第 1 章

回归分析概论

1.1 什么是计量经济学

“计量经济学太数学化了，这是我的好朋友不读经济学专业的原因。”

“最好不要看两样东西的产生过程：香肠和计量经济学估计量。”[1]

“计量经济学可以被定义为分析现实经济现象的定量分析方法。”[2]

“我认为‘经济手段’只不过是证实研究者在开始研究之前就已经相信的结论的依据而已。”

很明显，不同的人对计量经济学有不同的看法。对于希望掌握这门格外有用的学科的初学者来说，计量经济学似乎是个难以逾越的障碍。对于持怀疑态度的观察者来说，只有在产生计量经济学结论的所有步骤都完全知晓的情况下，他们才认为所得到的结论是可信的，而对计量经济学领域的专业人士来说，他们认为计量经济学是一套可以用于度量和分析经济现象并预测未来经济趋势的迷人的技术。

也许有人认为这么多的观点就像盲人摸象一般，但这种看法并不完全正确。事实上，计量经济学既有严密的定义又有更为广阔的背景。即使能够轻易地记住定义，但要全面了解计量经济学知识，就需要理解计量经济学的常用方法和其他方法。

也就是说，我们需要一个严密的定义。从字面意思上讲，“经济度量”是指对实际经济和商业现象进行数量度量和分析。它致力于量化经济现实，并沟通抽象世界的经济理论和现实世界的人类活动。对许多学生而言，理论与现实两者之间存在巨大的差异。一方面，经济学家通过严格设定的边际成本和边际收益推导出均衡价格；另一方面，许多企业在运作过程中似乎并不怎么涉及上述概念。计量经济学允许我们通过数据来量化企业、消费者和政府行为。这些度量有多种用途，考察这些用途是理解计量经济学的第一步。

---

1 Ed. Leamer, “Let’s take the Con out of Econometrics,” *American Economic Review*, Vol.73.

2 Paul A. Samuelson, T. C, Koopmans, and J. R, Stone “Report of the Evaluative Committee for Econometrica,” *Econometrica*, 1954, p.141.

1.1.1 计量经济学的用途

计量经济学有以下三方面用途。

（1）描述经济现实。

（2）检验经济理论假设。

（3）预测未来经济活动走势。

计量经济学最简单的用途就是描述经济现实。因为计量经济学能够估计出数值，并将其放入原先只有抽象符号的方程中，这样一来，就可以采用计量经济学来量化经济现实。例如，特定商品的消费需求可以看做需求量（Q）与商品价格（P）、替代品价格（P_S）、可支配收入（Y_d）之间的关系。对绝大多数商品而言，可以认为消费量和可支配收入之间的关系是正向的，因为随着可支配收入的增加，消费量也会增加。计量经济学能够基于过去的消费量、收入和价格的数据，估计上述经济关系。换句话说，它能够把一般性的纯理论函数关系

$$Q = f(P, P_S, Y_d) \tag{1-1}$$

表述成更为明确的表达式

$$Q = 27.7 - 0.11P + 0.03P_S + 0.23Y_d \tag{1-2}$$

比较方程（1-1）和方程（1-2）就会发现，利用计量经济学能够更明确、更生动地描述函数关系。[3] 在预期消费量会随着可支配收入的增加而增加的基础上，方程（1-2）还能够给出一个确定的预期增长量（可支配收入每增加1个单位，预期的消费量增加0.23个单位）。在这里，数值0.23称为估计出的回归参数。正因为能估计出这种参数，计量经济学才显得很有价值。

计量经济学的第二个用途是假设检验，也许是最常用的。假设检验采用量化了的证据来对备选理论进行评价。许多经济学科都涉及构建理论模型，然后用现实证据来检验，在这个过程中，假设检验起着举足轻重的作用。例如，要检验这样的假设：方程（1-1）中的商品是正常商品（需求量随着可支配收入增加而增加的商品）。可以通过应用各种各样的统计方法，对估计出的方程（1-2）中的可支配收入（Y_d）的参数（0.23）进行假设检验。乍看之下，参数估计值为0.23这个结果似乎支持假设，因为参数的符号是正的，但是在下结论之前，必须通过参数估计值的“统计显著性”检验。即使参数估计值为正，与预期相同，但它可能与零之间没有明显差异，并不能确信真实参数为正。

计量经济学的第三个用途，也是最难掌握的用途，是基于已经发生的事件去预测或者推测下一季度、明年或者更远的将来会发生什么。例如，经济学家可以采用计量经济学模型来预测销售量、利润、国内生产总值（GDP）、通货膨胀率等。这类预测的准确性，很大程度上取决于过去对未来的决定程度。商业领袖和政治家之所以格外看重计量经济学的这个用途，是因为他们需要依靠它来对未来进行决策，如果决策失误，后果会非常严重（企业破产或候选人政治失败）。由于计量经济学能够揭示政策效果，所以，商界和政府领导人能够更好地使用它来做出决策。例如，对于某个公司在销售方程（1-1）中提及的产品，公司总裁很想知道是否应该涨价，于是，就可以通过预测涨价前后的销售量进行计算和比较，做出是否应该涨价的决策。

1.1.2 其他计量经济学方法

定量研究有很多不同的方法。例如，与生物学、心理学和物理学都会面临数量问题一样，经济管理学也会遇到数量问题。然而，不同学科面临的问题不同，所以，采用的方法也不同。

---

3 方程（1-2）是一个鸡肉需求模型的估计结果，我们将在6.1节中进行更详细的讨论。当然，在构建鸡肉的需求模型时不考虑鸡肉的供给是不行的。然而，在学习单方程模型的估计之前，学习联立方程模型的估计非常困难。因此，我们将关于联立方程模型的讨论放在第14章。到那时，我们会遇到从理论观点来看并非是“独立”的解释变量。

比如，通常意义下，经济学是一门观察性学科而不是一门实验性学科。“我们之所以需要一个称为计量经济学的特殊的学科领域和相关的教科书，是因为人们普遍认为经济学家使用的经济数据拥有一些固有性质，而这些性质在统计学基础教科书中并没有考虑或者没有充分强调。[4]

在经济学领域内，不同的方法具有不同的意义。选用哪种计量经济学方法部分取决于计量经济学方程的用途，例如，仅仅用于描述目的的模型就可能和预测模型不同。

为了更好地了解这些方法，应该先弄清楚非实验定量研究方法的步骤：

（1）设定模型或者待研究的经济关系；

（2）收集用于量化模型的数据；

（3）使用数据量化模型。

第1步中用到的表达式和第2步中用到的方法在学科内外都是大相径庭的。为给定的模型选择最好的表达式这一理论方法，经常被称为计量经济的“艺术”。对同一个方程存在多种不同的量化方法，而不同的方法一般来说会产生不同的结果。方法的选择是计量经济学使用者的事情，不过每个使用者应该给出选择理由。

本书将重点讨论一种特定的计量经济学方法：单方程线性回归分析。本书的大部分内容都集中于讨论回归分析，但是，对于每一个计量经济学家来说，重要的是不要忘记回归分析只不过是计量经济学量化方法中的一种。

批判性评价的重要性再怎么强调也不过分。优秀的计量经济学家能够诊断出错误的地方并做出修正。任何一个使用回归分析及其结论的人都应该对回归分析方法的局限性有足够的认识。由于存在缺省数据或数据不准确、错误设定经济关系、错误选择估计方法或者统计检验程序不恰当等可能性，所以，在分析回归分析结果时需要格外小心。

1.2 什么是回归分析

计量经济学家通常采用回归分析，来对之前完全理论化的经济关系进行数量估计。说到底，每个人都可以声称需求量会随着价格的降低而升高（假定其他条件不变），但并没有多少人能够在方程中给出确切的数值，估计出光盘价格每下降1美元后需求量会增加多少。为了推断出变动的方向，就需要掌握经济学的理论知识和特定商品的性质。为了推断出变动量，就需要一组数据样本和一种方法来估计经济关系。在计量经济学中，使用最多的用于估计此类经济关系的方法就是回归分析。

1.2.1 被解释变量、解释变量和因果关系

回归分析（regression analysis）是一种统计方法，它通过对一个方程的量化来“解释”被解释变量（dependent variable）如何随着一系列解释变量（independent variable）的变动而变动。例如，在方程（1-1）中

$$Q = f(P, P_S, Y_d) \tag{1-1}$$

Q是被解释变量，P、P_S、Y_d是解释变量。回归分析对经济学家来说是一种很自然的分析工具，这是因为绝大部分（尽管不是所有的）经济命题都能表述为这种单方程的函数形式。例如，需求量（被解释变量）是价格、替代品价格和收入（解释变量）的函数。

许多经济学和工商管理学科都会触及因果关系命题。如果商品价格升高1单位，那么需求量下降的平均量则取决于需求的价格弹性（即价格变动1%引起的需求量变动的百分比）。与此类

4 Cliver Granger, “A Review of Some Recent Textbooks of Econometrics,” *Journal of Economic Literature*, Vol. 32, No.1, p. 117.

似，如果资本投入量增加1单位，那么产出量增加的平均量就称为资本的边际产出。诸如此类的命题提出了“如果……那么……”关系或因果关系，其在逻辑上假定被解释变量的变动是由一系列确定的解释变量的变动引起的。

> 千万不要被“被解释变量”和“解释变量”的字面意思所误导。虽然，很多经济关系由于其自身的性质而具有因果关系，但是要记住，无论统计上多么显著的回归结果都不能证明因果关系。数量分析所能做的仅仅是检验显著的数量关系是否存在。对因果关系的判断必须建立在一定的经济学理论和基本常识的基础之上。例如，一个顾客进入花店买花之前，花店的门铃响了，这并不能说明铃声导致了购买行为。如果事件A和事件B在统计上是有关联的，那么可能是A导致了B，也可能是B导致了A，还有可能是一些忽略掉的因素影响了两者，或者是两者之间存在相关关系。

因果关系时常表现得非常微妙，以至于误导了最为杰出的经济学家。例如，在19世纪晚期，英国经济学家Stanley Jevons曾假设太阳黑子导致经济活动增加。为了检验这个理论，他收集了国民产出的数据（被解释变量）和太阳黑子活动的数据（解释变量），发现两者存在显著的正相关关系。正是这个结论，使得他和另外一些人得出结论：太阳黑子确实能够促使产出增加。显然，这种结论是不合理的，因为回归分析不能确保因果关系，它只能检验相关数量关系的强度和方向。

1.2.2 单方程线性模型

最简单的单方程线性模型为：

$$Y = \beta_0 + \beta_1 X \tag{1-3}$$

方程（1-3）指出被解释变量Y是解释变量X的线性函数。这个模型是一个单方程模型，因为它是唯一被指定的方程。之所以说这个模型是线性的，是因为如果绘出方程（1-3）的图形，就会发现它是一条直线而不是曲线。

β为参数，它决定了回归直线上每一点的坐标。其中，β_0是常数项或截距项，表示X为零时，Y对应的值；β_1是回归直线的斜率，它表示X每变动1单位，Y的变动量。图1-1中的实线说明了回归方程的参数和图形之间的关系。正如从图中看到的一样，方程（1-3）确实是线性的。

斜率参数（slope coefficient）β_1表示的是解释变量X每增加1单位，被解释变量Y随之而变动的量。回归分析重点关注β_1的斜率参数。如图1-1所示，当解释变量X从X_1增加到X_2（增加量为ΔX）时，方程（1-3）中被解释变量Y的值也会从Y_1增加到Y_2（增加量为ΔY）。对于线性（例如直线方程）回归模型来说，被解释变量Y的预测值随着解释变量X的变动而变动的量等于斜率参数β_1，且恒定不变：

$$\frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \frac{\Delta Y}{\Delta X} = \beta_1$$

式中，Δ代表变量的变动量。有些读者将其理解为纵轴之差（ΔY）除以横轴之差（ΔX）。对线性模型而言，斜率在函数中恒定不变。

如果要把线性回归方法运用到方程中，那么，就要求这个方程必须是线性的。如果根据X和Y绘制出的函数图像是一条直线，那么这个方程就是线性的（linear）。例如方程（1-3）：

$$Y = \beta_0 + \beta_1 X \tag{1-3}$$

是线性的，而方程（1-4）

$$Y = \beta_0 + \beta_1 X^2 \tag{1-4}$$

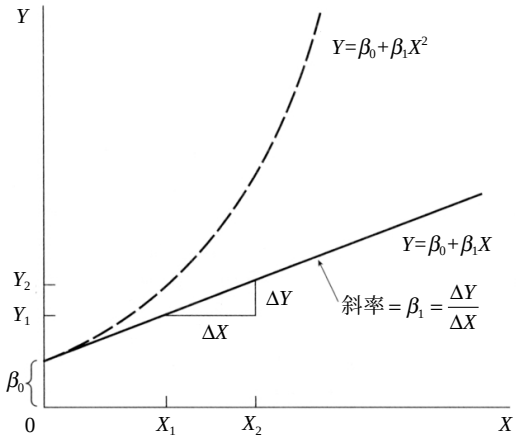

图1-1 回归直线参数的图形表示

注：方程$Y=\beta_0+\beta_1X$的图形是线性的，其斜率为$\beta_1=\Delta Y/\Delta X$，而方程$Y=\beta_0+\beta_1X^2$的图形是非线性的，其斜率是递增的（假设$\beta_1>0$）。

就是非线性的，这是因为根据方程（1-4）所描绘出来的函数图像是二次曲线形的，不是一条直线。可以从图1-1中看出这种差别。[5]如果回归分析必须要求方程是线性的话，那么，应该怎样处理诸如方程（1-4）之类的非线性方程呢？办法是将大多数非线性方程进行适当变换，使之成为线性的。例如，可以通过新设一个变量来代替方程（1-4）中X的平方项，这样就可以把方程（1-4）转化成线性方程。假定：

$$Z=X^2 \tag{1-5}$$

将方程（1-5）代入方程（1-4）有：

$$Y=\beta_0+\beta_1Z \tag{1-6}$$

现在这个重新定义的方程就是线性的[6]，并可以用回归分析方法进行估计了。

1.2.3 随机误差项

被解释变量（Y）的变化除了受解释变量（X）的影响外，还受到许多其他因素的影响。这些因素部分是来自遗漏了的解释变量（例如，X_2和X_3）。然而，就算将这些遗漏的变量引入方程，对被解释变量Y来说，仍然会存在一些无法被模型完全解释的误差。[7]这些误差可能来自一些已忽略的影响因素、数据的测量误差、错误的函数形式、纯粹的随机因素或完全无法预料的未知因素。在这里，随机（random）是指变量的取值完全是偶然的。

---

5 为了便于比较，图1-1中方程（1-3）和方程（1-4）的β_0是相等的。如果这两个方程应用于相同的数据中，那么估计值β_0会有所不同。同样，不难发现β_1的估计值也不一样。

6 事实上，在第7章将会学到，这个方程既是关于参数β_0和β_1的线性方程，又是关于变量Y和Z的线性方程，但却是关于Y和X的非线性方程。第7章会讲解将回归分析方法应用于变量非线性的方程。然而，将回归分析方法应用于参数非线性的方程会困难得多。

7 极端例外的情况是，数据被某些物理定律解释并被精确地测量。数据能被一些物理定律解释并被完全精确地测量的例外非常少见。这里，连续的误差表明存在一个被忽视的解释变量。天文学中也会遇到类似的问题，新的行星往往是通过记录已知星球的轨道偏差而发现的，因为这种偏差只能是另外天体的引力所导致的。由于缺乏这样的物理定律，经济学和商业领域的研究者就不能武断地认为，被解释变量的所有变化都可以由回归模型来解释，因为无论采用何种方法来测量一种行为关系，总会有误差存在。

计量经济学家承认存在这种内在的无法被解释的误差（偏差），因而直接在回归模型中引入一个随机误差项。

随机误差项（stochastic error term）是回归方程中除已有解释变量X之外，代表其他所有影响被解释变量Y的因素。它也是计量经济学家无法将所有被解释变量的变动因素用模型表示出来的象征。误差项（有时也叫干扰项）通常用符号ε（读作epsilon）表示。有时也用其他符号（诸如u或v）表示。

在方程（1-3）加入了随机误差项后，就形成了一个典型的回归方程：

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1-7}$$

可以看出方程（1-7）由两部分组成：确定部分和随机或偶然部分。表达式$\beta_0 + \beta_1 X$是回归方程的确定部分，因为它表示的Y值由给定的X的值所决定，这个过程是非随机的。这个确定部分也可看做是给定X值的情况下，Y的期望值（expected value）。Y的均值（期望值）与特定的X值有关。例如，假设13岁女孩的平均身高是1.52米，那么，1.52米就是一个女孩在13岁时身高的期望值。这个方程的确定部分可以表述为：

$$E(Y \mid X) = \beta_0 + \beta_1 X \tag{1-8}$$

它表示的是在给定X值的情况下，Y的期望值，记作$E(Y \mid X)$，它是一个（或多个）解释变量的线性函数。[8] 然而，现实中Y的观测值不可能恰好等于确定的期望值$E(Y \mid X)$。毕竟，不是所有13岁女孩都是1.52米高。因此，必须在方程中加入随机因素（ε）：

$$Y = E(Y \mid X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon \tag{1-9}$$

之所以必须把随机误差项引入到回归方程中，是因为除了解释变量X包含的误差，还存在至少4种导致被解释变量Y出现误差的因素。

（1）许多对Y有微小影响的因素没有包含在方程中（例如，无法获取的数据）。

（2）被解释变量的某些测量误差在实质上是无法避免的。

（3）潜在的理论方程可能与选定用做回归分析的方程具有不一样的函数形式（或形状）。例如，潜在的理论方程可能是非线性的。

（4）对人类行为的一般模型化表述通常都包含一些不可预测的或纯随机的因素。

为了加深对随机误差项的这些构成部分的理解，考察消费函数（总消费是总可支配收入的函数）。第一，由于未来经济中存在着不确定因素，特定年份的消费量可能会小于该年份应该达到的水平。然而这种不确定性是很难测量的，所以，在方程中很可能没有用来衡量消费者不确定性因素的变量。在这种情况下，被遗漏了的变量（消费者的不确定性因素）的影响很可能就表现在随机误差项中。第二，由于国民收入账户在衡量消费水平的时候会出现一定程度的误差（例如，样本误差），所以，特定年份的消费量的观测值可能与真实的消费水平不同。第三，隐含的消费函数可能是非线性的，但却用线性消费函数来估计（从图1-2中可以看出，错误的函数形式将导致多大的误差）。第四，消费函数用于描述人类行为，但是人类行为本身就包含一些不可预测的因素。在任何时候，一些随机事件也许会以一种不可重复也不可预测的方式增加或减少总消费量。

8 在满足古典假设（第4章会讲解）的情况下，有$E(\varepsilon|X) = 0$（在给定X的值的情况下，ε的期望值为0）。可以很直观地认为$E(\varepsilon)$是ε的均值，但是从技术层面来讲，期望算子E是函数所有取值的加权和，权重是每个取值出现的概率。常数的期望是常数，而变量和的期望等于变量期望的和。

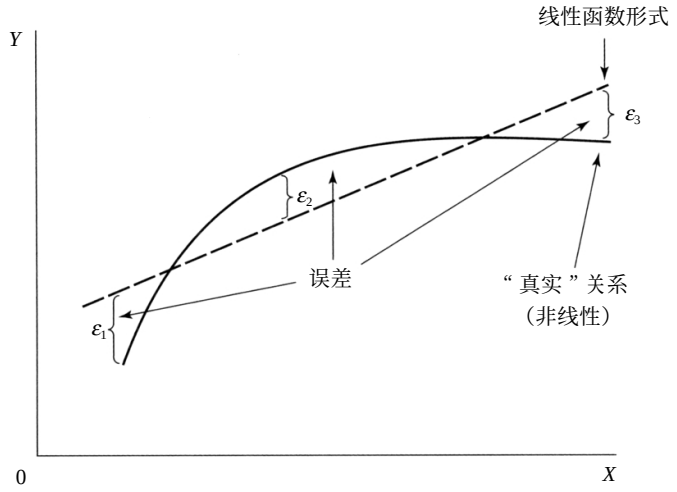

图1-2 使用线性函数模拟非线性关系造成的误差

注：使用错误的函数形式是产生随机误差的原因之一。例如，在隐含关系是非线性的情况下，采用了线性函数形式，系统误差（εs）便产生了。这种非线性因素仅仅是随机误差项的其中一个构成部分而已，其他部分包括遗漏了的变量、测量误差和纯随机误差等。

这在一定程度上解释了Y的观测值与从方程确定部分得出的期望值$E(Y \mid X)$之间存在差别的原因。在后面的章节中，将对这些误差的来源做更详尽的讲解，目前只要意识到在计量经济学研究中总会存在一些随机或偶然的因素就够了，正因为如此，必须要将误差项纳入到回归方程中。

1.2.4 符号的拓展

回归分析所用到的符号需要拓展，用于反映多个解释变量的情况和样本观测值的数目。个人、年份或者国家都可以成为典型的观测对象（或分析单元）。例如，在一个1985年开始的基于年度观测值的序列中，可以用Y_1表示Y在1985年的观测值，用Y_2表示1986年的观测值，依此类推。如果观测值的数目是确定的，那么单方程线性回归模型可以写为：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (i = 1, 2, \cdots, N) \tag{1-10}$$

式中，Y_i 代表被解释变量的第i个观测值；X_i 代表解释变量的第i个观测值；ε_i代表随机误差项的第i个观测值；β_0, β_1代表回归参数；N代表样本观测值的数目。

这实际上代表了N个方程，每组观测值对应一个方程，即

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ Y_3 &= \beta_0 + \beta_1 X_3 + \varepsilon_3 \\ &\vdots \\ Y_N &= \beta_0 + \beta_1 X_N + \varepsilon_N \end{aligned}$$

因此，假定这个回归模型适用于任何一组观测值。参数值不会随着观测值的改变而改变，但Y，X和ε的值会随着观测值的改变而改变。

第二种符号拓展是为了使回归方程能够容纳多个解释变量，因为被解释变量很可能受到一个以上解释变量的影响，所以，就要求方程中的符号能表示出额外加入的解释变量。定义：

X_{1i} —— 第1个解释变量的第i个观测值
X_{2i} —— 第2个解释变量的第i个观测值
X_{3i} —— 第3个解释变量的第i个观测值

这三个变量都可以看做Y的决定因素。

这样一来，方程就称为多元（multivariate，不止一个解释变量）线性回归模型：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \qquad (1\text{-}11)$$

方程中的回归参数β_1的经济意义（meaning of the regression coefficient β_1）是在X_2和X_3的值不变的情况下，X_1每增加1单位对被解释变量Y的影响。同理，β_2给出在X_1和X_3的值不变的情况下，X_2每增加1单位对被解释变量的影响。

多元回归参数（multivariate regression coefficient）用于分离某个变量与其他变量对Y的影响（这在性质上与微积分中的偏导数非常相似）。这是可行的，因为在估计参数X_1的时候，多元回归会考虑到X_2和X_3的变化。这样得到的结果与控制实验使每次只变动一个变量所得到的结果非常相似。

然而，在现实生活中，要进行控制性经济实验非常困难，[9] 因为许多经济因素的变动是同步的，且变动方向往往不一致。因此，回归分析能在其他解释变量的影响不变的条件下，测量出其中某个解释变量对被解释变量的影响，这是它非常突出的优点。值得注意的是，如果某个变量没有包含在方程中，那么，在估计其他回归参数的时候，它的影响将会发生变化。关于这一点，本书将在第6章做进一步分析。

上面讲到的内容比较抽象，接下来通过一个实例来说明。或许有人会认为在某个行业中存在工资歧视的现象，因而想了解一下该行业的工资是如何决定的。此时，若工人的工资是被解释变量（*WAGE*），那合适的解释变量应该是哪些呢？在一个给定的行业中，哪些变量能够影响一个工人的工资呢？事实上，可能存在数十种合理的因素，但最常见的有三种，分别是工作经验（*EXP*）、受教育程度（*EDU*）和性别（*GEND*），因而就使用这三个变量。为了构造一个包含这些变量的回归方程，重新定义方程（1-11）来满足此处的情况：

Y —— *WAGE*（工人的工资）
X_1 —— *EXP*（工人的从业年限）
X_2 —— *EDU*（工人所受高中以上教育的年数）
X_3 —— *GEND*（工人的性别（1=男性，0=女性））

最后一个变量*GEND*是不寻常的，它只有两个值 —— 0或1。这种变量称为虚拟变量（dummy variable），它在量化定性变量（如性别）时非常有用。关于虚拟变量更为深入的讨论将会在第3.1节和第7.4节中涉及。

将以上重新定义的变量代入方程（1-11），有：

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \varepsilon_i \qquad (1\text{-}12)$$

方程（1-12）指出工人的工资是工作经验、受教育程度和性别的函数。在这个方程中，β_1的经济意义是什么呢？有些读者认为β_1表示的是工作经验每增加1年的情况下，平均工资的增长量。然而，这种看法忽视了方程中另外两个解释变量对工资的影响。正确的说法是：β_1表示的是在受教育程度和性别不变的情况下，工作经验每增加1年对工资的影响。以上两种说法之间存

9 这种实验操作起来很困难，但并非不可能，请参考第16.1节的内容。

在显著的差别，因为后者使研究者在没有进行控制实验的情况下，控制住了特定复杂因素对被解释变量的影响。

在结束本节之前，有必要了解一下包含K个解释变量的多元回归模型的一般形式：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \varepsilon_i \tag{1-13}$$

式中，i从1取到N，表示观测期。

如果样本中包含年度或月度序列（称做时间序列（time series）），此时，通常用标记时间的下标t来取代下标i。[10]

1.3 回归方程的估计

一旦特定方程被设定之后，就必须对其进行量化。量化理论上的回归方程的过程被称为回归方程估计（estimated regression equation），它是通过一组包含X和Y的真实值的样本数据实现的。虽然理论上的回归方程在本质上是抽象的方程：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1-14}$$

但估计出回归方程为：

$$\hat{Y}_i = 103.40 + 6.38 X_i \tag{1-15}$$

这样，方程中就有了确定的数值，参数估计值103.40和6.38是用X和Y的实际观测值计算出来的。这些估计值用于决定Y的估计值或拟合值 $\hat{Y}$（读作Y尖）。

观察理论上的回归方程和估计出的回归方程之间的差别，就会发现：首先，方程（1-14）中理论上的回归参数β_0和β_1被方程（1-15）中的参数估计值103.40和6.38取代了。事实上，回归参数的真实值[11]是不能观察到的。因此，通常情况下，采用从样本数据中计算出来的估计值来代替。回归参数的估计值（estimated regression coefficients）通常记作 $\hat{\beta}_0$ 和 $\hat{\beta}_1$（读作贝塔尖），它们是经验上对回归参数真值的最佳推测，并且是通过包含X和Y的样本数据计算出来的。通常情况下，表达式

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1-16}$$

与理论上的回归方程（1-14）是相互对应的。方程（1-15）中计算出的参数估计值是回归参数估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的特例，每组样本都可以计算出一组特定的回归参数估计值。

$\hat{Y}_i$ 是Y_i 的估计值，表示的是将解释变量的第i组观测值代入估计出的回归方程后计算出来的Y值。因此，$\hat{Y}_i$ 是从回归方程中得出的$E(Y_i \mid X_i)$的预测值。$\hat{Y}$ 与样本中的Y之间越接近，表明方程拟合得越好。（这里的“拟合”一词所表达的意思与衣服合身中“合身”一词的意思差不多。）

被解释变量的估计值（$\hat{Y}_i$）与被解释变量的真值（Y_i）之间的差称为残差（residual，e_i）。

$$e_i = Y_i - \hat{Y}_i \tag{1-17}$$

---

10 只要正确的定义被表达出来，下标的顺序并不那么重要。本书将第一个变量写为(X_{1i})，因为这更便于计量经济学的初学者理解。然而，当读者接触到矩阵代数和计算机电子表格时，通常情况下将第一个变量写为X_{i1}。观测值的下标有时会省略掉，读者应该知道方程中包含了样本中的所有观测值。

11 本书中对于“真实”一词的使用是有所保留的。许多哲学家认为真实的概念只有在涉及科学研究问题时才有用。许多经济学家赞同这一点，并指出在经济学上对一代人是真理的东西对另一代人来说可能是谬误。对计量经济学来说，真实的参数是通过对总体进行回归分析所计算出来的参数值。因此，读者可以用“总体参数”来代替“真实参数”，两者的意思是一样的。

注意方程（1-17）中的残差与误差项之间的区别：

$$\varepsilon_i = Y_i - E(Y_i|X_i) \tag{1-18}$$

残差是Y的观测值与Y的估计值之间的差，而误差项是Y的观测值和真实回归方程（Y的期望值）之间的距离。值得注意的是，误差项是一个不能被观察到的理论概念，而残差是在回归分析中可以被每组观测值计算出来的实实在在的数值。残差可以看做是误差项的估计值，此时，e表示$\hat{\varepsilon}$。大多数回归方法不仅要计算残差，还要计算$\hat{\beta}_0$和$\hat{\beta}_1$，并使残差尽可能得小。残差越小，拟合程度越高，$\hat{Y}$与Y之间越接近。

以上所有的概念都可以在图1-3中看到。数组（X, Y）表示坐标轴中的点，真实的回归方程（不能应用于现实中）和估计出的回归方程也可以在图中看到。值得注意的是，估计出的回归方程的图像与真实回归方程的图像非常接近，但并不重合，这是很常见的。

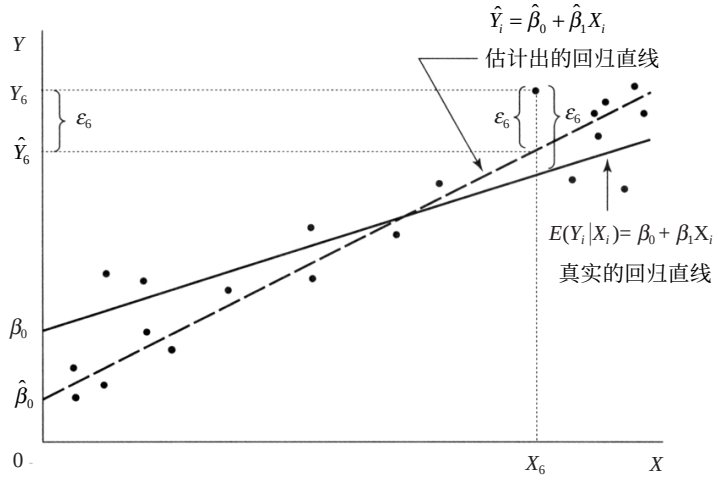

图1-3 真实的回归直线与估计出的回归直线

注：通常情况下，X与Y之间的真实关系（图中的实线）是不能被观察到的，但估计出的回归直线（图中的虚线）却能被观察到。观测点（例如i = 6）与真实回归直线之间的距离就是随机误差项的值（ε_6）。观测值Y_6与回归直线上的估计值$\hat{Y}_6$之间的差就是残差的观测值e_6。

在图1-3中，$\hat{Y}_6$是基于第6对观测值的Y的计算值，位于估计出的回归直线（图中的虚线）上，它与第6对中Y的实际观测值不同。观测值与估计值之间的差为残差，记为e_6。除此以外，虽然通常情况下误差项的观测值无法获取，但还是可以画出一条假定的真实的回归线（图中的实线）来观察第6对观察值所对应的误差项ε_6，即真实的回归线与Y的观测值Y_6之间的距离。

表1-1 真实回归方程与估计出的回归方程中用到的符号

真实的回归方程	估计出的回归方程
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
ε_i	e_i

表1-1总结了真实回归方程与估计出的回归方程中所用到的符号。

在方程的右边加入新的解释变量，可以将估计出的回归模型拓展为不止一个解释变量的情况。对应于方程（1-13）估计出的多元回归模型为：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} \tag{1-19}$$

需要注意的是，我们无法画出超过两个解释变量的多元回归方程的图像，即便只有两个解释变量，要画出对应的图像也非常困难。

1.4 回归分析实例

考虑一个回归分析的简单例子。假设你的暑期工作是在一家名叫Magic Hill的游乐园为游客猜测体重。游客先支付2美元，如果你的猜测精确到10磅以内，那么你将获得这笔钱。如果你没能做到，那就要退回2美元并送给游客一个你用3美元从Magic Hill买来的小礼物。好在Magic Hill的管理员在游客背后的墙上安排了一组标记，这样一来，你就能较为准确地测量游客的身高。不过，在你和游客之间隔着一座5英尺高的矮墙，所以，除了身高和性别外，你很难获取其他信息。

第一天，你的表现很不理想，亏损了2美元，因此，第二天你决定收集一些数据，用回归分析来估计身高和体重的关系。因为大多数参与者是男性，所以你决定把样本限定为男性。你假设出下面的理论关系：

$$Y_i = f(\overset{+}{X}_i) + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1-20}$$

式中，Y_i代表第i个游客的体重（单位：磅）；X_i代表第i个游客的身高（单位：英寸，高于5英尺的部分）；ε_i代表第i个游客的随机误差项的值。

在这里，身高和体重的理论关系被认为是正相关（用一般理论方程中X_i上面的正号表示），然而，这种关系需要得到量化，只有这样，才能在给定身高的情况下估计出体重。为此，你需要搜集数据，并将回归分析运用于搜集到的数据。

第二天，你搜集到了一些数据（见表1-2），并采用Magic Hill的计算机做了回归分析，得到了以下估计值：

$$\hat{\beta}_0 = 103.40 \qquad \hat{\beta}_1 = 6.38$$

这意味着方程

估计出的体重=103.40+6.38×身高（单位：英寸，高于5英尺的部分） (1-21)

该方程可以作为你猜测游客体重的一种替代方法。这个估计身高的方程以常数项103.40磅为基础，并随着身高（高于5英尺）每增加1英寸而增加6.38磅。$\hat{\beta}_1$的符号与预期一致，是正的。

表1-2 猜测体重方程的数据和结果

观测值 i (1)	身高（高于5英尺的部分）X_i (2)	体重 Y_i (3)	身高的预测值 $\hat{Y}_i$ (4)	残差 e_i (5)	利润和亏损 (6)
1	5.0	140.0	135.3	4.7	+2.00
2	9.0	157.0	160.8	-3.8	+2.00
3	13.0	205.0	186.3	18.7	-3.00
4	12.0	198.0	179.9	18.1	-3.00
5	10.0	162.0	167.2	-5.2	+2.00
6	11.0	174.0	173.6	0.4	+2.00
7	8.0	150.0	154.4	-4.4	+2.00
8	9.0	165.0	160.8	4.2	+2.00
9	10.0	170.0	167.2	2.8	+2.00
10	12.0	180.0	179.9	0.1	+2.00
11	11.0	170.0	173.6	-3.6	+2.00
12	9.0	162.0	160.8	1.2	+2.00
13	10.0	165.0	167.2	-2.2	+2.00
14	12.0	180.0	179.9	0.1	+2.00
15	8.0	160.0	154.4	5.6	+2.00

(续)

观测值 i (1)	身高（高于5英尺的部分）X_i (2)	体重 Y_i (3)	身高的预测值 $\hat{Y}_i$ (4)	残差 e_i (5)	利润和亏损 (6)
16	9.0	155.0	160.8	-5.8	+2.00
17	10.0	165.0	167.2	-2.2	+2.00
18	15.0	190.0	199.1	-9.1	+2.00
19	13.0	185.0	186.3	-1.3	+2.00
20	11.0	155.0	173.6	-18.6	-3.00
					总计：25美元

注：这个数据集，以及本书中涉及的其他数据集都可以在http://www.pearsonhighered.com/studenmund上找到，有4种格式。这个数据集基于EViews格式，文件名为HTWT1。

这个方程的效果怎么样呢？要回答这个问题，你需要计算出方程（1-21）的残差项（$Y_i - \hat{Y}_i$），看有多少个残差大于10。从表1-2的最后一列可以看出，如果你把回归方程运用到这20个人中，你不会因此而变得富有，但是至少可以让你赚到25美元而不是亏损2美元。图1-4不仅给出了方程（1-21）的图像，还给出了样本中20个游客的体重和身高数据。方程（1-21）有助于刚开始从事体重猜测工作的人，它可以通过加入新的变量或扩大样本容量的方式来改进。这样的方程是符合实际的，似乎每个成功猜测体重的人都会不经意地用到它。

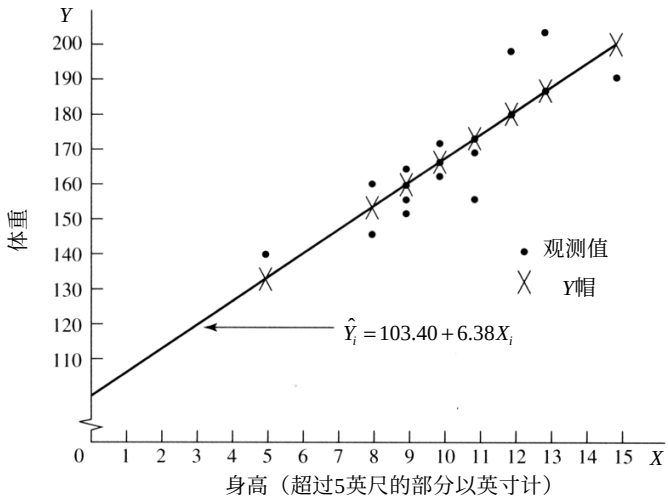

图1-4 体重猜测方程

注：如果把猜测体重案例中的数据和估计出的回归直线描绘在坐标轴中，你就会发现，除了三个观测值外，估计值 $\hat{Y}_S$ 与观测值 Y_S 之间非常接近。在图中找到一个男性的身高和体重，看看这个回归方程的效果如何。

这里的目标是通过搜集数据（见表1-2）计算出方程（1-21）中的估计值，以此量化体重和身高的理论方程（1-20）。正如随机误差项的观测值是不可知的，尽管我们无法知道真实回归方程，但还是可以得到估计出的回归方程，其 $\hat{\beta}_1$ 的符号符合预期，这一方程有助于工作的开展。在你试图在Magic Hill以猜测体重为生之前，还有很多关于回归分析的东西要学。

1.5 应用回归分析解释住房价格

尽管在游乐园猜测体重是一件有意思的事情，但这并不是应用回归分析的典型案例。类似于将回归分析运用到这样奇妙的主题，它还可以运用于其他大量主题，如描述GDP对货币供应量增加的反应，采用新数据来验证经济理论，预测价格变化对企业销售量的影响等。

为了使案例更贴近现实，接下来考虑房地产定价模型。购买住房可能是一个人一生中最重要的财务决策，而影响决策最重要的因素之一是房地产估价。如果高估了房价，那么可能会带来数千美元的损失；如果低估了房价，那么有人可能出价更高而买走住房。

如果住房是像玉米或黄金那样的同质产品，有广为人知的市场价格与特定要价作比较，那么就不会存在太大的问题。但在房地产市场上，情况完全不一样。因此，房地产估价就成为是否买房的重要因素。许多房地产估价师运用回归分析来开展工作。

假如你打算在南加利福尼亚州买一套住房，但你觉得房东的要价太高了。房东认为，230 000美元的要价是合理的，因为大约一年前隔壁一套稍大一些的住房就卖了230 000美元。你不能确定比较两套不同面积、在不同时期购买的住房价格的做法是否合理，那么你如何才能决定是否支付这230 000美元呢？

你决定搜集过去几周在当地出售的所有住房的数据，并建立一个以房价为被解释变量、住房面积为解释变量的回归模型。[12]这个数据集是截面的（cross-sectional），因为所有的观测值都来自同一个时间点，并代表不同的独立经济体（例如国家和地区，本例中是指住房）。

为了测量面积对价格的影响，你把住房面积作为回归方程的解释变量，把价格作为被解释变量。你预期面积的参数符号为正，因为大住房比小住房的建造成本更多，且大住房比小住房更受欢迎。因此，理论模型为：

$$PRICE_i = f(\overset{+}{SIZE}_i) + \varepsilon_i = \beta_0 + \beta_1 SIZE_i + \varepsilon_i \tag{1-22}$$

式中，$PRICE_i$ 代表第i套住房的价格（单位：1 000美元）；$SIZE_i$ 代表第i套住房的面积（单位：平方英尺）；ε_i 代表第i套住房的随机误差项。

你在搜集了最近所有的房地产交易记录后发现，过去4周，当地有43套住房售出，于是，你采用这43个观测值估计出了回归方程：

$$\widehat{PRICE}_i = 40.0 + 0.138 SIZE_i \tag{1-23}$$

这些参数估计值表示什么呢？其中最重要的参数是 $\hat{\beta}_1$ =0.138。这个参数表示住房面积每增加1平方英尺，住房价格将增加0.138千美元，即138美元。因为回归分析的目的在于弄清住房面积对住房价格的影响，因此，$\hat{\beta}_1$ 表示的是1单位面积变化所带来的价格变化。如图1-5所示，它是回归直线的斜率。

$\hat{\beta}_0$ =40.0表示什么呢？$\hat{\beta}_0$ 是常数项或截距项的估计值。在方程中，它表示住房面积为0时，住房的价格为40.0千美元，即40 000美元。如图1-5所示，估计出的回归直线过价格轴的40.0处。也许有人会说一块空地的价格怎么会是40 000美元呢，出于很多原因这个结论都是不合理的，第7.1节将对此做详细讨论。$\hat{\beta}_0$ =40.0可以更为妥当地解释为它仅仅是$SIZE_i$ =0时的估计价格。

---

12 对一个经济学家来说，在建立一个价格模型时如果不在方程右边写上度量单位，这很不寻常。这种以商品属性为函数的商品价格模型称为hedonic模型，第11.8节将对此做深入探讨。在继续这个案例之前，有兴趣的读者可以先浏览一下第11.8节的部分内容。

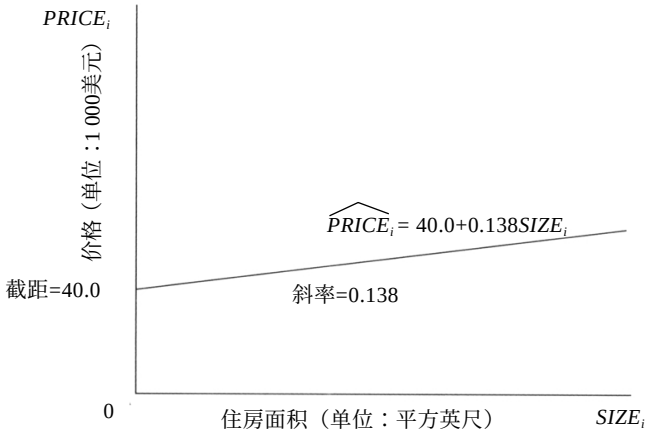

图1-5 房价的截面模型

注：对应方程（1-23），回归方程以南加利福尼亚州的房价为被解释变量、住房面积为解释变量，其截距项为40.0，斜率为0.138。

$\hat{\beta}_1$ =0.138表示什么呢？$\hat{\beta}_1$ 是方程（1-22）中*SIZE*的参数估计值，同时它也是图1-5中直线斜率的估计值。它表示住房面积每增加1单位会导致住房价格增加0.138千美元，即138美元。分析估计出的斜率参数是否讲得通是一个良好的习惯，正如所预料的一样，$\hat{\beta}_1$的符号为正。那么，参数的量纲呢？在解释参数的时候，必须要考虑其度量单位。在这个例子中，138美元/平方英尺合理吗？这很难确信，但一定比1.38美元/平方英尺或13 800美元/平方英尺更合理。

你怎么运用估计出的回归方程来帮助你决定是否为这套住房支付230 000美元呢？先计算出与你想买的住房面积（1 600平方英尺）相同的住房的 $\hat{Y}$ （价格的预测值），然后比较 $\hat{Y}$ 和房东的要价230 000。为此，用1 600替换方程（1-23）中的$SIZE_i$，得到：

$$\widehat{PRICE}_i = 40.0+0.138\times 1\,600=40.0+220.8=260.8\text{（千美元）}$$

这套住房看样子还比较划算，房东对面积值260 800美元的住房只收取230 000美元！也许你开始认为房价太高的想法是对整个南加利福尼亚州房价过高的总体感觉，而不是针对这个具体价格。

另一方面，影响房价的不仅仅是面积。（毕竟，南加利福尼亚州的住房如果没有游泳池或空调，又能好到哪里去呢？）这些多变量模型是计量经济学的核心，第11.8节再次探讨房价例子时，将会在方程（1-23）中加入更多的解释变量。

小结

1. 计量经济学，从字面意思上讲叫“经济度量”，属于经济学的一个分支学科，主要致力于量化理论关系。回归分析虽然仅仅是计量经济学分析方法的一种，却是目前最常用的一种方法。

2. 计量经济学最主要的用途是描述经济关系、假设检验和预测。根据研究的需要，特定的计量经济学方法会有所不同。

3. 尽管回归分析设定被解释变量是一个或多个解释变量的函数，但回归分析本身并不能证明或隐含因果关系。

4. 回归方程中必须加入随机误差项，用于度量被解释变量没有完全被解释变量解释而形成的误差。随机误差项的组成部分有：①遗漏或省略的变量；②数据的测量误差；③ 隐含理论的

函数形式与回归方程不同；④纯随机误差或不可预知的事件。

5. 估计出的回归方程是真实回归方程的近似替代，是从包含X和Y的样本数据中计算出来的。由于真实回归方程不可知，所以，计量经济分析注重的是估计出的回归方程和回归参数的参数估计值。被解释变量的特定观测值与其估计值之差被称为残差。

习题

（习题2的答案见附录A。）

1. 不查阅书本（或笔记），给出下列术语的定义，然后与本书所给出的定义做比较。
   a. 随机误差项　　f. 期望（值）
   b. 回归分析　　g. 残差
   c. 线性　　h. 时间序列数据
   d. 斜率系数　　i. 截面数据
   e. 多变量回归模型
2. 利用计算机软件，以及表1-2中重量(Y)和身高(X)的数据，对方程（1-21）重新进行估计。这里有三种录入数据的方法：直接录入数据；在EViews中打开数据文件HTWT1；从课程网站：www.pearsonhigh ered.com/ studenmund[⊖]中下载数据文件HTWT1（以表格、数据或ASCⅡ形式），下载好数据后，运行$Y=f(X)$，得出的结果应该与方程（1-21）一致。运用不同的程序对方程进行回归时，需要不同的指令。关于EViews和Stata的使用方法，详见附录A中的答案。
3. 下面每组解释变量与被解释变量之间的关系是正相关、负相关还是不确定？解释原因。
   a. 指定年份里美国的净投资总量和GDP之间的关系。
   b. 男性教授的头发数量与他的年龄之间的关系。
   c. 某个季节里种植小麦的亩数与季初小麦的价格之间的关系。
   d. 在同一年份里，某国的净投资总量与该国实际年利率之间的关系。
   e. 某一年份，GDP的年增长率与该年份头发的平均长度之间的关系。
   f. 金枪鱼罐头的需求量与其价格之间的关系。
4. 1.4节中关于身高和体重的例子。
   a. 回到数据集中，找出其中与估计出的回归直线偏离较大的三位消费者。如果将这三位消费者从样本中剔除掉，是否能得到更优的估计出的回归模型？
   b. 测量一位男性的身高，并将其输入方程（1-21）。得出的回归结果与真实值之差是否在10磅以内？如果不是，为什么？为什么对于相同身高的人，回归方程预测的体重结果一样，但是很明显，这些人的体重实际上并不完全一样？
   c. 仔细思考所选取的样本是不是随机选取？在某些方面，样本看起来是否有些反常？（提示：所抽取样本中的顾客是否具有随机性？）如果样本不是随机的，对回归方程的估计和体重的估计值有没有影响？
   d. 考虑除身高之外，至少一个对体重有影响的变量，并将其纳入到回归模型。如何获取这个变量的数据？如果将这个变量加入方程，该变量的符号预期是怎样的？
5. 下面继续讨论身高与体重的例子。假设收集了29位男性的身高和体重数据，估计得到方程：

$$\hat{Y}_i = 125.1 + 4.03X_i \qquad (1\text{-}24)$$

式中，Y_i代表第i个人的体重（单位：磅）；X_i代表第i个人的身高（单位：英寸，高于5英尺的部分）。

a. 为什么方程（1-24）中的各参数估计值

⊖ 打开该网站，点击网站上本书下面的Data Sets，即可进入数据下载页面，读者可以根据自己的需要下载相关数据，后面再提到数据文件时，都可以登录该网站下载。——译者注

与方程（1-21）中对应的各值不相同？

b. 比较方程（1-24）和方程（1-21）中对应的各参数估计值。哪个方程中身高和体重具有更陡峭的估计关系？哪个方程的截距更大？两个方程相交在哪一点？

c. 利用方程（1-24）和表1-2给出的身高数据，“预测”这20个人的体重。通过方程（1-24）估计得到的体重数据与原始数据相比，有多少次预测误差的绝对值在10磅以上？方程（1-24）和方程（1-21）相比，哪个更优？在此之前，能否预见这个结果？

d. 假设要做估计体重的工作，你将采用怎样的方程进行估计？（提示：答案不唯一。）

6. 并不是所有的参数估计值都是正数。例如，Jaime Diaz发表在《体育画报》上面的一篇论文，研究了美国职业高尔夫球协会（PGA）巡回赛中不同距离的推杆次数。[13] 论文中建立了推杆进洞次数百分比（P_i）关于推杆距离（L_i，英尺）的函数关系式。推杆距离越长，即使是专业的高尔夫球员，进洞的可能性也越小。所以可以预测，在用L_i解释P_i的方程中，L_i的参数为负。利用论文中的数据估计得到的方程如下：

$$\hat{P}_i = f(\bar{L}_i) = 83.6 - 4.1L_i \qquad (1\text{-}25)$$

a. 认真写出L_i参数的含义。

b. 假设另外一个人通过论文中的数据估计得到以下方程：

$$P_i = 83.6 - 4.1L_i + e_i$$

这个方程与方程（1-25）是否相同？如果是，那利用什么定义可以将该方程还原成方程（1-25）？

c. 利用方程（1-25）估计一个PGA高尔夫球员10英尺推杆进球的次数百分比。结果是否现实？再分别估计1英尺和25英尺的情况，结果是否现实？

d. 上一题的答案表明，这些数据应用于线性回归分析时存在一个问题，这个问题是什么？（提示：如果没有思路，先画出P_i关于L_i的函数图形，再在同一坐标中绘制方程（1-25）。）

7. 对于1.5节的住房价格模型，考察方程：

$$\widehat{SIZE}_i = -290 + 3.62PRICE_i \qquad (1\text{-}26)$$

式中，$SIZE_i$代表第i套住房的面积（单位：平方英尺）；$PRICE_i$代表第i套住房的价格（单位：1 000美元）。

a. 认真解释各回归参数估计值的意义。

b. 假设该方程解释了住房面积变动80%以上的原因。方程（1-26）是否表明，价格高的住房，其面积也大？如果不是，该方程表明了什么？

c. 如果价格单位不是1 000美元，而是美元，方程的参数估计值会有什么变化？详细阐述。

8. 如果方程不只有一个解释变量，则特别要注意回归参数的含义。比如，思考怎样建立一个方程，来解释不同州的公共教育支出花费在每个小学生身上的钱。一个州的收入越多，公共教育支出可能也越多。但是学生数量增加更快，导致花费在每个学生身上的钱更少。因此，一个合理的方程至少应该包括两个变量——收入和学生数量增长率：

$$S_i = \beta_0 + \beta_1 Y_i + \beta_2 G_i + \varepsilon_i \qquad (1\text{-}27)$$

式中，S_i代表第i州花费在每个公立学校学生身上的教育经费；Y_i代表第i州的资本收入；G_i代表第i州公立学校学生的增长率。

a. 说明变量Y与变量G的参数的经济意义。（提示：注意保持其他变量的影响不变。）

b. 如果要估计方程（1-27），你预期的变量Y和变量G的符号各是什么？请说明理由。

c. Silva和Sonstelie估计了以下关于各个州的每个学生教育经费的截面模型，方程（1-28）与方程（1-27）非常相似：[14]

$$\hat{S}_i = -183 + 0.142\,2Y_i - 5\,926G_i \qquad (1\text{-}28)$$
$$N = 49$$

方程中的估计参数与你预期的一致吗？请用常识来解释方程（1-28）。

d. 变量G是用小数来衡量的，因此，当一个

13 Jaime Diaz, “Perils of Putting,” *Sports Illustrated*, April 3, 1989, pp. 76-79.

14 Fabio Silva and Jon Sonstelie, “Did Serrano Gause a Decline in School Spending?” *National Tax Review*,Vol.48,No.2,pp.199-215.作者还将第i个州花费在每个学生身上的税收收入作为一个变量。

州的招生人数增加了10%时，G等于0.1。如果变量G用百分比的形式来衡量，那么当一个州的招生人数增加了10%时，G等于10，方程（1-28）会如何变化？（提示：写出参数估计值的真实数据。）

9. 你的朋友在大学里有一份工作，即向毕业的校友打电话，寻求他们对学校年度基金的捐献。她想要知道自己打去的电话是否能起到作用。为了检验学生的电话对基金捐献的影响，她搜集了50个校友的数据，并估计了方程（1-29）：

$$\widehat{GIFT}_i = 2.29 + 0.001INCOME_i + 4.62CALLS_i \quad (1\text{-}29)$$

式中，$GIFT_i$代表第i位校友对2008年年度基金的捐献（单位：美元）；$INCOME_i$代表估计的第i位校友在2008年的收入（单位：美元）；$CALLS_i$代表为鼓励校友捐献，给第i位校友打电话的次数。

a. 详细说明每一个参数估计值的意义。估计参数的符号与你预期的一致吗？

b. 为什么方程中左边的变量是$\widehat{GIFT}_i$而不是$GIFT_i$？

c. 在估计出的方程中，你的朋友并没有加入随机误差项。这种做法对吗？说明理由。

d. 假如你的朋友决定将变量$INCOME$的单位从“美元”变为“1 000美元”。方程中的参数估计值会发生什么变化？请详细说明。

e. 如果方程中可以再加入一个变量，应该加入什么变量？请说明理由。

10. 房价模型可以用时间序列数据或者是截面数据来估计。如果你研究了住房价格的时间序列数据（数据及资料来源见表1-3），得到估计方程：

$$\hat{P}_t = f(\overset{+}{GDP}) = 12\,928 + 17.08Y_t$$
$$N = 38\text{（样本区间：1970～2007年）}$$

式中，P_t代表第t年在美国一栋新的独立住宅的名义中间价格；Y_t代表第t年美国的GDP（单位：10亿美元）。

表1-3 关于住房价格的时间序列模型的数据

t	年度	价格(P_t)	GDP(Y_t)	t	年度	价格(P_t)	GDP(Y_t)
1	1970	23 400	1 038.5	20	1989	120 000	5 484.4
2	1971	25 200	1 127.1	21	1990	122 900	5 803.1
3	1972	27 600	1 238.3	22	1991	120 000	5 995.9
4	1973	32 500	1 382.7	23	1992	121 500	6 337.7
5	1974	35 900	1 500.0	24	1993	126 500	6 657.4
6	1975	39 300	1 638.3	25	1994	130 000	7 072.2
7	1976	44 200	1 825.3	26	1995	133 900	7 397.7
8	1977	48 800	2 030.9	27	1996	140 000	7 816.9
9	1978	55 700	2 294.7	28	1997	146 000	8 304.3
10	1979	62 900	2 563.3	29	1998	152 500	8 747.0
11	1980	64 600	2 789.5	30	1999	161 000	9 268.4
12	1981	68 900	3 128.4	31	2000	169 000	9 817.0
13	1982	69 300	3 255.0	32	2001	175 200	10 128.0
14	1983	75 300	3 536.7	33	2002	187 600	10 469.6
15	1984	79 900	3 933.2	34	2003	195 000	10 960.8
16	1985	84 300	4 220.3	35	2004	221 000	11 685.9
17	1986	92 000	4 462.8	36	2005	240 900	12 421.9
18	1987	104 500	4 739.5	37	2006	246 500	13 178.4
19	1988	112 500	5 103.8	38	2007	247 900	13 807.5

注：P_t代表第t年在美国一栋新的独立住宅的名义中间价格。（资料来源：The Statistical Abstract of the U.S.）Y_t代表第t年美国的GDP（单位：10亿美元）。（资料来源：The Economic Report of the President。）本表数据文件名为HOUSE1。

a. 详细说明估计参数的经济意义。
b. 一般情况下Y_t通常都出现在方程的左边，在这个方程中Y_t是在方程的右边，它在方程中起到什么作用？
c. 价格和GDP这两个变量都是以名义值（或是当前值，而不是实际值或是通货膨胀调整值）来衡量的。因此，方程中房价上升最主要的影响来自于1970～2007年巨大的通货膨胀（变量P_t的变化有99%可以由变量Y_t的变化来解释）。怎样才能消除通货膨胀在方程中的影响？
d. 将GDP包含在方程中，不仅仅是为了衡量通货膨胀的影响。变量GDP还代表了其他什么因素？有没有什么更好的变量可以替代GDP？
e. 为了加深你对时间序列数据和截面数据之间区别的理解，请将你在d问中提出的变量与你在方程（1-22）中要加入的变量进行比较。两个方程中的被解释变量都是房价，你能在两个方程中同时加入一个相同的变量吗？请说明理由。

11. 随机误差项与残差的区别是本章的一个难点。
a. 列出至少三个随机误差项与残差之间的区别。
b. 通常情况下，残差项是观察不到的，但是假设我们知道了估计参数的真实值，误差项就可以计算出来了。假设β_0的真实值为0，β_1的真实值为1.5，估计方程为$\hat{Y}_i = 0.48 + 1.32X_i$，分别计算下表中6组观测值的残差和误差项的值。

Y_i	2	6	3	8	5	4
X_i	1	4	2	5	3	4

（提示：为了回答上述问题，你需要先求出方程（1-14）中的ε。）
本表数据文件名为EX1。

12. 回到1.2节中决定工资水平的例子中，建立某一特定行业中第i位工人工资的模型，将其作为工作经验、教育水平和工人性别的函数。

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \varepsilon_i \quad (1\text{-}12)$$

式中，Y_i代表$WAGE_i$，即第i位工人的工资；X_{1i}代表EXP_i，即第i位工人的工龄；X_{2i}代表EDU_i，即第i位工人继高中之后的受教育年限；X_{3i}代表$GEND_i$，即第i位工人的性别（1=男性，0=女性）。
a. β_2在现实世界中的含义是什么？（提示：如果你不知如何解决，复习1.2节。）
b. β_3在现实世界中的含义是什么？（提示：注意性别是虚拟变量。）
c. 假定为了确定不同肤色工人的工资是否有差别，在模型中添加一个变量。如何定义这个变量？请详细说明。
d. 假定可以在方程中加入另一个变量，下面哪个变量更合理？请解释原因。
Ⅰ. 第i位工人的年龄。
Ⅱ. 第i位工人在这个行业中的工作年限。
Ⅲ. 这个行业的平均工资水平。
Ⅳ. 第i位工人的受雇佣月份数。
Ⅴ. 第i位工人拥有的孩子数。

13. 你听说过评师网站吗？在这个网站上，学生们对教授的总体教学能力和一系列其他因素进行评价。然后，基于学生从教授课堂上的收获，网站对这些学生评价进行总结。

网站筛选出来的两个影响教授评级高低最重要的因素是“容易度”（通过学习负荷量、成绩来衡量）和“热度”（假定为教授自身的吸引力）。一篇最近发表的论文[15]指出，“热度”大小对于教授评级高低的影响要大于“容易度”。为了对上述结论进行验证，我们建立了以下基于评师网站数据的模型：

$$RATING_i = \beta_0 + \beta_1 EASE_i + \beta_2 HOT_i + \varepsilon_i \quad (1\text{-}30)$$

式中，$RATING_i$代表第i位教授的评级总分（最好=5分）；$EASE_i$代表第i位教授的“容易度”评分（最容易=5分）；HOT_i代表虚拟变量，如果第i位教授被认为有很强的吸引力则为1，否则为0。

为了估计方程（1-30），需要以上三个变量的数据，表1-4提供了评师网站上

15 James Otto,Douglas Sanford, and Douglas Ross,“Does RateMyProfessors.com Really Rate My Professor?” *Assessment and Evaluation in Higher Education*, August 2008, pp.355-368.

任意选取的25位教授的数据。利用这些数据，得到下面的回归结果：

$$\widehat{RATING}_i = 3.23+0.01EASE_i+0.59HOT_i \quad (1\text{-}31)$$

a. 观察方程（1-31），相应的参数估计值是否符合预期？请解释理由。

b. 检验一下自己能否独立做出上述回归。利用表1-4中的数据，使用EViews软件、Stata或你自己的回归程序来进行回归估计。如果你进行回归的步骤正确，那得出的结果应该能够验证方程（1-31）中的估计结果。（如果你不确定该问题从何处着手解决，可参照附录A中的答案。）

c. 这个模型包含了两个解释变量，但教授评级的高低是否仅与这两个变量有关？是否存在其他重要的因素？

d. 假定可以将你认为重要的其他变量加入方程（1-31），变量*EASE*和*HOT*的参数会发生什么变化？你期望这些参数发生变化吗？请解释原因。

（选做）在评师网站上，任意选取25位观测者（见表1-4），估计方程（1-30）。将回归结果与利用方程（1-31）得到的结果进行比较，方程（1-30）中的参数估计值是否与方程（1-31）中的相同？为什么？

表1-4 评师网站的数据

OBS	评级总分	热度	容易度
1	2.8	3.7	0
2	4.3	4.1	1
3	4.0	2.8	1
4	3.0	3.0	0
5	4.3	2.4	0
6	2.7	2.7	0
7	3.0	3.3	0
8	3.7	2.7	0
9	3.9	3.0	1
10	2.7	3.2	0
11	4.2	1.9	1
12	1.9	4.8	0
13	3.5	2.4	1
14	2.1	2.5	0
15	2.0	2.7	1
16	3.8	1.6	0
17	4.1	2.4	0
18	5.0	3.1	1
19	1.2	1.6	0
20	3.7	3.1	0
21	3.6	3.0	0
22	3.3	2.1	0
23	3.2	2.5	0
24	4.8	3.3	0
25	4.6	3.0	0

注：数据文件名为RATE1。