第 1 章

回归分析概述

为了在系统学习回归分析之前对该课程的思想方法、主要内容、发展现状等有个概括的了解，本章将由变量间的统计关系引申出社会经济与自然科学等现象中的相关与回归问题，并扼要介绍“回归”名称的由来及近代回归分析的发展、回归分析研究的主要内容，以及建立回归模型的步骤与建模过程中应注意的问题。

1.1 变量间的相关关系

社会经济与自然科学等现象之间的相互联系和制约是一个普遍规律。例如社会经济的发展总是与一定的经济变量的数量变化紧密联系着。社会经济现象不仅同和它有关的现象构成一个普遍联系的整体，而且在它的内部存在着许多彼此关联的因素，在一定的社会环境、地理条件、政府决策影响下，一些因素推动或制约另外一些与之联系的因素发生变化。这种状况表明，在经济现象的内部和外部联系中存在着一定的相关性，人们往往利用这种相关关系来制定有关的经济政策，以指导、控制社会经济活动的发展。要认识和掌握客观经济规律就必须探求经济现象中经济变量的变化规律，变量间的统计关系是经济变量变化规律的重要特征。

互有联系的经济现象及经济变量间关系的紧密程度各不一样。一种极端的情况是一个变量的变化能完全决定另一个变量的变化。例如，一家保险公司承保汽车 5 万辆，每辆保费收入为 1 000 元，则该保险公司汽车承保总收入为 5 000 万元。如果把承保总收入记为 y，承保汽车辆数记为 x，则 $y = 1\,000x$。x 与 y 两个变量间完全表现为一种确定性关系，即函数关系，如图 1-1 所示。

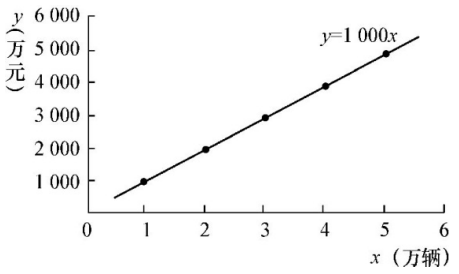

图 1-1　函数关系图

又如，银行的一年期存款利率为 2.55%，存入的本金用 x 表示，到期的本息用 y 表示，则 $y = x+2.55\%x$。这里 y 与 x 仍表现为一种函数关系。对于任意两个变量间的函数关系，可以表述为下面的数学形式

$$y = f(x)$$

再如，工业企业的原材料消耗总额用 y 表示，生产量用 x_1 表示，单位产量消耗用 x_2 表示，原材料价格用 x_3 表示，则

$$y = x_1 x_2 x_3$$

这里的 y 与 x_1，x_2，x_3 仍是一种确定性的函数关系，但它们显然不是线性函数关系。我们可以将变量 y 与 p 个变量 $x_1, x_2, \cdots, x_p$ 之间存在的某种函数关系用下面的形式表示

$$y = f(x_1, x_2, \cdots, x_p)$$

经济问题中还有很多函数关系的例子。物理学中的自由落体距离公式、初等数学中的许多计算公式等表示的都是变量间的函数关系。

然而，现实世界中还有不少情况是两事物之间有着密切的联系，但它们密切的程度并没有到由一个可以完全确定另一个的地步，下面举几个例子。

(1)我们都知道某种高档消费品的销售量与城镇居民的收入密切相关，居民收入高，这种消费品的销售量就大。但是由居民收入 x 并不能完全确定某种高档消费品的销售量 y，因为这种高档消费品的销售量还受人们的消费习惯、心理因素、其他商品的吸引程度及价格的高低等诸多因素的影响。这样变量 y 与变量 x 就是一种非确定的关系，如图 1-2 所示。

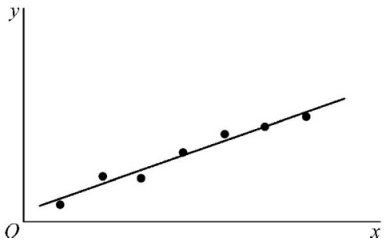

图 1-2　y 与 x 非确定性关系图

(2)粮食产量 y 与施肥量 x 之间有密切的关系，在一定的范围内，施肥量越多，粮食产量就越高。但是，施肥量并不能完全确定粮食产量，因为粮食产量还与其他因素有关，如降雨量、田间管理水平等。因此粮食产量 y 与施肥量 x 之间不存在确定的函数关系。

(3)储蓄额与居民的收入密切相关，但是由居民收入并不能完全确定储蓄额。因为影响储蓄额的因素很多，如通货膨胀、股票价格指数、利率、消费观念、投资意识等。因此尽管储蓄额与居民收入有密切的关系，但它们之间并不存在一种确定性关系。

再如广告费支出与商品销售额、保险利润与保费收入、工业产值与用电量等。这方面的例子不胜枚举。

以上变量间关系的一个共同特征是尽管密切，但却是一种非确定性关系。由于经济问题的复杂性，有许多因素因为我们的认识以及其他客观原因的局限，并没有包含在内，或者由于试验误差、测量误差以及其他种种偶然因素的影响，使得另外一个或一些变量的取值带有一定的随机性。因此当一个或一些变量取定值后，不能以确定值与之对应。

从图 1-1 看到确定性的函数关系，各对应点完全落在一条直线上。而由图 1-2 看到，各对应点并不完全落在一条直线上，即有的点在直线上，有的点在直线的两侧。这种对应点不能分布在一条直线上的变量间的关系，也就是变量 x 与 y 之间有一定的关系，但是又没有密切到可以通过 x 唯一确定 y 的程度，这种关系正是统计学研究的

重要内容。在推断统计中，我们把上述变量间具有密切关联而又不能由某一个或某一些变量唯一确定另外一个变量的关系称为变量间的统计关系或相关关系。这种统计关系的规律性是统计学中研究的主要对象，现代统计学中关于统计关系的研究已形成两个重要的分支，它们叫回归分析和相关分析。

回归分析和相关分析都是研究变量间关系的统计学课题。在应用中，两种分析方法经常相互结合和渗透，但它们研究的侧重点和应用面不同。它们的差别主要有以下几点：一是在回归分析中，变量 y 称为因变量，处在被解释的特殊地位。在相关分析中，变量 y 与变量 x 处于平等的地位，即研究变量 y 与变量 x 的密切程度与研究变量 x 与变量 y 的密切程度是一回事。二是相关分析中所涉及的变量 y 与 x 全是随机变量。而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。通常的回归模型中，我们总是假定 x 是非随机的确定变量。三是相关分析的研究主要是为刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

由于回归分析与相关分析研究的侧重点不同，它们的研究方法也大不相同。回归分析已成为现代统计学中应用最广泛、研究最活跃的一个独立分支。

1.2 “回归”思想及名称的由来

回归分析是处理变量 x 与 y 之间的关系的一种统计方法和技术。这里所研究的变量之间的关系就是上述的统计关系，即当给定 x 的值，y 的值不能确定，只能通过一定的概率分布来描述。于是，我们称给定 x 时 y 的条件数学期望

$$f(x)=E(y\mid x) \tag{1.1}$$

为随机变量 y 对 x 的回归函数，或称为随机变量 y 对 x 的均值回归函数。式(1.1)从平均意义上刻画了变量 x 与 y 之间的统计规律。

在实际问题中，我们把 x 称为自变量，y 称为因变量。如果要由 x 预测 y，就是要利用 x，y 的观察值，即样本观测值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) \tag{1.2}$$

来建立一个函数，当给定 x 值后，代入此函数中算出一个 y 值，这个值就称为 y 的预测值。如何建立这个函数？这就要从样本观测值 (x_i, y_i) 出发，观察 (x_i, y_i) 在平面直角坐标系上的分布情况，图 1-2 就是居民收入与商品销售量的散点图。由这个图可看出样本点基本上分布在一条直线的周围，因而要确定商品销售量 y 与居民收入 x 的关系，可考虑用一个线性函数来描述。图 1-2 中的直线即线性方程

$$E(y\mid x)=\alpha+\beta x \tag{1.3}$$

方程式(1.3)中的参数 α，β 尚不知道，这就需要由样本数据(1.2)去进行估计。具

体如何估计参数 α,β，我们将在第 2 章中详细介绍。

当我们由样本数据(1.2)估计出 α,β 的值后，用估计值 $\hat{\alpha},\hat{\beta}$ 分别代替式(1.3)中的 α,β，得方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x \tag{1.4}$$

方程式(1.4)就称为回归方程。这里因为因变量 y 与自变量 x 呈线性关系，故称式(1.4)为 y 对 x 的线性回归方程。又因式(1.4)的建立依赖于观察或试验积累的数据(1.2)，所以又称式(1.4)为经验回归方程。相对这种叫法，我们把式(1.3)称为理论回归方程。理论回归方程是设想把所研究问题的总体中每一个体的(x, y)值都测量了，利用其全部测量结果而建立的回归方程，这在实际中是做不到的。理论回归方程中的 α 是方程式(1.3)所画出的直线在 y 轴上的截距，β 为直线的斜率，它们分别称为回归常数和回归系数。而方程式(1.4)中的参数 $\hat{\alpha},\hat{\beta}$ 称为经验回归常数和经验回归系数。

回归分析的基本思想和方法以及“回归”(regression)名称的由来归功于英国统计学家 F.高尔顿(F.Galton，1822—1911)。高尔顿和他的学生、现代统计学的奠基者之一 K.皮尔逊(K.Pearson，1856—1936)在研究父母身高与其子女身高的遗传问题时，观察了 1078 对夫妇，以每对夫妇的平均身高作为 x，而取他们的一个成年儿子的身高作为 y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为

$$\hat{y}=33.73+0.516x \tag{1.5}$$

这种趋势及回归方程总的表明父母平均身高 x 每增加一个单位，其成年儿子的身高 y 平均增加 0.516 个单位。这个结果表明，虽然高个子父辈的确有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。反之，矮个子父辈的确有生矮个子儿子的趋势，但父辈身高减少一个单位，儿子身高仅减少半个单位左右。通俗地说，一群特高个子父辈(例如排球运动员)的儿子们在同龄人中平均仅为高个子，一群高个子父辈的儿子们在同龄人中平均仅为略高个子；一群特矮个子父辈的儿子们在同龄人中平均仅为矮个子，一群矮个子父辈的儿子们在同龄人中平均仅为略矮个子，即子代的平均高度向中心回归了。正是因为子代的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。这个例子生动地说明了生物学中“种”的概念的稳定性。正是为了描述这种有趣的现象，高尔顿引进了“回归”这个名词来描述父辈身高 x 与子辈身高 y 的关系。尽管“回归”这个名称的由来具有其特定的含义，而在人们研究的大量问题中，其变量 x 与 y 之间的关系并不总是具有这种“回归”的含义，但仍借用这个名词把研究变量 x 与 y 间统计关系的量化方法称为“回归”分析，也算是对高尔顿这位伟大的统计学家的纪念。

1.3　回归分析的主要内容及其一般模型

1.3.1　回归分析研究的主要内容

回归分析研究的主要对象是客观事物变量间的统计关系，它是建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的统计方法。回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态及进行模型预测的一种有效的工具。

回归分析方法在生产实践中的广泛应用是其发展和完善的根本动力。如果从 19 世纪初(1809 年)高斯(Gauss)提出最小二乘法算起，回归分析的历史已有 200 多年。从经典的回归分析方法到近代的回归分析方法，它们所研究的内容已非常丰富。如果按研究的方法来划分，回归分析研究的范围大致如下：

- 回归分析
  - 线性回归
    - 一元线性回归
    - 多元线性回归
    - 多个因变量与多个自变量的回归
  - 回归诊断
    - 讨论如何从数据推断回归模型基本假设的合理性
    - 当基本假设不成立时如何对模型进行修正
    - 判定回归方程拟合的效果
    - 选择回归函数的形式
  - 回归变量的选择
    - 自变量选择的准则
    - 逐步回归分析方法
  - 参数估计方法的改进
    - 岭回归
    - 主成分回归
    - 偏最小二乘法
  - 非线性回归
    - 一元非线性回归
    - 分段回归
    - 多元非线性回归
  - 含有定性变量的回归
    - 自变量含定性变量的情况
    - 因变量是定性变量的情况

1.3.2　回归模型的一般形式

如果变量 x_1, x_2,…, x_p 与随机变量 y 之间存在着相关关系，通常就意味着每当 x_1, x_2,…, x_p 取值确定后，y 便有相应的概率分布与之对应。随机变量 y 与相关变量 x_1, x_2,…, x_p 之间的模型为

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon \tag{1.6}$$

式中，随机变量 y 称为被解释变量(因变量)；$x_1, x_2,\cdots, x_p$ 称为解释变量(自变量)。在计量经济学中，也称因变量为内生变量，自变量为外生变量。$f(x_1, x_2,\cdots, x_p)$ 为一般变量 $x_1, x_2,\cdots, x_p$ 的确定性关系；ε 为随机误差。正是因为随机误差项 ε 的引入，才将变量之间的关系描述为一个随机方程，使得我们可以借助随机数学方法研究 y 与 $x_1, x_2,\cdots, x_p$ 的关系。由于客观经济现象是错综复杂的，一种经济现象很难用有限个因素来准确说明，随机误差项可以概括表示由于人们的认识以及其他客观原因的局限而没有考虑的种种偶然因素。随机误差项主要包括下列因素的影响：

(1)由于人们认识的局限或时间、费用、数据质量等的制约未引入回归模型但又对回归被解释变量 y 有影响的因素。

(2)样本数据的采集过程中变量观测值的观测误差。

(3)理论模型设定的误差。

(4)其他随机因素。

模型式(1.6)清楚地表达了变量 $x_1, x_2,\cdots, x_p$ 与随机变量 y 的相关关系，它由两部分组成：一部分是确定性函数关系，由回归函数 $f(x_1, x_2,\cdots, x_p)$ 给出；另一部分是随机误差项 ε。由此可见模型式(1.6)准确地表达了相关关系既有联系又不确定的特点。

当模型式(1.6)中回归函数为线性函数时，即有

$$y = \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{1.7}$$

式中，$\beta_0, \beta_1, \beta_2,\cdots, \beta_p$ 为未知参数，常称为回归系数。线性回归模型的“线性”是针对未知参数 $\beta_i (i = 0, 1, 2,\cdots, p)$ 而言的。回归解释变量的线性是非本质的，因为解释变量是非线性时，常可以通过变量的替换把它转化成线性的。

如果 $(x_{i1}, x_{i2},\cdots, x_{ip}; y_i)$ $(i = 1, 2,\cdots, n)$ 是式(1.7)中变量 $(x_1, x_2,\cdots, x_p; y)$ 的一组观测值，则线性回归模型可表示为

$$y_i = \beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i, \quad i = 1, 2,\cdots, n \tag{1.8}$$

为了估计模型参数的需要，古典线性回归模型通常应满足以下几个基本假设。

(1)解释变量 $x_1, x_2,\cdots, x_p$ 是非随机变量，观测值 $x_{i1}, x_{i2},\cdots, x_{ip}$ 是常数。

(2)等方差及不相关的假定条件为

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j \\ 0, \quad i\neq j\end{cases} \quad i,j=1,2,\cdots,n \end{cases}$$

这个条件称为高斯–马尔柯夫(Gauss-Markov)条件，简称 G-M 条件。在此条件下，便可以得到关于回归系数的最小二乘估计及误差项方差 σ^2 估计的一些重要性质，如回归系数的最小二乘估计是回归系数的最小方差线性无偏估计等。

(3)正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), \quad i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\text{相互独立} \end{cases}$$

在此条件下便可得到关于回归系数的最小二乘估计及 σ^2 估计的进一步结果，并且可以进行回归的显著性检验及区间估计。

(4)通常为了便于数学上的处理，还要求 $n>p$，即样本量的个数要多于解释变量的个数。

在整个回归分析中，线性回归的统计模型最为重要。一方面是因为线性回归的应用最广泛；另一方面是只有在回归模型为线性的假定下，才能得到比较深入和一般的结果；此外，有许多非线性的回归模型可以通过适当的变换转化为线性回归问题处理。因此，线性回归模型的理论和应用是本书研究的重点。

对线性回归模型通常要研究的问题如下。

(1)如何根据样本 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)$ $(i = 1, 2, \cdots, n)$ 求出 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 及方差 σ^2 的估计。

(2)对回归方程及回归系数的种种假设进行检验。

(3)如何根据回归方程进行预测和控制，以及如何进行实际问题的结构分析。

1.4　回归模型的建立过程

在实际问题的回归分析模型的建立和分析中有几个重要的阶段，为了给读者一个整体印象，我们以经济模型的建立为例，先用逻辑框图表示回归模型的建立过程(见图 1-3)。

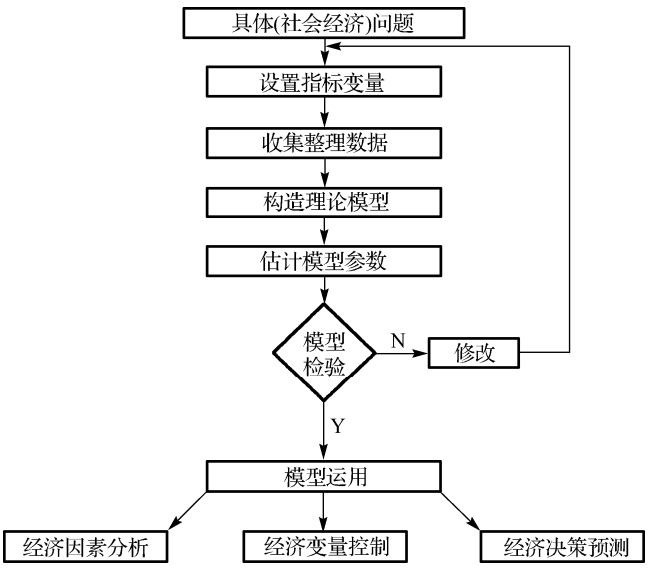

图 1-3　回归建模步骤流程图

下面按逻辑框图顺序叙述每个阶段要做的工作以及应注意的问题。

1.4.1 根据目的设置指标变量

回归分析模型主要是揭示事物间相关变量的数量联系。首先要根据所研究问题的目的设置因变量 y，然后再选取与 y 有统计关系的一些变量作为自变量。

通常情况下，我们希望因变量与自变量之间具有因果关系。尤其是在研究某种经济活动或经济现象时，必须根据具体的经济现象的研究目的，利用经济学理论，从定性角度来确定某种经济问题中各因素之间的因果关系。当把某一经济变量作为“果”之后，接着更重要的是正确选择作为“因”的变量。在经济问题回归模型中，前者被称为“内生变量”或“被解释变量”，后者被称为“外生变量”或“解释变量”。正确选择变量的关键在于能否正确把握所研究的经济活动的经济学内涵。这就要求研究者对所研究的经济问题及其背景有足够的了解。例如，要研究中国通货膨胀问题，必须懂得一些金融理论。通常把全国零售物价总指数作为衡量通货膨胀的重要指标，那么，全国零售物价总指数作为被解释变量，影响全国零售物价总指数的有关因素就作为解释变量。

对一个具体的经济问题，当研究目的确定之后，被解释变量就容易确定下来，被解释变量一般直接表达研究的目的。而对被解释变量有影响的解释变量的确定就不太容易：一是由于我们的认识有限，可能并不知道对被解释变量有重要影响的因素；二是为了保证模型参数估计的有效性，设置的解释变量之间应该是不相关的，而我们很难确定哪些变量是相关的，哪些是不相关的，因为在经济问题中很难找到影响同一结果的相互独立的因素。这就看我们如何在多个变量中确定几个重要且不相关的变量；三是从经济关系角度考虑，非常重要的变量应该引进，但是在实际中并没有这样的统计数据。这一点，在我国建立经济模型时经常会遇到。这时，可以考虑用相近的变量代替，或者由其他几个指标复合成一个新的指标。

在选择变量时要注意与一些专门领域的专家合作。研究金融模型，就要与金融专家和具体业务人员合作；研究粮食生产问题，就要与农业部门的专家合作；研究医学问题，就要与医学专家密切合作。这样做可以帮助我们更好地确定模型变量。

另外，不要认为一个回归模型所涉及的解释变量越多越好。一个经济模型，如果把一些主要变量漏掉肯定会影响模型的应用效果，但如果影响细枝末节的变量一起进入模型也未必就好。当引入的变量太多时，可能选择了一些与问题无关的变量，还可能由于一些变量的相关性很强，它们所反映的信息有较大的重叠，从而出现共线性问题。当变量太多时，计算工作量太大，计算误差也大，估计出的模型参数精度自然不高。

总之，回归变量的确定是一个非常重要的问题，是建立回归模型最基本的工作。一般并不能一次完全确定，通常要经过反复试算，最终找出最适合的一些变量。这在计算机和相关的统计软件的帮助下，已变得不太困难。

1.4.2　收集、整理数据

回归模型的建立基于回归变量的样本统计数据。当确定好回归模型的变量之后，就要对这些变量收集、整理统计数据。数据的收集是建立经济问题回归模型的重要一环，是一项基础性工作。样本数据的质量如何，对回归模型的水平有至关重要的影响。

常用的样本数据分为时间序列数据和横截面数据。

顾名思义，时间序列数据就是按时间顺序排列的统计数据。如新中国建立以来历年的工农业总产值、国民收入、发电量、钢产量、粮食产量等都是每年有一个对应的数据，那么到 2017 年每种指标就有 67 个按时间顺序排列的数据，它们都是时间序列数据。研究宏观经济问题，这方面的时间序列数据来自国家统计局或专业部委的统计年鉴。如果研究微观经济现象，如研究某企业的产值与能耗，数据就要在这个企业的计划统计科获取。

对于收集到的时间序列资料，要特别注意数据的可比性和数据的统计口径问题。如历年的国民收入数据，是否按可比价格计算。中国在改革开放前，几十年物价不变，而从 20 世纪 80 年代初开始，物价几乎是直线上升。那么你所获得的数据是否具有可比性？这就需认真考虑。如在宏观经济研究中，国内生产总值(GDP)与国民生产总值(GNP)二者在内容上是一致的，但在计算口径上不同。国民生产总值按国民原则计算，反映一国常住居民当期在国内外所从事的生产活动；国内生产总值则以国土为计算原则，反映一国国土范围内所发生的生产活动量。对于没有可比性和统计口径不一致的统计数据要作认真调整，这个调整过程就是数据整理过程。

时间序列数据容易产生模型中随机误差项的序列相关，这是因为许多经济变量的前后期之间总是有关联的。如在建立需求模型时，人们的消费习惯、商品短缺程度等具有一定的延续性，它们对相当一段时间的需求量有影响，这样就产生随机误差项的序列相关。对于具有随机误差项序列相关的情况，就要通过对数据的某种计算整理来消除序列相关性。最常用的处理方法是差分法，我们将在后面的章节中详细介绍。

横截面数据即在同一时间截面上的统计数据。如同一年在不同地块上测得的施肥量与小麦产量试验的统计数据就是截面数据。又如某一年的全国人口普查数据、工业普查数据、同一年份全国 35 个大中城市的物价指数等都是截面数据。当用横截面数据作样本时，容易产生异方差性。这是因为一个回归模型往往涉及众多解释变量，如果其中某一因素或一些因素随着解释变量观测值的变化而对被解释变量产生不同影响，就产生异方差性。如在研究城镇居民收入与购买消费品的关系时，用 x_i 表示第 i 户的收入量，y_i 表示第 i 户的购买量，购买回归模型为

$$y_i = \beta_0+\beta_1 x_i+\varepsilon_i, \qquad i = 1, 2,\cdots, n \tag{1.9}$$

在此模型中，随机项 ε_i 就具有不同的方差。因为在购买行为中，低收入的家庭购买的差异性比较小，大多购买生活必需品；高收入的家庭购买行为差异很大，高档消费品

很多，他们的选择余地很大，这样购买物品所花费用的差异就较大。因而，用随机获取的样本数据来建立回归模型，它的随机项 ε_i 就具有异方差性。

对于具有异方差性的建模问题，数据整理就要注意消除异方差性，这常与模型参数估计方法结合起来考虑。我们将在后面的章节中详细介绍。

不论是时间序列数据还是横截面数据的收集，样本量的多少一般要与设置的解释变量数目相匹配。为了使模型的参数估计更有效，通常要求样本量 n 大于解释变量个数 p。当样本量的个数小于解释变量数目时，普通的最小二乘估计方法失效。n 与 p 到底应该有怎样一个比例？英国统计学家M.肯德尔(M.Kendall)在他的《多元分析》一书中指出，样本量 n 应是解释变量个数 p 的 10 倍。如果 p 较大，按肯德尔的说法 n 就很大，这在许多经济问题中是办不到的，尤其新中国才建国 60 多年，统计数据不全是普遍现象。但由肯德尔的观点我们看到，样本量应比解释变量个数大一些才好，这告诉我们在收集数据时应尽可能多地收集一些样本数据。

统计数据的整理中不仅要把一些变量数据进行折算、差分，甚至要把数据对数化、标准化等，有时还需注意剔除个别特别大或特别小的“野值”。在统计数据质量不高时，经常会碰到这种情况。当然，有时还需利用插值的方法把空缺的数据补齐。

1.4.3 确定理论回归模型

当收集到所设置的变量的数据之后，就要确定适当的数学形式来描述这些变量之间的关系。绘制变量 y_i 与 $x_i(i = 1, 2, \cdots, n)$ 的样本散点图是选择数学模型形式的重要一环。一般我们把 (x_i, y_i) 所对应的点在平面直角坐标系上画出来，看看散点图的分布状况。如果 n 个样本点大致分布在带状区域，可考虑用线性回归模型去拟合这 n 个样本点，即选择线性回归模型。如果 n 个样本点的分布大致在一条指数曲线的周围，就可选择指数形式的理论回归模型去描述它。

经济回归模型的建立，通常要依据经济理论和一些数理经济学结果。数理经济学中已对投资函数、生产函数、需求函数、消费函数给出了严格的定义，并把它们分别用公式表示出来。借用这些理论，我们在它们的公式中增加随机误差项，就可把问题转化为用随机数学工具处理的回归模型。如数理经济学中最有名的生产函数 C-D 生产函数是 20 世纪 30 年代初美国经济学家查尔斯·W·柯布(Charles W.Cobb)和保罗·H·道格拉斯(Paul H.Douglas)根据历史统计数据建立的，资本 K 和劳动 L 与产出被确切地表达为

$$y = AK^{\alpha}L^{\beta} \tag{1.10}$$

式中，α, β 分别为 K 和 L 对产出 y 的弹性。C-D 生产函数指出了厂商行为的一种模式，在函数中变量之间的关系是准确实现的。但是由计量经济学的观点，变量之间的关系并不符合数理经济学所拟定的准确关系模式，而是有随机偏差的。因而给 C-D 生产函数增加一个随机项 U，将变量之间的关系描述为一个随机模型，然后用随机数学方法

加以研究，以得出非确定的概率性结论，这更能反映出经济问题的特点。随机模型为

$$y = AK^{\alpha}L^{\beta}U \tag{1.11}$$

或

$$\ln y = \ln A + \alpha \ln K + \beta \ln L + \ln U \tag{1.12}$$

式(1.11)是一个非线性的回归模型；式(1.12)是一个对数线性回归模型。我们在研究工业生产和农业生产问题时就可考虑用上述理论模型。

有时候，我们无法根据所获信息确定模型的形式，这时可以采用不同的形式进行计算机模拟，对于不同的模拟结果，选择较好的一个作为理论模型。

尽管模型中待估的未知参数要到参数估计、检验之后才能确定，但在很多情况下可以根据所研究的经济问题对未知参数的符号以及大小范围事先给予确定。如 C-D 生产函数式(1.11)中的待估参数 A，α，β 都应为正数。

1.4.4　模型参数的估计

回归理论模型确定之后，利用收集、整理的样本数据对模型的未知参数给出估计是回归分析的重要内容。未知参数的估计方法中最常用的是普通最小二乘法，它是经典的估计方法。对于不满足模型基本假设的回归问题，人们给出了种种新方法，如岭回归、主成分回归、偏最小二乘估计等。但它们都是以普通最小二乘法为基础的，这些具体方法是我们后边一些章节研究的重点。这里要说明的是，当变量及样本较多时，参数估计的计算量很大，只有依靠计算机才能得到可靠的准确结果。现在这方面的计算机软件很多，如 MINITAB，SPSS，SAS，R 等都是计算参数估计结果的基本软件。本书的计算实现主要运用 R 软件。

1.4.5　模型的检验与改进

当模型的未知参数估计出来后，就初步建立了一个回归模型。建立回归模型的目的是应用它来研究经济问题，但如果马上就用这个模型去作预测、控制和分析，显然是不够慎重的。因为这个模型是否真正揭示了被解释变量与解释变量之间的关系，必须通过对模型的检验才能确定。一般需要进行统计检验和模型经济意义的检验。

统计检验通常是对回归方程的显著性检验，以及回归系数的显著性检验，还有拟合优度的检验、随机误差项的序列相关检验、异方差性检验、解释变量的多重共线性检验等。这些内容都将在后边的章节中详细讨论。

在经济问题回归模型中，往往还碰到回归模型通过了一系列统计检验，可就是得不到合理的经济解释的情形。例如，国民收入与工农业总产值之间应该是正相关关系，回归模型中工农业总产值变量前的系数应该为正，但有时候由于样本量的限制或数据质量的问题，可能估计出的系数是负的。如此这般，这个回归模型就没有意义，也就谈不上进一步应用了。可见，回归方程经济意义的检验同样是非常重要的。

如果一个回归模型没有通过某种统计检验，或者通过了统计检验而没有合理的经济意义，就需要对其进行修改。模型的修改有时要从设置变量是否合理开始，是不是把某些重要的变量忘记了，变量间是否具有很强的依赖性，样本量是不是太少，理论模型是否合适。譬如某个问题本应用曲线方程去拟合，而我们误用直线方程去拟合，当然通不过检验。这就要重新构造理论模型。

模型的建立往往要反复修改几次，特别是建立一个实际经济问题的回归模型，要反复修正才能得到一个理想模型。

1.4.6 回归模型的应用

当一个经济问题的回归模型通过了各种统计检验，且模型具有合理的经济意义时，就可以运用这个模型来进一步研究经济问题了。

经济变量的因素分析是回归模型的一个重要应用。应用回归模型对经济变量之间的关系作出度量，从模型的回归系数可发现经济变量的结构关系，给出政策评价的一些量化依据。

既然回归模型揭示经济变量间的因果关系，那么可以考虑给定被解释变量值来控制解释变量值。比如把某年的通货膨胀指标定为全国零售物价指数增长5%以下，那么，根据通货膨胀的回归模型可以确定货币的发行量、银行的存款利率等。这就是对经济变量的一种控制。

进行经济预测是回归模型的另一个重要应用。比如我国 2020 年的国民收入是多少？通过建立国民经济的宏观经济模型就可以对未来作出预测。用回归模型进行经济预测在我国已有不少成功的例子。

在回归模型的运用中，我们还强调定性分析和定量分析的有机结合。这是因为数理统计方法只是从事物的数量表面去研究问题，不涉及事物质的规定性。单纯的表面上的数量关系是否反映事物的本质？这本质究竟如何？必须依靠专门学科的研究才能下定论。所以，在经济问题的研究中，我们不能仅凭样本数据估计的结果就不加分析地说长道短，必须把参数估计的结果和具体经济问题以及现实情况紧密结合，这样才能保证回归模型在经济问题研究中的正确运用。

1.5 回归分析应用与发展简评

从高斯提出最小二乘法算起，回归分析已有 200 多年的历史。回归分析的应用非常广泛，我们大概很难找到不用它的领域，这也正是 200 多年来其经久不衰、生命力强大的根本原因。

这里仅介绍回归分析在经济领域的广泛应用。我们知道计量经济学是现代经济学中影响最大的一门独立学科，诺贝尔经济学奖获得者萨缪尔森曾经说过，第二次世界

大战后的经济学是计量经济学的时代。然而，计量经济学中的基本计量方法就是回归分析，计量经济学的一个重要理论支柱是回归分析理论。

自 1969 年设立诺贝尔经济学奖以来，已有 80 多位学者获奖，其中绝大部分获奖者是统计学家、计量经济学家、数学家。从大多数获奖者的论著看，他们对统计学及回归分析方法的应用都有娴熟的技巧，这足以说明统计学方法在现代经济研究中的重要作用。

矩阵理论和计算机技术的发展为回归分析模型在经济研究中的应用提供了极大的方便。国民经济是一个错综复杂的系统，一个宏观经济问题常常需要涉及几十个甚至几千个变量和方程，如果没有先进的计算机和求解线性方程组的矩阵计算理论，要研究复杂的经济问题是不可想象的。比如一个20阶的线性方程组要用克莱姆法则去求解，就需要 10^{22} 次乘法运算，这可是一个天文数字。然而，用矩阵变换的方法只需 6 000 次乘法运算。也正是由于计算方法的改进和现代计算机的发展，过去不可想象的事情变成了现实。计量经济学研究中涉及的变量和方程也越来越多，例如英国剑桥大学的多部门动态模型涉及多达 2 759 个方程、7 484 个变量；由诺贝尔经济学奖获得者克莱因发起的国际连接系统，使用了 7 447 个方程和 3 368 个外生变量。

模型技术在经济问题研究中的应用在我国也盛行起来。自 20 世纪 80 年代初期以来，每年都有许多国家级和省部级鉴定的计量经济应用成果诞生。特别是在一些省级以上的重点经济课题和经济学学位论文中，如果没有模型技术的应用，给人的印象总是分量不足。这些足以说明模型技术的应用在我国备受重视。这里要强调说明的是，回归分析方法是模型技术中最基本的内容，众多的计量经济模型都是在回归模型基础上衍生的。

回归分析的理论和方法研究 200 多年来也得到不断发展，统计学中的许多重要方法都与回归分析有着密切的联系，如时间序列分析、判别分析、主成分分析、因子分析、典型相关分析等。这些都极大地丰富了统计学方法的宝库。

回归分析方法自身的完善和发展至今是统计学家研究的热点课题。例如自变量的选择、稳健回归、回归诊断、投影寻踪、分位回归、非参数回归模型等近年仍有大量研究文献出现。

在回归模型中，当自变量代表时间、因变量不独立并且构成平稳序列时，这种回归模型的研究就是统计学中的另一个重要分支——时间序列分析。它提供了一系列动态数据的处理方法，帮助人们科学地研究分析所获得的动态数据，从而建立描述动态数据的统计模型，以达到预测、控制的目的。

在前面的回归模型式(1.7)中，当因变量 y 和自变量 x 都是一维时，称它为一元回归模型；当 x 是多维，y 是一维时，则它为多元回归模型；若 x 是多维，y 也是多维，则称它为多重回归模型。特别是当因变量观察矩阵 $\boldsymbol{Y}$ 的诸行向量假定是独立的，而列向量假定是相关的，就称为半相依回归方程系统。

对于满足基本假设的回归模型，它的理论已经成熟，但对于违背基本假设的回归模型的参数估计问题近年仍有较多研究。

在实际问题的研究应用中，人们发现经典的最小二乘估计的结果并不总是令人满意，统计学家从多方面进行努力，试图克服经典方法的不足。例如，为了克服设计矩阵的病态性，提出了以岭估计为代表的多种有偏估计。斯泰因(Stein)于 1955 年证明了当维数 p 大于 2 时，正态均值向量最小二乘估计的不可容性，即能够找到另一个估计在某种意义上一致优于最小二乘估计。从此之后，人们提出了许多新的估计，其中主要有岭估计、压缩估计、主成分估计、Stein 估计，以及特征根估计。这些估计的共同点是有偏，即它们的均值并不等于待估参数，于是人们把这些估计称为有偏估计。当设计矩阵 $\boldsymbol{X}$ 呈病态时，这些估计都改进了最小二乘估计。

为了解决自变量个数较多的大型回归模型的自变量的选择问题，人们提出了许多关于回归自变量选择的准则和算法；为了克服最小二乘估计对异常值的敏感性，人们提出了各种稳健回归；为了研究模型假设条件的合理性及样本数据对统计推断影响的大小，产生了回归诊断；为了研究回归模型式(1.7)中未知参数非线性的问题，人们提出了许多非线性回归方法，这其中有利用数学规划理论提出的非线性回归参数估计方法、样条回归方法、微分几何方法等；为了分析和处理高维数据，特别是高维非正态数据，产生了投影寻踪回归、切片回归等。

近年来，新的研究方法不断出现，如非参数统计、自助法、刀切法、经验贝叶斯估计等方法都对回归分析起着渗透和促进作用。

由此看来，回归模型技术随着它自身的不断完善和发展以及应用领域的不断扩大，必将在统计学中占有更重要的位置，也必将为人类社会的发展发挥它独到的作用。

思考与练习

1.1 变量间统计关系和函数关系的区别是什么？

1.2 回归分析与相关分析的区别与联系是什么？

1.3 回归模型中随机误差项 ε 的意义是什么？

1.4 线性回归模型的基本假设是什么？

1.5 回归变量设置的理论根据是什么？在设置回归变量时应注意哪些问题？

1.6 收集、整理数据包括哪些内容？

1.7 构造回归理论模型的基本根据是什么？

1.8 为什么要对回归模型进行检验？

1.9 回归模型有哪几个方面的应用？

1.10 为什么强调运用回归分析研究经济问题要定性分析和定量分析相结合？