第 1 章　导论

1.1　什么是计量经济学

计量经济学(econometrics)，是运用概率统计方法对经济变量之间的(因果)关系进行定量分析的科学。

由于实验数据的缺乏，计量经济学常常不足以确定经济变量之间的因果关系。

但大多数实证分析的目的恰恰正是要确定变量之间的因果关系(即 X 是否导致 Y)，而非仅仅是相关关系。

例(相关关系) 你看到街上的人们带雨伞，于是预测今天要下雨。这只是一种相关关系，“人们带伞”并不导致“下雨”。

例(相关关系) 根据与流感相关的海量词条搜索记录，谷歌公司通过分析大数据(big data)，可以很快地预测流行病的地域传播。这也只是相关关系，上网搜索流感信息并不导致流感的传播。

如果只对预测感兴趣，相关关系就足够了。

如果要推断变量之间的因果关系，则计量分析必须建立在经济理论的基础之上，即在理论上存在X导致Y的作用机制。

但即使有理论基础，因果关系常常依然不好分辨。

首先，可能存在**逆向因果关系**(reverse causality)或“双向因果关系”。

例(逆向因果)　FDI(外商直接投资)促进经济增长，但 FDI 也被吸引到快速增长的地区。

例(逆向因果)　收入增加引起消费增长，而消费增长也拉动收入增加。

例(逆向因果)　经济萧条可能引起内战，但内战也会导致经济停滞。

其次，被遗漏的第三个变量(Z)也可能对这两个变量(X, Y)同时起作用，参见图 1.1。

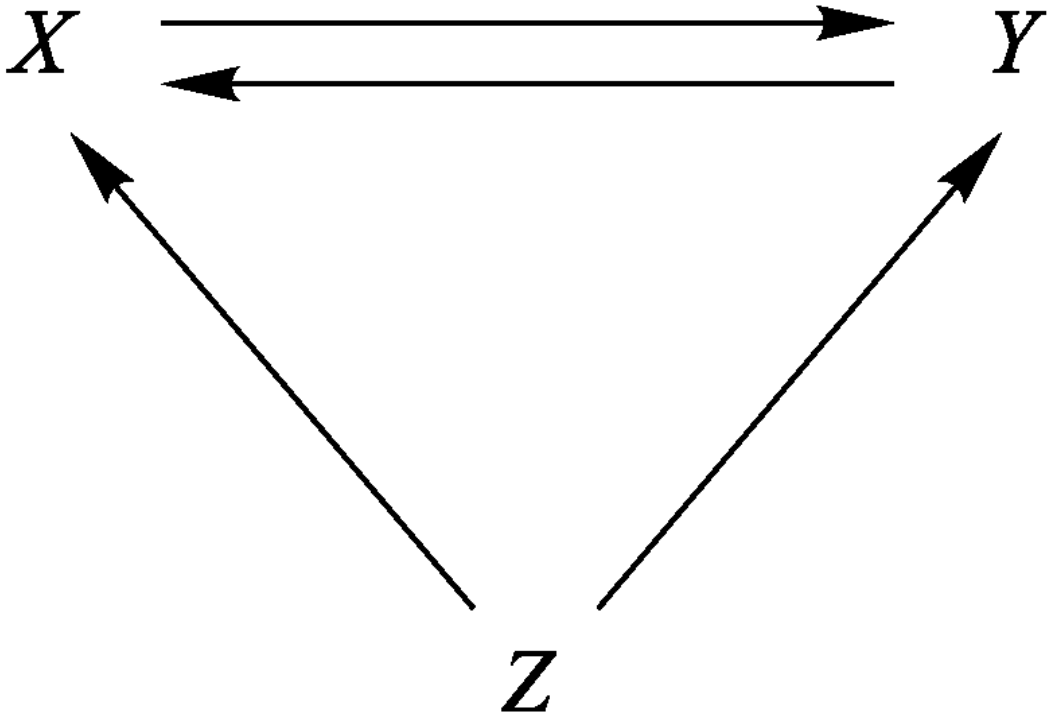

图 1.1　可能的因果关系

例(遗漏变量)　挪威西海岸蚊子的数量与该地区的游客数量高度相关。

例(遗漏变量)　某外星人来到地球，发现人类会死亡，十分不解。于是开始在全球广泛观察死亡现象，并收集了大量的数据。

结果发现，许多人类躺在医院病床(X)之后死去(Y)，故推断医院病床是死亡的原因。

外星人认为，由于躺在医院病床上，总是发生于死亡之前，故不可能存在逆向因果关系。

外星人于是将研究报告投稿发表于某顶尖经济学期刊，并在文末给出政策建议“珍爱生命，远离病床”。

例(遗漏变量)　考虑决定教育投资回报率(returns to schooling)的因素：

$$\ln w_i = \alpha + \beta s_i + \varepsilon_i \qquad (1.1)$$

$\ln w_i$(工资对数)：**被解释变量**(dependent variable)

s_i (schooling, 教育年限)：**解释变量**(explanatory variable, regressor)、**自变量**(independent variable)或**协变量**(covariate)

ε_i：不可观测(unobservable)的**误差项**(error term)或**随机扰动项**(stochastic disturbance)，包括所有除s_i以外对$\ln w_i$有影响的因素，以及人类行为的随机性。

下标 i：表示第 i 个观测值(即个体 i)。

截距项α与斜率β为待估参数。

β的经济含义为教育投资的回报率，即多上一年学，未来工资能增加百分之几。

如果用数据估计此回归方程，其结果一般会显示，对数工资与教育年限显著正相关，而且教育投资回报率β还挺高。

然而，一个人的工资收入也与能力有关，但能力一般不能直接观测，而能力高的人通常选择接受更多教育。

在此简单的回归中，教育的高回报其实包含了对能力的回报。

影响工资收入的因素还可能包括工作经验、毕业学校、人种、性别、外貌等。

需尽可能多地引入**控制变量**(control variables)，也就是多元回归的方法，才能较准确地估计我们**感兴趣的参数**(parameters of interest)，即本例中的教育投资回报率β。

现实中总有某些相关的变量无法观测，即存在**遗漏变量**(omitted variables)，而遗漏变量统统被纳入到随机扰动项ε_i中。

随机扰动项ε_i中还可能包含哪些其他因素呢？如果真实模型(true model)为

$$\ln w_i = \alpha + \beta s_i + \gamma s_i^2 + \varepsilon_i \qquad (1.2)$$

则γs_i^2也被纳入到扰动项中(可视为广义的遗漏变量)。

如果变量测量得不准确，则测量误差也被放入扰动项中。

扰动项就像是一个“垃圾桶”，所有你不想要、无法把握的东西都往里面扔。

但我们又希望扰动项拥有很好的性质。在很多情况下，这是自相矛盾的。

计量经济学的很多玄妙之处就在于扰动项。

如果真正理解了扰动项，也就加深了对计量经济学的理解。

1.2 经济数据的特点与类型

由于在经济学中通常无法像自然科学那样做**控制实验**(controlled experiment)，故经济数据一般不是**实验数据**(experimental data)，而是自然发生的**观测数据**(observational data)；比如，统计局所收集的数据。

由于个人行为的随机性，所有经济变量原则上都是随机变量。

在有些本科计量教材中，为简单起见，有时假设解释变量是非随机的、固定的(fixed regressors)。

这只是教学法上的权宜之计，给更深入的理论探讨带来了不便。如果解释变量为非随机，则无法考虑其与扰动项的相关性。

在本教程中，所有变量都是随机的(即使非随机的常数，也可视为退化的随机变量)。

经济数据按照其性质，可大致分成以下三种类型。

(1) **横截面数据**(cross-sectional data，简称截面数据)：

指多个经济个体的变量在同一时点上的取值。

比如，2022 年中国各省的 GDP，参见表 1.1。

表 1.1　2022 年中国分省 GDP (亿元)

省份	GDP
北京	41611
天津	16311.3
河北	42370.4
山西	25642.6
内蒙古	23158.7
辽宁	28975.1
吉林	13070.2
黑龙江	15901
上海	44652.8
江苏	122875.6

浙江	77715.4
安徽	45045
福建	53109.9
江西	32074.7
山东	87435.1
河南	61345.1
湖北	53734.9
湖南	48670.4
广东	129118.6
广西	26300.9
海南	6818.2
重庆	29129
四川	56749.8

贵州	20164.6
云南	28954.2
西藏	2132.6
陕西	32772.7
甘肃	11201.6
青海	3610.1
宁夏	5069.6
新疆	17741.3

数据来源：国家统计局网站(http://data.stats.gov.cn)

(2) **时间序列数据**(time series data)：

指某个经济个体的变量在不同时点上的取值。

比如，2003—2022 年山东省每年的 GDP，参见表 1.2。

表 1.2　2003-2022 年山东省 GDP (亿元)

年份	GDP
2003	10903.2
2004	13308.1
2005	15947.5
2066	18967.8
2007	22718.1
2008	27106.2

2009	29540.8
2010	33922.5
2011	39064.9
2012	42957.3
2013	47344.3
2014	50774.8
2015	55288.8
2016	58762.5
2017	63012.1
2018	66648.9
2019	70540.5
2020	72798.2
2021	82875.2
2022	87435.1

数据来源：国家统计局网站(http://data.stats.gov.cn)

(3) 面板数据(panel data)：

指多个经济个体的变量在不同时点上的取值。

比如，2003—2022 年中国各省每年的 GDP，参见表 1.3。

表 1.3　2003-2022 年中国分省 GDP (亿元)

省份	年份	GDP
北京	2003	5267.2
北京	2004	6252.5
⋮	⋮	⋮
北京	2021	41045.6
北京	2022	41611

天津	2003	2257.8
天津	2004	2621.1
⋮	⋮	⋮
天津	2021	15685.1
天津	2022	16311.3
⋮	⋮	⋮
新疆	2003	1889.2
新疆	2004	2170.4
⋮	⋮	⋮
新疆	2021	16311.6
新疆	2022	17741.3

数据来源：国家统计局网站(http://data.stats.gov.cn)

本书介绍的计量经济理论包括以上三种数据类型，

将使用国际上最为流行的 Stata 计量软件进行数据处理(Stata 17 版本，2021 年发布)。

第 2 章介绍 Stata 软件。

第 3 章将回顾相关数学知识，并引入一些新概念(比如，均值独立、迭代期望定律)。

第 4 章将正式进入计量经济学的理论部分。