新编21世纪经济学系列教材

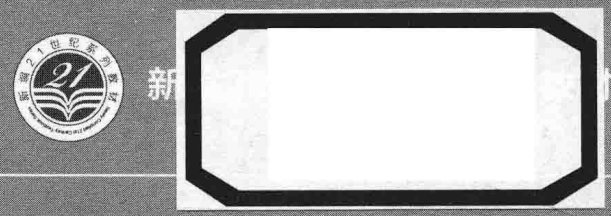
应用计量经济学

第二版

Applied Econometrics

赵国庆 编著

中国人民大学出版社



应用计量经济学

第二版

Applied Econometrics

赵国庆 编著

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

应用计量经济学/赵国庆编著. —2版. —北京：中国人民大学出版社，2017.1
新编21世纪经济学系列教材
ISBN 978-7-300-23706-0

Ⅰ.①应… Ⅱ.①赵… 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（2016）第286244号

新编21世纪经济学系列教材
应用计量经济学（第二版）
赵国庆 编著
Yingyong Jiliang Jingjixue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京密兴印刷有限公司	**版　　次**	2011年9月第1版
规　　格	185mm×260mm　16开本		2017年1月第2版
印　　张	13.5	**印　　次**	2017年1月第1次印刷
字　　数	311 000	**定　　价**	28.00元

第二版说明

《应用计量经济学》一书从2011年9月第一版发行至今已近6年，有必要对教材进行修订。结合读者的意见，与第一版比较，第二版主要在以下几个方面做了修改与调整。

1. 增加新的第一章“数据分析基础”，第一版中的第一章至第七章分别调整为第二章至第八章。

2. 根据教学实践对部分例题进行调整。

3. 修订部分公式与表述中存在的问题。

4. 修改参考文献与统计附表。

在本书的定稿过程中，作者用EViews计量分析软件对书中的例题和习题重新进行了计算，中国人民大学信息学院范红岗博士，山东师范大学尹慧博士，中国人民大学数量经济学专业研究生李本钊、惠炜、文韬、姚青松等参与了部分工作，值此书付梓之际，向他们表示感谢。

从本书的第一版开始，一直得到中国人民大学经济学院的支持，本书为经济学院规划教材，中国人民大学出版社王美玲编辑也为本书的出版付出了辛勤的劳动，在此表示由衷的感谢。

赵国庆

2016年12月

引　言

构造计量经济模型是研究经济现象的一种方法，它使人们能够客观地解释经济现象背后的形成机制，其主要目的在于经济理论的数量化研究。近年来，计算机的日益更新和计量软件的多样化发展也给计量经济分析方法的利用与普及提供了强有力的支撑。《应用计量经济学（第二版）》一书旨在使读者掌握常用的计量经济理论与方法，强调以解决实际问题为导向，训练读者的数量化基本技能。众所周知，计量经济习题的计算对于理解计量经济学是必不可少的重要环节，在计算过程中要用到 EViews 软件，基于这样的考虑，本书每章都配备了大量的习题，并给出了习题解答，在习题的答案中详细给出了 EViews 软件的输出结果和部分程序。我们希望通过对问题的分析与估算逐步培养读者对经济问题的理解力，使读者能够更好地运用计量经济分析方法解决实际经济应用问题，在此基础上，达到对经济变化趋势进行预测和政策评价之目的。

本书共由八章构成。第一章为数据分析基础。第二章为一元线性回归模型，对于两个变量 X 和 Y 之间关系的度量，在 X 为原因、Y 为结果关系已知的情况下，回归分析比两个变量间的相关分析更为有效。本章将讨论两变量回归分析中最小二乘法的含义、计算方法、估计结果的评价和检验等基本内容。第三章为多元线性回归分析，是第一章两变量分析方法的扩展，将讨论包含 k 个变量的线性回归模型，主要包括模型的参数估计、估计结果的评价指标，以及对估计结果的检验等内容，同时，对实证分析中遇到的多重共线性问题加以说明。第四章讨论模型中误差项假定的诸问题，如果误差项的经典假设条件不成立，会给最小二乘估计与检验结果带来很大影响，本章将讨论误差项不相关和同方差的假设条件不成立时，如何进行估计的问题，同时给出序列相关和异方差性的检验方法。第五章是对线性模型的扩展，第二至四章主要研究线性模型的推断问题。本章将扩展前面有关线性模型的讨论，主要内容包括模型的类型与变换、解释变量中虚拟变量的使用、经济结构变化的检验、分布滞后变量对模型的影响等内容。第六章为联立方程组模型的估计，一

般经济变量之间都存在着相互依存关系。对这种相互依存关系的处理，要借助于联立方程组模型，联立方程组模型对经济预测和政策评价来说都是非常重要的工具。本章在讨论联立方程组模型的偏误的基础上，解释方程组模型存在的识别问题，重点介绍模型估计方法TSLS。第七章为模型估计方法的扩展，本章介绍近年来在实证分析中经常使用的一些模型，包括：二元选择模型、受限因变量模型、面板数据模型，以及SUR方程的估计问题。第八章讨论如何利用EViews对经济数据进行建模，EViews是国际上通用的计量经济分析软件包，是初学者掌握计量方法的有效工具，本章主要介绍EViews软件包的一些基本功能，同时给出运用EViews对经济模型进行估计和预测的一些简单实例，本章的学习最好结合前面几章的习题解答中EViews估计结果一并进行。

过去对于计量经济分析方法的学习，建模者通常是在充分掌握计量经济理论后，再对模型进行估计推断。这种学习计量经济学的时代已经过去，今天是在进行模型估算的同时学习必要的计量分析理论，对计量经济学本身而言，这可能是一个必然的发展趋势。不需要难懂的计量分析理论，掌握必要的计量分析方法与计量软件的时代已经来临。《应用计量经济学（第二版）》一书正是基于这一理念编写的。

内容简介

本书旨在使读者掌握常用的计量经济理论与方法，强调以解决实际问题为导向。习题的计算是理解计量经济学必不可少的重要环节，本书每章都配备了大量的习题，并给出了详细的解答。本书主要内容如下：两变量回归分析中，最小二乘法的计算、估计结果评价与检验；k 变量线性回归模型的估计、模型评价指标及检验统计量；误差项不相关与同方差假设条件不成立时的估计与检验；模型中虚拟变量的使用、经济结构变化的检验、分布滞后模型的估计；联立方程组模型的偏误、识别及 TSLS 估计量。二元选择、受限因变量、面板数据以及 SUR 模型的估计与检验；本书最后给出利用 EViews 对经济问题进行建模的部分案例。

作者简介

赵国庆，中国人民大学经济学院教授、博士生导师，日本京都大学经济学博士。主要研究方向为计量经济学理论与应用。

在 *International Journal of Production Economics*，*Japanese Economic Review*，*China & World Economy*，《经济学季刊》，《金融研究》，《数量经济 & 技术经济研究》，《统计研究》等国内外专业杂志发表论文 60 余篇。负责和承担过国家省部委等多个科研项目。浙江大学、华侨大学兼职教授，日本关西学院大学客座教授。中国数量经济学会学术委员会副主任。

主要成果：

1. Unit Root Analyses of the Causality between Japanese Money and Income, *Japanese Economic Review*, 1997. (with Morimune. K.)
2. Heuristics for Replenishment with Linear Decreasing Demand, *International Journal of Production Economics*, 2001. (with Yang. J. and Rand G. K.)
3. Uncovering the Relationship between FDI, Human Capital and Technological Progress in Chinese High-technology Industries, *China & World Economy*, 2010. (with Zhang Z.)

目　录

第一章

数据分析基础

一、本章主要内容

1. 均值与方差
2. 相关分析
3. 假设检验

二、习题

三、习题解答与提示

本章介绍数据分析的基础知识，包括样本均值与方差的计算、相关分析与假设检验等基本内容。本章内容构成下面各章的预备知识。

1.1 均值与方差

1. 样本均值

设 $\{x_1, x_2, \cdots, x_n\}$ 为变量 x 的 n 个观测值，x 的样本均值定义如下

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1—1}$$

对变量 x 的全部观测值求和之后，除以样本容量 n 得到 $\bar{x}$，$\bar{x}$ 表示 n 个观测值 $\{x_1, x_2, \cdots, x_n\}$ 的中心位置。

在统计学中，也使用中位数代表观测值的中心位置。把观测值 $\{x_1, x_2, \cdots, x_n\}$ 按从小到大的顺序排成一列后，中心位置的值称为中位数。如果从小到大排列的观测值为 {3,7,

8,12,20}，那么样本均值为

$$\frac{1}{5}(3+7+8+12+20)=10$$

由于中位数是中心位置的值，所以中位数为8。如果没有最后的观测值20，那么样本均值变为7.5。而中位数是7和8的平均值7.5。此时，样本均值与中位数一致。当观测值的个数为偶数时，因为中心位置的值无法确定，故用中心位置的两个数的平均作为中位数。

可以发现从数据中删除20的结果是，样本均值从10变到7.5，减少得很大。但是中位数却仅从8变到7.5，减少得很小。由此可见，中位数受特别大的值或特别小的值的影响很小，比较稳定。当观测值个数比较多时，这个性质会变得更加明显，这一性质也称为中位数的稳健性。

加权平均也是一种求平均数的方法。加权平均是对每个观测值给予不同的权重，用权重来计算平均。权重取值非负，权重之和为1。例如，对于观测值{3,7,8,12,20}，其加权平均如下

$$\frac{1}{9}\times 3+\frac{2}{9}\times 7+\frac{3}{9}\times 8+\frac{2}{9}\times 12+\frac{1}{9}\times 20=\frac{85}{9}=9.44$$

式中每个观测值都乘以它的权重，对离中心位置近的值给予比较大的权重，但权重的总和为1。一般来说，加权平均使用非负权重w_i，定义如下：

$$\sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{1—2}$$

式中权重w_i之和为1。权重w_i也可以是0，权数为0时，表明对应的观测值被忽略。不难发现样本均值是具有相同权重的加权平均。

2. *样本方差*

样本方差是以样本均值为中心的表示数据分散程度的指标，其定义如下

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1—3}$$

式中的每一项为观测值x_i与样本均值$\bar{x}$之差的平方，故样本方差表示x_i与$\bar{x}$之差的平方的平均。如果数据集中在均值的周围，则样本方差的值较小。如果数据离均值较远，则样本方差的值较大。注意到式（1—3）中的分母取值为$(n-1)$，本书中样本方差用此定义。也存在分母取值为n的情形，当n较大时，两个结果基本上没有差异。

3. *样本方差的分解公式*

计算样本方差可以由下面的公式简化：

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \tag{1—4}$$

证明：上式左端为

$$(x_1^2-2x_1\bar{x}+\bar{x}^2)+(x_2^2-2x_2\bar{x}+\bar{x}^2)+\cdots+(x_n^2-2x_n\bar{x}+\bar{x}^2)$$

整理括号内的第 1 项、第 2 项、第 3 项

$$\begin{aligned}&(x_1^2+x_2^2+\cdots+x_n^2)-2(x_1+x_2+\cdots+x_n)\bar{x}+(\bar{x}^2+\bar{x}^2+\cdots+\bar{x}^2)\\=&(x_1^2+x_2^2+\cdots+x_n^2)-2n\bar{x}^2+n\bar{x}^2\\=&(x_1^2+x_2^2+\cdots+x_n^2)-n\bar{x}^2\end{aligned}$$

得到式（1—4）的右端。与左端相比，右端的计算比较容易。

根据样本均值的定义，样本对于均值的离差之和为 0。

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \tag{1—5}$$

利用这一性质，可以给出另外的证明。

离差的平方和可以分解为

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2&=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\sum_{i=1}^{n}[(x_i-\bar{x})x_i-(x_i-\bar{x})\bar{x}]\\&=\sum_{i=1}^{n}(x_i-\bar{x})x_i-\sum_{i=1}^{n}(x_i-\bar{x})\bar{x}\end{aligned}$$

式中右端的第 1 项：

$$\sum_{i=1}^{n}(x_i-\bar{x})x_i=\sum_{i=1}^{n}(x_i^2-\bar{x}x_i)=\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i$$

式中右端的第 2 项：

$$\sum_{i=1}^{n}(x_i-\bar{x})\bar{x}=(x_1-\bar{x})\bar{x}+\cdots+(x_n-\bar{x})\bar{x}=\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})$$

根据式（1—5），该项为 0。

例如，数据集合 {3,7,8,12,20} 的样本均值为 10。该组数据的方差根据对均值 10 的离差平方求和来计算：

$$\frac{1}{4}(7^2+3^2+2^2+2^2+10^2)=41.5$$

或者由观测值的平方和减去均值平方的 5 倍来计算：

$$\frac{1}{4}\{(3^2+7^2+8^2+12^2+20^2)-5\times10^2\}=41.5$$

如果不是减平均的 5 倍，而是减总和平方的$\frac{1}{5}$，也可以得到相同的答案：

$$\frac{1}{4}\{(3^2+7^2+8^2+12^2+20^2)-\frac{1}{5}\times50^2\}=41.5$$

4. 样本标准差

样本方差正的平方根 s_x 称为样本标准差。样本方差是观测值和样本均值之差的平方的

平均值。故样本方差是以样本均值为中心的。对于观测值的分布情况，一般无法从样本方差得到直观的信息。同时，度量单位是原单位的平方，因此有必要以样本标准差来表示数据的性质。

样本标准差的值，在理解以均值为中心的数据的分散程度这一概念时非常有用。样本标准差常用 σ 表示，下面要说明的 σ 区间是直观理解数据分散程度的重要工具。例如以均值为中心，1 σ 区间

（均值－样本标准差，均值＋样本标准差）

内含有一半以上的观测值，2 σ 区间

（均值－2×样本标准差，均值＋2×样本标准差）

内可以理解为含有大部分的观测值。更进一步地，通常认为 3 σ 区间

（均值－3×样本标准差，均值＋3×样本标准差）

内含有所有的观测值。

对于数据集合 {3,7,8,12,20}，其标准差是 6.4，因此 1 σ 区间是

$$(10-6.4,10+6.4)=(3.6,16.4)$$

5 个观测值中有 3 个位于此区间内。所有的观测值都包含于 2 σ 区间中。

当观测对象是来自服从正态分布的总体时，包含于 1 σ 区间内的观测值约占 70%，2 σ 区间内的观测值约占 95%，3 σ 区间内的观测值占 99%以上。

5. 切比雪夫不等式

对于判断数据的分散程度，切比雪夫不等式虽然比较粗略，但却是一个经常使用的标准。k 为正整数，以均值为中心的 $k\sigma$ 区间由

$$(\bar{x}-k\times\text{样本标准差},\bar{x}+k\times\text{样本标准差}) \tag{1—6}$$

来确定。在所有的观测值中，不包含于此区间的观测值所占的比例在 $(1/k^2)$ 以下。切比雪夫不等式虽然可以说对所有的数据都适用，但是对比如超出 2 σ 区间的观测值所占的比例，根据切比雪夫不等式所得结果是 1/4＝0.25，因此，它只能够提供像小于 25%这样非常粗略的信息。同样的，超出 3 σ 区间的观测值所占比例小于 1/9。超出 1 σ 区间的观测值所占比例小于 1，这个结果基本上没有意义。虽然这个不等式较为粗略，但对于观测值分布的直观印象而言，却非常有用。

例 1—1 18 个国家（地区）2002 年度的寿命与收入数据见表 1－1，其中 G 表示收入，L 表示寿命。讨论数据的分散程度。

解：18 个国家（地区）的平均收入为

$$\frac{210.6}{18}=11.7(\text{千美元})$$

样本方差为

$$\frac{1}{17}\left(4\ 251.8-\frac{1}{18}\ 210.6^2\right)\approx 105.16(\text{千美元}^2)$$

表 1-1　　人均收入与平均寿命

国家（地区）	G	G^2	L	L^2	$G\times L$
老挝	1.5	2.3	54	2 916	83
孟加拉国	1.7	2.7	61	3 721	101
印度	2.4	5.7	63	3 969	151
印度尼西亚	2.8	8.1	66	4 356	187
巴西	7.3	53.6	67	4 489	490
菲律宾	4.2	17.8	69	4 761	291
土耳其	7.0	49.4	69	4 761	485
泰国	6.3	40.1	69	4 761	437
中国	3.9	15.5	70	4 900	276
伊朗	5.9	34.8	71	5 041	419
马来西亚	8.4	69.9	72	5 184	602
沙特阿拉伯	11.1	122.1	72	5 184	796
韩国	17.3	300.7	73	5 329	1 266
美国	34.3	1 173.7	77	5 929	2 638
新加坡	25.0	623.5	78	6 084	1 948
以色列	19.3	373.3	78	6 084	1 507
中国香港	25.7	658.4	80	6 400	2 053
日本	26.5	700.1	81	6 561	2 143
总和	210.6	4 251.8	1 270	90 430	15 871

注：数据有四舍五入。

G：通过购买力平价（PPP）核算的人均 NGI（单位：1 000 美元）；L：平均寿命。

资料来源：世界银行，《世界发展报告》，2003 年。

标准差为 10.25（千美元）＝10 250 美元。注意，方差的单位是千美元的平方。类似计算得到，平均年龄是 70.6 岁，样本方差是

$$\frac{1}{17}\left(90\ 430-\frac{1}{18}\ 1\ 270^2\right)\approx 48.5\ (\text{岁}^2)$$

标准差是 7.0 岁。收入的 $1\hat{\sigma}$ 区间为

$$(11\ 700-10\ 250,11\ 700+10\ 250)=(1\ 450\ \text{美元},21\ 950\ \text{美元})$$

18 个国家（地区）中有 14 个国家（地区）包含于此区间。$2\hat{\sigma}$ 区间为

$$(11\,700-2\times10\,250,11\,700+2\times10\,250)=(-8\,800\text{美元},32\,200\text{美元})$$

在该区间内含有负收入。18个国家（地区）中只有美国超出了此区间。关于年龄的$1\hat{\sigma}$区间为

$$(70.6-7,70.6+7)=(63.6\text{岁},77.6\text{岁})$$

有11个国家（地区）包含于此区间。$2\hat{\sigma}$区间为

$$(70.6-2\times7,\ 70.6+2\times7)=(56.6\text{岁},\ 84.6\text{岁})$$

只有老挝不在此区间。无论哪一组数据，不包含于$2\hat{\sigma}$区间的观测值所占比例只有1/18，也就是5.6%。这个比例当然要比1/4小。

6. *波动率*

波动率是近年在金融统计和金融工程中频繁使用的用语之一，它的意思是金融资产的不稳定性。不稳定性也叫风险，经常使用标准差作为将其定量化的一个方法。图1-1是日经平均股票价格（每日的日经225价格指数）的收益率的轨迹，y轴是以标准差（σ）为刻度的，根据图形，收益率在64日中有5次超过了2σ的范围。这是比5%要高的比率。

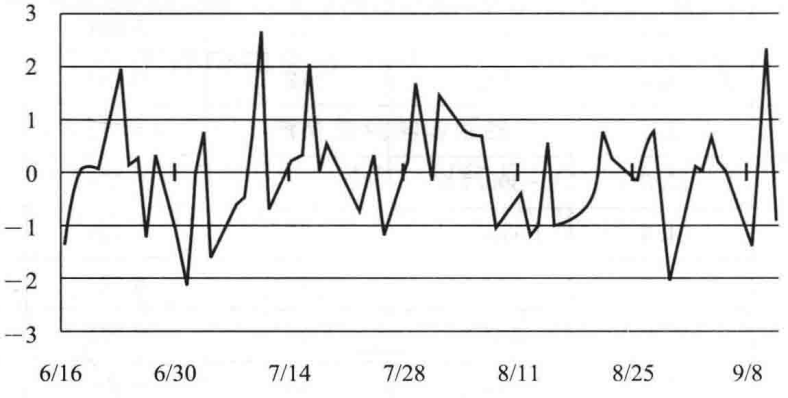


图1-1　日经225的收益率

7. *数据的标准化*

如果求出数据集合$\{x_1,x_2,\cdots,x_n\}$的样本均值为$\bar{x}$和标准差s_x，就可以对该组数据进行标准化。标准化就是对各个观测值x，减去样本均值并除以标准差的运算。

$$z_1=\frac{x_1-\bar{x}}{s_x},\cdots,z_n=\frac{x_n-\bar{x}}{s_x} \tag{1—7}$$

数据的标准化是统计分析的基础，数据标准化后具有以下性质：

数据标准化后，样本均值为0，标准差为1。以均值为中心的$k\sigma$区间为$(-k,k)$。

证明：标准化后数据的样本均值为0，样本方差是

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}z_i^2&=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)^2\\&=\frac{1}{s_x^2}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\\&=1\end{aligned} \tag{1—8}$$

关于 $k\sigma$ 区间，由不等式

$$\bar{x}-ks_x<x_i<\bar{x}+ks_x$$

的两边减去 $\bar{x}$，再除以 s_x 而得到。

$$-k<z_i<k$$

这样就得到所要求的不等式。

例 1—2 在表 1-2 中，第 3 行（z）是对寿命数据（L）进行标准化后的数据。老挝（国家代码 1）以外国家和地区寿命的数据分布于从 -1.5σ 到 1.5σ 的区间内。18 个国家和地区中有 11 个分布于 1σ 区间（-1，1）内。2σ 区间（-2，2）内有 17 个。图 1-2 是根据表 1-2 得到的直方图。

表 1-2　　寿命（L）和标准化数值（z）

国家和地区	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
L	54	61	63	66	67	69	69	69	70	71	72	72	73	77	78	78	80	81
z	−2.4	−1.4	−1.1	−0.7	−0.5	−0.2	−0.2	−0.2	−0.1	0.1	0.2	0.2	0.4	0.9	1.1	1.1	1.4	1.5

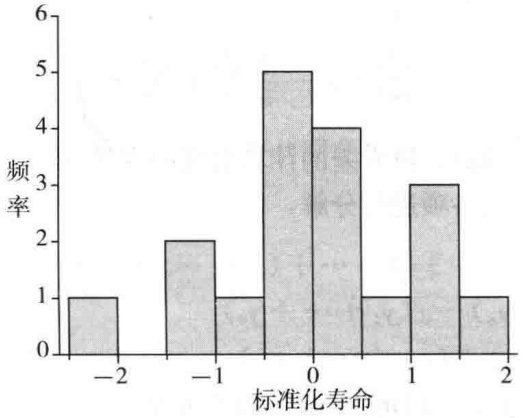


图 1-2　标准化寿命 z 的直方图

因为观测的个数只有 18 个，所以并未形成左右对称的直方图。通常在观测值的个数较多时，直方图呈左右对称形状的情况比较常见。

1.2　相关分析

1. 样本协方差

当观测值表示观测对象的多个特征时，不仅仅需要计算均值和方差，协方差和相关系数也是描述统计特征的重要指标。例如有 n 个人的身高 x 和体重 y 的数据：

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$$

其中，(x_1,y_1) 表示第 1 个人的观测值，下标符号是表示观测对象的序号。如果得到 n 个人的这两个特征的数据，就可以分别计算身高、体重的样本均值和样本标准差。身高和体重的样本均值和样本标准差分别由 $\bar{x},\bar{y},s_x,s_y$ 来表示。

均值和方差可以表示单个变量的数据性质，然而却忽视了 x 和 y 之间的关系。换一种说法，在计算 $\bar{x}$ 和 s_x 时，只需要数据 $\{x_1,x_2,\cdots,x_n\}$，没有使用变量 y 所带来的信息。同样，在计算 $\bar{y}$ 和 s_y 时，也忽略了变量 x 所带来的信息。此时，表示身高 x 和体重 y 这两个变量相关联的数字特征是 x 与 y 的样本协方差，定义如下：

$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{1—9}$$

如果 x 和 y 之间没有关联，样本协方差与 0 非常接近。注意式中的求和为：

$$(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y}) \tag{1—10}$$

在这一计算中，先分别求出 x 和 y 关于其均值的离差，然后相乘求和。与式（1—4）的计算类似，在求样本协方差时，也可以进行下面的分解：

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y} \tag{1—11}$$

$$=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i \tag{1—12}$$

利用式（1—11）与式（1—12），协方差的计算会变得容易。

证明：对式（1—10）的各项进行分解，得到

$$\begin{aligned}&(x_1y_1-\bar{x}y_1-\bar{y}x_1+\bar{x}\bar{y})+\cdots+(x_ny_n-\bar{x}y_n-\bar{y}x_n+\bar{x}\bar{y})\\=&(x_1y_1+\cdots+x_ny_n)-\bar{x}(y_1+\cdots+y_n)\\&-\bar{y}(x_1+\cdots+x_n)+(\bar{x}\bar{y}+\cdots+\bar{x}\bar{y})\\=&(x_1y_1+\cdots+x_ny_n)-\bar{x}(n\bar{y})-\bar{y}(n\bar{x})+n\bar{x}\bar{y}\end{aligned}$$

或者直接使用式（1—7）得到

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})&=\sum_{i=1}^{n}(x_i-\bar{x})y_i-\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})\\&=\sum_{i=1}^{n}(x_iy_i-\bar{x}y_i)\\&=\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i\end{aligned}$$

例 1—3　利用数据 $\{(3,12),(7,7),(8,5),(12,2),(20,4)\}$ 计算 x 和 y 之间的协方差。

解：通过计算各数据关于均值的离差来求协方差：

$$\frac{1}{4}\{(-7)\times6+(-3)\times1+(-2)\times(-1)+2\times(-4)+10\times(-2)\}=-17.75$$