存在调节效应。如果 β_3 显著不为零，则拒绝原假设，认为存在调节效应。

注意事项：1. 即使 β_3 显著，也应结合简单斜率分析进行解释 2. 主效应 β_1 和 β_2 的解释依赖于变量的编码或中心化方式 3. 调节效应的大小应结合变量的测量尺度进行评估

13.6.2 调节效应的简单斜率检验步骤

简单斜率检验步骤：1. 选择一个或多个有理论意义的调节变量取值点（如均值、均值 ±1 标准差等）2. 计算每个点上的简单斜率：$\theta = \beta_1 + \beta_3 W_0$ 3. 计算简单斜率的方差：$Var(\theta) = Var(\beta_1) + W_0^2 Var(\beta_3) + 2W_0 Cov(\beta_1, \beta_3)$ 4. 构建 t 统计量：$t = \frac{\theta}{\sqrt{Var(\theta)}}$ 5. 进行显著性检验

简单斜率的置信区间：$\theta \pm t_{df,1-\alpha/2} \times \sqrt{Var(\theta)}$

13.6.3 如何正确解读与报告调节效应结果

调节效应的解读要点：1. 方向：调节效应是增强型（β_3 与 β_1 同号）还是削弱型（β_3 与 β_1 异号）？2. 大小：调节效应的实际大小是多少？（考虑变量的测量单位）3. 范围：在调节变量的哪些取值范围内，X 对 Y 的影响是显著的？4. 理论意义：调节效应如何支持或扩展现有理论？

结果报告应包括：1. 包含和不包含交互项的模型结果 2. 交互项系数的估计值、标准误和显著性 3. 简单斜率分析结果 4. Johnson-Neyman 显著性区域（如适用）5. 调节效应图示

13.6.4 调节效应中的多重共线性问题与处理

多重共线性问题：在包含交互项的模型中，X、W 和 X×W 之间通常存在高度相关，导致：1. 系数估计不稳定 2. 标准误增大 3. 统计检验功效降低

处理方法：1. 中心化：对连续自变量和调节变量进行中心化 - 减少 X 与 X×W、W 与 X×W 之间的相关 - 提高系数的可解释性 2. 标准化：将变量标准化为 z 分数 - 便于比较不同变量的效应大小 - 减少多重共线性 3. 岭回归或 **LASSO**：在严重多重共线性时考虑使用正则化方法 4. 增加样本量：更大的样本量可以缓解多重共线性的影响

中心化后的模型：

$$Y = \beta_0 + \beta_1(X - \bar{X}) + \beta_2(W - \bar{W}) + \beta_3(X - \bar{X})(W - \bar{W}) + \epsilon$$

此时，β_1 表示当 W 取均值时，X 对 Y 的效应；β_2 表示当 X 取均值时，W 对 Y 的效应。

13.7 整合模型：有调节的中介与有中介的调节

13.7.1 有调节的中介模型：中介路径受调节

有调节的中介模型检验中介效应是否受到调节变量的影响。即，中介路径（X→M 或 M→Y）的强度是否随调节变量 W 的变化而变化。

模型设定：有调节的中介模型有多种形式，最常见的是：1. 第一阶段调节：调节变量 W 调节 X→M 路径

$$M = a_0 + a_1X + a_2W + a_3XW + e_M$$
$$Y = b_0 + c'X + b_1M + e_Y$$

此时，中介效应为 $(a_1 + a_3W) \times b_1$，它随 W 的变化而变化。

2. 第二阶段调节：调节变量 W 调节 M→Y 路径

$$M = a_0 + a_1X + e_M$$
$$Y = b_0 + c'X + b_1M + b_2W + b_3MW + e_Y$$

此时，中介效应为 $a_1 \times (b_1 + b_3W)$。

3. 两阶段调节：W 同时调节 X→M 和 M→Y 路径

检验方法：使用 Bootstrap 法检验在不同 W 取值下的条件间接效应。

13.7.2 被中介的调节效应模型：调节作用通过中介实现

被中介的调节模型检验调节效应是否通过中介变量传递。即，X 与 W 的交互效应是否通过 M 影响 Y。

模型设定：

$$M = a_0 + a_1X + a_2W + a_3XW + e_M$$
$$Y = b_0 + c'X + b_2W + b_3XW + b_1M + e_Y$$

被中介的调节效应：如果 a_3 和 b_1 都显著，且 b_3 变得不显著或减小，则调节效应被 M 中介。

效应分解：1. 直接调节效应：b_3 2. 被中介的调节效应：$a_3 \times b_1$

13.7.3 整合模型的构建、估计与检验策略

整合模型的类型：根据 Edwards 和 Lambert（2007），整合模型可以分为：1. 第一阶段调节模型 2. 第二阶段调节模型 3. 两阶段调节模型

估计方法：1. 使用结构方程模型同时估计所有路径 2. 使用分层回归或路径分析 3. 使用 Bootstrap 法进行推断

检验策略：1. 检验调节效应：交互项系数是否显著？2. 检验中介效应：间接效应是否显著？3. 检验有调节的中介：条件间接效应在不同 W 水平下是否不同？4. 检验被中介的调节：交互效应是否通过中介变量传递？

13.7.4 整合模型的应用实例与理论贡献

整合模型的理论贡献：1. 提供更精细的理论解释 2. 揭示更复杂的因果关系模式 3. 整合不同理论视角

应用实例：例如，研究领导风格（X）对员工绩效（Y）的影响：- 中介变量：员工工作投入（M）- 调节变量：工作复杂性（W）

可以检验：1. 领导风格是否通过工作投入影响绩效？（中介）2. 这种中介效应是否受工作复杂性的调节？（有调节的中介）3. 领导风格与工作复杂的交互效应是否通过工作投入传递？（被中介的调节）

13.8 应用实践、常见误区与稳健性讨论

13.8.1 Stata 与 R 中的操作命令与流程示例

13.8.2 中介与调节分析中的常见误用与误读

常见误区：1. 忽视内生性：在核心因果关系或中介变量存在内生性时进行中介分析 2. 错误的时序：中介变量测量时间晚于结果变量 3. 过度解读：将统计上的中介效应等同于理论上的机制 4. 忽略检验前提：未检验中介分析的前提假设 5. 多重比较问题：在探索性分析中测试多个中介模型而不校正显著性水平 6. 样本量不足：在小样本中进行复杂的中介或调节分析

正确做法：1. 首先确保核心因果关系的识别 2. 基于理论选择中介和调节变量 3. 检验分析的前提假设 4. 使用适当的方法（如 Bootstrap）进行检验 5. 进行敏感性分析 6. 谨慎解释结果，考虑替代解释

13.8.3 内生性问题的挑战：中介/调节变量本身的内生性

中介变量的内生性：中介变量 M 可能存在内生性，原因包括：1. 遗漏变量同时影响 M 和 Y 2. M 的测量误差 3. M 与 Y 之间的双向因果关系

后果：中介效应估计有偏，可能：1. 高估中介效应 2. 低估中介效应 3. 错误地识别不存在的中介效应

解决方法：1. 工具变量法：为中介变量寻找工具变量 2. 固定效应模型：如果有面板数据，可以控制个体固定效应 3. 实验操纵：在实验中直接操纵中介变量 4. 敏感性分析：评估结论对未观测混淆的稳健性

13.8.4 机制分析的稳健性检验与替代解释排除

稳健性检验方法：1. 不同模型设定：尝试不同的函数形式或控制变量组合 2. 不同估计方法：比较不同方法（如 SEM、Bootstrap）的结果 3. 子样本分析：在不同子样本中检验结果的稳健性 4. 安慰剂检验：使用理论上不应有影响的变量进行” 伪中介” 分析

排除替代解释：1. 反向因果：检验 Y 对 M 的影响是否可能 2. 共同原因：寻找可能的第三变量同时影响 X、M 和 Y 3. 测量误差：评估关键变量的测量质量 4. 选择偏误：检查样本选择是否可能导致偏误

透明度要求：研究报告应：1. 明确说明所有分析的前提假设 2. 报告所有稳健性检验的结果 3. 讨论分析的局限性 4. 提供足够的信息让读者可以重复分析

本章总结与因果推断模块回顾

本章总结

本章系统学习了在确认主效应后，探究其作用机制与边界条件的方法。必须再次强调，机制分析的质量上限由核心因果关系的识别质量决定。中介与调节分析是强大的理论检验工具，但其结论的稳健性依赖于严苛的假设、精良的测量以及对未观测混淆的持续警惕。

中介效应分析帮助我们理解 X 如何通过 M 影响 Y，但面临中介变量内生性的严峻挑战。传统方法（如 Baron & Kenny 三步法）简单易用但统计功效有限，而基于反事实框架的因果中介分析提供了更严谨但假设更强的替代方案。

调节效应分析帮助我们理解 X 对 Y 的影响何时更强或更弱，通过检验 X 与 W 的交互作用来实现。正确解释调节效应需要结合简单斜率分析和 Johnson-Neyman 技术。

整合模型（如有调节的中介、被中介的调节）让我们能够检验更复杂的理论命题，但对数据质量和样本量有更高要求。

因果推断模块回顾

回顾第 6-13 章，我们完成了从理解内生性问题、掌握多种因果识别策略（IV, DID, RDD, PSM, SCM, RC），到深化因果解释（中介、调节）的完整训练。

第 **6** 章建立了因果推断的基本框架——反事实模型，并系统阐述了内生性问题及其来源。

第 **7-12** 章提供了解决内生性问题的工具箱：- 第 **7** 章：工具变量法，解决测量误差和双向因果问题 - 第 **8** 章：倾向得分匹配，解决可观测选择偏误 - 第 **9** 章：双重差分法，利用政策实施前后的变化 - 第 **10** 章：断点回归，利用制度断点创造局部随机性 - 第 **11** 章：合成控制法，为单一处理单元构造反事实 - 第 **12** 章：回归控制法，通过回归模型预测反事实

第 **13** 章：在前述方法识别出可靠因果关系的基础上，进一步探究机制（中介）和边界条件（调节）。

方法论启示

因果推断并非应用一套公式，而是基于理论、数据与方法的不断对话。研究者应像侦探一样，运用不同的工具寻找证据，同时始终保持对证据局限性的清醒认识，从而在不确定性中做出最合理的因果判断。

关键原则：1. 没有免费的午餐：每种方法都有其前提假设和局限性 2. 透明度至上：明确报告所有假设、检验和局限性 3. 稳健性检验：通过多种方法检验结果的稳健性 4. 理论指导：方法选择应由研究问题和理论指导，而非数据驱动 5. 谦虚态度：认识到因果推断的固有不确定性

通过这 8 章的学习，希望读者不仅掌握了各种因果推断方法的技术细节，更重要的是培养了严谨的因果思维习惯——在面对任何因果主张时，都会本能地追问：识别策略是什么？关键假设是什么？这些假设合理吗？有哪些证据支持或反对这些假设？这种思维习惯是进行严谨社会科学研究的核心素养。

III 理论与算法

14 大样本理论

13.1 大样本理论的基本动机

13.1.1 有限样本推断的局限性

在经典计量经济学中，我们通常基于有限样本性质（finite-sample properties）对估计量进行评价，如无偏性、有效性等。然而，有限样本理论存在以下局限性：

1. 分布假设的强依赖性：有限样本性质通常需要严格的分布假设（如正态性假设）
2. 小样本偏误：某些估计量在小样本下可能存在显著偏误
3. 精确分布难以推导：除少数简单情况外，大多数估计量的精确分布难以获得

13.1.2 渐近理论的作用与意义

大样本理论（large sample theory）或渐近理论（asymptotic theory）研究当样本容量 $n \to \infty$ 时统计量的性质，主要优势包括：

1. 放松分布假设：只需较弱的正则条件
2. 提供近似分布：通过中心极限定理获得渐近正态分布
3. 统一分析框架：适用于广泛的估计量和检验统计量

13.1.3 经济学中大样本分析的常见场景

1. 横截面数据：当样本量足够大时（通常 $n > 100$）
2. 时间序列数据：当时间跨度足够长时
3. 面板数据：当横截面维度或时间维度较大时

13.2 随机序列的收敛性

13.2.1 依概率收敛

定义 **13.1**（依概率收敛）：设 $\{X_n\}$ 是随机变量序列，X 是一个随机变量。如果对于任意 $\epsilon > 0$，有：

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

则称 X_n 依概率收敛于 X，记作 $X_n \xrightarrow{p} X$。

性质 **13.1**：若 $X_n \xrightarrow{p} a$，$Y_n \xrightarrow{p} b$，且 $g(\cdot)$ 在 (a,b) 处连续，则：1. $X_n + Y_n \xrightarrow{p} a+b$ 2. $X_nY_n \xrightarrow{p} ab$ 3. $g(X_n) \xrightarrow{p} g(a)$

13.2.2 几乎必然收敛

定义 **13.2**（几乎必然收敛）：如果：

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

则称 X_n 几乎必然收敛于 X，记作 $X_n \xrightarrow{a.s.} X$。

定理 **13.1**：几乎必然收敛强于依概率收敛，即：

$$X_n \xrightarrow{a.s.} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

13.2.3 均方收敛

定义 **13.3**（均方收敛）：如果：

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0$$

则称 X_n 均方收敛于 X，记作 $X_n \xrightarrow{m.s.} X$。

定理 **13.2**：均方收敛强于依概率收敛，即：

$$X_n \xrightarrow{m.s.} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

13.2.4 收敛关系总结

a 几乎必然收敛 (a.s. convergence);

b 均方收敛 (m.s. convergence);

c 依概率收敛 (p. convergence);

d 分布收敛 (d. convergence);

a -> c; b -> c; c -> d;

13.3 分布收敛与渐近分布

13.3.1 分布收敛的定义

定义 **13.4**（分布收敛）：设 $\{X_n\}$ 的累积分布函数为 $F_n(x)$，X 的累积分布函数为 $F(x)$。如果对于 $F(x)$ 的所有连续点 x，有：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

则称 X_n 依分布收敛于 X，记作 $X_n \xrightarrow{d} X$ 或 $X_n \rightsquigarrow X$。

13.3.2 连续映射定理

定理 **13.3**（连续映射定理，CMT）：如果 $X_n \xrightarrow{d} X$，且函数 $g(\cdot)$ 连续，则：

$$g(X_n) \xrightarrow{d} g(X)$$

更一般地，对于随机向量，如果 $(X_n, Y_n) \xrightarrow{d} (X, Y)$，且 $g(\cdot,\cdot)$ 连续，则：

$$g(X_n, Y_n) \xrightarrow{d} g(X, Y)$$

13.3.3 渐近分布的核心性质

定义 **13.5**（渐近分布）：如果 $\sqrt{n}(X_n-\theta)\xrightarrow{d}N(0,\sigma^2)$，则称 X_n 的渐近分布为：

$$X_n\sim AN\left(\theta,\frac{\sigma^2}{n}\right)$$

其中 AN 表示" 渐近正态"。

性质 **13.2**：若 $X_n\sim AN(\theta,\sigma^2/n)$，则：1. X_n 是 θ 的一致估计量 2. $\sqrt{n}(X_n-\theta)/\sigma\xrightarrow{d}N(0,1)$

13.3.4 例子：样本均值的渐近正态性

设 $X_1,\dots,X_n\sim i.i.d.(\mu,\sigma^2)$，样本均值 $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$。

由中心极限定理：

$$\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}N(0,\sigma^2)$$

因此：

$$\bar{X}_n\sim AN\left(\mu,\frac{\sigma^2}{n}\right)$$

13.4 中心极限定理及其扩展

13.4.1 Lindeberg-Lévy 中心极限定理

定理 **13.4**（Lindeberg-Lévy CLT）：设 $X_1,\dots,X_n$ 是独立同分布随机变量，$E[X_i]=\mu$，$Var(X_i)=\sigma^2<\infty$，则：

$$\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}N(0,\sigma^2)$$

等价地：

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}\xrightarrow{d}N(0,1)$$

13.4.2 Lindeberg-Feller 中心极限定理

定理 **13.5**（Lindeberg-Feller CLT）：设 $X_1,\dots,X_n$ 是独立随机变量，$E[X_i]=\mu_i$，$Var(X_i)=\sigma_i^2$。记：

$$s_n^2=\sum_{i=1}^n\sigma_i^2,\quad \bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i,\quad \bar{\mu}_n=\frac{1}{n}\sum_{i=1}^n\mu_i$$

如果满足 Lindeberg 条件：对于任意 $\epsilon > 0$，

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} E\left[(X_i - \mu_i)^2 I(|X_i - \mu_i| > \epsilon s_n)\right] = 0$$

则：

$$\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{s_n} \xrightarrow{d} N(0,1)$$

13.4.3 多元中心极限定理

定理 **13.6**（多元 CLT）：设 $\{X_i\}_{i=1}^n$ 是 k 维独立同分布随机向量，$E[X_i] = \mu$，$Cov(X_i) = \Sigma$ 正定，则：

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N_k(0, \Sigma)$$

其中 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$，$N_k(0,\Sigma)$ 表示 k 维多元正态分布。

13.4.4 在回归模型中的应用

考虑线性回归模型：

$$y_i = x_i'\beta + u_i, \quad i = 1, \dots, n$$

假设 $\{(x_i, u_i)\}$ 独立同分布，$E[u_i|x_i] = 0$，$E[u_i^2|x_i] = \sigma^2$。

OLS 估计量：

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i$$

在正则条件下：

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$$

其中 $Q = E[x_i x_i']$。

13.5 Slutsky 定理及其应用

13.5.1 Slutsky 定理的表述

定理 **13.7**（Slutsky 定理）：如果 $X_n \xrightarrow{d} X$，$Y_n \xrightarrow{p} c$（常数），则：1. $X_n + Y_n \xrightarrow{d} X + c$ 2. $X_n Y_n \xrightarrow{d} cX$ 3. 若 $c \neq 0$，则 $X_n / Y_n \xrightarrow{d} X/c$

更一般地，对于连续函数 $g(\cdot,\cdot)$：

$$g(X_n, Y_n) \xrightarrow{d} g(X, c)$$

13.5.2 估计量组合的渐近性质

例 **13.1**：设 $\hat{\theta}_n \xrightarrow{p} \theta$，$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$，则 t 统计量：

$$t_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{\sigma}_n} \xrightarrow{d} N(0,1)$$

证明：由 CLT 知 $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$，即：

$$Z_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \xrightarrow{d} N(0,1)$$

而 $\hat{\sigma}_n/\sigma \xrightarrow{p} 1$，由 Slutsky 定理：

$$t_n = \frac{Z_n}{\hat{\sigma}_n/\sigma} \xrightarrow{d} N(0,1)$$

13.5.3 渐近方差的计算与估计

Delta 方法：设 $\hat{\theta}_n$ 满足 $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma)$，$g : \mathbb{R}^k \to \mathbb{R}^m$ 在 θ 处可微，记 $G(\theta) = \frac{\partial g(\theta)}{\partial \theta'}$，则：

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} N(0, G(\theta)\Sigma G(\theta)')$$

渐近方差的估计：

$$\widehat{Avar}(g(\hat{\theta}_n)) = G(\hat{\theta}_n)\hat{\Sigma}G(\hat{\theta}_n)'/n$$

其中 $\hat{\Sigma}$ 是 Σ 的一致估计。

13.6 大样本理论在 OLS 估计中的应用

13.6.1 OLS 估计量的渐近性质

考虑线性模型：

$$y_i = x_i'\beta + u_i, \quad i = 1, \ldots, n$$

假设 **13.1**（正则条件）：1. $\{(x_i, u_i)\}_{i=1}^n$ 独立同分布 2. $E[u_i|x_i] = 0$（外生性）3. $E[u_i^2|x_i] = \sigma^2$（条件同方差）4. $Q = E[x_i x_i']$ 非奇异 5. $E[||x_i u_i||^2] < \infty$

定理 **13.8**：在假设 13.1 下：1. 一致性：$\hat{\beta}_{OLS} \xrightarrow{p} \beta$ 2. 渐近正态性：$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$ 3. 渐近有效性：$\hat{\beta}_{OLS}$ 在满足条件 1-4 的线性无偏估计类中是渐近有效的

13.6.2 异方差稳健标准误

当条件同方差不成立时，$E[u_i^2|x_i] = \sigma_i^2$。此时：

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, Q^{-1}\Omega Q^{-1})$$

其中 $\Omega = E[u_i^2 x_i x_i']$。

Eicker-Huber-White 三明治估计量：

$$\widehat{Avar}(\hat{\beta}) = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 x_i x_i'\right) \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}$$

其中 $\hat{u}_i = y_i - x_i'\hat{\beta}$。

13.6.3 渐近分布的应用：置信区间

基于渐近正态性，β_j 的 $(1-\alpha)100\%$ 渐近置信区间为：

$$\hat{\beta}_j \pm z_{1-\alpha/2} \times \widehat{se}(\hat{\beta}_j)$$

其中 $\widehat{se}(\hat{\beta}_j) = \sqrt{\widehat{Avar}(\hat{\beta}_j)/n}$，$z_{1-\alpha/2}$ 是标准正态分布的 $1-\alpha/2$ 分位数。

13.7 大样本假设检验

13.7.1 三大渐近检验

考虑检验 $H_0 : g(\theta) = 0$ vs $H_1 : g(\theta) \neq 0$，其中 $g : \mathbb{R}^k \to \mathbb{R}^m$。

定义 **13.6**：

1. 无约束估计量：$\hat{\theta}$ 最大化无约束对数似然 $\ell(\theta)$

2. 约束估计量：$\tilde{\theta}$ 最大化受约束于 $g(\theta) = 0$ 的对数似然

Wald 检验

$$W = n \cdot g(\hat{\theta})' \left[G(\hat{\theta}) \hat{I}(\hat{\theta})^{-1} G(\hat{\theta})' \right]^{-1} g(\hat{\theta}) \xrightarrow{d} \chi^2_m$$

其中 $\hat{I}(\hat{\theta})$ 是信息矩阵的估计。

似然比检验

$$LR = 2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \xrightarrow{d} \chi^2_m$$

拉格朗日乘数检验

$$LM = \left. \frac{\partial \ell(\theta)}{\partial \theta} \right|'_{\theta=\tilde{\theta}} \hat{I}(\tilde{\theta})^{-1} \left. \frac{\partial \ell(\theta)}{\partial \theta} \right|_{\theta=\tilde{\theta}} \xrightarrow{d} \chi^2_m$$

13.7.2 线性约束的 **Wald** 检验

考虑线性约束 $H_0 : R\beta = r$，其中 R 是 $m \times k$ 矩阵。

Wald 统计量：

$$W = (R\hat{\beta} - r)' [R\widehat{Avar}(\hat{\beta})R']^{-1} (R\hat{\beta} - r) \xrightarrow{d} \chi^2_m$$

特别地，当 $m = 1$ 时：

$$t = \frac{R\hat{\beta} - r}{\sqrt{R\widehat{Avar}(\hat{\beta})R'}} \xrightarrow{d} N(0,1)$$

13.7.3 模型设定检验的渐近性质

例 **13.2**（RESET 检验）：检验线性设定是否正确。

步骤：1. 估计原模型：$y = X\beta + u$ 2. 获得拟合值 $\hat{y} = X\hat{\beta}$ 3. 估计扩展模型：$y = X\beta + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + v$ 4. 检验 $H_0 : \delta_1 = \delta_2 = 0$ 使用 Wald 或 F 检验

在 H_0 下，nR^2 从辅助回归中 $\xrightarrow{d} \chi^2_2$。

13.8 自助法与大样本近似

13.8.1 自助法的基本思想

自助法（**Bootstrap**）通过重抽样来近似统计量的抽样分布。

算法 **13.1**（非参数自助法）：1. 从原始样本 $\{z_1, \dots, z_n\}$ 中有放回地抽取 n 个观测，得到自助样本 $\{z_1^*, \dots, z_n^*\}$ 2. 计算自助统计量 $\hat{\theta}^* = s(z_1^*, \dots, z_n^*)$ 3. 重复步骤 1-2 共 B 次，得到 $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ 4. 用 $\{\hat{\theta}_b^*\}$ 的经验分布近似 $\hat{\theta}$ 的抽样分布

13.8.2 参数自助法

算法 **13.2**（参数自助法）：1. 估计模型参数 $\hat{\theta}$ 2. 从分布 $F(\cdot; \hat{\theta})$ 中生成 n 个观测 $z_1^*, \dots, z_n^*$ 3. 计算 $\hat{\theta}^*$ 4. 重复 B 次

13.8.3 自助法的渐近合理性

定理 **13.9**（自助法的一致性）：在正则条件下，如果 $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$，则自助分布满足：

$$\sup_x \left| P^*(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \le x) - P(\sqrt{n}(\hat{\theta} - \theta) \le x) \right| \xrightarrow{p} 0$$

其中 P^* 表示给定原始样本下的自助分布概率。

13.8.4 自助置信区间

1. 百分位数区间：

   $$[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}]$$

   其中 $\hat{\theta}^*_{(q)}$ 是自助统计量的 q 分位数。

2. 偏差校正区间：

   $$[\hat{\theta}^*_{(\alpha_1)}, \hat{\theta}^*_{(\alpha_2)}]$$

   其中 $\alpha_1 = \Phi(2z_0 + z_{\alpha/2})$，$\alpha_2 = \Phi(2z_0 + z_{1-\alpha/2})$，$z_0 = \Phi^{-1}(\hat{F}(\hat{\theta}))$。

3. 自助 **t** 区间：

   $$\hat{\theta} \pm t^*_{1-\alpha/2} \cdot \widehat{se}(\hat{\theta})$$

   其中 $t^*_{1-\alpha/2}$ 是自助 t 统计量的 $1-\alpha/2$ 分位数。

13.9 大样本理论的局限与注意事项

13.9.1 渐近性质与实际样本量

有限样本偏差：即使估计量是一致的，小样本下仍可能有显著偏差。

例 **13.3**：动态面板数据的 Arellano-Bond 估计量：- 理论：当 T 固定，$n \to \infty$ 时一致 - 实际：当 T 较小（如 $T = 5$）时，即使 n 很大，仍可能有显著偏差

13.9.2 大样本近似的质量

收敛速度：不同估计量的收敛速度不同：- OLS 估计量：$\sqrt{n}$-收敛 - 非参数估计量：通常慢于 $\sqrt{n}$-收敛

Edgeworth 展开：用于改进渐近近似：

$$P\left(\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \leq x\right) = \Phi(x) + \frac{\phi(x)}{\sqrt{n}} g(x) + O\left(\frac{1}{n}\right)$$

其中 $g(x)$ 包含偏度和峰度信息。

13.9.3 适用条件的检验与诊断

1. 样本量足够大的判断：
   - 经验法则：$n \geq 30$ 可应用 CLT，但取决于问题复杂度
   - 模拟研究：通过蒙特卡洛模拟检查有限样本性质
2. 依赖结构的检验：
   - 时间序列：检验自相关、平稳性
   - 横截面：检验空间相关性、异方差性
3. 重尾分布的诊断：
   - 峰度系数：$\hat{\kappa} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 / \hat{\sigma}^4$
   - 若 $\hat{\kappa} > 3$（正态分布的峰度），收敛可能较慢

13.9.4 稳健推断方法

1. 异方差和自相关稳健（**HAC**）估计：

$$\hat{\Omega}_{HAC} = \sum_{j=-m}^{m} k\left(\frac{j}{m}\right) \hat{\Gamma}(j)$$

其中 $\hat{\Gamma}(j) = \frac{1}{n}\sum_{i=|j|+1}^{n} \hat{u}_i \hat{u}_{i-|j|} x_i x'_{i-|j|}$，$k(\cdot)$ 是核函数。

2. 聚类稳健标准误：当数据存在聚类结构时（如面板数据、调查数据）：

$$\widehat{Avar}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{g=1}^{G} X'_g \hat{u}_g \hat{u}'_g X_g \right) (X'X)^{-1}$$

其中 G 是聚类数。

13.9.5 实践建议

1. 报告稳健标准误：在实证研究中，应同时报告普通标准误和稳健标准误
2. 检查敏感性：对不同的渐近方差估计方法进行比较
3. 使用自助法验证：当渐近理论条件存疑时，使用自助法作为补充
4. 结合经济理论：统计显著性需结合经济意义进行解释
5. 样本量透明度：明确报告样本量，讨论其对推断可靠性的影响

本章总结

大样本理论为计量经济学提供了在有限样本分布难以获得时的推断基础。本章系统介绍了：

1. 收敛性概念：依概率收敛、几乎必然收敛、均方收敛、分布收敛及其关系
2. 核心定理：中心极限定理、Slutsky 定理、连续映射定理
3. 渐近分布理论：Delta 方法、渐近正态性
4. 应用：OLS 估计量的渐近性质、假设检验的渐近分布
5. 现代方法：自助法及其渐近合理性
6. 实践考量：大样本近似的局限性、诊断方法和稳健推断

大样本理论的重要性体现在：- 为大多数计量经济推断提供理论基础 - 允许在相对弱的条件下进行统计推断 - 支持现代计量方法的发展（如 GMM、半参数估计）

然而，研究者必须清醒认识：- 渐近性质是近似，实际样本量下可能不精确 - 收敛速度因问题和估计量而异 - 需要结合稳健方法和诊断工具

掌握大样本理论不仅有助于理解计量方法的内在逻辑，更能指导实证研究中方法的选择和结果的解释，是计量经济学理论素养的重要组成部分。

附录：关键定理证明概要

中心极限定理的直观理解

设 $X_i \sim i.i.d.(\mu, \sigma^2)$，特征函数为 $\varphi_X(t) = E[e^{itX}]$。

$\bar{X}_n$ 的特征函数：

$$\varphi_{\bar{X}_n}(t) = \left[\varphi_X\left(\frac{t}{n}\right)\right]^n = \left[1 + i\mu\frac{t}{n} - \frac{\sigma^2 + \mu^2}{2}\frac{t^2}{n^2} + o\left(\frac{1}{n^2}\right)\right]^n$$

对于 $\sqrt{n}(\bar{X}_n - \mu)$：

$$\varphi_{\sqrt{n}(\bar{X}_n - \mu)}(t) = e^{-i\mu t\sqrt{n}}\varphi_{\bar{X}_n}(\sqrt{n}t) \to e^{-\frac{1}{2}\sigma^2 t^2}$$

即正态分布的特征函数。

Slutsky 定理的证明思路

以 $X_n + Y_n \xrightarrow{d} X + c$ 为例：1. 对于任意 $\epsilon > 0$，$P(X_n + Y_n \le x) \le P(X_n \le x - c + \epsilon) + P(|Y_n - c| > \epsilon)$ 2. 取极限：$\limsup P(X_n + Y_n \le x) \le F(x - c + \epsilon)$ 3. 类似可得下界 4. 令 $\epsilon \to 0$，利用 F 的连续性得证

练习与思考题

1. 证明：若 $X_n \xrightarrow{p} X$，$Y_n \xrightarrow{p} Y$，则 $X_n + Y_n \xrightarrow{p} X + Y$。
2. 设 $\hat{\beta}_{OLS}$ 是线性回归的 OLS 估计量，推导其渐近分布，并讨论异方差情况下的调整。
3. 比较 Wald 检验、LR 检验和 LM 检验的优缺点及适用场景。
4. 设计一个蒙特卡洛实验，考察 OLS 估计量在小样本下的有限样本性质与渐近性质的差异。
5. 讨论在大数据时代（n 很大但 p 也很大）大样本理论面临的挑战。

15 最大似然估计理论

本章导读

最大似然估计法是现代计量经济学的核心方法论之一，它为参数估计和统计推断提供了一个统一而强大的理论框架。本章将系统介绍最大似然估计的基本原理、统计性质、计算方法及其在计量经济学中的重要应用。

通过本章学习，您将能够：1. 理解最大似然估计的基本思想与哲学基础 2. 掌握似然函数和对数似然函数的构建方法 3. 理解 MLE 的大样本性质及其证明思路 4. 掌握基于 MLE 的三大假设检验方法 5. 了解 MLE 在非线性计量模型中的应用 6. 能够使用统计软件实现 MLE 估计并解释结果

本章需要读者具备概率论、数理统计和矩阵代数的基本知识，特别是关于概率分布、期望、方差、协方差和矩阵求导等内容。

14.1 最大似然估计的基本原理

14.1.1 直观思想

最大似然估计的基本思想可以用一个简单的例子说明：假设我们有一个硬币，想要估计它正面朝上的概率 p。我们抛掷 10 次，观察到 7 次正面。那么，什么样的 p 值最有可能产生这样的观察结果呢？

形式上，对于参数 θ 和观测数据 $y = (y_1, y_2, \ldots, y_n)$，我们寻找使得观测数据出现概率最大的参数值：

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L(\theta; y)$$

其中 Θ 是参数空间，$L(\theta; y)$ 是似然函数。

14.1.2 似然函数与对数似然函数

设随机变量 Y 的概率密度函数（连续情形）或概率质量函数（离散情形）为 $f(y;\theta)$，其中 θ 是未知参数向量。对于独立同分布的样本 $y_1, y_2, \dots, y_n$，似然函数定义为：

$$L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)$$

在实际计算中，我们通常使用对数似然函数：

$$\ell(\theta; y) = \ln L(\theta; y) = \sum_{i=1}^{n} \ln f(y_i; \theta)$$

取对数的原因：1. 将乘积转化为求和，简化计算 2. 许多分布的对数形式更简单 3. 不改变极值点的位置（因为对数函数是单调递增的）

14.1.3 一个简单例子：正态分布均值的 **MLE**

假设 $Y_i \sim N(\mu, \sigma^2)$，$\sigma^2$ 已知，$i = 1, \dots, n$。似然函数为：

$$L(\mu; y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

对数似然函数为：

$$\ell(\mu; y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

最大化 $\ell(\mu; y)$ 等价于最小化 $\sum_{i=1}^{n}(y_i - \mu)^2$，得到：

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

这个结果与我们的直觉一致：样本均值是总体均值的最佳估计。

14.2 MLE 的求解与计算方法

14.2.1 一阶条件与似然方程

MLE 估计量 $\hat{\theta}_{MLE}$ 满足一阶条件：

$$\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MLE}} = 0$$

这个方程组称为似然方程或得分方程。其中，

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$$

称为得分函数（Score Function）。

14.2.2 信息矩阵

Fisher 信息矩阵衡量了似然函数的曲率，定义为：

$$I(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}\right]$$

在 i.i.d. 样本下，$I(\theta) = n \cdot \mathcal{I}(\theta)$，其中 $\mathcal{I}(\theta)$ 是单个观测的信息矩阵。

观测信息矩阵为：

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}$$

14.2.3 数值优化方法

对于大多数计量经济学模型，似然方程没有解析解，需要数值方法求解：

1. 牛顿**-**拉夫森法（**Newton-Raphson Method**）

迭代公式：

$$\theta^{(k+1)} = \theta^{(k)} - \left[\left. \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} \right|_{\theta=\theta^{(k)}} \right]^{-1} \left. \frac{\partial \ell(\theta)}{\partial \theta} \right|_{\theta=\theta^{(k)}}$$

2. 得分算法（**Method of Scoring**）

使用期望信息矩阵代替观测信息矩阵：

$$\theta^{(k+1)} = \theta^{(k)} + \left[I(\theta^{(k)}) \right]^{-1} s(\theta^{(k)})$$

3. BHHH 算法（**Berndt-Hall-Hall-Hausman**）

基于外积估计信息矩阵：

$$\theta^{(k+1)} = \theta^{(k)} + \lambda_k \left[\sum_{i=1}^{n} s_i(\theta^{(k)}) s_i(\theta^{(k)})' \right]^{-1} s(\theta^{(k)})$$

其中 $s_i(\theta) = \frac{\partial \ln f(y_i;\theta)}{\partial \theta}$ 是第 i 个观测的得分。

14.2.4 收敛准则与初始值选择

数值优化需要设定收敛准则：1. 参数变化：$\|\theta^{(k+1)} - \theta^{(k)}\| < \epsilon_1$ 2. 函数值变化：$|\ell(\theta^{(k+1)}) - \ell(\theta^{(k)})| < \epsilon_2$ 3. 梯度范数：$\|s(\theta^{(k)})\| < \epsilon_3$

初始值 $\theta^{(0)}$ 的选择至关重要，常用方法包括：- 使用简单的矩估计作为初始值 - 使用简化模型的估计结果 - 网格搜索法

14.3 MLE 的统计性质

14.3.1 正则条件

为保证 MLE 具有良好的大样本性质，需要以下正则条件：

1. 识别条件：不同的 θ 值对应不同的分布
2. 紧参数空间：Θ 是紧集
3. 连续性：$\ln f(y;\theta)$ 关于 θ 连续

4. 可微性：$\ln f(y;\theta)$ 关于 θ 三阶连续可微
5. 可积性：期望 $E[\ln f(y;\theta)]$ 存在且有限
6. 信息矩阵正定：$I(\theta)$ 有限且正定

14.3.2 一致性

在正则条件下，MLE 估计量具有一致性：

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0 \quad \text{当} \quad n \to \infty$$

其中 θ_0 是真实参数值。

14.3.3 渐近正态性

MLE 估计量具有渐近正态性：

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

等价地，

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \frac{1}{n} I(\theta_0)^{-1}\right)$$

在实践中，我们用估计的信息矩阵代替 $I(\theta_0)$：

$$\widehat{Var}(\hat{\theta}_{MLE}) = \left[-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} \bigg|_{\theta=\hat{\theta}_{MLE}} \right]^{-1}$$

14.3.4 渐近有效性（**Cramér-Rao** 下界）

MLE 达到了 Cramér-Rao 下界，即在所有一致且渐近正态的估计量中，MLE 的渐近方差最小。对于任意无偏估计量 $\tilde{\theta}$：

$$Var(\tilde{\theta}) \geq I(\theta)^{-1}$$

MLE 的方差恰好等于这个下界（渐近意义上）。

14.3.5 不变性原理

MLE 具有不变性：如果 $\hat{\theta}$ 是 θ 的 MLE，且 $g(\cdot)$ 是一对一函数，那么 $g(\hat{\theta})$ 是 $g(\theta)$ 的 MLE。即使 $g(\cdot)$ 不是一对一函数，这个性质在一定条件下仍然成立。

14.4 基于 **MLE** 的假设检验

14.4.1 三大检验方法

1. 似然比检验（**Likelihood Ratio Test, LRT**）

比较有约束模型和无约束模型的似然函数最大值：

$$LR = 2[\ell(\hat{\theta}_u) - \ell(\hat{\theta}_r)] \sim \chi^2(q)$$

其中：- $\ell(\hat{\theta}_u)$：无约束模型的对数似然最大值 - $\ell(\hat{\theta}_r)$：有约束模型的对数似然最大值 - q：约束条件的个数

2. 沃尔德检验（**Wald Test**）

直接检验约束条件 $H_0 : R\theta = r$：

$$W = (R\hat{\theta} - r)'[R\widehat{Var}(\hat{\theta})R']^{-1}(R\hat{\theta} - r) \sim \chi^2(q)$$

优点：只需估计无约束模型。

3. 拉格朗日乘子检验（**Lagrange Multiplier Test, LM**）或得分检验（**Score Test**）

基于约束模型下的得分函数：

$$LM = s(\tilde{\theta})'I(\tilde{\theta})^{-1}s(\tilde{\theta}) \sim \chi^2(q)$$

其中 $\tilde{\theta}$ 是在 H_0 约束下的 MLE。优点：只需估计约束模型。

14.4.2 三种检验的比较

检验方法	需要估计的模型	计算复杂度	小样本性质	对重新参数化的不变性
LRT	无约束和约束模型	高	较好	不变
Wald	仅无约束模型	低	一般	可变
LM	仅约束模型	中等	一般	不变

14.4.3 模型选择准则

对于非嵌套模型，使用信息准则：

1. **Akaike** 信息准则（**AIC**）：
$$AIC = -2\ell(\hat{\theta}) + 2k$$
其中 k 是参数个数。

2. 贝叶斯信息准则（**BIC**）：
$$BIC = -2\ell(\hat{\theta}) + k\ln n$$

选择 AIC 或 BIC 最小的模型。

14.5 MLE 在计量经济学中的应用

14.5.1 经典线性回归模型

假设 $y_i = x_i'\beta + \varepsilon_i$，$\varepsilon_i \sim N(0, \sigma^2)$，则对数似然函数为：

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

MLE 估计量为：

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2$$

注意：$\hat{\sigma}^2_{MLE}$ 是有偏估计，通常使用 $\frac{n}{n-k}\hat{\sigma}^2_{MLE}$ 作为无偏估计。

14.5.2 离散选择模型

Logit 模型：

假设 $P(y_i = 1|x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$，对数似然函数为：

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln \Lambda(x_i'\beta) + (1 - y_i) \ln(1 - \Lambda(x_i'\beta))]$$

其中 $\Lambda(z) = \frac{\exp(z)}{1+\exp(z)}$。

Probit 模型：

假设 $P(y_i = 1|x_i) = \Phi(x_i'\beta)$，其中 $\Phi(\cdot)$ 是标准正态分布函数。

14.5.3 计数数据模型

泊松回归模型：

假设 $y_i|x_i \sim Poisson(\lambda_i)$，$\lambda_i = \exp(x_i'\beta)$，对数似然函数为：

$$\ell(\beta) = \sum_{i=1}^{n} [y_i(x_i'\beta) - \exp(x_i'\beta) - \ln(y_i!)]$$

负二项回归：

用于处理过度离散问题，方差大于均值的情况。

14.5.4 受限因变量模型

Tobit 模型（**Type I**）：

适用于归并数据（censored data）：

$$y_i^* = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \le 0 \end{cases}$$

对数似然函数由两部分组成：

$$\ell(\beta, \sigma) = \sum_{y_i=0} \ln \Phi\left(-\frac{x_i'\beta}{\sigma}\right) + \sum_{y_i>0} \left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right]$$

14.5.5 时间序列模型

ARMA(p,q) 模型：

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \quad \varepsilon_t \sim N(0, \sigma^2)$$

使用条件似然或精确似然方法估计。

GARCH 模型：

用于波动率建模：

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

14.6 实践中的问题与扩展

14.6.1 准最大似然估计（QMLE）

当分布假设错误时，MLE 可能不一致。但若条件均值设定正确，QMLE 仍可得到条件均值参数的一致估计（在广义线性模型框架下）。此时需要使用稳健标准误（Huber-White 标准误）。

14.6.2 数值问题

1. 局部极大值：对数似然函数可能有多个极值点
2. 平坦区域：信息矩阵接近奇异，估计不精确

3. 边界解：估计值落在参数空间边界
4. 收敛失败：迭代算法不收敛

应对策略：尝试不同初始值、重新参数化、使用全局优化算法。

14.6.3 缺失数据与 **EM** 算法

当数据存在缺失时，可以使用期望最大化（EM）算法：1. **E** 步：计算完全数据对数似然的条件期望 2. **M** 步：最大化这个期望

14.6.4 贝叶斯方法与 **MLE** 的关系

贝叶斯估计将参数视为随机变量，使用后验分布进行推断。当先验分布是均匀分布时，后验众数等于 MLE 估计量。在大样本下，贝叶斯后验分布近似正态，中心在 MLE 估计量处。

14.7 应用案例：工资方程的 **MLE** 估计

本章总结

最大似然估计法是计量经济学中最重要的估计方法之一，具有坚实的理论基础和广泛的应用价值。本章系统地介绍了：

1. 基本原理：MLE 通过最大化似然函数寻找最可能产生观测数据的参数值，其核心是似然函数和对数似然函数。

2. 计算方法：对于简单模型有解析解，复杂模型需要数值优化算法（牛顿-拉夫森法、BHHH 算法等）。

3. 统计性质：在正则条件下，MLE 具有一致性、渐近正态性和渐近有效性，达到了 Cramér-Rao 下界。

4. 假设检验：基于 MLE 的三大检验方法（似然比检验、沃尔德检验、得分检验）为模型设定检验提供了系统工具。

5. 应用领域：MLE 是离散选择模型、计数模型、受限因变量模型、时间序列模型等非线性计量模型的标准估计方法。

6. 实践问题：需要关注分布假设的合理性、数值计算的稳定性、模型设定的正确性等问题。

MLE 的魅力在于它提供了一个统一的框架来处理各种复杂的计量经济学问题。然而，在实际应用中，研究者必须谨慎对待其前提假设，正确解释估计结果，并理解各种检验方法的适用条件。

随着计算技术的发展，MLE 的应用范围不断扩大，特别是在处理高维数据、复杂数据结构和非标准模型方面。掌握 MLE 不仅有助于理解经典计量方法，也为学习更高级的计量经济学方法（如广义矩方法、半参数和非参数方法）奠定了坚实基础。

关键术语

- 最大似然估计（Maximum Likelihood Estimation, MLE）
- 似然函数（Likelihood Function）
- 对数似然函数（Log-Likelihood Function）
- 得分函数（Score Function）
- 信息矩阵（Information Matrix）
- 渐近正态性（Asymptotic Normality）
- 似然比检验（Likelihood Ratio Test）
- 沃尔德检验（Wald Test）
- 得分检验（Score Test）
- Cramér-Rao 下界（Cramér-Rao Lower Bound）
- 准最大似然估计（Quasi-MLE）
- EM 算法（Expectation-Maximization Algorithm）

思考与练习

1. 证明正态分布方差 σ^2 的 MLE 估计量是有偏的，并推导其偏差。
2. 比较 MLE 与矩估计法（MM）的优缺点。
3. 推导 Logit 模型的得分函数和信息矩阵。
4. 在 Tobit 模型中，解释系数 β 的经济含义与线性回归模型有何不同？
5. 当对数似然函数有多个局部极大值时，如何寻找全局最大值？
6. 讨论 MLE 在小样本下的性质及其改进方法。

16 广义矩估计法

本章导读

计量经济学的演进如同一棵知识之树，从最初的最小二乘法这一主干，逐渐生长出处理内生性的工具变量法、基于分布假设的最大似然法等多个分支。然而，这些看似迥异的方法背后，隐藏着深刻的统一性逻辑。1982 年，拉尔斯 • 彼得 • 汉森提出的广义矩方法（GMM）正是揭示这一统一性的关键框架，它将各种估计方法置于共同的矩条件基础之上。

广义矩方法的精妙之处在于其哲学思辨：任何参数估计问题本质上都是寻找使样本矩条件接近总体矩条件的参数值。这一思想不仅统一了传统方法，更为处理复杂的经济计量问题——从资产定价到动态面板，从宏观时间序列到微观因果推断——提供了灵活而强大的工具。

本章将引领您完成一次从具体到抽象、再从抽象回到具体的思维旅程。我们将首先以全新视角重新审视 OLS、2SLS 和 MLE，揭示它们共有的矩条件本质；然后系统构建 GMM 的一般理论框架；接着探讨其在各类模型中的应用与实践挑战；最后展望前沿发展。通过本章学习，您将掌握：

1. 将传统估计方法统一表述为 GMM 特例的能力
2. GMM 估计的完整实施流程与统计推断方法
3. 在实际研究中恰当运用 GMM 解决内生性、动态性等问题的技能
4. 对估计方法演进逻辑的深刻理解

让我们开始这次统一性探索之旅。

15.1 回顾：传统估计方法的矩条件视角

OLS 的最小二乘条件：正交性的矩表达

考虑经典线性回归模型：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n$$

OLS 估计量 $\hat{\boldsymbol{\beta}}_{OLS}$ 通过最小化残差平方和获得，其一阶条件为：

$$\sum_{i=1}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0$$

这一条件可重述为样本矩条件：

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0$$

对应的总体矩条件为：

$$E[\mathbf{x}_i\varepsilon_i] = E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}_0)] = 0$$

关键洞见：OLS 本质上是求解 k 个矩条件的系统，每个条件对应一个解释变量与误差项的正交性要求。当 $E[\mathbf{x}_i\varepsilon_i] = 0$ 成立时，我们获得了参数的一致估计。

2SLS 的工具变量条件：外生性的矩约束

当解释变量存在内生性时，工具变量法应运而生。设内生模型：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad E[\mathbf{x}_i\varepsilon_i] \neq 0$$

引入工具变量 $\mathbf{z}_i$ 满足：1. 相关性：$Cov(\mathbf{z}_i, \mathbf{x}_i) \neq 0$ 2. 外生性：$Cov(\mathbf{z}_i, \varepsilon_i) = 0$

外生性条件可表述为矩条件：

$$E[\mathbf{z}_i\varepsilon_i] = E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}_0)] = 0$$

设 $\dim(\mathbf{z}_i) = L \geq k = \dim(\boldsymbol{\beta})$，则我们有 L 个矩条件。当 $L = k$ 时，系统恰好识别；当 $L > k$ 时，系统过度识别，需要特殊处理来平衡这 L 个条件。

统一性洞察：2SLS 可视为特定权重矩阵下的 GMM 估计量。定义矩条件函数 $g(\mathbf{w}_i, \boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$，选择权重矩阵 $\mathbf{W}_n = (\frac{1}{n}\sum \mathbf{z}_i\mathbf{z}_i')^{-1}$，则 GMM 解即为 2SLS 估计量。

MLE 的 score 函数条件：似然框架的矩表述

设观测数据 $\mathbf{w}_i$ 来自参数分布 $f(\mathbf{w}_i; \boldsymbol{\theta})$。最大似然估计最大化对数似然函数：

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{w}_i; \boldsymbol{\theta})$$

一阶条件（score 函数）为：

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

这等价于样本矩条件：

$$\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{w}_i;\boldsymbol{\theta}) = 0, \quad s(\mathbf{w}_i;\boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

对应的总体矩条件为：

$$E[s(\mathbf{w}_i;\boldsymbol{\theta}_0)] = 0$$

这一条件在模型设定正确时必然成立，因为 score 函数的期望为零。

深刻联系：MLE 可视为使用 score 函数作为矩条件、并以信息矩阵的逆作为最优权重矩阵的 GMM。当矩条件来自 score 函数且使用最优权重时，GMM 达到与 MLE 相同的渐近效率。

三种方法的矩条件统一表述

方法	矩条件函数 $g(\mathbf{w}_i,\boldsymbol{\theta})$	矩条件数 q	参数数 p	识别状态	关键假设
OLS	$\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$	k	k	恰好识别	$E[\mathbf{x}_i\varepsilon_i] = 0$
2SLS	$\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$	L	k	$L \geq k$	$E[\mathbf{z}_i\varepsilon_i] = 0$
MLE	$\frac{\partial \ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$	p	p	恰好识别	分布 $f(\cdot;\boldsymbol{\theta})$ 正确设定

统一性证明：对于恰好识别情形（$q = p$），三类方法均可表示为求解方程组：

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{w}_i,\boldsymbol{\theta}) = 0$$

解的唯一性保证了估计的一致性。对于过度识别情形（如 $L > k$ 的 2SLS），GMM 通过最小化加权二次型来平衡多个矩条件。

教学启示：这一统一视角揭示了计量估计的本质——寻找满足特定矩条件的参数值。不同方法的区别仅在于矩条件的选择和数量，而非根本原理。这种理解为我们构建更一般的估计框架奠定了基础。

15.2 广义矩方法的基本框架

矩条件的一般形式：从特殊到一般的升华

广义矩方法始于一组矩条件函数：

$$g(\mathbf{w}_i, \boldsymbol{\theta}) = \begin{pmatrix} g_1(\mathbf{w}_i, \boldsymbol{\theta}) \\ \vdots \\ g_q(\mathbf{w}_i, \boldsymbol{\theta}) \end{pmatrix}$$

其中 $\mathbf{w}_i$ 为第 i 个观测值，$\boldsymbol{\theta}$ 为 $p \times 1$ 的未知参数向量，g 为 $q \times 1$ 的向量函数。

总体矩条件假设在参数真值 $\boldsymbol{\theta}_0$ 处满足：

$$E[g(\mathbf{w}_i, \boldsymbol{\theta}_0)] = 0$$

样本矩条件为其经验对应：

$$\bar{g}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{w}_i, \boldsymbol{\theta})$$

根据大数定律，当 $n \to \infty$ 时，$\bar{g}_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$。

识别维度分析：- 恰好识别：$q = p$，方程有唯一解 - 过度识别：$q > p$，通常无精确解，需” 平衡” 条件 - 识别不足：$q < p$，无法唯一确定参数

GMM 估计量的定义：统一框架的构建

在过度识别情形下，GMM 通过最小化矩条件的加权二次型来估计参数：

目标函数：

$$J_n(\boldsymbol{\theta}) = \bar{g}_n(\boldsymbol{\theta})' \mathbf{W}_n \bar{g}_n(\boldsymbol{\theta})$$

其中 $\mathbf{W}_n$ 为 $q \times q$ 的对称正定权重矩阵。

GMM 估计量：

$$\hat{\boldsymbol{\theta}}_{GMM} = \arg\min_{\boldsymbol{\theta} \in \Theta} J_n(\boldsymbol{\theta})$$

权重矩阵的三重作用：1. 标准化：平衡不同量纲的矩条件 2. 效率化：通过最优选择实现最小渐近方差 3. 数值稳定化：改善优化问题的条件数

关键性质：在恰好识别时（$q = p$），只要 $\mathbf{W}_n$ 可逆，估计量不依赖于权重矩阵的选择，因为解满足 $\bar{g}_n(\hat{\boldsymbol{\theta}}) = 0$，从而 $J_n(\hat{\boldsymbol{\theta}}) = 0$。

识别条件：估计一致性的基石

阶条件（必要条件）

矩条件数量不少于参数数量：$q \geq p$

秩条件（充分条件）

矩条件函数的雅可比矩阵在真值处列满秩。定义：

$$\mathbf{G}(\boldsymbol{\theta}) = E\left[\frac{\partial g(\mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]$$

要求在 $\boldsymbol{\theta}_0$ 处，$\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_0)$ 为 $q \times p$ 矩阵，且 $\text{rank}(\mathbf{G}) = p$。

全局与局部识别

- 局部识别：在 $\boldsymbol{\theta}_0$ 的邻域内唯一性
- 全局识别：在整个参数空间 Θ 内的唯一性

GMM 理论通常要求局部识别，而经济解释需要全局识别。非线性模型可能只满足局部识别条件。

弱识别问题

当 $\mathbf{G}(\boldsymbol{\theta})$ 接近降秩时，即使满足秩条件，有限样本性质也可能很差。这在工具变量较弱时尤为常见，需要专门的诊断和稳健推断方法。

统一框架下的传统方法再阐释

OLS 的 **GMM** 表述

矩条件：$g^{OLS}(\mathbf{w}_i, \boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$

识别状态：恰好识别（$q = k = p$）

解：$\hat{\boldsymbol{\beta}}_{GMM} = (n^{-1}\sum \mathbf{x}_i\mathbf{x}_i')^{-1}(n^{-1}\sum \mathbf{x}_i y_i) = \hat{\boldsymbol{\beta}}_{OLS}$

2SLS 的 **GMM** 表述

矩条件：$g^{2SLS}(\mathbf{w}_i, \boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$

识别状态：$L = k$ 时恰好识别，$L > k$ 时过度识别

权重矩阵：$\mathbf{W}_n = (n^{-1}\sum \mathbf{z}_i\mathbf{z}_i')^{-1}$（对应传统 2SLS）

MLE 的 **GMM** 表述

矩条件：$g^{MLE}(\mathbf{w}_i, \boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

识别状态：恰好识别（$q = p$）

最优权重：$\mathbf{W}_n^* = [Var(g(\mathbf{w}_i, \boldsymbol{\theta}_0))]^{-1} = \mathcal{I}(\boldsymbol{\theta}_0)^{-1}$

教学洞见：GMM 框架的威力在于其包容性。它不替代传统方法，而是提供一个统一视角来理解它们。这种理解有助于学生在面对新问题时，能够灵活构造适当的矩条件，而非机械套用现成方法。

15.3 GMM 的统计性质

一致性：大样本下的确定性

在以下正则条件下，GMM 估计量是一致的：

1. 参数识别：$\boldsymbol{\theta}_0$ 是唯一满足 $E[g(\mathbf{w}_i, \boldsymbol{\theta})] = 0$ 的参数值
2. 紧参数空间：Θ 是紧集
3. 矩条件连续性：$g(\mathbf{w}, \boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 连续
4. 一致收敛：$\sup_{\boldsymbol{\theta}\in\Theta} \|\bar{g}_n(\boldsymbol{\theta}) - E[g(\mathbf{w}_i, \boldsymbol{\theta})]\| \xrightarrow{p} 0$
5. 权重矩阵收敛：$\mathbf{W}_n \xrightarrow{p} \mathbf{W}$，$\mathbf{W}$ 正定

在这些条件下：

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0$$

证明思路：1. 样本矩条件一致收敛于总体矩条件（均匀大数定律）2. 目标函数一致收敛：$J_n(\boldsymbol{\theta}) \xrightarrow{p} J(\boldsymbol{\theta}) = E[g(\mathbf{w}_i, \boldsymbol{\theta})]'\mathbf{W}E[g(\mathbf{w}_i, \boldsymbol{\theta})]$ 3. $J(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 处唯一最小（识别条件）4. 应用极值估计量一致性定理

渐近正态性：分布形态的刻画

附加光滑性条件后，GMM 估计量具有渐近正态性：

定理（**GMM** 渐近分布）：假设 1. $\boldsymbol{\theta}_0$ 位于 Θ 内部 2. $g(\mathbf{w},\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 邻域内连续可微 3. $\mathbf{G}(\boldsymbol{\theta}) = E\left[\frac{\partial g(\mathbf{w}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]$ 在 $\boldsymbol{\theta}_0$ 处连续 4. 中心极限定理适用：$\sqrt{n}\bar{g}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0,\boldsymbol{\Omega})$

则：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0,\mathbf{V})$$

其中：

$$\mathbf{V} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}$$

$\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_0)$，$\boldsymbol{\Omega} = \lim_{n\to\infty} Var(\sqrt{n}\bar{g}_n(\boldsymbol{\theta}_0))$

证明要点：1. 一阶条件泰勒展开：

$$0 = \frac{\partial J_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \approx 2\mathbf{G}_n(\tilde{\boldsymbol{\theta}})'\mathbf{W}_n\bar{g}_n(\boldsymbol{\theta}_0) + 2\mathbf{G}_n(\tilde{\boldsymbol{\theta}})'\mathbf{W}_n\mathbf{G}_n(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

2. 重新整理：

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\mathbf{G}_n(\tilde{\boldsymbol{\theta}})'\mathbf{W}_n\mathbf{G}_n(\tilde{\boldsymbol{\theta}})]^{-1}\mathbf{G}_n(\tilde{\boldsymbol{\theta}})'\mathbf{W}_n\sqrt{n}\bar{g}_n(\boldsymbol{\theta}_0)$$

3. 应用 Slutsky 定理和中心极限定理

效率与最优 **GMM**：方差最小化的追求

最优权重矩阵理论

定理（最优权重）：在所有使用相同矩条件的 GMM 估计量中，选择 $\mathbf{W}^* = \boldsymbol{\Omega}^{-1}$ 可得到最小渐近方差：

$$\mathbf{V}_{opt} = (\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}$$

证明：对任意 $\mathbf{W}$，考虑差矩阵：

$$\mathbf{V} - \mathbf{V}_{opt} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}$$

其中 $\boldsymbol{\Sigma} = \boldsymbol{\Omega} - \mathbf{G}(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}\mathbf{G}'$ 半正定。通过代数运算可证 $\mathbf{V} - \mathbf{V}_{opt}$ 半正定。

两步骤 **GMM** 实现

由于 $\mathbf{\Omega}$ 未知，实践中采用两步骤法：

步骤 **1**：获取初步估计 $\tilde{\boldsymbol{\theta}}$，使用初始权重 $\mathbf{W}_n^{(1)}$（如 $\mathbf{I}_q$ 或 $(n^{-1}\sum \mathbf{z}_i\mathbf{z}_i')^{-1}$）

步骤 **2**：计算 $\mathbf{\Omega}$ 的估计：

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{w}_i, \tilde{\boldsymbol{\theta}})g(\mathbf{w}_i, \tilde{\boldsymbol{\theta}})'$$

以 $\hat{\mathbf{W}}_n = \hat{\mathbf{\Omega}}^{-1}$ 重新估计，得到 $\hat{\boldsymbol{\theta}}_{GMM}$

迭代与连续更新 **GMM**

- 迭代 **GMM**：反复执行步骤 2 直至收敛
- 连续更新 **GMM (CUE)**：同时优化参数和权重矩阵：

$$\hat{\boldsymbol{\theta}}_{CUE} = \arg\min_{\boldsymbol{\theta}} \bar{g}_n(\boldsymbol{\theta})'\hat{\mathbf{\Omega}}(\boldsymbol{\theta})^{-1}\bar{g}_n(\boldsymbol{\theta})$$

其中 $\hat{\mathbf{\Omega}}(\boldsymbol{\theta}) = n^{-1}\sum g(\mathbf{w}_i, \boldsymbol{\theta})g(\mathbf{w}_i, \boldsymbol{\theta})'$

假设检验：模型设定的评估

过度识别检验（**Hansen's J** 检验）

检验所有 q 个矩条件是否成立：

原假设：$H_0 : E[g(\mathbf{w}_i, \boldsymbol{\theta}_0)] = 0$

检验统计量：

$$J_n = n\bar{g}_n(\hat{\boldsymbol{\theta}}_{GMM})'\hat{\mathbf{\Omega}}^{-1}\bar{g}_n(\hat{\boldsymbol{\theta}}_{GMM})$$

渐近分布：$J_n \xrightarrow{d} \chi^2_{q-p}$ under H_0

解释：大 J_n 值表明矩条件可能不成立，但无法指出具体哪些条件有问题。

矩条件子集检验（**C** 统计量）

将矩条件分为 $g = (g_1', g_2')'$，检验 $H_0 : E[g_2(\mathbf{w}_i, \boldsymbol{\theta}_0)] = 0$，已知 $E[g_1(\mathbf{w}_i, \boldsymbol{\theta}_0)] = 0$

C 统计量：

$$C_n = J_n^{UR} - J_n^R \xrightarrow{d} \chi^2_{q_2}$$

其中 J_n^{UR} 使用所有矩条件，J_n^R 仅使用 g_1，$q_2 = \dim(g_2)$

参数约束检验（**Wald** 检验）

检验线性约束 $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$：

Wald 统计量：

$$W_n = n(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})'[\mathbf{R}\hat{\mathbf{V}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}) \xrightarrow{d} \chi^2_s$$

其中 $s = \text{rank}(\mathbf{R})$

统一视角下的传统方法性质

在 GMM 框架下，传统方法的渐近性质获得统一表述：

OLS 的渐近方差

$$\mathbf{V}_{OLS} = (E[\mathbf{x}_i\mathbf{x}_i'])^{-1}E[\mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2](E[\mathbf{x}_i\mathbf{x}_i'])^{-1}$$

同方差时简化为 $\sigma^2(E[\mathbf{x}_i\mathbf{x}_i'])^{-1}$

2SLS 的渐近方差

$$\mathbf{V}_{2SLS} = (E[\mathbf{z}_i\mathbf{x}_i']'\mathbf{W}E[\mathbf{z}_i\mathbf{x}_i'])^{-1}E[\mathbf{z}_i\mathbf{x}_i']'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}E[\mathbf{z}_i\mathbf{x}_i'](E[\mathbf{z}_i\mathbf{x}_i']'\mathbf{W}E[\mathbf{z}_i\mathbf{x}_i'])^{-1}$$

其中 $\mathbf{W} = (E[\mathbf{z}_i\mathbf{z}_i'])^{-1}$，$\boldsymbol{\Omega} = E[\mathbf{z}_i\mathbf{z}_i'\varepsilon_i^2]$

MLE 的渐近方差

$$\mathbf{V}_{MLE} = \mathcal{J}(\boldsymbol{\theta}_0)^{-1}$$

这是 Cramér-Rao 下界，体现了 MLE 在正确设定下的最优性。

总结：GMM 不仅统一了估计量的构造，也统一了它们的渐近性质。所有估计量的一致性都源于矩条件的正确设定，渐近正态性都来自中心极限定理，效率差异则源于权重矩阵的选择。这种统一视角极大简化了计量理论的学习和理解。

15.4 GMM 的具体应用

线性模型的 GMM 估计：从传统到一般

OLS 的异方差稳健形式

当存在异方差时，传统 OLS 标准误失效。在 GMM 框架下，我们使用更一般的协方差矩阵估计：

矩条件：$g(\mathbf{w}_i, \boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$

最优权重矩阵：

$$\hat{\boldsymbol{\Omega}}^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\varepsilon}_i^2 \right)^{-1}$$

其中 $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{OLS}$。对应的协方差估计为：

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\varepsilon}_i^2 \right) \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}$$

这正是 Eicker-Huber-White 异方差稳健标准误。

2SLS 的最优 GMM 形式

传统 2SLS 对应权重矩阵 $\mathbf{W}_n = (n^{-1} \sum \mathbf{z}_i \mathbf{z}_i')^{-1}$。当存在异方差时，最优 GMM 使用：

$$\hat{\boldsymbol{\Omega}}^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' \hat{\varepsilon}_i^2 \right)^{-1}$$

其中 $\hat{\varepsilon}_i$ 来自第一步 2SLS 残差。这产生了比传统 2SLS 更有效的估计。

非线性模型的 **GMM** 估计：超越线性框架

消费资本资产定价模型（**CCAPM**）

CCAPM 的欧拉方程提供天然矩条件。对于资产 j：

$$E\left[\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{j,t+1} - 1 \middle| \mathcal{I}_t\right] = 0$$

基于工具变量 $\mathbf{z}_t \in \mathcal{I}_t$，构造矩条件：

$$g(\mathbf{w}_t, \boldsymbol{\theta}) = \left[\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \mathbf{R}_{t+1} - \mathbf{1}\right] \otimes \mathbf{z}_t$$

其中 $\boldsymbol{\theta} = (\delta, \gamma)'$，$\mathbf{R}_{t+1}$ 为资产回报向量，$\otimes$ 为 Kronecker 积。

估计步骤：1. 选择工具变量（消费增长和回报的滞后项）2. 构造样本矩条件 3. 应用 GMM 估计 (δ, γ) 4. J 检验评估模型设定

动态面板数据的 **GMM**：时间维度的挑战

考虑动态面板：

$$y_{it} = \alpha y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \eta_i + \varepsilon_{it}$$

一阶差分 **GMM**（**Arellano-Bond**）：差分消除固定效应：

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta \varepsilon_{it}$$

矩条件：对 $t = 2, \ldots, T$，$s \geq 2$，

$$E[y_{i,t-s} \Delta \varepsilon_{it}] = 0$$

矩条件数量：$\frac{T(T-1)}{2}$，随 T 快速增长。

系统 **GMM**（**Blundell-Bond**）：结合水平方程和差分方程矩条件，提高效率。

水平方程矩条件：对 $t = 2, \ldots, T$，

$$E[\Delta y_{i,t-1}(\eta_i + \varepsilon_{it})] = 0$$

需假设初始条件 $E[\Delta y_{i1} \eta_i] = 0$。

时间序列 **GMM**：序列相关的处理

异方差自相关一致（**HAC**）估计

当矩条件存在序列相关时，$\mathbf{\Omega} = \sum_{j=-\infty}^{\infty} \Gamma_j$，其中 $\Gamma_j = E[g_t g_{t-j}']$。

Newey-West 估计量：

$$\hat{\mathbf{\Omega}}_{HAC} = \hat{\Gamma}_0 + \sum_{j=1}^{m} w(j,m)(\hat{\Gamma}_j + \hat{\Gamma}_j')$$

常用核函数：- Bartlett：$w(j,m) = 1 - \frac{j}{m+1}$ - Parzen：$w(j,m) = \begin{cases} 1 - 6\left(\frac{j}{m}\right)^2 + 6\left(\frac{j}{m}\right)^3, & 0 \le j \le m/2 \\ 2(1-j/m)^3, & m/2 < j \le m \end{cases}$

- Quadratic Spectral：$w(j,m) = \frac{25}{12\pi^2(j/m)^2}\left[\frac{\sin(6\pi j/5m)}{6\pi j/5m} - \cos(6\pi j/5m)\right]$

带宽选择：$m = \lfloor 4(n/100)^{2/9} \rfloor$（Newey-West 建议）

长面板与短面板的不同策略

短面板（T 固定，$N \to \infty$）：- 关注截面相关 - 使用截面聚类标准误：$\hat{\mathbf{\Omega}} = \sum_{i=1}^{N} g_i g_i'$

长面板（N 固定，$T \to \infty$）：- 关注时间序列性质 - 使用 HAC 标准误 - 可能面临结构变化问题

应用实例解析

实例 **1**：教育回报估计（**Card, 1995**）

- 内生变量：教育年限
- 工具变量：大学 proximity
- 矩条件：$E[proximity_i \cdot (lnwage_i - \beta_0 - \beta_1 educ_i - \mathbf{x}_i'\boldsymbol{\beta}_2)] = 0$
- 扩展：加入更多工具变量（父母教育等）形成过度识别系统

实例 **2**：货币政策反应函数估计

- 泰勒规则：$i_t = \alpha + \beta\pi_t + \gamma y_t + \varepsilon_t$
- 内生性：利率与通胀、产出相互影响
- 工具变量：通胀和产出的滞后项、外生冲击
- 矩条件：$E[\mathbf{z}_t(i_t - \alpha - \beta\pi_t - \gamma y_t)] = 0$

实践启示：GMM 的应用关键在于矩条件的合理构造。好的矩条件应同时满足：1. 经济理论合理性 2. 统计识别能力 3. 外生性保障 4. 计算可行性

从简单模型开始，逐步增加复杂性，是应用 GMM 的明智策略。

15.5 实践中的 GMM：问题与对策

弱工具变量问题：识别不足的挑战

弱工具变量指 $\mathbf{z}_i$ 与 $\mathbf{x}_i$ 相关性微弱，这导致：

有限样本问题：1. 估计量偏误大：即使渐近无偏，小样本偏误可能严重 2. 近似正态性差：分布高度非正态，尤其当 L 大时 3. 标准误低估：常规推断严重扭曲

诊断工具：1. 第一阶段 **F** 统计量：经验法则要求 $F > 10$ 2. **Shea's partial R²**：度量每个内生变量的识别强度 3. **Cragg-Donald** 统计量：检验弱识别的正规方法 4. **Stock-Yogo** 临界值：基于最大相对偏误或 Wald 检验扭曲的临界值

改进估计量：1. **LIML**（有限信息最大似然）：对弱 IV 更稳健 2. **Fuller** 修正估计量：$\hat{\boldsymbol{\beta}}_{Fuller} = (1 - c/(n-L))\hat{\boldsymbol{\beta}}_{LIML}$，$c$ 为常数 3. 连续更新 **GMM**：对弱识别更稳健 4. **Jackknife IV**：消除许多弱 IV 的偏误

稳健推断方法：- **Anderson-Rubin** 检验：对弱 IV 稳健，检验 $\beta = \beta_0$ - 条件似然比检验：在弱识别下保持正确大小 - 识别鲁棒置信区间：通过逆检验构造，如：

$$CI = \beta_0 : AR(\beta_0) \leq \chi^2_{1-\alpha}(1)$$

权重矩阵估计：效率与稳定的平衡

一步与两步 GMM 的比较

一步 **GMM**（使用固定权重）：- 优点：计算简单，避免第一步估计误差传播 - 缺点：通常非最优，效率损失 - 适用：大样本，计算资源有限

两步 **GMM**（使用估计的最优权重）：- 优点：渐近最优，效率高 - 缺点：有限样本偏误可能更大，尤其当 q 大时 - 适用：样本量足够大，追求效率

迭代 **GMM** 的实践

迭代至收敛的过程：1.$\hat{\boldsymbol{\theta}}^{(0)}$ = 一步 GMM 估计 2. For $k = 1, 2, ...$: - $\hat{\boldsymbol{\Omega}}^{(k)} = n^{-1}\sum g(\mathbf{w}_i, \hat{\boldsymbol{\theta}}^{(k-1)})g(\mathbf{w}_i, \hat{\boldsymbol{\theta}}^{(k-1)})'$ - $\hat{\boldsymbol{\theta}}^{(k)} = \arg\min_{\boldsymbol{\theta}} \bar{g}_n(\boldsymbol{\theta})'[\hat{\boldsymbol{\Omega}}^{(k)}]^{-1}\bar{g}_n(\boldsymbol{\theta})$ 3. 当 $\|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k-1)}\| < \epsilon$ 时停止

收敛性：通常 3-5 次迭代足够。迭代 GMM 与两步 GMM 渐近等价，但有限样本性质可能更好。

高维权重矩阵问题

当 q 很大时，$\hat{\boldsymbol{\Omega}}$ 的估计可能不稳定。解决方法：

1. 收缩估计：
$$\hat{\boldsymbol{\Omega}}_{shrink} = \lambda\hat{\boldsymbol{\Omega}} + (1-\lambda)\mathbf{I}_q$$
2. 因子结构：假设 $\boldsymbol{\Omega} = \mathbf{F}\mathbf{F}' + \mathbf{D}$，其中 $\mathbf{D}$ 为对角阵
3. 正则化：加入惩罚项 $\rho\|\boldsymbol{\Omega}^{-1}\|_*$，其中 $\|\cdot\|_*$ 为核范数

矩条件选择：数量与质量的权衡

冗余矩条件问题

定义：矩条件 g_j 冗余，如果存在函数 h 使 $g_j = h(g_1, \dots, g_{j-1}, g_{j+1}, \dots, g_q)$

影响：- 不改变一致性 - 增加渐近方差 - 恶化有限样本性质 - 使权重矩阵估计不稳定

检测方法：1. 秩检验：检验 $\boldsymbol{\Omega}$ 是否满秩 2. 特征值分析：小特征值对应的矩条件可能冗余 3. 逐步选择：基于信息准则增加/删除矩条件

矩条件数量优化

偏差-方差权衡：- 矩条件少：方差大，但偏误小（对错误设定稳健）- 矩条件多：方差小（渐近），但有限样本偏误大，对错误设定敏感

选择准则：1. **Hansen's J** 准则：选择使 J 统计量最小的子集（需调整自由度）2. 信息准则：

$$IC(q) = J_n(q) + q \cdot \text{penalty}(n)$$

如：BIC penalty $= \ln n$，AIC penalty $= 2$ 3. 交叉验证：将样本分为训练集和验证集

降维技术

1. 主成分 **GMM**：对矩条件进行 PCA，保留主要成分

$$\tilde{g}_i = \mathbf{V}_r' g_i$$

   其中 $\mathbf{V}_r$ 为前 r 个特征向量

2. 因子 **GMM**：假设 $g_i = \square f_i + u_i$，使用因子得分作为新矩条件

3. 分组平均：将相关矩条件分组平均，减少数量

数值优化问题

GMM 估计需要数值优化，可能遇到：

局部最小值：目标函数 $J_n(\boldsymbol{\theta})$ 可能非凸

解决策略：1. 多起点搜索：从不同初始值开始，选择最小结果 2. 全局优化算法：模拟退火、遗传算法 3. 参数变换：将约束优化转为无约束优化

梯度信息利用：解析梯度加速收敛：

$$\frac{\partial J_n}{\partial \boldsymbol{\theta}} = 2\mathbf{G}_n(\boldsymbol{\theta})' \mathbf{W}_n \bar{g}_n(\boldsymbol{\theta})$$

收敛准则：1. 参数变化：$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\| < \epsilon_1$ 2. 函数值变化：$|J_n^{(k)} - J_n^{(k-1)}| < \epsilon_2$ 3. 梯度大小：$\|\partial J_n / \partial \boldsymbol{\theta}\| < \epsilon_3$

软件实践建议

1. 从简单开始：先用 OLS/2SLS 获得初始值
2. 监控收敛：记录每次迭代的参数值和目标函数值
3. 敏感性分析：检查不同权重矩阵、不同矩条件选择的结果稳定性
4. 诊断检验：必须报告 J 检验、第一阶段 F 统计量等
5. 稳健标准误：总是报告异方差/自相关稳健的标准误

黄金法则：如果 GMM 结果与简单方法差异巨大，应深入探究原因，而非简单接受 GMM 结果。

15.6 GMM 的扩展与前沿

经验似然方法：非参数似然的视角

经验似然（Empirical Likelihood, EL）提供了一种非参数似然框架，与 GMM 有深刻联系。

基本思想

在满足矩条件的约束下，最大化非参数似然：

$$\max_{p_1,\dots,p_n} \prod_{i=1}^{n} p_i$$

约束：1. $p_i \geq 0$，$\sum p_i = 1$ 2. $\sum_{i=1}^{n} p_i g(\mathbf{w}_i, \boldsymbol{\theta}) = 0$

拉格朗日函数：

$$\mathcal{L} = \sum_{i=1}^{n} \ln p_i - \mu(\sum p_i - 1) - n\boldsymbol{\lambda}' \sum p_i g(\mathbf{w}_i, \boldsymbol{\theta})$$

解得：

$$p_i = \frac{1}{n} \cdot \frac{1}{1 + \boldsymbol{\lambda}' g(\mathbf{w}_i, \boldsymbol{\theta})}$$

其中 $\boldsymbol{\lambda}$ 满足：

$$\frac{1}{n} \sum_{i=1}^{n} \frac{g(\mathbf{w}_i, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}' g(\mathbf{w}_i, \boldsymbol{\theta})} = 0$$

与 **GMM** 的关系

一阶等价性：经验似然估计量 $\hat{\boldsymbol{\theta}}_{EL}$ 与连续更新 GMM 估计量 $\hat{\boldsymbol{\theta}}_{CUE}$ 一阶渐近等价。

二阶优势：经验似然有二阶效率性质（Bartlett 可修正性）：1. 置信区间覆盖精度更高 2. 不需要估计 $\boldsymbol{\Omega}$ 3. 自动产生正权重 p_i

数值实现：双重优化问题：

$$\hat{\boldsymbol{\theta}}_{EL} = \arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \ln(1 + \boldsymbol{\lambda}' g(\mathbf{w}_i, \boldsymbol{\theta}))$$

扩展形式

1. 指数倾斜经验似然：使用 KL 散度惩罚

2. 广义经验似然：包含 CUE、EL 等作为特例
3. 贝叶斯经验似然：结合先验信息

局部识别与弱识别：渐近理论的扩展

弱收敛理论框架

当识别强度随样本量衰减时，需要新的渐近理论。

弱识别设定：设 $\mathbf{G}(\boldsymbol{\theta}_0) = \mathbf{G}_0/\sqrt{n}$，其中 $\mathbf{G}_0$ 固定。

此时，传统渐近理论失效：1. GMM 估计量不一致 2. 收敛速度为 $\sqrt{n}$，但极限分布非正态 3. 标准检验扭曲严重

Staiger-Stock 近似：在弱 IV 下，2SLS 的近似分布：

$$\hat{\beta}_{2SLS} - \beta_0 \approx \frac{\boldsymbol{\pi}'\mathbf{Z}'\varepsilon/n}{\boldsymbol{\pi}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\pi}/n} + \text{非正态项}$$

稳健推断方法

Anderson-Rubin 检验：原假设 $H_0 : \beta = \beta_0$

$$AR(\beta_0) = \frac{(y - \mathbf{X}\beta_0)'\mathbf{P}_Z(y - \mathbf{X}\beta_0)/q}{(y - \mathbf{X}\beta_0)'(\mathbf{M}_Z)(y - \mathbf{X}\beta_0)/(n-q)}$$

在 H_0 下，$AR(\beta_0) \xrightarrow{d} \chi^2_q/q$，即使存在弱 IV。

条件似然比检验：

$$CLR(\beta_0) = \frac{1}{2}\left[AR(\beta_0) - rk + \sqrt{(AR(\beta_0) - rk)^2 + 4 \cdot LR(\beta_0) \cdot rk}\right]$$

其中 rk 为 Cragg-Donald 统计量，LR 为似然比统计量。

识别鲁棒置信区间：通过逆检验构造：

$$CI_{1-\alpha} = \beta_0 : \text{test}(\beta_0) \leq c_{1-\alpha}$$

常用检验包括 AR、Kleibergen、CLR 等。

许多弱工具变量

当 L 很大但每个工具都很弱时：

正则化方法：1. 岭回归第一阶段：$\hat{\boldsymbol{\pi}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}'\mathbf{X}$ 2. 主成分 **IV**：使用 **Z** 的主成分作为新工具 3. **LASSO** 选择：选择相关工具变量

Jackknife IV：

$$\hat{\beta}_{JIVE} = \frac{\sum_i \mathbf{x}'_{(i)} y_i}{\sum_i \mathbf{x}'_{(i)} \mathbf{x}_i}$$

其中 $\mathbf{x}_{(i)}$ 使用除 i 外所有观测估计的第一阶段预测值，避免" 自身预测" 偏误。

高维 GMM：大 q 时代的挑战

当 q 很大，可能 $q > n$ 时：

正则化 **GMM**

1. 弹性网络惩罚：

$$\min_{\boldsymbol{\theta}} \bar{g}_n(\boldsymbol{\theta})' \mathbf{W}_n \bar{g}_n(\boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2$$

2. 稀疏 **GMM**：假设只有少量矩条件重要，使用 L1 惩罚选择：

$$\min_{\boldsymbol{\theta}} \bar{g}_n(\boldsymbol{\theta})' \mathbf{W}_n \bar{g}_n(\boldsymbol{\theta}) + \lambda \sum_{j=1}^{q} |\theta_j|$$

3. 两步选择：

   - 第一步：用 LASSO 选择活跃矩条件
   - 第二步：用选定矩条件进行 GMM 估计

去偏推断

高维下，直接推断可能偏误。去偏（debiased）GMM：

$$\hat{\boldsymbol{\theta}}^{db} = \hat{\boldsymbol{\theta}} - \hat{\square} \bar{g}_n(\hat{\boldsymbol{\theta}})$$

其中 $\hat{\square}$ 为 **G** 的估计的广义逆。

渐近分布：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{db} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \square\boldsymbol{\Omega}\square')$$

自助法推断

高维下渐近近似可能不准确，可使用：

1. 配对自助法：重采样 $(\mathbf{w}_i, \mathbf{z}_i)$ 对
2. 残差自助法：固定 $\mathbf{X}, \mathbf{Z}$，重抽样残差
3. 子抽样：使用小子样本计算分布

机器学习与 **GMM** 的结合

基于机器学习的矩条件

1. 神经网络矩条件：用神经网络学习矩条件函数

$$g_{NN}(\mathbf{w}_i, \boldsymbol{\theta}) = \phi(\mathbf{w}_i; \boldsymbol{\omega}) \cdot (y_i - m(\mathbf{x}_i; \boldsymbol{\theta}))$$

   其中 ϕ 为神经网络，$\boldsymbol{\omega}$ 为网络参数

2. 随机森林 **IV**：用随机森林预测内生变量

3. 深度学习 **GMM**：用深度学习模型构建矩条件

双重机器学习

1. 用机器学习估计倾向得分或条件期望
2. 构造基于估计量的矩条件
3. 应用 GMM 估计结构参数

示例：处理效应估计：

$$g(\mathbf{w}_i, \theta) = \frac{D_i(Y_i - \hat{\mu}_1(\mathbf{X}_i))}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1 - D_i)(Y_i - \hat{\mu}_0(\mathbf{X}_i))}{1 - \hat{\pi}(\mathbf{X}_i)} + \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) - \theta$$

其中 $\hat{\pi}, \hat{\mu}_0, \hat{\mu}_1$ 由机器学习估计。

因果推断中的 **GMM**

1. 双重稳健估计：结合倾向得分和结果回归
2. 动态处理效应：使用序列矩条件
3. 分位数处理效应：基于分位数矩条件

计算前沿：高效算法与软件

现代优化算法

1. 随机梯度下降：适用于大规模问题

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla J_n(\boldsymbol{\theta}_t)$$

2. 自适应矩估计（**Adam**）：结合动量与自适应学习率
3. 二阶方法：拟牛顿法（BFGS）、共轭梯度法

分布式计算

对于海量数据：1. 分块 **GMM**：将数据分块，分别计算矩条件，再合并 2. **MapReduce** 实现：mapper 计算个体矩条件，reducer 加总 3. 随机化算法：使用子样本加速计算

软件进展

1. 专用包：`gmm` (R), `linearmodels` (Python), `ivreg2` (Stata)
2. 自动微分：使用 `JAX`、`PyTorch` 等计算精确梯度
3. **GPU** 加速：利用 GPU 并行计算矩条件

未来方向：GMM 框架将继续融合机器学习、高维统计、分布式计算等技术，成为处理复杂经济计量问题的核心工具。

本章总结

广义矩方法代表了计量经济学估计思想的集大成与统一。通过本章的学习，我们应建立起以下核心认知体系：

一、统一性认知

GMM 不是孤立的估计技术，而是统一的理论框架：1. 方法统一：OLS、2SLS、MLE 都是 GMM 的特例，区别仅在于矩条件的选择和数量 2. 理论统一：所有估计量的一致性源于矩条件的正确设定，渐近正态性来自中心极限定理 3. 推断统一：假设检验都基于相同的渐近分布理论

这种统一性极大简化了计量经济学的理论体系，使学习者能够” 以简驭繁”。

二、实践性智慧

应用 GMM 需要平衡多个维度：1. 假设与效率：更强的假设（更多矩条件）带来潜在效率增益，但也增加误设风险 2. 有限与无限样本：渐近最优性在有限样本下可能不成立，需关注弱工具变量等问题 3. 简洁与丰富：模型应足够丰富以捕捉重要特征，又足够简洁以避免过拟合

实用准则：- 从简单模型（OLS/2SLS）开始，作为基准 - 逐步增加矩条件，监控 J 检验和估计值稳定性 - 报告多种标准误（传统、异方差稳健、聚类稳健等）- 进行敏感性分析和稳健性检验

三、前沿性视野

GMM 仍在不断发展中：1. 理论前沿：弱识别、高维 GMM、非标准渐近理论 2. 方法前沿：与机器学习结合、因果推断应用、贝叶斯 GMM 3. 计算前沿：分布式算法、自动微分、GPU 加速

这些发展使 GMM 能够应对日益复杂的经济数据和问题。

四、批判性思考

尽管强大，GMM 并非” 银弹”：1. 矩条件的质量决定一切：垃圾进，垃圾出 2. 有限样本性质可能不佳：尤其当工具变量弱或矩条件多时 3. 计算复杂性：可能需要专门优化算法 4. 解释透明性：过度复杂的矩条件可能难以解释

五、学习建议

1. 夯实基础：深入理解 OLS、2SLS、MLE 的矩条件本质
2. 循序渐进：从恰好识别到过度识别，从同方差到异方差
3. 重视实践：通过实际数据分析掌握 GMM 的应用技巧
4. 关注前沿：了解 GMM 的最新发展，但不必盲目追求复杂方法

最终启示：GMM 的精髓不在于复杂的数学，而在于其统一的思想——将经济理论转化为可检验的矩条件，用数据验证理论，用理论解释数据。这一思想将伴随您整个计量经济学学习与研究历程。

进一步阅读

经典文献

1. 奠基之作：
   - Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029-1054.
2. 权威教材：
   - Hayashi, F. (2000). *Econometrics*. Princeton University Press. (第 3 章)
   - Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press. (第 8、14 章)
   - Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. (第 6 章)
3. 应用指南：
   - Baum, C. F. (2006). *An Introduction to Modern Econometrics Using Stata*. Stata Press.
   - Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

前沿研究

1. 弱识别与推断：
   - Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4), 518-529.
   - Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11, 727-753.
2. 高维 **GMM**：
   - Caner, M., & Zhang, H. H. (2014). Adaptive elastic net for generalized method of moments. *Journal of Business & Economic Statistics*, 32(1), 30-47.
   - Chang, J., Chen, S. X., & Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1), 283-304.
3. 机器学习结合：
   - Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.

软件资源

1. **R** 包：
   - `gmm`: 通用 GMM 估计
   - `ivreg`: 工具变量回归
   - `AER`: 应用计量包，包含 GMM 函数
   - `lavaan`: 结构方程模型（GMM 的特例）
2. **Python** 库：
   - `linearmodels`: 线性计量模型，包括 IV、GMM
   - `statsmodels`: 统计模型，包含 GMM 基础功能
   - `econml`: 微软经济机器学习库
3. **Stata** 命令：
   - `ivregress`: 工具变量回归
   - `gmm`: 广义矩估计
   - `ivreg2`: 增强的 IV 估计

在线课程

1. Coursera: “Econometrics: Methods and Applications” (Erasmus University)
2. MIT OpenCourseWare: “Econometrics” (课程 14.381)
3. 中国大学 MOOC: “高级计量经济学” (清华大学、厦门大学等)

思考与练习

理论推导

1. 统一性证明：
   - 证明当 $q = p$ 时，GMM 估计量不依赖于权重矩阵 $\mathbf{W}_n$ 的选择
   - 推导 2SLS 作为 GMM 特例的具体条件
   - 证明在正确分布设定下，MLE 的渐近方差达到 Cramér-Rao 下界
2. 渐近性质：
   - 推导 GMM 估计量的渐近方差公式
   - 证明最优权重矩阵为 $\mathbf{\Omega}^{-1}$
   - 推导 Hansen’s J 检验的渐近分布

实证分析

1. 数据练习：
    - 使用 Card (1995) 数据，用 GMM 估计教育回报
    - 比较 OLS、2SLS、不同矩条件的 GMM 结果
    - 进行弱工具变量诊断和过度识别检验
2. 模型扩展：
    - 构造动态面板数据的 GMM 估计
    - 应用经验似然方法估计 CCAPM 参数
    - 实现高维情况下的正则化 GMM

研究设计

1. 矩条件构造：
    - 为劳动供给弹性估计设计矩条件
    - 为资产定价模型设计时间序列矩条件
    - 为处理效应评估设计双重稳健矩条件
2. 敏感性分析：
    - 设计方案评估弱工具变量的影响
    - 比较不同权重矩阵估计方法的表现
    - 分析矩条件数量对估计结果的影响

批判性思考

1. 方法比较：
    - GMM 与传统方法在哪些情况下差异显著？为什么？
    - 有限样本下，何时应优先使用简单方法而非 GMM？
    - 如何权衡矩条件的数量与质量？
2. 应用伦理：
    - 如何避免” 数据挖掘” 式地选择矩条件？
    - 在政策评估中，如何透明报告 GMM 的不确定性？
    - 如何处理冲突的矩条件检验结果？

学习目标：通过这些练习，您应能不仅理解 GMM 的数学原理，更能掌握其在实际研究中的恰当应用，培养出对计量方法选择的敏锐判断力。

17 蒙特卡洛法与自助法

本章导读

在传统计量经济学中，参数的统计推断主要依赖于大样本渐近理论。然而，在实际应用中，研究者常常面临小样本、模型设定复杂、分布假设不满足等问题，此时传统渐近理论的适用性受到限制。随着计算能力的飞速发展，基于计算机模拟的统计方法已成为现代计量经济学不可或缺的工具。

本章系统介绍两类核心的模拟方法：蒙特卡洛法和自助法。蒙特卡洛法通过随机抽样解决确定性计算问题，特别适用于高维积分和复杂期望的计算。自助法则通过重抽样技术，仅基于观测数据即可构造统计量的经验分布，为统计推断提供了一种灵活的数据驱动方法。此外，本章还将深入探讨马尔可夫链蒙特卡洛方法及其变体，这些方法在贝叶斯计量经济学和高维模型估计中发挥着关键作用。

本章的学习目标是：理解各类模拟方法的基本原理；掌握它们在计量经济学中的应用场景；能够根据研究问题选择合适的方法；并正确解释模拟结果。通过本章的学习，读者将获得一套强大的工具，用于处理传统方法难以解决的复杂计量问题。

16.1 引言：模拟方法在计量经济学中的作用

16.1.1 传统方法的局限性

传统计量经济学推断主要建立在渐近理论基础之上。考虑线性回归模型：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n$$

其中 $\varepsilon_i \sim i.i.d.(0, \sigma^2)$。普通最小二乘估计量 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 在满足经典假设下具有良好性质：无偏性、一致性，且渐近服从正态分布：

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2\mathbf{Q}^{-1})$$

其中 $\mathbf{Q} = \text{plim}(n^{-1}\mathbf{X}'\mathbf{X})$。基于此渐近分布，我们可以构造置信区间和假设检验。

然而，这种渐近推断在实际应用中面临诸多挑战：

1. 小样本问题：当样本量有限时，渐近近似可能不准确，特别在模型非线性或存在弱工具变量时
2. 分布假设的敏感性：许多传统方法对误差项分布有严格要求（如正态性），而实际数据常违反这些假设
3. 复杂统计量的分布：对于中位数、分位数、最大似然估计量等复杂统计量，其精确分布难以推导
4. 模型不确定性：模型设定误差对传统推断方法的影响难以量化

16.1.2 计算机模拟的优势

模拟方法通过计算机生成人工数据来研究统计量的性质，主要优势体现在：

1. 有限样本性质研究：直接评估统计量在有限样本下的表现，不依赖于大样本近似
2. 分布自由：无需对数据分布做出严格假设
3. 灵活性：适用于各种复杂模型和估计方法
4. 直观性：通过可视化的方式展示统计量的抽样分布

模拟方法可分为两大类：基于设计的蒙特卡洛研究和基于数据的自助法。前者需要设定数据生成过程，主要用于方法评估和比较；后者直接基于观测数据，主要用于实际数据的统计推断。

16.1.3 方法分类概览

本章将系统介绍以下三类核心方法：

1. 经典蒙特卡洛法：基于已知概率分布的随机抽样，用于计算积分、期望和复杂统计量的性质
2. 自助法：通过有放回重抽样从原始数据中生成伪样本，用于估计统计量的抽样分布
3. 马尔可夫链蒙特卡洛：通过构造马尔可夫链从复杂目标分布中抽样，特别适用于贝叶斯推断

这些方法共同构成了现代计量经济学家的工具箱，极大地扩展了我们处理复杂问题的能力。

16.2 蒙特卡洛方法基础

16.2.1 蒙特卡洛方法的基本原理

蒙特卡洛方法的核心思想是利用随机抽样解决确定性计算问题。其数学基础是大数定律和中心极限定理。

定义 **16.1 (**蒙特卡洛积分**)** 考虑计算期望值 $\mu = E[g(X)]$，其中 X 是随机变量，$g(\cdot)$ 是已知函数。若能从 X 的分布 F_X 中生成独立同分布的样本 $x_1, x_2, \dots, x_N$，则蒙特卡洛估计量为：

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} g(x_i)$$

由强大数定律，$\hat{\mu}_N \xrightarrow{a.s.} \mu$。由中心极限定理：

$$\sqrt{N}(\hat{\mu}_N - \mu) \xrightarrow{d} N(0, \sigma_g^2)$$

其中 $\sigma_g^2 = \mathrm{Var}[g(X)]$。

蒙特卡洛误差的标准差为 $\sigma_g/\sqrt{N}$，以 $O(N^{-1/2})$ 的速度收敛，这一收敛速率与维度无关，使其在高维积分中特别有优势。

16.2.2 简单蒙特卡洛积分

考虑一般形式的积分问题：

$$I = \int_{\mathcal{D}} f(\mathbf{x})d\mathbf{x}$$

其中 $\mathcal{D} \subseteq \mathbb{R}^d$。若能将积分改写为期望形式，即可应用蒙特卡洛方法。

例 **16.1 (**概率计算**)** 设 $\mathbf{X} \sim p(\mathbf{x})$，要计算 $P(\mathbf{X} \in A) = \int_A p(\mathbf{x})d\mathbf{x}$。定义示性函数 $I_A(\mathbf{x}) = 1$ 当 $\mathbf{x} \in A$，否则为 0。则：

$$P(\mathbf{X} \in A) = E[I_A(\mathbf{X})] \approx \frac{1}{N}\sum_{i=1}^{N} I_A(\mathbf{x}_i)$$

其中 $\mathbf{x}_i \sim p(\mathbf{x})$。

16.2.3 重要性抽样

当从目标分布 $p(\mathbf{x})$ 直接抽样困难时，重要性抽样是一种有效的方差缩减技术。

算法 **16.1 (**重要性抽样**)** 1. 选择提议分布 $q(\mathbf{x})$，使其满足：当 $p(\mathbf{x}) > 0$ 时，$q(\mathbf{x}) > 0$ 2. 从 $q(\mathbf{x})$ 中生成独立样本 $\mathbf{x}_1, \dots, \mathbf{x}_N$ 3. 计算重要性权重 $w_i = p(\mathbf{x}_i)/q(\mathbf{x}_i)$ 4. 估计期望值：$E_p[g(\mathbf{X})] \approx \frac{\sum_{i=1}^{N} w_i g(\mathbf{x}_i)}{\sum_{i=1}^{N} w_i}$

归一化权重估计量是渐近无偏的。最优提议分布为 $q^*(\mathbf{x}) \propto |g(\mathbf{x})|p(\mathbf{x})$，可使方差最小化。

16.2.4 反变换法与接受-拒绝法

反变换法适用于一维分布。若 $F(x)$ 是累积分布函数，$U \sim U(0,1)$，则 $X = F^{-1}(U)$ 的分布函数为 F。

接受-拒绝法适用于已知分布密度 $p(x)$ 但难以直接抽样的情况。算法步骤如下：

1. 找到包络函数 $Cq(x)$ 满足 $Cq(x) \geq p(x)$ 对所有 x
2. 从 $q(x)$ 生成候选样本 x^*
3. 生成 $u \sim U(0,1)$
4. 若 $u \leq p(x^*)/[Cq(x^*)]$，则接受 x^*；否则拒绝

接受概率为 $1/C$，效率取决于包络函数的紧致性。

16.2.5 蒙特卡洛方法在计量经济学中的应用

在计量经济学中，蒙特卡洛方法主要有三个应用方向：

1. 有限样本性质研究：评估估计量在小样本下的偏误、方差和分布形态
2. 检验功效分析：计算假设检验在不同备择假设下的拒绝概率
3. 模型比较与选择：通过模拟比较不同模型的预测性能

例 **16.2 (**工具变量法的有限样本偏误**)** 考虑模型：

$$y_i = \beta x_i + u_i, \quad x_i = \pi z_i + v_i$$

其中 $(u_i, v_i) \sim N(0, \Sigma)$。工具变量估计量为 $\hat{\beta}_{IV} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}$。通过蒙特卡洛模拟可以研究：- 弱工具变量 ($\pi \approx 0$) 下的估计量偏误 - 有限样本分布与渐近正态近似的差异 - 不同识别强度下的检验水平扭曲

16.3 自助法

16.3.1 自助法的基本思想

自助法（bootstrapping）由 Efron(1979) 提出，其核心是通过对原始样本的有放回重抽样来近似统计量的抽样分布。

设 $\mathbf{X}_n = (X_1, \dots, X_n)$ 是来自分布 F 的独立同分布样本，$\theta = \theta(F)$ 是感兴趣的参数，$\hat{\theta} = s(\mathbf{X}_n)$ 是其估计量。我们关心 $\hat{\theta}$ 的抽样分布 $G_n(x) = P_F(\hat{\theta} \le x)$。

算法 **16.2 (**非参数自助法**)** 1. 从原始样本 $\mathbf{X}_n$ 中有放回地抽取 n 个观测，得到自助样本 $\mathbf{X}_n^{*1}$ 2. 计算自助统计量 $\hat{\theta}^{*1} = s(\mathbf{X}_n^{*1})$ 3. 重复步骤 1-2 共 B 次，得到 $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ 4. 用 $(\hat{\theta}^{*b}\,(b = 1, \dots, B))$ 的经验分布近似 $\hat{\theta}$ 的抽样分布

自助法有效性的理论基础是经验过程理论。经验分布函数 F_n 以速率 $O_p(n^{-1/2})$ 收敛于真实分布 F，因此当 $s(\cdot)$ 是平滑函数时，G_n^*（基于 F_n 的分布）能很好地近似 G_n（基于 F 的分布）。

16.3.2 非参数自助法

非参数自助法不对总体分布 F 做任何参数假设，直接使用经验分布函数 F_n 作为 F 的估计。

标准误的自助估计：

$$\hat{\mathrm{se}}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\theta}^*)^2}$$

其中 $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$。

百分位置信区间：令 $\hat{\theta}^*_{(\alpha)}$ 表示自助统计量的 α 分位数，则 $100(1-\alpha)\%$ 置信区间为：

$$[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}]$$

16.3.3 参数自助法

当对总体分布有参数假设 $F = F_{\boldsymbol{\phi}}$ 时，可使用参数自助法：

1. 基于原始样本估计参数 $\hat{\boldsymbol{\phi}}$
2. 从 $F_{\hat{\boldsymbol{\phi}}}$ 中生成 B 组样本
3. 对每组样本计算 $\hat{\theta}^{*b}$

参数自助法在模型假设正确时效率更高，但对模型误设更敏感。

16.3.4 各种自助法变体

残差自助法：适用于回归模型 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ 1. 拟合模型得到残差 $\hat{\varepsilon}_i$ 和参数估计 $\hat{\boldsymbol{\beta}}$ 2. 从 $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n\}$ 中有放回抽样得到 ε_i^* 3. 生成 $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \varepsilon_i^*$ 4. 重新估计模型得到 $\hat{\boldsymbol{\beta}}^*$

块自助法：针对时间序列数据的依赖性 1. 将时间序列划分为长度为 l 的重叠块：$B_1 = (y_1, \dots, y_l)$, $B_2 = (y_2, \dots, y_{l+1})$, … 2. 从这些块中有放回抽样，拼接成自助样本 3. 最优块长 $l = O(n^{1/3})$，由数据依赖性决定

对偶自助法：适用于异方差模型生成 $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_i v_i^*$，其中 v_i^* 独立同分布，满足 $E[v_i^*] = 0$, $\text{Var}[v_i^*] = 1$

16.3.5 自助法的统计性质

定理 **16.1 (**自助法的一致性**)** 设 $\hat{\theta}_n$ 是 θ 的估计量，若：1. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ 2. $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$ 3. $s(\cdot)$ 在适当意义下平滑

则自助分布一致地近似真实抽样分布：

$$\sup_x |P_*(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \le x) - P(\sqrt{n}(\hat{\theta}_n - \theta) \le x)| \xrightarrow{p} 0$$

其中 P_* 表示给定原始样本的条件概率。

自助法常能提供高阶准确性。对于平滑函数模型，自助置信区间有覆盖误差 $O(n^{-1})$，而基于正态近似的区间仅有 $O(n^{-1/2})$ 的覆盖误差。

16.3.6 计量经济学应用

异方差稳健推断：在存在异方差时，传统 OLS 标准误失效。自助法（特别是对偶自助法）能提供有效的推断。

工具变量法：在弱工具变量情况下，传统检验水平扭曲严重。自助 Anderson-Rubin 检验能提供更可靠的推断。

分位数回归：分位数估计量的渐近方差涉及密度估计，计算复杂。自助法直接提供标准误和置信区间。

模型选择：通过自助法估计模型预测误差，用于比较不同模型的样本外预测能力。

16.4 马尔可夫链蒙特卡洛方法

16.4.1 MCMC 的基本框架

MCMC 用于从复杂目标分布 $\pi(\mathbf{x})$ 中生成样本，其中 $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$。基本思想是构造一个马尔可夫链，使其平稳分布等于目标分布。

定义 **16.2 (**马尔可夫链**)** 随机序列 $\mathbf{X}^{(t)}_{t=0}^{\infty}$ 称为马尔可夫链，若满足：

$$P(\mathbf{X}^{(t+1)} \in A|\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, \dots, \mathbf{X}^{(0)} = \mathbf{x}^{(0)}) = P(\mathbf{X}^{(t+1)} \in A|\mathbf{X}^{(t)} = \mathbf{x}^{(t)})$$

转移核 $P(\mathbf{x}, A) = P(\mathbf{X}^{(t+1)} \in A|\mathbf{X}^{(t)} = \mathbf{x})$ 完全刻画了链的演化。

定义 **16.3 (**平稳分布**)** 分布 π 称为转移核 P 的平稳分布，若满足：

$$\pi(A) = \int_{\mathcal{X}} P(\mathbf{x}, A)\pi(\mathbf{x})d\mathbf{x}, \quad \forall A \subseteq \mathcal{X}$$

细致平衡条件是充分条件：若存在转移核 $p(\mathbf{x}, \mathbf{y})$ 满足：

$$\pi(\mathbf{x})p(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})p(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$$

则 π 是平稳分布。

16.4.2 Metropolis-Hastings 算法

算法 **16.3 (Metropolis-Hastings)** 给定当前状态 $\mathbf{x}^{(t)}$：1. 从提议分布 $q(\cdot|\mathbf{x}^{(t)})$ 生成候选值 $\mathbf{x}^*$ 2. 计算接受概率：

$$\alpha(\mathbf{x}^{(t)}, \mathbf{x}^*) = \min\left\{1, \frac{\pi(\mathbf{x}^*)q(\mathbf{x}^{(t)}|\mathbf{x}^*)}{\pi(\mathbf{x}^{(t)})q(\mathbf{x}^*|\mathbf{x}^{(t)})}\right\}$$

3. 以概率 α 接受 $\mathbf{x}^{(t+1)} = \mathbf{x}^*$，否则 $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$

MH 算法的转移核为：

$$p_{\text{MH}}(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})\alpha(\mathbf{x}, \mathbf{y}) + \delta_{\mathbf{x}}(\mathbf{y})\left[1 - \int q(\mathbf{z}|\mathbf{x})\alpha(\mathbf{x}, \mathbf{z})d\mathbf{z}\right]$$

满足细致平衡条件，因此 π 是平稳分布。

随机游走 **MH**：取 $q(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}-\mathbf{x})$，其中 f 是对称密度（如多元正态）。此时接受概率简化为：

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\}$$

16.4.3 Gibbs 抽样

Gibbs 抽样适用于高维分布，当满条件分布易于抽样时特别高效。

算法 **16.4 (Gibbs 抽样)** 设 $\mathbf{x} = (x_1, \ldots, x_d)$，目标分布为 $\pi(\mathbf{x})$：1. 初始化 $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$ 2. 对 $t = 0, 1, \ldots$，依次更新：

$$\begin{aligned}
x_1^{(t+1)} &\sim \pi(x_1 | x_2^{(t)}, \ldots, x_d^{(t)}) \\
x_2^{(t+1)} &\sim \pi(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_d^{(t)}) \\
&\vdots \\
x_d^{(t+1)} &\sim \pi(x_d | x_1^{(t+1)}, \ldots, x_{d-1}^{(t+1)})
\end{aligned}$$

Gibbs 抽样是 MH 算法的特例，其中提议分布取为满条件分布，接受概率恒为 1。

定理 **16.2 (Gibbs 抽样的收敛性)** 若满条件分布几乎处处正，则 Gibbs 链不可约、非周期，且以 π 为唯一平稳分布。

16.4.4 MCMC 的收敛诊断

MCMC 产生的样本序列是相关的，需要判断链是否收敛到平稳分布。

燃烧期：丢弃链的初始部分，消除初始值影响。通常丢弃前 $10-50\%$ 的迭代。

自相关分析：计算样本自相关系数：

$$\hat{\rho}_k = \frac{\sum_{t=1}^{T-k}(x^{(t)} - \bar{x})(x^{(t+k)} - \bar{x})}{\sum_{t=1}^{T}(x^{(t)} - \bar{x})^2}$$

高自相关意味着有效样本量减少。

Gelman-Rubin 统计量：运行 m 条独立链，每条链长度 $2T$，丢弃前半部分。定义：- 链内方差：$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$ - 链间方差：$B = \frac{T}{m-1}\sum_{j=1}^{m}(\bar{x}_j - \bar{x})^2$ - 合并方差估计：$\hat{V} = \frac{T-1}{T}W + \frac{1}{T}B$

统计量 $\hat{R} = \sqrt{\hat{V}/W}$，当 $\hat{R} \approx 1$ 时表明收敛。

16.4.5 贝叶斯计量经济学中的 MCMC

在贝叶斯框架下，参数 $\boldsymbol{\theta}$ 的后验分布为：

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

其中 L 是似然函数，p 是先验分布。MCMC 用于从后验分布抽样。

例 **16.3 (**贝叶斯线性回归**)** 模型：$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ 先验：$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{V}_0)$，$\sigma^2 \sim \text{IG}(a_0, b_0)$

满条件分布：

$$\begin{aligned}\boldsymbol{\beta}|\sigma^2, \mathbf{y} &\sim N(\boldsymbol{\beta}_n, \mathbf{V}_n)\\ \sigma^2|\boldsymbol{\beta}, \mathbf{y} &\sim \text{IG}(a_n, b_n)\end{aligned}$$

其中：

$$\begin{aligned}\mathbf{V}_n^{-1} &= \mathbf{V}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\\ \boldsymbol{\beta}_n &= \mathbf{V}_n\left(\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{y}\right)\\ a_n &= a_0 + \frac{n}{2}\\ b_n &= b_0 + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned}$$

可直接应用 Gibbs 抽样。

16.5 高级 MCMC 方法

16.5.1 哈密尔顿蒙特卡洛

HMC 结合了物理系统的哈密尔顿动力学，能有效探索高维参数空间。

考虑扩展状态空间 $(\mathbf{q}, \mathbf{p})$，其中 $\mathbf{q}$ 是位置变量（感兴趣参数），$\mathbf{p}$ 是动量变量。定义哈密尔顿函数：

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p})$$

其中 $U(\mathbf{q}) = -\log \pi(\mathbf{q})$ 是势能，$K(\mathbf{p}) = \frac{1}{2}\mathbf{p}'\mathbf{M}^{-1}\mathbf{p}$ 是动能，$\mathbf{M}$ 是质量矩阵。

哈密尔顿方程：

$$\frac{d\mathbf{q}}{dt} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p}$$
$$\frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial \mathbf{q}} = -\nabla U(\mathbf{q})$$

算法 **16.5 (**蛙跳算法**)** 给定当前状态 $(\mathbf{q}, \mathbf{p})$ 和步长 ε、步数 L：1. 动量刷新：$\mathbf{p} \sim N(0, \mathbf{M})$ 2. 蛙跳积分：对 $l = 1, \ldots, L$

$$\mathbf{p} \leftarrow \mathbf{p} - \frac{\varepsilon}{2}\nabla U(\mathbf{q})$$
$$\mathbf{q} \leftarrow \mathbf{q} + \varepsilon \mathbf{M}^{-1}\mathbf{p}$$
$$\mathbf{p} \leftarrow \mathbf{p} - \frac{\varepsilon}{2}\nabla U(\mathbf{q})$$

3. 以概率 $\min 1, \exp(-H(\mathbf{q}^*, \mathbf{p}^*) + H(\mathbf{q}, \mathbf{p}))$ 接受 $(\mathbf{q}^*, -\mathbf{p}^*)$

HMC 的关键优势是能产生远离当前状态的提议，接受率高，特别适用于高维、相关参数。

16.5.2 No-U-Turn 采样器

NUTS 是 HMC 的自适应变体，自动调整步长 ε 和步数 L。

核心思想是通过递归构建二叉树，当轨迹开始" 回转"（新位置与初始位置的点积为负）时停止模拟。算法自动确定最优积分时间，避免手动调参。

16.5.3 切片抽样

切片抽样通过引入辅助变量实现从目标分布的抽样。

算法 **16.6 (**切片抽样**)** 目标分布 $\pi(x) \propto f(x)$：1. 给定当前 x，在 $[0, f(x)]$ 上均匀抽取 u 2. 从水平集 $x' : f(x') \geq u$ 中均匀抽取新 x'

水平集可通过 stepping-out 和 shrinkage 方法有效抽样。

16.5.4 可逆跳 MCMC

RJ-MCMC 用于模型选择问题，允许在不同维度的参数空间间跳跃。

设模型 M_k 有参数 $\boldsymbol{\theta}_k \in \mathbb{R}^{d_k}$，后验概率为 $p(M_k, \boldsymbol{\theta}_k|\mathbf{y})$。从模型 M_k 跳转到 $M_{k'}$ 时，需要维度匹配：引入随机向量 $\mathbf{u} \sim q(\mathbf{u})$ 和 $\mathbf{u}' \sim q'(\mathbf{u}')$，建立双射：

$$(\boldsymbol{\theta}_{k'}, \mathbf{u}') = g_{k \to k'}(\boldsymbol{\theta}_k, \mathbf{u})$$

接受概率为：

$$\alpha = \min\left\{1, \frac{p(M_{k'}, \boldsymbol{\theta}_{k'}|\mathbf{y})}{p(M_k, \boldsymbol{\theta}_k|\mathbf{y})} \frac{q'(\mathbf{u}')}{q(\mathbf{u})} \left|\frac{\partial g_{k\to k'}(\boldsymbol{\theta}_k, \mathbf{u})}{\partial(\boldsymbol{\theta}_k, \mathbf{u})}\right|\right\}$$

16.6 方法比较与选择

16.6.1 自助法 vs 蒙特卡洛法 vs MCMC

特性	经典蒙特卡洛	自助法	MCMC
数据要求	已知分布假设	仅需观测数据	已知分布形式
计算目标	积分/期望计算	抽样分布近似	复杂分布抽样
样本性质	独立同分布	近似独立	序列相关
收敛速度	$O(N^{-1/2})$	$O(n^{-1})$ (高阶)	依赖混合时间
主要应用	方法评估、模拟研究	频率推断、置信区间	贝叶斯推断、高维问题
实现难度	低	中	高

16.6.2 小样本性能

在小样本情况下，不同方法的表现差异显著：

1. 自助法：当 $n < 50$ 时，非参数自助法可能不稳定，特别是对于非平滑统计量
2. 参数自助法：在模型正确设定下表现良好，但对模型误设敏感
3. **MCMC**：小样本下后验分布可能受先验影响大，需要谨慎选择先验
4. 贝叶斯自助法：结合自助法与贝叶斯思想，为小样本推断提供稳健方法

16.6.3 计算成本考量

选择方法时需权衡计算成本与统计效率：

1. 时间复杂度：
    - 自助法：$O(B \cdot C(n))$，其中 $C(n)$ 是估计量计算成本
    - MCMC：$O(T \cdot C_{\text{iter}})$，$T$ 为迭代次数，C_{iter} 为单次迭代成本
2. 存储需求：MCMC 需要存储完整链，内存需求随维度线性增长

3. 并行化潜力：

    - 自助法：天然并行，不同自助样本可独立计算
    - MCMC：序列相关限制了并行化，但可运行多条独立链

4. 收敛验证：MCMC 需要诊断收敛，增加了计算和人力成本

实践建议：- 对于简单模型的频率推断，优先考虑自助法 - 对于高维贝叶斯模型，MCMC 是必要工具 - 当计算资源有限时，考虑重要性抽样等方差缩减技术 - 始终进行敏感性分析，检验方法选择对结论的影响

16.7 综合案例分析

- **16.7.1** 案例一：线性回归模型的稳健推断
    - 问题背景：存在异方差时的 OLS 推断
    - 方法应用：残差自助法 vs 对偶自助法
    - 实现步骤：R/Python 代码演示
    - 结果比较：与传统稳健标准误的对比
- **16.7.2** 案例二：时间序列模型的自助推断
    - 问题背景：ARMA 模型参数的不确定性
    - 方法应用：块自助法实现
    - 关键问题：块长度的选择
    - 应用扩展：预测区间构造
- **16.7.3** 案例三：贝叶斯逻辑回归的 **MCMC** 估计
    - 模型设定：二元选择模型
    - 方法选择：Metropolis-Hastings vs Gibbs 抽样
    - 实现细节：先验选择与收敛诊断
    - 结果解释：后验分布与可信区间
- **16.7.4** 案例四：高维 **VAR** 模型的哈密尔顿蒙特卡洛
    - 问题挑战：宏观经济 VAR 的参数估计

  - 方法优势：HMC 在高维空间的效率
  - 实施步骤：Stan/PyMC3 实现
  - 经济解释：脉冲响应函数的不确定性
- **16.7.5** 案例五：工具变量法的自助法检验
  - 问题背景：弱工具变量问题
  - 方法应用：自助法 Anderson-Rubin 检验
  - 比较分析：与传统检验方法的优劣
  - 实践建议：实际研究中的应用指南

本章小结

本章系统介绍了三类核心的模拟方法：蒙特卡洛法、自助法和 MCMC。这些方法为现代计量经济学提供了强大的计算工具，极大地扩展了我们处理复杂问题的能力。

关键要点总结：

1. 蒙特卡洛法是基于随机抽样的数值计算方法，其核心是利用大数定律和中心极限定理。重要性抽样等方差缩减技术能显著提高计算效率。
2. 自助法通过重抽样技术，仅基于观测数据即可近似统计量的抽样分布。非参数自助法灵活稳健，参数自助法在模型正确时效率更高，各种变体（块自助法、对偶自助法等）适应不同数据结构。
3. **MCMC** 方法通过构造马尔可夫链从复杂分布中抽样。Metropolis-Hastings 算法是最一般的形式，Gibbs 抽样在满条件分布易于抽样时高效，哈密尔顿蒙特卡洛特别适合高维问题。
4. 方法选择需要综合考虑问题性质、数据特征、计算资源和统计目标。自助法适合频率推断，MCMC 适合贝叶斯分析，经典蒙特卡洛适合方法评估。

未来发展方向：- 大数据场景下的高效算法 - 深度学习与模拟方法的结合 - 不确定性量化的新方法 - 自动化模型选择与推断

模拟方法已成为现代计量经济学不可或缺的部分。掌握这些工具不仅有助于解决传统方法难以处理的问题，也为探索新的方法论提供了可能。在实际应用中，研究者应当理解各种方法的假设和局限性，根据具体问题选择合适的方法，并结合领域知识进行合理解释。

18 数值优化与矩阵方法

本章导读

计量经济学的理论模型，无论是极大似然估计、广义矩估计，还是非线性最小二乘法，最终都需要通过数值计算转化为具体的参数估计值。本章深入探讨这一转化过程所依赖的两大计算支柱：数值线性代数与数值优化算法。前者为高效、稳定地处理数据与模型提供了基础数学工具，后者则利用这些工具，通过系统化的搜索策略求解最优化问题。

我们将揭示：矩阵分解如何成为构建稳健计算流程的” 基石”，而优化算法如何作为使用这些基石构造解决方案的” 建筑蓝图”。理解这两层架构，将使研究者从被动的软件使用者转变为能够洞察计算本质、诊断数值问题并针对特定问题选择适当方法的实证分析专家。这种能力对于应对高维数据、复杂模型和大规模计算等现代计量经济学的挑战至关重要。

本章学习目标：1. 掌握核心矩阵分解的原理及其在计量计算中的应用场景 2. 理解主要优化算法的数学基础、收敛性质与适用条件 3. 学会诊断和处理常见的数值稳定性问题 4. 能够为特定计量问题设计合理的数值计算策略

17.1 引言：从理论估计量到数值实现

17.1.1 计量估计的计算本质

计量经济学中的大多数估计问题最终都可以归结为以下两类数值问题：

1. 方程求解问题：寻找参数向量 $\boldsymbol{\theta} \in \mathbb{R}^p$ 使得一组矩条件成立：

$$\boldsymbol{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{z}_i, \boldsymbol{\theta}) = \mathbf{0}$$

   其中 $\boldsymbol{g}(\cdot)$ 是矩函数向量，$\boldsymbol{z}_i$ 是第 i 个观测值。

2. 函数优化问题：寻找参数 $\boldsymbol{\theta}$ 最小化（或最大化）某个准则函数：

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n(\boldsymbol{\theta})$$

其中 $Q_n(\boldsymbol{\theta})$ 是样本准则函数。例如：

- 在极大似然估计中，$Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \ln f(\boldsymbol{z}_i;\boldsymbol{\theta})$
- 在广义矩估计中，$Q_n(\boldsymbol{\theta}) = \boldsymbol{g}_n(\boldsymbol{\theta})'\boldsymbol{W}_n\boldsymbol{g}_n(\boldsymbol{\theta})$
- 在非线性最小二乘中，$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\boldsymbol{x}_i;\boldsymbol{\theta})]^2$

这两种问题在本质上相互关联。一方面，优化问题的一阶条件通常是一个方程组；另一方面，许多方程求解问题可以通过构造适当的优化问题来更稳定地求解。

17.1.2 数值计算的三个核心关切

在实际实现计量估计时，我们需要同时关注三个相互关联又可能冲突的目标：

1. 数值稳定性：算法对数据扰动、舍入误差和病态问题的不敏感性。不稳定的算法可能在小样本或病态条件下给出荒谬的结果。

2. 计算效率：算法的时间和空间复杂度。随着数据维度 p 和样本量 n 的增长，计算成本可能成为瓶颈。

3. 统计精度：数值解与理论统计性质的吻合程度。即使是渐近无偏的估计量，也可能因数值误差而在有限样本中产生偏误。

这三者构成一个权衡三角（见图 @ref(fig:tradeoff-triangle)）。例如，奇异值分解（SVD）通常比 Cholesky 分解更稳定，但计算成本更高；牛顿法收敛速度快但可能数值不稳定；梯度下降法稳定但收敛缓慢。

17.1.3 本章的结构逻辑

本章按照从基础到应用、从通用到专用的逻辑展开：

1. 基础工具层（第 17.2 节）：介绍核心矩阵分解方法，这是所有高级计算的基础。
2. 核心算法层（第 17.3-17.4 节）：系统讲解主要优化算法的原理、性质和适用条件。
3. 应用策略层（第 17.5 节）：展示如何将基础工具与优化算法结合，解决实际计量问题。
4. 前沿展望（第 17.6 节）：探讨大规模计算、自动微分等现代发展。

17.2 数值线性代数基础：核心矩阵分解

矩阵分解是将复杂矩阵运算分解为简单、稳定、高效运算的数学技术。在计量计算中，它不仅是实现工具，更是理解数值稳定性的关键。

17.2.1 LU 分解：通用线性系统求解器

数学定义：对于任意非奇异方阵 $A \in \mathbb{R}^{n\times n}$，LU 分解将其表示为下三角矩阵 L 和上三角矩阵 U 的乘积：

$$A = LU$$

其中 L 是单位下三角矩阵（对角线元素为 1），U 是上三角矩阵。

为了保证数值稳定性，实际中通常使用带行交换的 LUP 分解：

$$PA = LU$$

其中 P 是置换矩阵。

计量应用：1. 线性方程组求解：求解 $A\boldsymbol{x} = \boldsymbol{b}$ 通过以下步骤：- 分解：$PA = LU$ - 求解：$L\boldsymbol{y} = P\boldsymbol{b}$（前向替代）- 求解：$U\boldsymbol{x} = \boldsymbol{y}$（后向替代）

2. 行列式计算：$\det(A) = \det(P^{-1})\det(U) = (-1)^s \prod_{i=1}^{n} u_{ii}$，其中 s 是置换的符号。

3. 矩阵求逆：通过求解 $AX = I$ 获得。

数值性质：- 计算复杂度：$\frac{2}{3}n^3 + O(n^2)$ 次浮点运算 - 稳定性：部分主元法（LUP）通常足够稳定 - 局限性：要求矩阵非奇异，对病态矩阵敏感

17.2.2 Cholesky 分解：对称正定系统的高效解法

数学定义：对于对称正定矩阵 A（即 $A = A'$ 且 $\boldsymbol{x}'A\boldsymbol{x} > 0$ 对所有 $\boldsymbol{x} \neq \boldsymbol{0}$），Cholesky 分解表示为：

$$A = LL'$$

其中 L 是下三角矩阵（对角线元素为正）。

存在性与唯一性：对称正定矩阵必有唯一的 Cholesky 分解，且 L 的对角线元素 $l_{ii} > 0$。

计量应用：1. **OLS** 估计的稳定计算：求解正规方程 $(X'X)\boldsymbol{\beta} = X'\boldsymbol{y}$：- 计算 $X'X$ 的 Cholesky 分解：$X'X = LL'$ - 求解 $L\boldsymbol{z} = X'\boldsymbol{y}$（前向替代）- 求解 $L'\boldsymbol{\beta} = \boldsymbol{z}$（后向替代）

相比直接求逆 $(X'X)^{-1}$，Cholesky 方法避免了显式计算逆矩阵，数值稳定性更好。

2. 多元正态分布的模拟：若 $\boldsymbol{z} \sim N(\mathbf{0}, I_n)$，则 $\boldsymbol{x} = \boldsymbol{\mu} + L\boldsymbol{z} \sim N(\boldsymbol{\mu}, LL' = A)$。

3. 似然计算：多元正态对数似然中的二次型和行列式：

$$(\boldsymbol{y} - \boldsymbol{\mu})' A^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \|L^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\|^2$$

$$\ln |A| = 2 \sum_{i=1}^{n} \ln l_{ii}$$

通过 Cholesky 分解可稳定计算。

数值性质：- 计算复杂度：$\frac{1}{3}n^3 + O(n^2)$ 次浮点运算，约为 LU 分解的一半 - 稳定性：对称正定条件下非常稳定 - 病态诊断：当 A 接近奇异时，l_{ii} 会变得很小

17.2.3 QR 分解：最小二乘问题的黄金标准

数学定义：对于任意矩阵 $A \in \mathbb{R}^{m \times n}$（$m \geq n$），QR 分解为：

$$A = QR$$

其中 $Q \in \mathbb{R}^{m \times m}$ 是正交矩阵（$Q'Q = QQ' = I_m$），$R \in \mathbb{R}^{m \times n}$ 是上三角矩阵。经济型 QR 分解为：

$$A = Q_1 R_1$$

其中 $Q_1 \in \mathbb{R}^{m \times n}$ 列正交（$Q_1'Q_1 = I_n$），$R_1 \in \mathbb{R}^{n \times n}$ 上三角。

计量应用：1. 线性回归的最小二乘解：考虑问题 $\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$：- 计算 X 的 QR 分解：$X = QR$ - 解 $\boldsymbol{\beta} = R^{-1}Q'\boldsymbol{y}$（实际通过回代求解）

关键优势：避免显式计算 $X'X$，从而避免因条件数平方而放大的数值误差。

2. 回归诊断：帽子矩阵 $H = X(X'X)^{-1}X' = QQ'$，其对角线元素（杠杆值）可直接从 Q 获得。

3. 秩亏回归：当 X 不满秩时，QR 分解可通过列旋转揭示秩缺陷。

稳定性分析：QR 分解的数值稳定性源于正交变换的范数保持性质。对于最小二乘问题，解 $\hat{\boldsymbol{\beta}}$ 的相对误差满足：

$$\frac{\|\Delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}}\|} \leq \kappa(X) \left(\frac{\|\Delta X\|}{\|X\|} + \frac{\|\Delta \boldsymbol{y}\|}{\|\boldsymbol{y}\|} \right) + O(\epsilon^2)$$

其中 $\kappa(X) = \|X\|\|X^+\|$ 是条件数，X^+ 是伪逆。这比基于正规方程的方法（条件数为 $\kappa(X)^2$）有显著改进。

17.2.4 奇异值分解：诊断与稳健计算的终极工具

数学定义：对于任意矩阵 $A \in \mathbb{R}^{m\times n}$，SVD 分解为：

$$A = U\Sigma V'$$

其中：- $U \in \mathbb{R}^{m\times m}$，$U'U = UU' = I_m$ - $V \in \mathbb{R}^{n\times n}$，$V'V = VV' = I_n$ - $\Sigma \in \mathbb{R}^{m\times n}$，对角矩阵，对角线元素 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ 为奇异值，$r = \text{rank}(A)$

计量应用：1. 条件数诊断：矩阵 A 的 2-范数条件数定义为：

$$\kappa_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\sigma_1}{\sigma_r}$$

当 $\kappa_2(A)$ 很大时（如 $> 10^3$），问题病态，OLS 估计可能不可靠。

2. 主成分回归：对于病态设计矩阵 X，可构造：

$$\hat{\boldsymbol{\beta}}_{\text{PCR}} = \sum_{i=1}^{k} \frac{\boldsymbol{u}_i'\boldsymbol{y}}{\sigma_i}\boldsymbol{v}_i, \quad k < r$$

其中截断小奇异值相当于施加平滑约束。

3. 广义逆计算：Moore-Penrose 伪逆为：

$$A^{+} = V\Sigma^{+}U' = \sum_{i=1}^{r} \frac{1}{\sigma_i}\boldsymbol{v}_i\boldsymbol{u}_i'$$

4. 降维技术：主成分分析本质上是协方差矩阵的 SVD。

数值性质：- 计算成本：$O(mn^2)$ 对 $m \geq n$，比 QR 分解昂贵 - 稳定性：非常稳定，可可靠计算秩和零空间 - 截断误差：秩 k 近似 $A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i\boldsymbol{v}_i'$ 满足：

$$\|A - A_k\|_2 = \sigma_{k+1}, \quad \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}$$

17.2.5 矩阵分解方法的选择策略

选择适当的矩阵分解需要综合考虑问题结构、数值要求和计算约束。图 @ref(fig:decomp-decision) 展示了基于问题特性的决策流程。

关键决策因素：1. 矩阵结构：是否对称？是否正定？2. 问题类型：线性方程组？最小二乘？特征值问题？3. 数值要求：是否需要最大稳定性？是否需要秩信息？4. 计算资源：矩阵规模？可用内存？

时间限制？

实用指南：- 对于对称正定线性系统：**Cholesky** 分解（高效稳定）- 对于一般线性最小二乘：**QR** 分解（稳定性与效率的平衡）- 对于病态或秩亏问题：**SVD** 分解（最大稳定性，完整诊断）- 对于大规模稀疏问题：考虑稀疏矩阵格式和专门分解

17.3 无约束优化算法：寻找函数的极值

17.3.1 优化问题的数学框架与最优性条件

考虑无约束优化问题：

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$$

其中 $f : \mathbb{R}^p \to \mathbb{R}$ 是二阶连续可微的目标函数。

一阶必要条件（驻点条件）：若 $\boldsymbol{\theta}^*$ 是局部极小点，则

$$\nabla f(\boldsymbol{\theta}^*) = \mathbf{0}$$

其中 $\nabla f(\boldsymbol{\theta}) = \left(\frac{\partial f}{\partial \theta_1}, \dots, \frac{\partial f}{\partial \theta_p} \right)'$ 是梯度向量。

二阶充分条件：若 $\boldsymbol{\theta}^*$ 满足：1. $\nabla f(\boldsymbol{\theta}^*) = \mathbf{0}$ 2. $\nabla^2 f(\boldsymbol{\theta}^*) \succ 0$（Hessian 矩阵正定）

则 $\boldsymbol{\theta}^*$ 是严格局部极小点。

收敛速度的度量：设迭代序列 $\{\boldsymbol{\theta}_k\}$ 收敛到 $\boldsymbol{\theta}^*$，定义收敛速率：- 线性收敛：$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\| \leq c \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|$，$0 < c < 1$ - 超线性收敛：$\lim_{k \to \infty} \frac{\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|} = 0$ - 二次收敛：$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\| \leq M \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2$，$M > 0$

17.3.2 迭代优化算法的通用架构

大多数迭代优化算法遵循以下模板：

1. 初始化：选择初始点 $\boldsymbol{\theta}_0$，设定收敛容差 $\epsilon > 0$
2. 迭代循环（$k = 0, 1, 2, \dots$）：
    a. 方向计算：确定搜索方向 $\boldsymbol{p}_k$
    b. 步长选择：确定步长 $\alpha_k > 0$
    c. 参数更新：$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \boldsymbol{p}_k$
    d. 收敛检验：若 $\|\nabla f(\boldsymbol{\theta}_{k+1})\| < \epsilon$ 或满足其他停止准则，则终止
3. 输出：返回近似解 $\boldsymbol{\theta}_{k+1}$

不同算法的区别主要在于方向 $\boldsymbol{p}_k$ 的计算方式。

17.3.3 梯度下降法：基础但重要的基准方法

算法原理：梯度下降法使用目标函数的负梯度作为搜索方向：

$$\boldsymbol{p}_k = -\nabla f(\boldsymbol{\theta}_k)$$

更新公式为：

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \nabla f(\boldsymbol{\theta}_k)$$

步长选择策略：1. 固定步长：$\alpha_k \equiv \alpha$，简单但不适应曲率变化 2. 精确线搜索：$\alpha_k = \arg\min_{\alpha>0} f(\boldsymbol{\theta}_k + \alpha \boldsymbol{p}_k)$，计算成本高 3. 回溯线搜索（Armijo 准则）：选择 α_k 使得：

$$f(\boldsymbol{\theta}_k + \alpha_k \boldsymbol{p}_k) \leq f(\boldsymbol{\theta}_k) + c\alpha_k \nabla f(\boldsymbol{\theta}_k)' \boldsymbol{p}_k$$

其中 $c \in (0,1)$，通常 $c = 10^{-4}$

收敛性质：- 对于强凸且 L-光滑函数（$\mu I \preceq \nabla^2 f(\boldsymbol{\theta}) \preceq LI$），梯度下降法线性收敛：

$$f(\boldsymbol{\theta}_k) - f(\boldsymbol{\theta}^*) \leq \left(1 - \frac{\mu}{L}\right)^k [f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}^*)]$$

- 收敛速率取决于条件数 $\kappa = L/\mu$，κ 越大收敛越慢 - 实际应用中常因条件数大而表现不佳

在计量经济学中的角色：虽然梯度下降法很少作为最终求解器，但它作为：1. 基准方法：用于对比更复杂算法的性能 2. 预处理步骤：在更精细算法前进行粗略优化 3. 随机变体：随机梯度下降是大规模机器学习的基础

17.3.4 牛顿法：利用曲率信息的快速方法

算法原理：牛顿法基于目标函数的二阶泰勒展开：

$$f(\boldsymbol{\theta}_k + \boldsymbol{p}) \approx f(\boldsymbol{\theta}_k) + \nabla f(\boldsymbol{\theta}_k)' \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}' \nabla^2 f(\boldsymbol{\theta}_k) \boldsymbol{p}$$

最小化该二次近似得到牛顿方向：

$$\boldsymbol{p}_k^{\text{Newton}} = -[\nabla^2 f(\boldsymbol{\theta}_k)]^{-1} \nabla f(\boldsymbol{\theta}_k)$$

算法特性：1. 收敛速度：在解附近，若 $\nabla^2 f(\boldsymbol{\theta}^*)$ 正定且 Lipschitz 连续，则牛顿法二次收敛：

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*\| \leq M \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|^2$$

2. 不变性：牛顿法在参数仿射变换下不变，而梯度下降法不变。

3. 计算需求：每步需要计算 Hessian 矩阵 $\nabla^2 f(\boldsymbol{\theta}_k)$ 并求解线性系统，复杂度 $O(p^3)$。

数值实现的关键问题：1. **Hessian** 正定性：牛顿方向是下降方向当且仅当 $\nabla^2 f(\boldsymbol{\theta}_k)$ 正定。在非凸区域可能不成立。2. 线性系统求解：需要稳定求解 $\nabla^2 f(\boldsymbol{\theta}_k)\boldsymbol{p} = -\nabla f(\boldsymbol{\theta}_k)$，通常使用：- Cholesky 分解（如果 Hessian 正定）- LU 分解（一般情况）- QR/SVD 分解（病态情况）3. 步长控制：纯牛顿步（$\alpha_k = 1$）可能不下降，需结合线搜索。

修正牛顿法：为保证下降方向和数值稳定性，常用修正策略：1. 阻尼牛顿法：$\boldsymbol{p}_k = -(\nabla^2 f(\boldsymbol{\theta}_k) + \mu_k I)^{-1}\nabla f(\boldsymbol{\theta}_k)$，$\mu_k \geq 0$ 2. 修改 **Cholesky** 分解：将 Hessian 分解为 LDL' 并确保 D 的对角元足够正

在计量经济学中的应用：牛顿法是极大似然估计的标准算法，因为：1. MLE 的渐近性质保证 Hessian 在解附近正定（等于信息矩阵的负值）2. 二次收敛意味着很少迭代即可达到高精度 3. 计算 Hessian 的额外成本常被快速收敛所抵消

17.3.5 拟牛顿法：平衡效率与稳定性的计量主力

核心思想：拟牛顿法构造 Hessian 矩阵的近似 $B_k \approx \nabla^2 f(\boldsymbol{\theta}_k)$ 或其逆 $H_k \approx [\nabla^2 f(\boldsymbol{\theta}_k)]^{-1}$，仅使用一阶信息（梯度）更新。

拟牛顿条件（割线方程）：

$$B_{k+1}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k) = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$$

记 $\boldsymbol{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$，$\boldsymbol{y}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$，则条件为：

$$B_{k+1}\boldsymbol{s}_k = \boldsymbol{y}_k$$

BFGS 公式（Broyden-Fletcher-Goldfarb-Shanno）：这是最成功的拟牛顿更新之一。逆 Hessian 近似 $H_k = B_k^{-1}$ 的 BFGS 更新为：

$$H_{k+1} = \left(I - \frac{\boldsymbol{s}_k \boldsymbol{y}_k'}{\boldsymbol{y}_k' \boldsymbol{s}_k}\right) H_k \left(I - \frac{\boldsymbol{y}_k \boldsymbol{s}_k'}{\boldsymbol{y}_k' \boldsymbol{s}_k}\right) + \frac{\boldsymbol{s}_k \boldsymbol{s}_k'}{\boldsymbol{y}_k' \boldsymbol{s}_k}$$

BFGS 的性质：1. 正定性保持：若 $H_k \succ 0$ 且 $\boldsymbol{y}_k'\boldsymbol{s}_k > 0$（曲率条件），则 $H_{k+1} \succ 0$ 2. 超线性收敛：在适当条件下，BFGS 超线性收敛 3. 自我校正：BFGS 更新能自动纠正近似误差 4. 计算效率：每步 $O(p^2)$ 操作，无需计算或存储 Hessian

L-BFGS（有限内存 BFGS）：对于大规模问题（p 很大），存储 $p \times p$ 矩阵 H_k 不可行。L-BFGS 只

保存最近的 m 组 $(\boldsymbol{s}_i, \boldsymbol{y}_i)$ 对（通常 $m = 5 \sim 20$），通过递归公式计算矩阵-向量乘积 $H_k \nabla f(\boldsymbol{\theta}_k)$。

双循环递归算法：L-BFGS 通过以下两步计算搜索方向 $\boldsymbol{p}_k = -H_k \nabla f(\boldsymbol{\theta}_k)$：1. 前向循环：利用最近的历史信息 2. 后向循环：对称地应用更新

复杂度为 $O(mp)$，内存需求 $O(mp)$。

SR1 公式（对称秩 1 更新）：另一种重要的拟牛顿更新：

$$B_{k+1} = B_k + \frac{(\boldsymbol{y}_k - B_k \boldsymbol{s}_k)(\boldsymbol{y}_k - B_k \boldsymbol{s}_k)'}{(\boldsymbol{y}_k - B_k \boldsymbol{s}_k)' \boldsymbol{s}_k}$$

SR1 不强制正定性，但有时能产生更好的 Hessian 近似，特别在非凸问题中。

拟牛顿法的计量应用：在计量经济学中，拟牛顿法尤其是 BFGS，通常是极大似然估计的首选算法，因为：1. 避免了 Hessian 的计算，对复杂模型特别有利 2. 保持了牛顿法的快速收敛特性 3. 对线搜索不敏感，实现相对简单 4. 软件包（如 Stata 的 `ml`、R 的 `optim`）常默认使用 BFGS 或其变体

17.4 稳健与专用优化策略

17.4.1 信任域法：鲁棒的牛顿类方法

基本思想：信任域法在每次迭代中，在当前点 $\boldsymbol{\theta}_k$ 周围定义一个信任域 $\|\boldsymbol{d}\| \leq \Delta_k$，在这个区域内优化局部模型。相比线搜索方法（先定方向再找步长），信任域法同时确定方向和步长。

数学表述：在第 k 步，求解子问题：

$$\min_{\boldsymbol{d} \in \mathbb{R}^p} m_k(\boldsymbol{d}) = f(\boldsymbol{\theta}_k) + \nabla f(\boldsymbol{\theta}_k)' \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}' B_k \boldsymbol{d}$$

$$\text{s.t.} \quad \|\boldsymbol{d}\| \leq \Delta_k$$

其中 B_k 是 Hessian 或其近似，$\|\cdot\|$ 通常是欧几里得范数或其对等范数。

算法框架：1. 模型选择：构造局部二次模型 $m_k(\boldsymbol{d})$ 2. 子问题求解：在信任域内最小化 $m_k(\boldsymbol{d})$ 3. 接受性检验：计算实际下降与预测下降的比率：

$$\rho_k = \frac{f(\boldsymbol{\theta}_k) - f(\boldsymbol{\theta}_k + \boldsymbol{d}_k)}{m_k(\boldsymbol{0}) - m_k(\boldsymbol{d}_k)}$$

4. 信任域调整：- 若 ρ_k 接近 1，接受步长，可能扩大 Δ_k - 若 ρ_k 很小，拒绝步长，缩小 Δ_k - 若 ρ_k 中等，接受步长，保持 Δ_k

子问题求解方法：1. 柯西点法：沿最速下降方向到信任域边界 2. 狗腿法：沿最速下降方向与牛顿方向的折衷路径 3. 截断共轭梯度法：在边界内应用共轭梯度法 4. 精确解：通过求解 $\nabla m_k(\boldsymbol{d}) = -\lambda \boldsymbol{d}$，

$\lambda \geq 0$

优势与应用场景：1. 鲁棒性强：对初始点和非凸区域不敏感 2. 处理不定 **Hessian**：信任域约束自然处理不定二次模型 3. 适用于非线性最小二乘：在 Gauss-Newton 和 Levenberg-Marquardt 算法中特别有效 4. 计量应用：常用于复杂的结构方程模型、非凸似然函数

Levenberg-Marquardt 算法：非线性最小二乘的信任域特例，其中：

$$B_k = J_k' J_k + \lambda_k I$$

J_k 是残差函数的雅可比矩阵，λ_k 控制信任域大小。

17.4.2 Nelder-Mead 单纯形法：无导数优化

算法思想：Nelder-Mead 法（下坡单纯形法）通过比较单纯形顶点的函数值，进行反射、扩张、收缩等几何操作，适用于导数不可用、不可靠或计算昂贵的情况。

算法步骤：设单纯形有 $p+1$ 个顶点 $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$，对应函数值 $f_0 \leq f_1 \leq \cdots \leq f_p$。

1. 排序：$f(\boldsymbol{\theta}_0) \leq f(\boldsymbol{\theta}_1) \leq \cdots \leq f(\boldsymbol{\theta}_p)$
2. 计算重心：$\bar{\boldsymbol{\theta}} = \frac{1}{p}\sum_{i=0}^{p-1} \boldsymbol{\theta}_i$（排除最差点 $\boldsymbol{\theta}_p$）
3. 反射：$\boldsymbol{\theta}_r = \bar{\boldsymbol{\theta}} + \alpha(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_p)$，$\alpha > 0$（通常 $\alpha = 1$）
4. 决策：
   - 若 $f_0 \leq f_r < f_{p-1}$：用 $\boldsymbol{\theta}_r$ 替换 $\boldsymbol{\theta}_p$
   - 若 $f_r < f_0$：扩张：$\boldsymbol{\theta}_e = \bar{\boldsymbol{\theta}} + \gamma(\boldsymbol{\theta}_r - \bar{\boldsymbol{\theta}})$，$\gamma > 1$（通常 $\gamma = 2$）
     - 若 $f_e < f_r$：接受 $\boldsymbol{\theta}_e$
     - 否则：接受 $\boldsymbol{\theta}_r$
   - 若 $f_r \geq f_{p-1}$：收缩：
     - 若 $f_r < f_p$：外收缩 $\boldsymbol{\theta}_c = \bar{\boldsymbol{\theta}} + \beta(\boldsymbol{\theta}_r - \bar{\boldsymbol{\theta}})$，$\beta \in (0,1)$（通常 $\beta = 0.5$）
     - 若 $f_r \geq f_p$：内收缩 $\boldsymbol{\theta}_c = \bar{\boldsymbol{\theta}} + \beta(\boldsymbol{\theta}_p - \bar{\boldsymbol{\theta}})$
     - 若 $f_c < \min(f_r, f_p)$：接受 $\boldsymbol{\theta}_c$
     - 否则：缩小单纯形，向最好点 $\boldsymbol{\theta}_0$ 收缩

算法特性：1. 无需求导：只依赖函数值比较 2. 适应性强：能处理不连续、不可微函数 3. 收敛性：理论上可能不收敛到驻点，实践中常有效 4. 维度限制：通常适用于 $p \leq 10$ 的中小规模问题

计量应用：1. 初始值生成：为梯度基方法提供好的起点 2. 非标准模型：目标函数不可微或导数难以计算时 3. 模型调试：快速获得参数的大致范围 4. 鲁棒估计：某些稳健估计量（如 LAD）的求解

17.4.3 EM 算法：潜变量与缺失数据问题的专用框架

问题背景：当观测数据 $\boldsymbol{y}$ 不完整，存在缺失数据或潜变量 $\boldsymbol{z}$ 时，直接最大化观测数据似然 $f(\boldsymbol{y};\boldsymbol{\theta})$ 可能困难。EM 算法通过引入完整数据 $(\boldsymbol{y},\boldsymbol{z})$ 简化问题。

算法框架：给定当前估计 $\boldsymbol{\theta}^{(t)}$，EM 算法迭代：1. **E** 步（期望步）：计算 Q 函数：

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}[\ln f(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})]$$

2. **M** 步（最大化步）：更新参数：

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

收敛性质：1. 单调性：观测数据似然不减：$\ell(\boldsymbol{\theta}^{(t+1)};\boldsymbol{y}) \geq \ell(\boldsymbol{\theta}^{(t)};\boldsymbol{y})$ 2. 收敛到驻点：在适当条件下，$\boldsymbol{\theta}^{(t)}$ 收敛到似然函数的驻点 3. 收敛速度：线性收敛，速度依赖于信息缺失比例

EM 作为优化算法：可将 EM 视为一种特殊的优化算法，其中：- 方向：由 Q 函数与当前似然的梯度差决定 - 步长：隐式由 E 步和 M 步确定

加速变体：1. **ECM**（期望条件最大化）：将 M 步分解为多个条件最大化，简化计算 2. **ECME**（期望条件最大化要么）：某些条件最大化直接针对观测似然而非 Q 函数 3. **PX-EM**（参数扩展 EM）：引入辅助参数加速收敛

计量应用：1. 混合模型：有限混合分布、隐马尔可夫模型 2. 面板数据：带有个体效应的非线性面板模型 3. 生存分析：包含删失数据的模型 4. 因子分析：潜变量结构方程模型

17.4.4 坐标下降法与近端梯度法：高维稀疏模型求解

坐标下降法原理：每次迭代只优化一个坐标（变量），固定其他坐标：

$$\theta_j^{(k+1)} = \arg\min_{\theta_j} f(\theta_1^{(k+1)},\ldots,\theta_{j-1}^{(k+1)},\theta_j,\theta_{j+1}^{(k)},\ldots,\theta_p^{(k)})$$

循环或随机遍历所有坐标。

收敛条件：1. 若 f 凸且可微，且每个坐标最小化有唯一解，则收敛到全局最优 2. 对于非凸问题，收敛到驻点

计算优势：1. 子问题简单：单变量优化可能有解析解 2. 内存效率：每步只更新一个变量 3. 并行潜力：某些变体可并行计算

LASSO 问题的坐标下降：考虑 LASSO 问题：

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

坐标更新公式为：

$$\beta_j^{\text{new}} = S\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j} x_{ik}\beta_k\right), \lambda\right)$$

其中 $S(z,\lambda) = \text{sign}(z)(|z| - \lambda)_+$ 是软阈值函数。

近端梯度法：适用于复合优化问题：

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$$

其中 g 可微，h 可能不可微但" 简单"（近端算子易计算）。

迭代格式：

$$\boldsymbol{\theta}^{(k+1)} = \text{prox}_{\alpha h}\left(\boldsymbol{\theta}^{(k)} - \alpha\nabla g(\boldsymbol{\theta}^{(k)})\right)$$

其中近端算子定义为：

$$\text{prox}_{\alpha h}(\boldsymbol{v}) = \arg\min_{\boldsymbol{\theta}}\left\{h(\boldsymbol{\theta}) + \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{v}\|^2\right\}$$

FISTA（快速迭代收缩阈值算法）：Nesterov 加速的近端梯度法，用于 LASSO 等问题，达到最优收敛速率 $O(1/k^2)$。

计量应用：1. 高维回归：LASSO、弹性网、稀疏组 LASSO 2. 结构方程模型：带稀疏约束的协方差矩阵估计 3. 时间序列：向量自回归的稀疏估计 4. 图形模型：高斯图模型的结构学习

17.5 综合应用：计量估计的数值实现策略

17.5.1 极大似然估计的完整数值流程

以 Logit 模型为例，展示 MLE 的系统化实现策略。

模型设定：二值选择模型 $y_i \in \{0,1\}$，条件概率：

$$P(y_i = 1|\boldsymbol{x}_i) = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

对数似然函数：

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \ln\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) + (1 - y_i)\ln(1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}))\right]$$

梯度与 **Hessian**：记 $p_i = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})$，则有：

$$\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - p_i)\boldsymbol{x}_i = X'(\boldsymbol{y} - \boldsymbol{p})$$

$$\nabla^2 \ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} p_i(1-p_i)\boldsymbol{x}_i\boldsymbol{x}_i' = -X'DX$$

其中 $D = \text{diag}\{p_i(1-p_i)\}$。

数值实现策略：

1. 初始值选择：
   - 使用线性概率模型：$\boldsymbol{\beta}^{(0)} = (X'X)^{-1}X'\boldsymbol{y}$
   - 或零向量：$\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$
2. 优化算法选择：
   - 牛顿法：需要计算和求逆 Hessian，由于 $-X'DX$ 半负定，牛顿方向：

     $$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (X'D^{(k)}X)^{-1}X'(\boldsymbol{y} - \boldsymbol{p}^{(k)})$$

     这是迭代加权最小二乘（IRLS）形式。
   - 拟牛顿法（**BFGS**）：避免构造和求逆 Hessian，内存效率高。
   - 信任域法：当 $X'DX$ 接近奇异时更稳定。
3. 收敛准则：
   - 梯度范数：$\|\nabla \ell(\boldsymbol{\beta}^{(k)})\| < \epsilon_1$
   - 参数变化：$\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\| < \epsilon_2(1 + \|\boldsymbol{\beta}^{(k)}\|)$
   - 函数值变化：$|\ell(\boldsymbol{\beta}^{(k+1)}) - \ell(\boldsymbol{\beta}^{(k)})| < \epsilon_3(1 + |\ell(\boldsymbol{\beta}^{(k)})|)$
4. 数值稳定性措施：
   - 概率裁剪：计算 p_i 时避免数值溢出，如 $p_i = \max(\epsilon, \min(1-\epsilon, \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})))$，$\epsilon = 10^{-8}$
   - 正则化：在 Hessian 中添加小扰动，$(X'DX + \delta I)^{-1}$，$\delta = 10^{-6}$
   - 重新参数化：对高度相关的协变量进行正交化
5. 标准误计算：信息矩阵估计：$\hat{I}(\hat{\boldsymbol{\beta}}) = -\nabla^2 \ell(\hat{\boldsymbol{\beta}}) = X'\hat{D}X$ 协方差矩阵：$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = [X'\hat{D}X]^{-1}$

   当 $X'\hat{D}X$ 病态时，使用：
   - Cholesky 分解加扰动
   - QR 分解
   - SVD 截断伪逆

17.5.2 病态问题的诊断与处理

病态性来源：1. 近似多重共线性：设计矩阵 X 列近似线性相关 2. 尺度差异：协变量量纲差异巨大 3. 分离或拟分离：在 Logit/Probit 模型中，某些协变量组合完美预测结果 4. 稀疏数据：某些协变量取值变化很小

诊断工具：1. 条件数：$\kappa(X) = \|X\|\|X^+\|$，$\kappa(X'X) = \kappa(X)^2$ - $\kappa < 10^2$：良态 - $10^2 \leq \kappa < 10^3$：轻度病态 - $\kappa \geq 10^3$：严重病态 2. 方差膨胀因子：$\text{VIF}_j = 1/(1 - R_j^2)$，$R_j^2$ 是 x_j 对其他协变量的回归 R^2 - VIF > 10 表明严重共线性 3. 奇异值分解：小奇异值 $\sigma_r/\sigma_1 < 10^{-6}$ 表明数值秩亏 4. 相关性矩阵：绝对值接近 1 的相关系数

处理策略：1. 变量选择：剔除高度相关的变量 2. 正则化：- 岭回归：$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ - LASSO：$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$ 3. 主成分回归：用 X 的主成分作为新设计矩阵 4. 重新参数化：- 中心化：$x_{ij} \leftarrow x_{ij} - \bar{x}_j$ - 标准化：$x_{ij} \leftarrow (x_{ij} - \bar{x}_j)/s_j$ - 正交多项式：对于多项式项 5. 增加数据：收集更多样本或设计实验打破共线性

数值稳定算法选择：1. 对于线性回归：**QR** 分解或 **SVD** 而非正规方程 2. 对于非线性最小二乘：**Levenberg-Marquardt**（带阻尼的 Gauss-Newton）3. 对于 MLE：信任域牛顿法或带正则化的拟牛顿法 4. 对于高维问题：坐标下降法或近端梯度法

17.5.3 收敛失败的原因与调试策略

常见收敛问题：1. 不收敛：迭代在有限步内未达到收敛准则 2. 收敛到错误点：局部最优而非全局最优 3. 收敛过慢：需要过多迭代 4. 数值溢出：函数值、梯度或 Hessian 中出现 NaN 或 Inf

诊断步骤：1. 检查梯度：计算有限差分梯度与解析梯度比较：

$$\frac{\|\nabla f_{\text{analytic}} - \nabla f_{\text{finite-diff}}\|}{\|\nabla f_{\text{analytic}}\| + 1} < 10^{-6}$$

2. 检查 **Hessian**：验证正定性，计算最小特征值 3. 轨迹分析：记录每次迭代的函数值、梯度范数、步长 4. 条件数检查：计算 Hessian 或设计矩阵的条件数

调试策略：1. 尝试不同初始值：使用网格搜索、随机抽样或简化模型获得初始值 2. 调整算法参数：- 线搜索参数（Wolfe 条件常数）- 信任域半径初始值和更新策略 - 正则化参数 3. 变换参数空间：- 对数变换：$\theta \leftarrow \exp(\phi)$ 对正参数 - Logit 变换：$\theta \leftarrow \Lambda(\phi)$ 对 $(0, 1)$ 内参数 - 标准化：使参数量级相近 4. 简化模型：先估计简化形式，逐步增加复杂度 5. 使用鲁棒算法：从单纯形法开始，然后切换到梯度基方法

软件实现提示：1. 梯度检查：大多数优化库提供梯度检查选项 2. 详细输出：请求输出每次迭代的信息 3. 多种算法尝试：比较不同算法的结果 4. 缩放选项：利用软件的自动缩放功能

17.6 前沿发展与展望

17.6.1 大规模优化：随机方法与分布式计算

随机梯度下降：对于样本量 n 很大的问题，计算全梯度 $\nabla f(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{\theta})$ 成本高。SGD 每次迭代使用单个或小批量样本：

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \nabla f_{i_k}(\boldsymbol{\theta}_k)$$

自适应方法：1. **AdaGrad**：为每个参数调整学习率 2. **RMSProp**：使用指数加权移动平均调整 3. **Adam**：结合动量和自适应学习率

分布式优化：1. 同步并行：参数服务器架构，所有工作节点同步更新 2. 异步并行：允许延迟更新，减少通信开销 3. 联邦学习：分散数据下的隐私保护优化

17.6.2 自动微分：精确高效求导

基本原理：自动微分通过计算图追踪运算，应用链式法则，提供精确到机器精度的导数。

两种模式：1. 前向模式：计算 $\dot{y} = \nabla f(\boldsymbol{x}) \cdot \dot{\boldsymbol{x}}$，适合输入维度低的情况 2. 反向模式：计算 $\bar{\boldsymbol{x}} = \nabla f(\boldsymbol{x})' \cdot \bar{y}$，适合输出维度低的情况（最常见）

优势：1. 比有限差分更精确 2. 比符号微分更高效 3. 方便实现高阶导数 4. 与现代机器学习框架（TensorFlow, PyTorch）集成

在计量经济学中的应用前景：1. 复杂结构模型的梯度计算 2. 基于梯度的贝叶斯计算（HMC, NUTS） 3. 高维模型的正则化路径计算

17.6.3 贝叶斯计算中的优化视角

最大后验估计：MAP 估计可视为带先验的 MLE：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} [\ell(\boldsymbol{\theta}; \boldsymbol{y}) + \ln p(\boldsymbol{\theta})]$$

变分推断：将后验分布 $p(\boldsymbol{\theta}|\boldsymbol{y})$ 近似为简单分布 $q(\boldsymbol{\theta}; \boldsymbol{\phi})$，通过优化证据下界（ELBO）：

$$\text{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_q[\ln p(\boldsymbol{y}, \boldsymbol{\theta}) - \ln q(\boldsymbol{\theta}; \boldsymbol{\phi})]$$

随机变分推断：结合随机梯度与自然梯度，处理大规模数据。

优化与抽样的结合：1. 哈密顿蒙特卡洛：使用梯度信息指导 MCMC 采样 2. 朗之万动力学：带噪声的梯度下降，连接优化与抽样 3. 模拟退火：从优化到抽样的温度调度

17.6.4 计算思维的培养

从封闭形式到数值解：传统计量教学强调存在解析解的特殊情况，但现实问题多需数值解。计算思维包括：1. 将理论估计量转化为可计算形式 2. 理解数值算法的假设与局限 3. 诊断和解决计算问题 4. 验证数值结果的可靠性

可重复计算实践：1. 代码文档化：记录算法选择、参数设置、收敛准则 2. 敏感性分析：检查结果对初始值、算法参数、数值容差的敏感性 3. 基准测试：与已知解或替代方法比较 4. 版本控制：跟踪代码和数据的变化

跨学科工具借鉴：1. 从数值分析借鉴稳定算法 2. 从优化理论借鉴收敛分析 3. 从计算机科学借鉴数据结构与算法 4. 从机器学习借鉴大规模优化方法

本章总结

本章系统构建了计量经济学数值实现的知识体系，涵盖了从基础矩阵分解到高级优化算法的完整链条。

核心要点回顾：

1. 矩阵分解是数值稳定性的基石：
    - **Cholesky** 分解 为对称正定系统提供高效求解
    - **QR** 分解 是线性最小二乘的黄金标准，避免条件数平方
    - **SVD** 分解 提供最完整的矩阵分析和稳健计算
    - 分解方法的选择应基于问题结构、数值需求和计算约束
2. 优化算法是参数估计的引擎：
    - 梯度下降法 是基础基准，适合大规模问题但收敛慢
    - 牛顿法 利用二阶信息快速收敛，是 MLE 的标准选择
    - 拟牛顿法（**BFGS**）平衡效率与稳定性，是计量实践的主力
    - 信任域法 更鲁棒，适合非凸问题和不定 Hessian
    - 坐标下降法 是高维稀疏模型的高效求解器
3. 专用算法解决特定问题：
    - **EM** 算法 处理缺失数据和潜变量
    - **Nelder-Mead** 法 在导数不可用时提供无导数优化
    - 近端梯度法 处理非光滑正则化项
4. 系统化实现策略：

- 从合理初始值开始
- 选择适合问题特性的算法
- 实施数值稳定性措施
- 建立全面的收敛诊断
- 验证结果的可靠性

关键启示：

计量经济学的数值实现不是简单的” 黑箱” 操作，而是需要深入理解的科学过程。成功的数值实现需要：

1. 算法与问题的匹配：没有一种算法适合所有问题。理解算法的假设、收敛性质和数值行为是选择合适算法的前提。
2. 稳定性优先于速度：在计量应用中，获得稳定、可靠的结果比快速计算更重要。有时需要牺牲一些效率来保证数值稳定性。
3. 诊断驱动的开发：实施系统化的诊断流程，包括梯度检查、条件数分析、收敛轨迹监控等。
4. 分层设计：从简单模型开始，逐步增加复杂性；从鲁棒算法开始，再切换到高效算法。

未来方向：

随着计量经济学问题日益复杂和数据规模不断增长，数值计算方法的重要性只会增加。值得关注的发展包括：

1. 自动化算法选择：基于问题特征自动推荐合适算法
2. 混合方法：结合不同算法的优势，如随机方法与二阶方法结合
3. 硬件感知计算：利用 GPU、TPU 等专用硬件加速
4. 不确定性量化：不仅提供点估计，还量化数值误差的影响

掌握本章介绍的工具和思维，将使研究者能够更自信地处理复杂的计量模型，更深入地理解软件输出背后的计算过程，并在面对新的计量挑战时设计有效的数值解决方案。

本章习题

理论习题

1. 矩阵分解比较：设 $X \in \mathbb{R}^{n\times p}$ 是列满秩设计矩阵，$n > p$。比较求解 OLS 估计 $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ 的三种方法：直接求逆、Cholesky 分解和 QR 分解。
   - 推导每种方法的计算复杂度（以浮点运算次数表示）
   - 分析每种方法的数值稳定性，特别是当 X 病态时
   - 说明在什么条件下应选择哪种方法

2. 收敛性分析：考虑梯度下降法应用于强凸且 L-光滑函数 $f:\mathbb{R}^p\to\mathbb{R}$，即存在 $\mu,L>0$ 使得 $\mu I\preceq\nabla^2 f(\boldsymbol{\theta})\preceq LI$ 对所有 $\boldsymbol{\theta}$ 成立。
   - 证明固定步长 $\alpha=1/L$ 的梯度下降法满足：

$$f(\boldsymbol{\theta}_k)-f(\boldsymbol{\theta}^*)\leq\left(1-\frac{\mu}{L}\right)^k[f(\boldsymbol{\theta}_0)-f(\boldsymbol{\theta}^*)]$$

   - 解释条件数 $\kappa=L/\mu$ 如何影响收敛速度
   - 对比梯度下降法与牛顿法在相同假设下的收敛速率
3. 拟牛顿条件与更新唯一性：设 B_k 是 Hessian 近似，$\boldsymbol{s}_k=\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k$，$\boldsymbol{y}_k=\nabla f(\boldsymbol{\theta}_{k+1})-\nabla f(\boldsymbol{\theta}_k)$。
   - 证明满足拟牛顿条件 $B_{k+1}\boldsymbol{s}_k=\boldsymbol{y}_k$ 且 $B_{k+1}-B_k$ 秩最小的更新是 SR1（对称秩 1）更新
   - 证明在 $B_{k+1}-B_k$ 秩为 2 且对称正定的约束下，BFGS 更新是唯一的
   - 讨论为何 BFGS 在实践中比 SR1 更常用
4. 信任域法的全局收敛：考虑信任域法应用于一般非线性函数 $f(\boldsymbol{\theta})$，局部模型为 $m_k(\boldsymbol{d})=f(\boldsymbol{\theta}_k)+\nabla f(\boldsymbol{\theta}_k)'\boldsymbol{d}+\frac{1}{2}\boldsymbol{d}'B_k\boldsymbol{d}$，其中 B_k 对称。
   - 定义柯西点 $\boldsymbol{d}_k^c$ 并证明它至少提供与最速下降方向成比例的下降量
   - 证明如果每次迭代选择的步长 $\boldsymbol{d}_k$ 至少提供与柯西点成比例的下降，且 B_k 一致有界，则算法全局收敛到驻点
   - 对比信任域法与线搜索方法在全局收敛性保证方面的差异

应用习题

5. **Logit 模型 MLE 的实现设计**：考虑二值 Logit 模型 $P(y_i=1|\boldsymbol{x}_i)=\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})$，样本量为 n，协变量维度为 p。
   - 设计基于牛顿法的完整实现方案，包括初始值选择、迭代格式、收敛准则和标准误计算
   - 讨论当 $X'DX$ 接近奇异时的处理策略，其中 $D=\text{diag}\{\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]\}$
   - 对比牛顿法与拟牛顿法（BFGS）在此问题上的计算复杂度和存储需求
6. 病态回归问题的诊断与处理：假设在线性回归 $y=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 中，设计矩阵 X 存在严重多重共线性。
   - 列出诊断病态性的数值方法，包括条件数、VIF、奇异值分析
   - 比较岭回归、主成分回归和 LASSO 在处理此问题上的优缺点
   - 设计一个系统流程，从数据检查到模型估计再到结果验证
7. 高维稀疏回归的算法选择：考虑高维线性回归 $p\gg n$，假设真实参数 $\boldsymbol{\beta}^*$ 是稀疏的（大多数元素为零）。
   - 解释为什么坐标下降法特别适合求解 LASSO 问题
   - 推导 LASSO 的坐标更新公式，并说明软阈值函数的作用
   - 讨论在什么情况下应使用近端梯度法或加速变体（如 FISTA）而非坐标下降法
8. 优化失败案例分析与调试：分析以下优化失败场景，提出诊断和解决策略：

- 牛顿法迭代中 Hessian 矩阵不正定
- 拟牛顿法收敛到明显错误的解
- 算法在达到收敛准则前停止，但梯度仍很大
- 函数值在迭代中不单调下降

综合项目

9. 完整计量模型的数值实现：选择一个中等复杂的计量模型（如 Tobit 模型、多层模型或动态面板模型），完成以下任务：
   - 推导对数似然函数、梯度和 Hessian 矩阵
   - 设计数值实现方案，包括初始值策略、优化算法选择和收敛准则
   - 讨论潜在的数值问题及应对措施
   - 设计模拟实验验证实现的正确性和效率
10. 算法性能比较研究：对一个具体的计量估计问题（如 MLE for Probit 模型），设计实验比较不同优化算法的性能：
    - 包括梯度下降法、牛顿法、BFGS、L-BFGS、信任域法和 Nelder-Mead 法
    - 性能指标：迭代次数、计算时间、最终精度、对初始值的敏感性
    - 在不同问题设置下测试（不同样本量、不同条件数、不同噪声水平）
    - 基于结果给出算法选择的实用建议

这些习题旨在巩固本章的核心概念，并培养将理论知识应用于实际计量问题的能力。理论习题强调数学推导和性质分析，应用习题侧重实践设计和问题解决，综合项目则提供完整的建模与实现体验。

本章介绍了计量经济学数值计算的核心方法。矩阵分解提供了稳定高效的基础运算，而优化算法则利用这些基础求解复杂的估计问题。理解这两者的原理和相互作用，是进行可靠计量实证研究的关键能力。随着计算技术的发展，这些数值方法将继续演化，但其中蕴含的稳定性、效率和精度权衡的基本原则将始终重要。

19 机器学习在计量中的应用

本章导读

在当今数据爆炸的时代，经济数据呈现出前所未有的复杂性：高维度、非线性、非结构化以及大规模特征。传统计量经济学方法在处理这些问题时，常常面临维数诅咒、模型设定偏误和过拟合等挑战。与此同时，机器学习方法在计算机科学、统计学等领域展现出强大的数据建模能力，特别是在预测和模式识别方面。然而，机器学习的预测导向与计量经济学的因果推断导向之间存在本质差异。本章旨在架起这两大领域的桥梁，系统介绍如何将机器学习方法有效、严谨地应用于计量经济学研究。

本章重点探讨机器学习技术如何服务于计量经济学的核心使命——因果识别与推断。我们将学习如何利用正则化方法处理高维控制变量，如何使用现代机器学习工具估计异质性处理效应，以及如何将无监督学习应用于经济数据的结构发现。特别地，我们将看到机器学习不仅不威胁计量经济学的因果推断传统，反而为克服传统方法的局限提供了新工具和新思路。

本章不要求读者具备深厚的机器学习背景，所有概念都将从计量经济学家的视角出发进行阐释。我们的目标是使读者能够理解这些方法的核心思想，掌握其适用条件，并能够在实际研究中审慎应用。

18.1 因果推断的新工具：机器学习下的处理效应估计

18.1.1 传统方法的局限与机器学习的优势

在因果推断中，我们通常关注处理效应（Treatment Effect）的估计。在 Rubin 的潜在结果框架下，个体 i 的处理效应定义为：

$$\tau_i = Y_i(1) - Y_i(0)$$

其中 $Y_i(1)$ 和 $Y_i(0)$ 分别表示个体 i 在接受处理和未接受处理时的潜在结果。传统方法如倾向得分

匹配、双重差分法等在处理高维协变量、非线性关系以及异质性处理效应时面临挑战。

机器学习方法的优势在于：1. 高维处理能力：能有效处理协变量维度 p 大于样本量 n 的情况 2. 非线性建模：能够自动捕捉变量间的复杂非线性关系和交互效应 3. 异质性识别：可以估计个体层面的处理效应，而非仅关注平均处理效应 4. 正则化：通过惩罚复杂模型防止过拟合，提高样本外预测能力

18.1.2 双重机器学习

Chernozhukov 等（2018）提出的双重机器学习（Double/Debiased Machine Learning）为在因果推断中应用机器学习提供了理论框架。考虑以下部分线性模型：

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + \zeta, \quad \mathbb{E}[\zeta|X, D] = 0 \\ D &= m_0(X) + V, \quad \mathbb{E}[V|X] = 0 \end{aligned}$$

其中 D 是处理变量，X 是高维协变量，θ_0 是我们关心的处理效应参数。双重机器学习估计量通过以下步骤获得：

1. 使用机器学习方法分别估计 Y 对 X 的回归函数 $l_0(X) = \mathbb{E}[Y|X]$ 和 D 对 X 的回归函数 $m_0(X) = \mathbb{E}[D|X]$
2. 构造残差：
$$\tilde{Y} = Y - \hat{l}_0(X), \quad \tilde{D} = D - \hat{m}_0(X)$$
3. 通过以下回归估计处理效应：
$$\hat{\theta}_0 = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{D}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{D}_i \tilde{Y}_i$$

该估计量具有 $\sqrt{n}$ 一致性，只要机器学习估计量以 $o(n^{-1/4})$ 的速率收敛。

18.1.3 异质性处理效应的识别：因果森林

对于条件平均处理效应（CATE）：

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

Athey 和 Wager（2019）提出了因果森林（Causal Forest），这是随机森林在因果推断中的扩展。因果森林通过以下步骤估计 CATE：

1. 将样本随机分成若干子集
2. 在每个子集上，通过以下目标函数进行分割：

$$\Delta(C_1, C_2) = \frac{|C_1||C_2|}{|C|^2} \left(\hat{\tau}(C_1) - \hat{\tau}(C_2)\right)^2$$

   其中 $\hat{\tau}(C)$ 是子集 C 内处理效应的估计
3. 对每个观测值 i，收集包含它的所有叶子节点的处理效应估计，取平均作为最终估计

因果森林的估计量具有渐近正态性：

$$\sqrt{n}\left(\hat{\tau}(x) - \tau(x)\right) \xrightarrow{d} N(0, \sigma^2(x))$$

18.1.4 实践案例：评估职业培训项目的异质性收益

考虑评估一项职业培训项目对参与者收入的影响。传统方法可能只提供平均处理效应，但实际影响可能因个体特征而异。使用因果森林，我们可以：

1. 收集数据：处理状态 D_i（是否参与培训），结果变量 Y_i（收入），协变量 X_i（年龄、教育、工作经验等）
2. 使用因果森林估计条件平均处理效应 $\hat{\tau}(x)$
3. 分析发现：培训项目对年轻、低教育水平的参与者效果显著，但对高教育水平参与者效果不显著

这种异质性分析为政策优化提供了重要依据。

18.2 高维控制与变量选择：从 **Lasso** 到正则化回归

18.2.1 高维数据下的" 维数诅咒" 与稀疏性假设

在高维数据中，当协变量数量 p 超过样本量 n 时，普通最小二乘法（OLS）不可行，因为设计矩阵不满秩。更一般地，当 p 与 n 可比拟时，OLS 估计量的方差很大，预测性能差。

解决高维问题的关键假设是稀疏性：尽管有大量潜在协变量，但只有少数对结果有实质性影响。形式上，假设真实模型为：

$$Y = X\beta^* + \epsilon$$

其中 β^* 是 p 维向量，但只有 $s \ll p$ 个非零元素。

18.2.2 核心方法：Lasso、岭回归与弹性网络

Lasso（Least Absolute Shrinkage and Selection Operator） Lasso 通过 L_1 惩罚实现变量选择和系数收缩：

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

其中 $\lambda > 0$ 是调节参数，$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$。$L_1$ 惩罚的几何特性使得某些系数恰好为零，从而实现变量选择。

岭回归（**Ridge Regression**）岭回归使用 L_2 惩罚：

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

其中 $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$。岭回归收缩系数但不将其设为零，适用于所有变量都有微小影响的情况。

弹性网络（**Elastic Net**）弹性网络结合了 L_1 和 L_2 惩罚：

$$\hat{\beta}^{enet} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2 \right) \right\}$$

其中 $\alpha \in [0, 1]$ 控制两种惩罚的混合比例。弹性网络在处理高度相关变量时比 Lasso 更稳定。

18.2.3 后选择推断

在 Lasso 进行变量选择后，直接对所选变量进行 OLS 推断会产生偏误，因为选择过程引入了数据窥视（data snooping）。后选择推断方法提供有效的置信区间：

去偏 **Lasso（Debiased Lasso）** 对于 Lasso 估计量 $\hat{\beta}$，构造去偏估计：

$$\hat{b} = \hat{\beta} + \frac{1}{n} \hat{\Theta} X^\top (Y - X\hat{\beta})$$

其中 $\hat{\Theta}$ 是精度矩阵 Σ^{-1} 的估计。在适当条件下，去偏估计量满足：

$$\sqrt{n}(\hat{b}_j - \beta_j^*) \xrightarrow{d} N(0, \sigma_j^2)$$

可用于构造置信区间。

18.2.4 应用：在收入决定模型中控制海量家庭特征

研究教育对收入的影响时，需要控制大量家庭背景变量。假设我们有 $p = 500$ 个潜在控制变量（包括家庭资产、父母教育、社区特征等），但样本只有 $n = 1000$。

1. 使用弹性网络选择相关控制变量
2. 得到稀疏模型，只包含约 30 个重要变量
3. 在控制这些变量后，估计教育回报率
4. 使用去偏 Lasso 计算教育回报率的置信区间

这种方法比主观选择控制变量更系统、更可靠。

18.3 结构识别与数据模式发现：异常检测与结构突变

18.3.1 无监督学习在计量经济中的角色

无监督学习不依赖标签数据，而是从数据本身发现结构。在计量经济学中，无监督学习主要用于：1. 数据探索和模式发现 2. 异常值和离群点检测 3. 结构变化和断点识别 4. 降维和特征提取

18.3.2 检测经济与金融异常值：孤立森林

孤立森林（Isolation Forest）是一种高效的异常检测算法。其核心思想是：异常点稀少且与正常点差异大，因此更容易被” 孤立”。

算法流程：1. 随机选择一个特征和分割点 2. 递归地分割数据，直到每个点被孤立或达到深度限制 3. 异常点具有较短的平均路径长度

对于观测值 x，异常分数定义为：

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

其中 $E(h(x))$ 是 x 在多次树中的平均路径长度，$c(n)$ 是平均路径长度的标准化因子。$s(x,n)$ 接近 1 表示很可能是异常值。

在经济金融中，孤立森林可用于：- 检测财务报表欺诈 - 识别金融市场异常交易 - 发现经济数据中的录入错误

18.3.3 识别经济关系的结构断点

考虑时间序列模型：

$$y_t = \begin{cases} f_1(x_t, \theta_1) + \epsilon_t, & t \leq \tau \\ f_2(x_t, \theta_2) + \epsilon_t, & t > \tau \end{cases}$$

其中 τ 是未知的结构断点。传统方法如 Bai-Perron 检验假设 f 是线性形式，机器学习方法可以处理非线性断点。

基于机器学习的断点检测：1. 将时间序列划分为多个窗口 2. 在每个窗口内训练预测模型 3. 比较相邻窗口模型的预测差异 4. 当预测差异超过阈值时，标记为潜在断点

对于非线性模型，定义断点统计量：

$$Q(\tau) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}_1(x_t) - \hat{f}_2(x_t)\right)^2$$

其中 $\hat{f}_1$ 和 $\hat{f}_2$ 分别是断点前后训练的模型。

18.3.4 应用：金融危机预警与政策体制转换识别

金融危机预警：1. 收集多种经济指标（股市波动率、信用利差、外汇储备等）2. 使用孤立森林识别异常时期 3. 发现这些异常时期往往领先于金融危机

货币政策体制转换：1. 使用断点检测方法分析中央银行利率政策 2. 识别从通胀目标制到非传统货币政策的转换点 3. 分析不同体制下货币政策传导机制的变化

18.4 面板数据的深化：机器学习与个体异质性建模

18.4.1 超越固定效应：在面板数据中引入非线性与交互效应

传统面板数据模型：

$$y_{it} = \alpha_i + x_{it}^{\top}\beta + \epsilon_{it}$$

假设个体效应 α_i 是加性的，且与 x_{it} 的关系是线性的。

机器学习扩展允许更灵活的设定：

$$y_{it} = g(x_{it}, \alpha_i) + \epsilon_{it}$$

其中 $g(\cdot)$ 可以是任意复杂函数，通过神经网络或树模型估计。

18.4.2 机器学习方法估计时变个体效应

考虑时变个体效应模型：

$$y_{it} = \alpha_{it} + x_{it}^{\top}\beta + \epsilon_{it}$$

使用矩阵分解方法：

$$\min_{\alpha,\beta}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \alpha_{it} - x_{it}^{\top}\beta)^2 + \lambda R(\alpha)$$

其中 $R(\alpha)$ 是惩罚项，鼓励 α 具有低秩或平滑结构。

例如，假设 α 可分解为：

$$\alpha_{it} = u_i^\top v_t$$

其中 $u_i \in \mathbb{R}^k$，$v_t \in \mathbb{R}^k$，$k \ll \min(N, T)$。这实质上是面板数据的因子模型。

18.4.3 交互固定效应模型与机器学习的结合

Bai（2009）的交互固定效应模型：

$$y_{it} = x_{it}^\top \beta + \lambda_i^\top f_t + \epsilon_{it}$$

机器学习方法可以估计更一般的形式：

$$y_{it} = h(x_{it}) + \lambda_i^\top f_t + \epsilon_{it}$$

其中 $h(\cdot)$ 通过机器学习方法估计。

估计步骤：1. 使用主成分分析估计因子结构：$\hat{f}_t$ 和 $\hat{\lambda}_i$ 2. 构造残差：$\tilde{y}_{it} = y_{it} - \hat{\lambda}_i^\top \hat{f}_t$ 3. 在残差上使用机器学习估计：$\hat{h}(x) = \arg\min_h \sum_{i,t} (\tilde{y}_{it} - h(x_{it}))^2 + \lambda J(h)$

18.4.4 应用：企业生产率分析中的异质性技术溢出效应

研究研发投入对企业生产率的影响：- 传统方法：估计平均弹性 - 机器学习方法：允许异质性影响

模型设定：

$$\ln(Prod_{it}) = h(RD_{it}, Controls_{it}) + \alpha_i + \gamma_t + \epsilon_{it}$$

其中 $h(\cdot)$ 通过梯度提升树估计。研究发现：1. 研发对生产率的促进作用呈非线性：边际效应递减 2. 异质性明显：对高科技企业影响更大 3. 存在互补性：研发与人力资本投资有协同效应

18.5 政策评估的强化：基于机器学习的合成控制与反事实构建

18.5.1 合成控制法的回顾及其机器学习扩展

传统合成控制法（Abadie 等，2010）用于评估处理单元（如实施政策的地区）的处理效应。对于处理单元 $i = 0$（$t \geq T_0$ 后接受处理），寻找权重 $w^* = (w_1, ..., w_J)$ 使得：

$$\min_w \|X_0 - X_c w\|_V \quad \text{s.t.} \quad w_j \geq 0, \sum_{j=1}^{J} w_j = 1$$

其中 X_0 是处理单元的处理前特征，X_c 是控制单元的特征矩阵，V 是权重矩阵。

合成控制法的机器学习扩展：1. 广义合成控制：放松凸组合约束，允许负权重和权重大于 1 2. 矩阵补全方法：将反事实预测视为矩阵补全问题 3. 正则化合成控制：加入惩罚项防止过拟合

18.5.2 矩阵补全方法与反事实预测

将面板数据视为矩阵 $Y \in \mathbb{R}^{N \times T}$，其中部分条目缺失（处理后的处理单元结果）。矩阵补全的目标是：

$$\min_M \sum_{(i,t) \in \Omega} (Y_{it} - M_{it})^2 + \lambda \|M\|_*$$

其中 Ω 是观测到的条目集合，$\|M\|_*$ 是核范数（奇异值之和），鼓励低秩结构。

对于处理单元 $i = 0$ 在 $t \geq T_0$ 的反事实预测：

$$\hat{Y}_{0t}(0) = \hat{M}_{0t}, \quad t \geq T_0$$

处理效应估计：

$$\hat{\tau}_{0t} = Y_{0t} - \hat{Y}_{0t}(0), \quad t \geq T_0$$

18.5.3 广义合成控制与正则化合成控制

广义合成控制（Xu，2017）：

$$\hat{w}^{GSC} = \arg\min_w \left\{ \sum_{t=1}^{T_0} (Y_{0t} - \sum_{j=1}^{J} w_j Y_{jt})^2 + \lambda \|w\|_2^2 \right\}$$

放松了非负和求和为 1 的约束，但增加了 L_2 惩罚。

正则化合成控制（Arkhangelsky 等，2021）：考虑更一般的因子模型：

$$Y_{it}(0) = \alpha_i + \beta_t + \lambda_i^\top f_t + \epsilon_{it}$$

使用矩阵补全方法估计缺失的反事实。

18.5.4 应用：评估大型区域性经济政策（如特区设立）的净效应

评估某经济特区设立对区域经济增长的影响：- 处理单元：设立特区的城市 - 控制单元：其他类似城市 - 结果变量：人均 GDP 增长率

传统合成控制法局限：1. 只能处理单一处理单元 2. 对处理前拟合要求高 3. 权重非负约束可能限制拟合效果

机器学习改进方法：1. 使用矩阵补全方法，同时估计多个特区的效应 2. 允许处理前拟合不完美，但保证模型复杂度受控 3. 得到处理效应的动态路径和置信区间

研究发现：特区政策在短期（1-3 年）内效应不明显，长期（5 年以上）显著促进经济增长，但存在区域异质性。

本章总结

本章系统探讨了机器学习方法在计量经济学中的五大核心应用领域，展示了这些现代数据科学工具如何丰富和扩展传统计量经济学方法论。

在因果推断方面，我们学习了双重机器学习和因果森林等方法，它们使得在高维数据中估计处理效应和识别异质性成为可能。这些方法严格建立在计量经济学因果推断框架内，为处理复杂观测数据提供了新工具。

在高维控制与变量选择方面，正则化方法如 Lasso、岭回归和弹性网络解决了” 维数诅咒” 问题，使得研究者能够系统地从大量潜在控制变量中选择相关变量，减少模型设定偏误。

在结构识别与模式发现方面，无监督学习方法如孤立森林为检测经济异常和识别结构变化提供了自动化工具，帮助经济学家从数据中发现新的经验规律。

在面板数据分析方面，机器学习方法允许更灵活地建模个体异质性和时间效应，特别是通过因子模型与机器学习的结合，能够捕捉复杂的个体-时间交互效应。

在政策评估方面，机器学习增强了合成控制法等反事实预测方法，通过矩阵补全和正则化技术，提高了政策效应估计的精度和稳健性。

需要特别强调的是，机器学习的引入不是要取代传统计量经济学，而是要弥补其在高维、非线性、复杂数据环境中的不足。成功的应用需要深刻理解计量经济学的因果推断逻辑，审慎选择机器学习工具，并正确解释结果。

未来，“计量机器学习” 这一交叉领域将继续蓬勃发展，可能的方向包括：1. 发展更适合经济数据特性的机器学习算法 2. 建立更完整的理论框架，理解机器学习方法的经济计量性质 3. 开发用户友好的软件包，降低方法应用门槛 4. 探索机器学习在结构计量模型中的应用

计量经济学与机器学习的融合，正推动着经验经济学研究向更严谨、更精细、更实用的方向发展。

本章练习题

1. 概念辨析：比较双重机器学习与传统工具变量法在解决内生性问题时的逻辑异同。双重机器学习如何处理不可观测的混杂变量？

2. 方法操作：考虑线性模型 $Y = X\beta + \epsilon$，其中 X 为 $n \times p$ 设计矩阵，$p > n$。推导 Lasso 估计量 $\hat{\beta}$ 的闭式解（当 $X^\top X = I$ 时），并解释 L_1 惩罚如何导致稀疏性。

3. 案例分析：设计一个研究方案，利用因果森林方法评估” 提高最低工资” 政策对不同规模、不同地区企业的就业效应差异。需详细说明：

    - 所需数据类型和来源
    - 核心变量定义和度量
    - 因果森林的具体设定和参数选择
    - 如何解释异质性处理效应结果
    - 可能的识别挑战和解决方案

4. 模型比较：在政策评估中，比较传统合成控制法、广义合成控制法和矩阵补全方法：

    - 各自的假设条件是什么？
    - 各适用于什么类型的数据和政策评估问题？
    - 如何从实证角度比较这些方法的优劣？

5. 综合论述：“机器学习虽然预测能力强，但对计量经济学追求的因果推断构成了威胁。” 请结合本章内容，对此观点进行评述。讨论：

    - 预测和因果推断的根本区别
    - 机器学习如何可能威胁因果推断（如过拟合、黑箱问题）
    - 如何正确使用机器学习辅助因果推断而非威胁它
    - 计量机器学习方法如何保持因果推断的严谨性