高等学校经济学类核心课程教材

计量经济学

Econometrics

陈诗一　陈登科　著

高等教育出版社

高等学校经济学类核心课程教材

计量经济学

Econometrics

陈诗一　陈登科　著

高等教育出版社·北京

内容简介

本书在统一的框架内系统介绍了计量经济学的基本原理、基本内容与基本方法。该框架围绕回归模型设定偏误与内生性这两个影响回归关系具有因果解释的因素展开。

本书主要特色显著体现在：首先，在讨论回归关系的同时，侧重介绍因果关系，更重要的是，还在二者之间搭建了一座桥梁；其次，统一介绍了基于可观测结果表述的经典回归框架和基于“反事实”结果表述的潜在结果框架；最后，强调计量理论知识在经济学上的直观解释，尽可能多地使用具有中国特色的经济案例与直观、精美的图形，并特略去了一些技术性较强但不影响理解计量经济学基本思想的内容。

除了作为高等学校本科计量经济学的入门教材，本书适合从事经济学、金融学、统计学、管理学、社会学以及政治学等相关科学研究的学者或教师作为参考书使用，也推荐给那些有意了解数理统计基础知识并对因果推断感兴趣的读者使用。

图书在版编目(CIP)数据

计量经济学/陈诗一，陈登科著. --北京 ：高等教育出版社，2019.4

ISBN 978-7-04-051565-7

Ⅰ. ①计… Ⅱ. ①陈… ②陈… Ⅲ. ①计量经济学-高等学校-教材 Ⅳ. ①F224.0

中国版本图书馆 CIP 数据核字(2019)第 041715 号

策划编辑 施春花	责任编辑 施春花	封面设计 王 鹏	版式设计 马敬茹
插图绘制 于 博	责任校对 刘娟娟	责任印制 韩 刚	

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	北京东君印刷有限公司	网上订购	http://www.hepmall.com.cn
			http://www.hepmall.com
开 本	787mm × 1092mm 1/16		http://www.hepmall.cn
印 张	12.75	版 次	2019 年 4 月第 1 版
字 数	300 千字	印 次	2019 年 4 月第 1 次印刷
购书热线	010-58581118	定 价	33.00 元

物 料 号 51565-00

前　言

计量经济学是运用数学和统计学对经济学进行定量分析的一门学科。正如马克思指出的，在一门学科中能够使用数学是该学科成熟的重要标志。如果说，1969 年诺贝尔经济学奖第一次颁奖标志着经济学成为一门科学，那么在经济学不断科学化的过程中，计量经济学起到了不可或缺的关键作用。今天，计量经济学已经成为现代经济分析的基本方法论，拥有分析充满不确定性和风险的各种经济现象以及经济特征事实的专门方法。可以说，计量经济学是各种经济理论通向经济现实（数据）的桥梁，是我们进行经济分析与解释、经济理论创新与经济政策评估的“虚拟实验室”。

作为高校经济学教学的专业基础课程，计量经济学的教材内容十分重要。本书专门为高校本科生计量经济学教学而编制，力图展现以下几方面特色：首先，现有计量经济学教材重点讲解统计上的回归关系，本书则在讨论统计上回归关系的同时，更加侧重介绍经济学上的因果关系，更为重要的是，我们还尝试在回归关系与因果关系之间架起一座桥梁，以帮助读者理解在什么条件下回归关系具有因果解释；其次，本书统一介绍了基于可观测结果表述的经典回归框架和基于“反事实”结果表述的潜在结果框架，重点探究了二者之间的内在联系；最后，为帮助读者更加深刻地理解计量经济学的本质，本书强调理论知识在经济学上的直观解释，同时，为了帮助读者更好地了解和分析当代中国的经济改革与发展，本书尽可能多地使用具有中国特色的经济案例与直观、精美的图形，并特意略去了一些技术性较强但不影响读者理解计量经济学基本思想的内容。

本书在一个统一的框架内，系统地介绍了计量经济学的基本原理、基本内容与基本方法。内容简介如下。

第一章绪论较为系统地介绍了变量之间的关系。首先，区分了确定性的变量关系与不确定性变量关系，并指出本书主要关注的是不确定性关系。然后，依次介绍了不确定性关系中的相关关系、回归（预测）关系以及因果关系，并着重对回归关系和因果关系进行了区分。此外，由于经济数据是计量经济分析的前提和基础，第一章还对经济数据进行了详细的介绍。综合而言，本章介绍的是计量经济学的研究对象（变量间的不确定性关系）与研究基础（经济数据）。

第二章回顾了计量经济学所涉及的数学基础知识，主要包括：概率论、矩阵基础以及数值方法初步。其中，概率论是本章重点介绍的内容，矩阵基础与数值方法初步可以作为选读内容，跳过它们不会影响本书余下章节的学习。本章重点介绍了条件期望及其性质，这在后续章节中将经常用到；还讨论了简单随机抽样，强调了样本的随机特征；进一步介绍了大数定律与中心极限定理，并创新性地辅以图形对其进行直观的展示，大数定律与中心极限定理的重要性在于，搭建了一座从样本参数通往总体参数的桥梁。

第三章介绍了普通最小二乘（OLS）回归这一经典计量经济学的核心方法论，并着重讨论了回归关系具有因果关系所要满足的两个前提条件。本章以一元回归模型为例，系统介绍了 OLS

回归的基本原理、参数估计方法、参数估计的性质以及参数假设检验。本章重点区分了样本回归参数与总体回归参数、回归模型与结构(因果)模型。尽管OLS回归模型在经典计量经济学中处于核心地位,但它所得到的回归系数往往不具有因果解释。为此,本章紧接着正式给出了回归关系具有因果解释所要满足的两个前提条件:不存在回归模型设定偏误与不存在内生性问题。这两个条件的重要意义在于,搭建了一座从回归关系到因果关系的桥梁。本书主要内容就是紧紧围绕这两个条件循序渐进地展开。此外,由于与一元回归模型相比,多元回归模型更加可能具有因果解释,本章还介绍了多元回归模型。

第四章主要介绍当条件期望不是变量的线性函数时,为使回归分析具有因果解释而应该使用的正确模型设定方法,它对应回归关系具有因果关系所要满足的第一个前提条件。首先,本章假设回归模型中不存在内生性问题,这使得因变量的条件期望具有因果解释,从而可以从条件期望入手考察回归分析中的正确模型设定;随后,在多种不同情形下,对条件期望可能具有的非线性函数形式以及相应的回归模型设定进行介绍,这主要包括:饱和回归模型设定、变量间存在交互作用以及因果关系存在突变时的模型设定、多项式模型、对数模型以及因变量取值受到限制时的回归模型设定。

第五章阐述了导致内生性问题产生的具体原因,对应回归关系具有因果关系所要满足的第二个前提条件,主要包括测量误差、遗漏变量、联立性(或者反向因果)、样本选择以及"自选择"等。其中,"自选择"所导致的内生性问题是本书关注的重点。一方面,"自选择"问题与人的行为密切相关;另一方面,以潜在结果框架为特征的现代计量经济学,主要围绕"自选择"问题所导致的内生性问题进行论述。第六章至第九章则依次介绍了处理内生性的随机化实验、匹配方法、工具变量法、双重差分法与断点回归设计。

第六章开始引入潜在结果框架,并在此基础上介绍处理内生性问题的两种基本方法:随机化实验和匹配方法。首先,介绍了潜在结果框架的基本内容;随后,介绍了回归框架下的潜在结果表述,并对回归框架与潜在结果框架进行了比较分析,强调二者在本质上存在着密切相关;最后,介绍了处理内生性问题的两种基本方法:随机化实验与匹配方法。随机化实验是因果推断的黄金法则,但由于存在着诸如伦理道德、非完美随机化以及外部有效性等问题,因而很难真正地实施;匹配方法能够解决基于可观测变量选择导致的内生性问题。不过,如果是基于不可观测变量选择所导致的内生性问题,那么匹配法就无法真正奏效,此时需要采用诸如工具变量、双重差分以及断点回归设计等方法。

第七章介绍工具变量法。首先,对同质性因果效应和异质性因果效应进行了区分。然后分别讨论了这两种情况下的内生性问题及相应解决方法。其中,在同质性因果效应部分,介绍了工具变量有效的两个基本条件、两阶段最小二乘回归、瓦尔德估计量、针对工具变量的两个相关检验以及控制函数方法;在异质性因果效应部分,则基于潜在结果框架介绍了局部平均处理效应,并介绍了它和第六章各类处理效应之间的不同。

第八章介绍双重差分法。本章首先在潜在结果框架下,对双重差分法的基本原理进行了介绍,突出了平行趋势假设在这一方法中的重要性。紧接着,基于回归分析对该方法的具体运用和相关拓展进行了说明。然后,讨论了平行趋势假设遭到违背的三种情形和相应的平行趋势检验方法。最后,在双重差分法不再适用的情况下,进一步介绍了三重差分法。

第九章是本书的最后一章,我们在此章分别从参数化(回归)和非参数化(潜在结果)视角介

绍了断点回归设计，并将断点回归设计与其他方法进行了比较分析。在参数化表述方式下，如果一旦在回归方程中控制了参考变量，那么内生性问题就消失了，但是却可能存在回归模型设定偏误问题；在非参数化表述方式下，不存在回归模型设定偏误问题，但却可能存在内生性问题，在该情形下断点回归设计利用断点左右两侧附近个体基本类似的特点来处理内生性问题。

除了可作为高等学校本科计量经济学的入门教材，本书也适合从事经济学、金融学、统计学、管理学、社会学以及政治学等相关科学研究的学者或教师作为参考书使用，也推荐给那些有意了解数理统计基础知识并对因果推断感兴趣的读者使用。本教材得以顺利出版，我们还要感谢高等教育出版社的施春花编辑，感谢上海市理论经济学Ⅰ类高峰计划的资助。在本书的成书过程中，我的博士研究生金浩、张建鹏、卢郁霖、刘婉琳以及吴超、吴友也提供了不少帮助，在此一并表示感谢。当然，文责自负。

作 者

2018 年秋于复旦园

目　录

第一章 绪 论

一般地，计量经济学是在已有数据的基础上，运用概率统计学工具来定量探讨经济学变量之间关系的科学。作为计量经济学学习的起点，本章主要介绍变量之间存在的不同类型的关系，重点区分了变量之间的回归关系与因果关系。此外，由于经济数据是开展计量经济分析的前提和基础，因此在本章中我们还对经济数据的类型与特征进行了简要说明。

1.1 变量之间的关系

在现实生活、科学研究以及政策制定中，人们往往对变量之间的特定关系感兴趣。在考察变量之间关系的过程中，首要的是明确所关心的是变量之间的哪种关系。总体而言，变量之间的关系具体可以划分为**确定性关系**（Certainty Relationship）与**不确定性关系**（Uncertainty Relationship），后者又通常被称为**统计关系**（Statistical Relationship）。

在经典物理学中，变量之间一般体现为确定性关系。例如，牛顿万有引力定律：若用 m_1，m_2 表示两个物体的质量，r 表示它们之间的距离，G 为万有引力常数，那么这两个物体之间的相互吸引力为 $F=G\cdot\frac{m_1m_2}{r^2}$。此时，若给定 m_1，m_2，r 以及 G 的具体取值，则能够完全确定引力 F 的大小。然而，在社会科学中，变量之间的关系一般是不确定的（随机的）。比如，只知道一个人的受教育程度，计量经济学家往往无法精确地推知这个人的收入水平。这是因为除了受教育程度外，还有大量其他因素会影响个人收入水平，比如性别、家庭背景、能力以及工作经验，等等。事实上，影响个人收入的因素如此之多，以至于无论考虑了多少变量，都无法完全精确地推知收入水平。

图 1.1 给出了确定性变量关系与不确定性变量关系的示意图。在图 1.1(a)中，给定 X 的任一取值，都有唯一的 Y 值与其对应，故而变量 X 与变量 Y 之间的关系是确定的；但是在图 1.1(b)中，给定 X 的取值无法确定 Y 的值，同样地，给定 Y 的值也无法确定 X 的值，故而变量 X 与 Y 之间的关系是不确定的。一般而言，确定性关系可以通过**函数**来描述，不确定性关系则采用**相关关系**、**预测关系**以及**因果关系**来刻画。[①] 计量经济学主要关注的是不确定性关系（或统计关系），特别是其中的预测关系与因果关系。在不确定性关系中所要处理的变量是随机变量。

---

① 值得指出的是，因果关系也常用来刻画确定性关系。

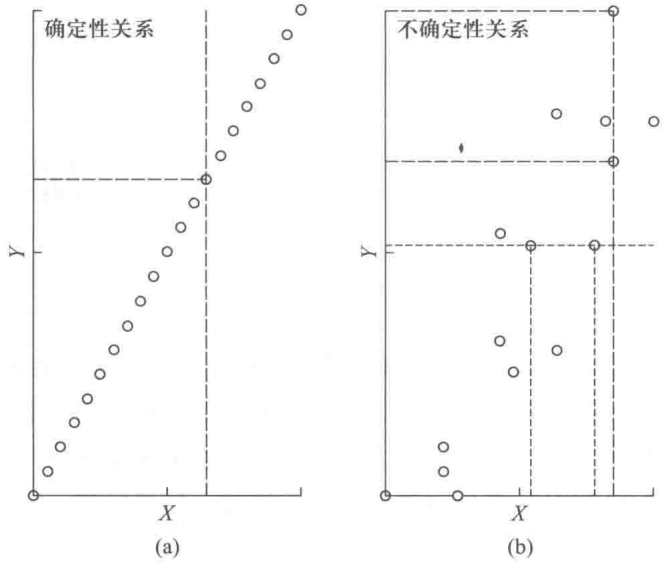

图 1.1 确定性关系与不确定性关系

一、确定性关系

若变量 X 与 Y 之间的关系是确定的，那么这种关系一般可以采用函数来描述，记为 $Y=f(X)$。尽管确定性关系不是本教材讨论的重点，但是了解几种常见函数的形态对后续章节的学习非常有益处。为此，图 1.2 给出了几类函数曲线示意图，本教材后续章节多处涉及了这些函

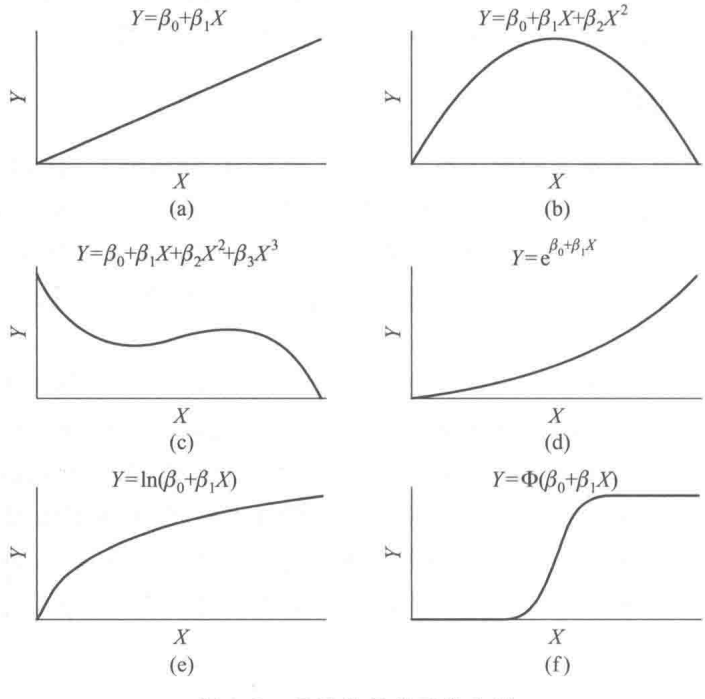

图 1.2 几种常见的函数曲线

注：$\Phi(\cdot)$表示正态分布累积分布函数。

数。从图形中可以看出，无论 $f(X)$ 的具体形式如何，在给定变量 X 的取值的条件下，都能够完全精确地计算出变量 Y 的数值。

二、不确定性关系

1. 相关关系

相关关系刻画的是变量之间的关联程度。在不确定性关系中，虽然无法通过一个变量精确地推知另一变量，但是我们通常会对变量之间的关联程度感兴趣。比如，受教育程度与收入水平的关联度，收入水平与消费水平的关联度，上市公司利润率与其股票收益率之间的关联度，经济增长与房价的关联度，以及货币供给与通货膨胀率的关联度，等等。变量间的关联程度可以采用**相关系数**（在第二章中还会更详细讨论相关系数）来度量。令 X 与 Y 表示两个变量，它们相关系数记为 $\rho_{X,Y}$，表达式为：

$$\rho_{X,Y}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{1.1}$$

相关系数 $\rho_{X,Y}$ 测度了 X 与 Y 这两个变量之间的线性关联程度，它不随变量刻度单位的变化而变化，并且取值在 -1 至 1 之间。图 1.3 展示了不同大小相关系数的示意图。值得指出的是，图 1.3(a)与图 1.3(i)给出了 $|\rho_{X,Y}|=1$ 的特殊情形，该种情形下，X 与 Y 呈现确定的线性函数关系：$Y=\beta_0+\beta_1 X$，其中 $\beta_1\neq 0$，若 $\beta_1>0$ 则有 $\rho_{X,Y}=1$；若 $\beta_1<0$ 则有 $\rho_{X,Y}=-1$。

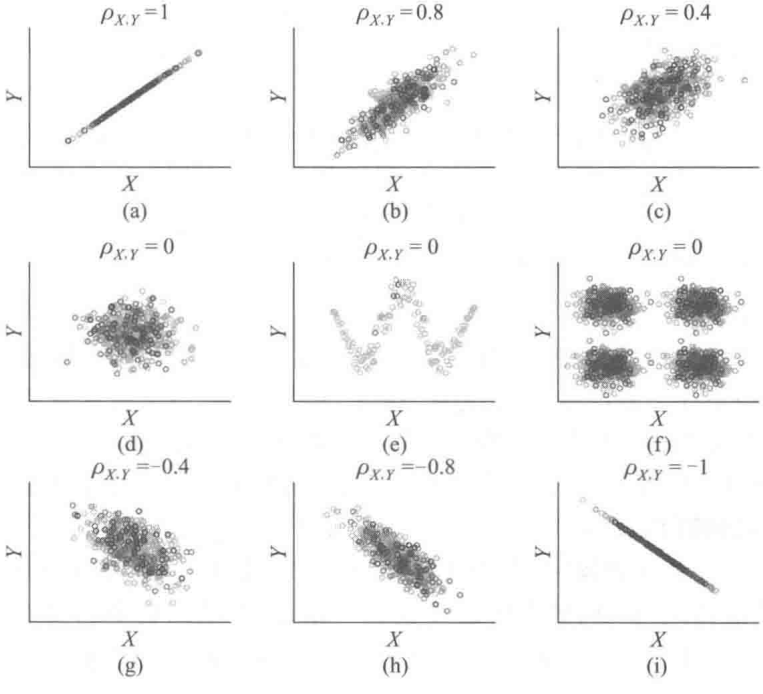

图 1.3 相关系数示意图

式(1.1)给出的是变量 X 与 Y 的总体相关系数。由于总体样本一般非常难以获得，相关系数通常基于样本来计算。一般地，若给定一组容量为 n 的样本 (X_i,Y_i) $(i=1,2,\cdots,n)$，则样本相关系数为：

$$\hat{\rho}_{X,Y}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{1.2}$$

其中，$\bar{X}$ 与 $\bar{Y}$ 分别表示变量 X 与 Y 的样本均值。一般而言，$\hat{\rho}_{X,Y}\neq\rho_{X,Y}$。然而，依据大数定律（第二章会对其进行详细介绍）可知，当样本量 n 很大时，样本相关系数 $\hat{\rho}_{X,Y}$ 可以以很高的概率收敛到总体相关系数 $\rho_{X,Y}$，图 1.4 直观地说明了这一点。具体而言，从总体相关系数 $\rho_{X,Y}$ 为 0.5 的二元正态分布中抽取不同容量的样本，并分别计算样本相关系数 $\hat{\rho}_{X,Y}$。从图 1.4 中可以清晰地看出，随着样本量 n 的增加，样本相关系数 $\hat{\rho}_{X,Y}$ 越来越靠近总体相关系数 $\rho_{X,Y}$ 的值 0.5。

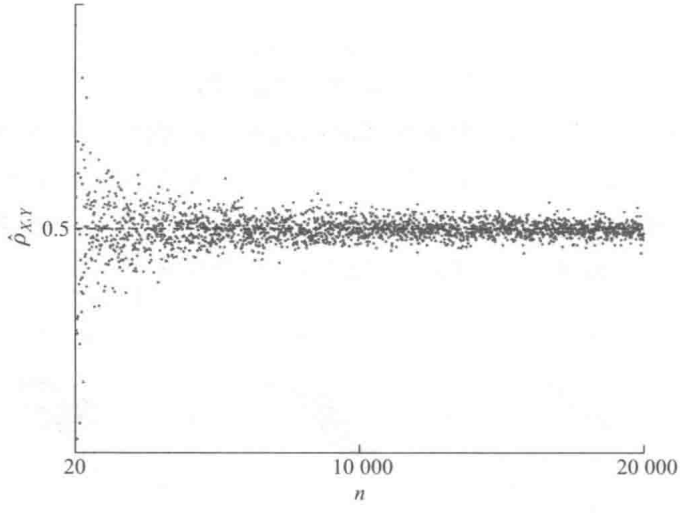

图 1.4 样本相关系数与总体相关系数

注：从总体相关系数 $\rho_{X,Y}$ 为 0.5 的二元正态分布中抽取不同容量的样本，并分别计算样本相关系数 $\hat{\rho}_{X,Y}$。

2. 预测（回归）关系

如前所述，在不确定性关系中，一般无法做到通过变量 X 精确地推知另一变量 Y。尽管如此，仍可在给定 X 的条件下最好地预测（估计、拟合）Y。[①] 比如，在分析受教育程度与收入水平关系的过程中，虽然无法通过一个人的受教育程度（X）来确定其收入水平（Y），但是可以根据受教育程度预测出最有可能出现的收入水平。给定 X 的取值对 Y 进行预测在计量经济学中被称为变量 Y 对变量 X 进行**回归**（Regression）。

在不同的准则下，“最好地预测”具有不同的含义，本教材主要讨论的是使预测误差平方的期望值最小。正式地，若令 $g(X)$ 表示变量 X 的任一函数，并采用它来近似地表示（估计、预测）Y，则有 $Y-g(X)$ 表示预测误差。在给定 X 的条件下最好地预测 Y 则意味着：

$$g^*(X)=\underset{g(X)}{\arg\min}\mathrm{E}[Y-g(X)]^2 \tag{1.3}$$

① 值得注意的是，这里的预测实际上是“找出”变量 Y 对变量 X 的某种依从的关系，而非通常意义上的已知当前状态对未来的预测。

即通过选择函数 $g(X)$ 使预测误差平方的期望值 $\mathrm{E}[Y-g(X)]^2$ 最小。图 1.5 比较直观地呈现了这一点，图形中的 $g^*(X)$ 是变量 X 所有函数中使 $\mathrm{E}[Y-g(X)]^2$ 最小的函数，因而也称函数 $g^*(X)$ 最好地拟合了变量 Y。

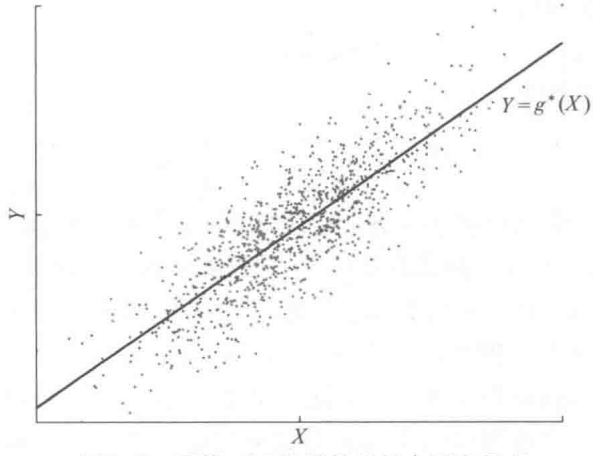

图 1.5 函数 $g^*(X)$ 最好地拟合了变量 Y

3. 因果关系

所谓因果关系是指，变量之间存在的相互作用关系，且其中一个变量被认为是导致另一变量变化的原因。关于不确定性关系，以上介绍了相关关系与预测（回归）关系。从逻辑上说，无论是相关关系还是回归关系本质上都是**统计关系**，它们并不必然意味着**因果关系**。因果关系是变量间最为本质的关系，它的判定需要先验知识或者经济理论。比如，将谷物收成对降雨量回归能够得到谷物收成如何依赖降雨量的统计关系；反过来，将降雨量对谷物收成回归则能够得到降雨量如何依赖谷物收成的统计关系。然而，先验知识告诉我们前者才是因果关系，而非后者。因为我们无法通过改变谷物收成的做法来改变降雨量。[①] 接下来，分别通过**控制实验**（Controlled Trial）与**"反事实"分析**（Counterfactual Analysis）这两种方式来介绍在计量经济学中如何来界定因果关系。

（1）**控制实验**。令 X 表示解释变量，Y 表示被解释变量（比如，可以将 X 与 Y 分别想象为前述例子中的受教育程度和收入水平），[②] ε 为除了 X 之外其他所有影响 Y 的因素（ε 也通常被称作误差项），则有：

$$Y=f(X,\varepsilon) \tag{1.4}$$

其中，无论是 X 还是 ε 对被解释变量 Y 的解释都具有独立的经济含义，该式在计量经济学中被称作**结构方程**（Structural Formula），ε 则称为**结构误差项**（Structural Error Term）。X 对 Y 的因果效应为：在控制了其他所有影响 Y 的因素后（或者给定其他所有影响 Y 的因素都不变）X 变化所引起的 Y 变化量。这实际上类似于自然科学中的控制实验，直观上，保持其他所有因素都固定不变时，Y 的变化只有可能是 X 的变化所引起，这就是因果效应。若令 τ 表示 X 对 Y 的因果效应，其表达式则为：

$$\tau=\mathrm{E}\left[\frac{\partial f(X,\varepsilon)}{\partial X}\right] \tag{1.5}$$

---

① 严格来讲，这里排除了人工降雨的情形，即谷物收成会影响到政府的人工降雨决策。这个时候降雨量对谷物收成回归所得到的统计关系则具有了因果解释。

② 第三章详细介绍了解释变量与被解释变量的具体含义。

其中，$\frac{\partial f(X,\varepsilon)}{\partial X}$随着 ε 的变化而变化，$\mathrm{E}(\cdot)$为 ε 的期望。τ 实际上表示的是 X 对 Y 的平均因果效应。

若 Y 是 X 的线性函数，则有：

$$Y=\alpha_0+\alpha_1 X+\varepsilon \tag{1.6}$$

在该情形下，X 对 Y 的因果效应为：

$$\tau=\mathrm{E}\left[\frac{\partial(\alpha_0+\alpha_1 X+\varepsilon)}{\partial X}\right]=\alpha_1 \tag{1.7}$$

(2)“反事实”分析。假设这里感兴趣的问题仍然是受教育程度对收入水平产生的影响。为方便理解，令变量 X_i 表示个体 i 的受教育程度，且只取 0 和 1 这两个值。其中，$X_i=1$ 表示个体 i 接受大学教育；$X_i=0$ 表示个体 i 未接受大学教育。假定 Y_{1i} 与 Y_{0i} 为个体 i 的潜在收入水平。其中，Y_{1i} 表示如果个体 i 接受大学教育的收入水平，Y_{0i} 表示如果个体 i 未接受大学教育的收入水平。对于每一个体而言，均同时存在 Y_{1i} 与 Y_{0i} 这两个潜在结果，但是却只能观测到其中的一个。对于接受大学教育的个体，只能观测到 Y_{1i}，无法观测到该个体如果没有接受大学教育的收入水平 Y_{0i}，此时，Y_{0i} 是“反事实”收入水平；反之，对于未接受大学教育的个体，只能观测到 Y_{0i}，无法观测到如果接受大学教育的收入水平 Y_{1i}，此时，Y_{1i} 是非大学生群体的“反事实”收入水平。在上述“反事实”框架中，[①]我们能够非常自然地定义因果关系。具体而言，接受大学教育对个体 i 收入水平的因果效应可以表示为：

$$\tau_i=Y_{1i}-Y_{0i} \tag{1.8}$$

对于接受大学教育的个体 i 来说，τ_i 表示其可观测收入 Y_{1i} 减去“反事实”收入 Y_{0i}；对于未接受大学教育的个体 i 而言，τ_i 表示其“反事实”收入 Y_{1i} 减去观测收入 Y_{0i}。

考虑到每一个体的因果效应可能存在差异，有时我们更加关心平均因果效应。它可以表示为：

$$\tau=\mathrm{E}(Y_{1i}-Y_{0i}) \tag{1.9}$$

本节最后，我们将变量间的上述关系总结在图 1.6 中。

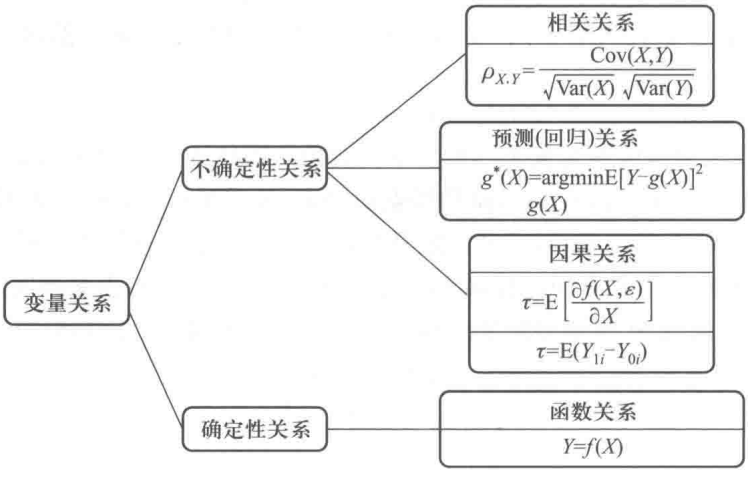

图 1.6 变量间的关系

① 在第六章中，我们还会对反事实框架进行更加细致的介绍。

1.2 经济数据:来源与结构特征

从来源或者获取方式的视角上来看,经济数据可以划分为**实验数据**(Experimental Data)和**观测数据**(Observational Data)。所谓实验数据是指那些通过类似于自然科学中的**控制实验**或者**随机试验**(Randomized Experiment)的方式所得到的数据。观测数据是指那些自然生成的数据。从数据结构特征的视角上来看,经济数据又可划分为**截面数据**(Cross-sectional Data)、**时间序列数据**(Time Series Data)、**面板数据**(Panel Data)以及**混合截面数据**(Pooled Data)四类。无论是实验数据还是观测数据均可以是这四类数据类型中的任意一种。

一、实验数据

一般而言,由于在经济学中无法像自然科学那样做实验,因而经济数据基本上是观测数据。尽管如此,获得并基于实验数据开展计量经济分析,已经越来越成为大家普遍接受的研究范式。为此,我们通过一个假想的例子来简要介绍实验数据。具体而言,假设政府提供一个就业培训项目,并通过抽签的方式来随机地选择哪些人参加该项目,哪些人不参加该项目。被选中参加项目的个人不能不参加项目,与此同时,没有被选中参加项目的个人也不能参加项目。在项目实施后,表 1.1 分组报告了两组人受教育年限和工资收入的数据。

表 1.1 实验数据(一个演示性例子)

参加就业培训组			未参加就业培训组		
序号	受教育年限(年)	工资收入(万元)	序号	受教育年限(年)	工资收入(万元)
1	12	9.61	1	10	7.06
2	6	6.81	2	15	9.13
3	14	10.37	3	11	6.70
4	11	8.73	4	10	8.75
5	8	12.54	5	12	11.11
6	9	12.32	6	23	7.36
7	11	8.69	7	10	8.36
8	10	11.56	8	12	6.56
9	9	8.44	9	8	8.09
10	8	9.82	10	14	8.26
⋮	⋮	⋮	⋮	⋮	⋮
5 000	6	9.09	5 000	8	7.12

直观上而言,由于个体是否参加就业培训项目是完全随机决定的,因此参加就业培训项目群体与未参加就业培训项目群体的特征应当基本类似。这可以通过图 1.7 看出来。图 1.7 分别绘制了参加就业培训群体与未参加就业培训群体受教育年限的直方图,可以发现,二者非常相近。

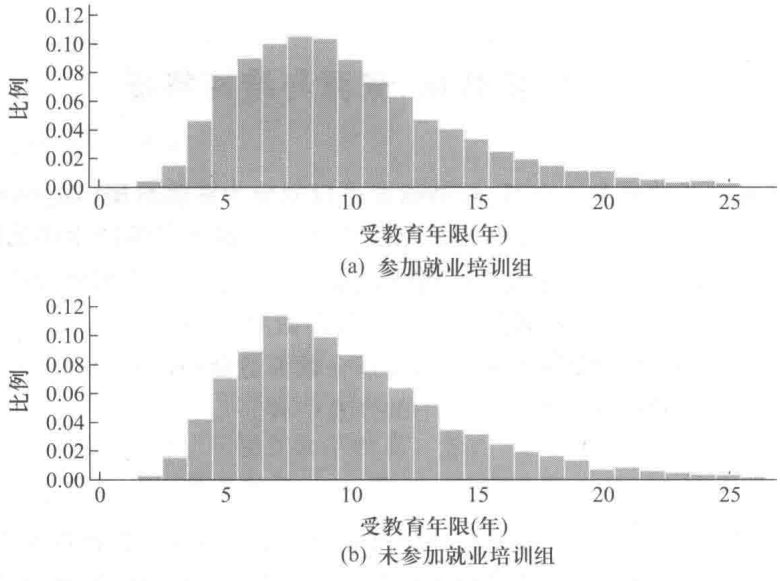

图 1.7 受教育年限直方图

此外,如果就业培训项目对个人工资收入真的具有影响,那么上述随机试验条件下,我们应该能够观测到,参加项目群体的工资收入应当高于未参加项目群体。图 1.8 演示了这一点。从图 1.8 中可以清楚地发现,与未参加就业培训群体相比,参加就业培训群体工资收入的分布明显整体偏右。值得指出的是,由于在随机试验条件下参加就业培训群体与未参加就业培训群体的特征基本类似,那么这两组人收入的差别就可以认为是由就业培训项目导致。这个时候我们说识别了就业培训项目对收入的因果效应。

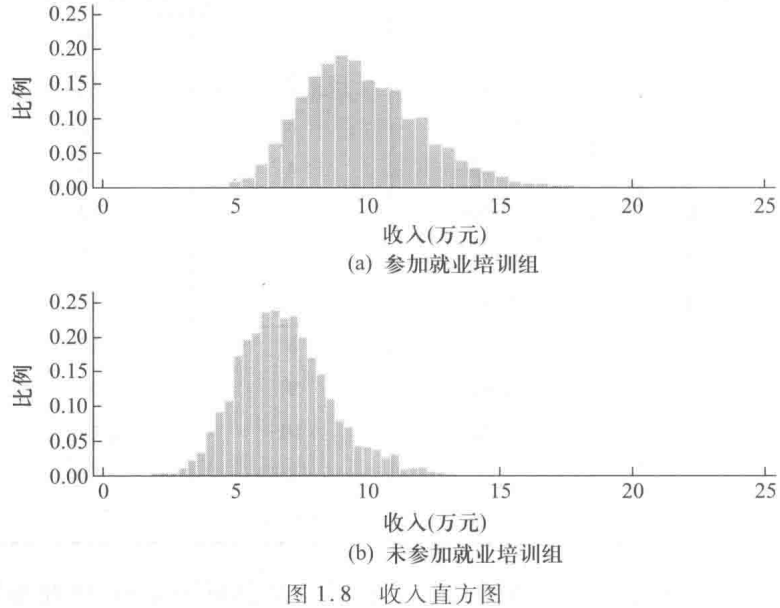

图 1.8 收入直方图

二、数据结构

如前所述,从数据结构特征的视角上来看,经济数据可以划分为截面数据、时间序列数据、面

板数据以及混合截面数据四类。

截面数据是指,变量在同一时间点不同个体上的取值。比如,2016 年我国部分地区 GDP 数据就是截面数据的典型例子(表 1.2)。

表 1.2 全国部分地区 2016 年 GDP 数据

地区	地区代码	GDP(十亿元)	地区	地区代码	GDP(十亿元)
北京	11	2 567	湖北	42	3 267
天津	12	1 789	湖南	43	3 155
河北	13	3 207	广东	44	8 085
山西	14	1 305	广西	45	1 832
内蒙古	15	1 813	海南	46	405
辽宁	21	2 225	重庆	50	1 774
吉林	22	1 478	四川	51	3 293
黑龙江	23	1 539	贵州	52	1 178
上海	31	2 818	云南	53	1 479
江苏	32	7 739	西藏	54	115
浙江	33	4 725	陕西	61	1 940
安徽	34	2 441	甘肃	62	720
福建	35	2 881	青海	63	257
江西	36	1 850	宁夏	64	317
山东	37	6 802	新疆	65	965
河南	41	4 047			

时间序列数据是指,变量在同一个体不同时间点上的取值。比如,1997—2016 年 GDP 数据就是典型的时间序列数据(表 1.3)。

表 1.3 1997—2016 年 GDP 数据

年份	GDP(十亿元)	年份	GDP(十亿元)
1997	7 972	2007	27 023
1998	8 520	2008	31 952
1999	9 056	2009	34 908
2000	10 028	2010	41 303
2001	11 086	2011	48 930
2002	12 172	2012	54 037
2003	13 742	2013	59 524
2004	16 184	2014	64 397
2005	18 732	2015	68 905
2006	21 944	2016	74 359

面板数据是指,变量在不同个体不同时间点上的取值。比如,全国部分地区 2015—2016 年 GDP 数据就是一个面板数据(表 1.4)。可以看出,面板数据是截面数据和时间序列数据的混合,它兼具截面数据和时间序列数据的特征。面板数据要求同一经济个体在不同年份重复出现。

表 1.4 全国部分地区 2015—2016 年 GDP 数据

地区名称	地区代码	年份	GDP(十亿元)
北京	11	2015	2 301
北京	11	2016	2 567
天津	12	2015	1 654
天津	12	2016	1 789
河北	13	2015	2 981
河北	13	2016	3 207
山西	14	2015	1 277
山西	14	2016	1 305
内蒙古	15	2015	1 783
内蒙古	15	2016	1 813
辽宁	21	2015	2 867
辽宁	21	2016	2 225
吉林	22	2015	1 406
吉林	22	2016	1 478
黑龙江	23	2015	1 508
黑龙江	23	2016	1 539
上海	31	2015	2 512
上海	31	2016	2 818
⋮	⋮	⋮	⋮
新疆	65	2015	932
新疆	65	2016	965

除了截面数据、时间序列数据以及面板数据外,另一类常见的数据是混合截面数据。与面板数据类似,该类型数据兼具截面数据和时间序列数据的特征,然而不同的是,它并不要求同一个体在不同年份重复出现。也就是说,在不同年份,数据中的个体可能完全不同。混合截面数据可以简单地理解为,不同时间截面数据的简单混合。

习 题

1.1 根据表 1.5 所给出的数据,计算我国 GDP 与社会消费品零售总额的相关系数。