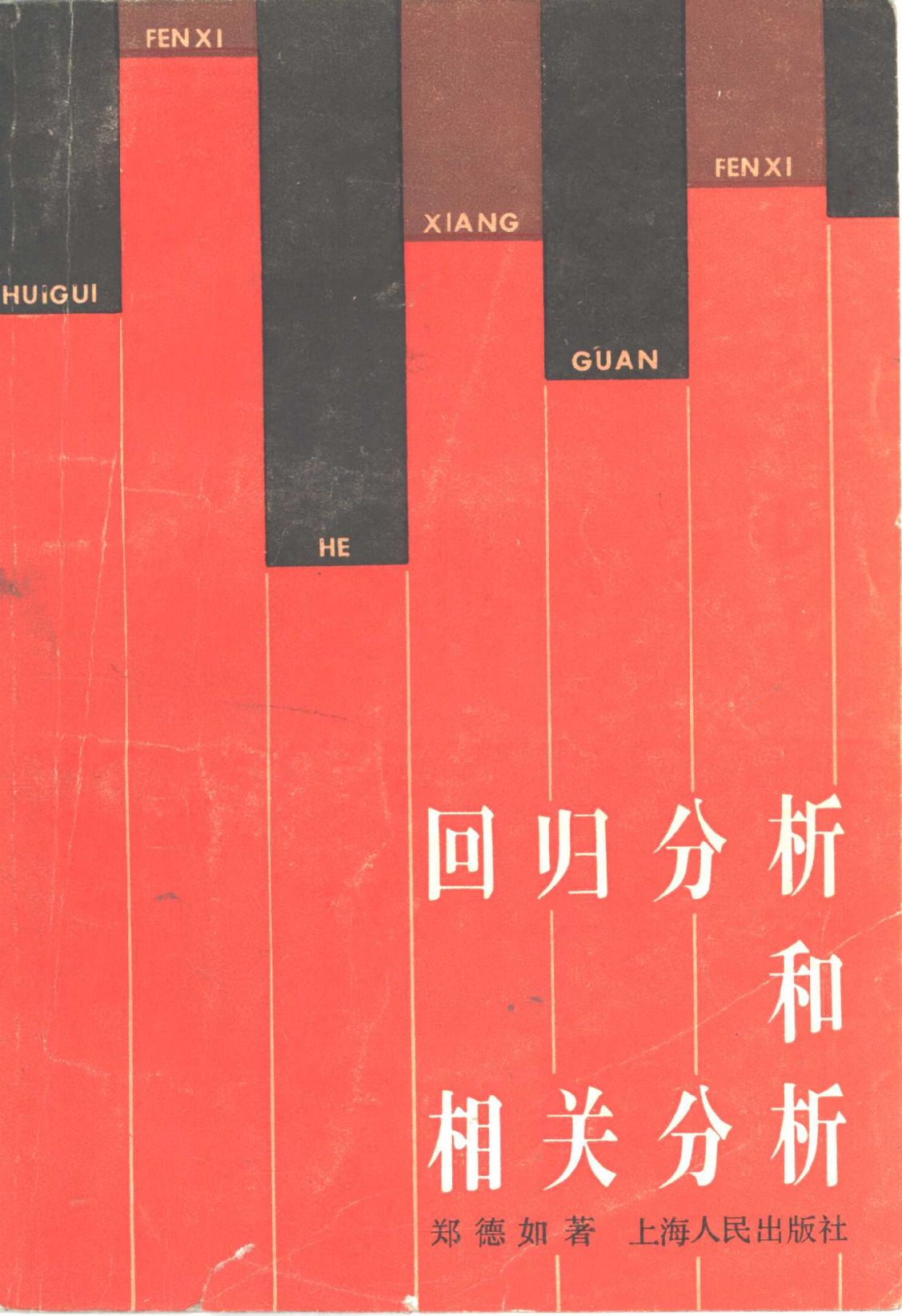

回归分析和相关分析

郑德如著

上海人民出版社

回归分析和相关分析

郑德如 著

上海人民出版社

封面装帧　杨德鸿

回归分析和相关分析

郑德如 著

上海人民出版社出版

（上海绍兴路54号）

新华书店上海发行所发行　常熟兴隆印刷厂印刷

开本 787×1092 1/32　印张 4.75　插页 1　字数 100,000

1984年12月第1版　1984年12月第1次印刷

印数 1—19,000

书号 4074·558　定价 0.64元

编者的话

回归分析和相关分析是数理统计学的一个重要组成部分，它在经济管理、工农业生产和科学研究等各个方面已得到广泛应用，它在社会经济领域中将有广阔的发展道路。

为了使回归分析和相关分析在社会经济中的应用取得更大成果，以适应社会主义现代化建设的需要，作者编写了这本《回归分析和相关分析》。

本书叙述了回归分析和相关分析的基本理论和方法，力求讲透基本原理，内容通俗易懂。书中广泛搜集了国民经济、工农业生产等方面的实例，详细列出其计算方法和步骤，使读者易于掌握其基本方法和实际应用。

本书可供高等财经院校、中等专业学校作为教学用书，也可供广大经济管理工作者、统计工作者、科技工作者自学参考之用。

本书承上海财经学院统计系主任贾宏宇教授审阅，提出了宝贵的意见，在此特致谢意。

一九八三年十月

目　录

引　言

社会经济现象是相互依存又相互联系的，对现象间相互联系的认识和分析，是人们改造客观世界的一个极其重要的方面。

对现象间的相互联系，可以从不同的角度和范畴进行研究分析。

回归分析和相关分析是分析现象间联系形态和密切程度的数学方法。

所谓回归分析，就是对具有相互联系的现象，根据其关系的形态，选择一个合适的数学模式，用来近似地表达变量间平均变化关系。这个数学模式，称为回归方程式。

相关分析是在回归分析的基础上，用一个指标，表明变量间相互依存关系的密切程度。这个表明关系密切程度的指标，称为相关指标。

回归分析和相关分析早在十九世纪后半期就已经开始应用，我们可以追溯它们的历史。

十九世纪后半期，随着工业生产和科学技术的发展，在生物学界、人类学界以及社会科学领域里需要回答这样一个问题：如何度量现象间的关系。在 1877 年以及 1892 年，美国有两个学者对此曾作过尝试。但较有成效的，要归功于英国的学者。

英国的遗传学者高尔登(Francis Galton)对遗传问题进行了大量的研究。在1877～1889年的十多年间，高尔登得出了一个数学公式，这个公式用来度量孩子们的身高与父母平均身高之间的关系。根据统计测定，假如父母的身高是在人类平均高度上下y英寸，则他们的子女的身高是在人类平均高度上下$\frac{2}{3}y$英寸。他发现了一个规律，即子女的平均高度有回复到人类总平均高度的倾向，这就是著名的"回归法则"，虽然$\frac{2}{3}y$这个数值并未最后作出定论。回归这一名词最初用于对血缘关系的研究，现已成为统计上研究事物间相互关系的通用语。我们取符号r表示相关，正是起源于回归"regression"这个名词。

1890年，高尔登的学生皮尔逊(Karl Pearson)初次创用"积矩相关系数"(product-moment coefficient of correlation)。其后，这个方法广泛应用于各个领域。例如，1901年霍克尔(R. H. Hooker)用积矩相关系数研究结婚率同贸易之间的关系；俞尔(G. U. Yule)用此方法研究出生率与死亡率同对外贸易间的关系。

近年来，回归与相关分析方法广泛应用于生物学、心理学、教育学、经济学、医学等各个方面。尤其是应用多元回归进行经济预测，已在生产实践、科学管理和科学研究中取得了一定成效。例如，产量与成本可以用线性回归方程式表示它们之间的关系，按照计划成本的要求达到控制一定数量的产量。铁路运输量的多少与工农业产值有密切的关系，应用多元回归分析，可以根据一定时期的工农业总产值预测运输量，作为运输部门进行计划调度的依据。应用多元线性回归，又可以对农业产量进行预测。实践证明，应用这种方法可以达

到预期的效果。回归和相关分析方法将随着国民经济的发展而得到更广泛的应用。

回归分析(regression analysis)与相关分析(correlation analysis)均为研究及测度两个或两个以上变量间关系的方法。如果研究两个变量间关系，则称为简单回归与简单相关分析；如果研究两个以上变量间的关系，则称为多元回归与多元相关分析(multiple regression and correlation)，不论变量的多少，我们在进行分析时，必须选择其中之一作为因变量(dependent variable)，而把其余的变量当作自变量(independent variable)。

回归分析是研究自变量与因变量之间的关系形式的分析方法。由回归分析求出的关系式，称为回归方程式(regression equation)。如果因变量为自变量的一次函数，则称为线性回归方程式(linear regression equation)，否则称为非线性回归方程式(nonlinear regression equation)。

相关分析是测度各个变量之间的关系密切程度的方法，它用指标数值表明变量之间关系的密切程度。可以按照关系变化的形态分为线性相关分析和非线性相关分析。在每种相关分析中，所计算的相关指标其名称又各不相同。

应用回归分析和相关分析，可以研究下列问题：

一、通过回归分析，观察变量之间是否有一定的联系。如存在着联系，选择合适的数学模式对变量之间的联系给以近似描述。

二、用统计指标说明变量之间关系的密切程度。这些统计指标还可以用来说明回归方程对观察值的拟合程度的好坏。

三、根据样本资料求得的现象之间的联系形式和密切程度，推断总体中现象之间的联系形式和密切程度。

四、根据自变量的数值，预测或控制因变量的数值，并应用统计推断方法，估计预测数值的可靠程度。

第一章　一元线性回归

一元线性回归是指一个因变量只与一个自变量有依从关系，它们之间关系的形态表现为具有直线趋势。在分析时，首先可以作散点图以判定变量之间的关系是否直线型的。如果是直线型的，再配合回归直线来表达变量间的平均变化关系。

第一节　一元线性回归的测定方法

若已知变量X与变量Y之间存在着某种相互关系。为了研究它们之间的关系，一个最简单的方法是作图。我们将X作为自变量，Y作为因变量，每一组的数值在图中以一个点表示，这个图称为散点图或相关图。从散点图中可以看出两个变量之间的大致关系。今以后面例 1-1 中之资料（参见第九页）绘成图 1-1 说明之。

根据经验，平均每亩使用肥料量和蔬菜产量之间有一定的联系。假定历年所施氮肥量与蔬菜产量的实际资料如表 1-1。我们将使用氮肥量作为自变量X，蔬菜产量作为因变量Y，然后根据实际资料先作散点图，确定每一组数值在图中的位置；再把各点依次连接起来，就如图 1-2 所示。从图中可以看出，在一定区间内它们之间大致成直线趋势，故可以配合一回归直线如图中虚线所示来表述两变量间的平均变化关系。

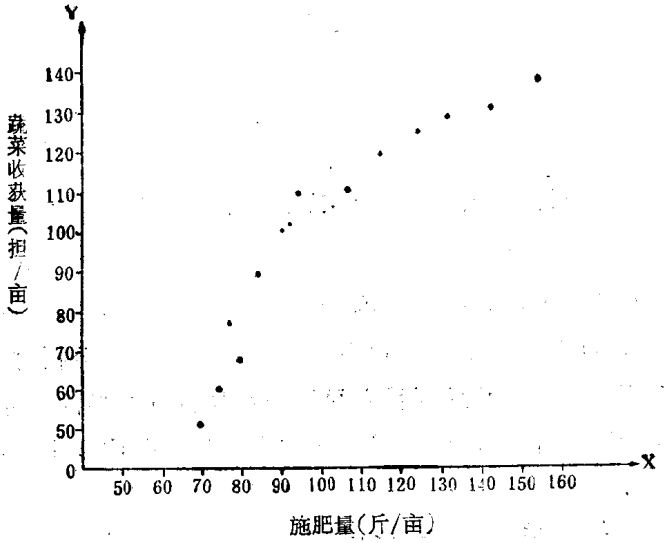

图 1-1　施肥量和蔬菜产量散点图

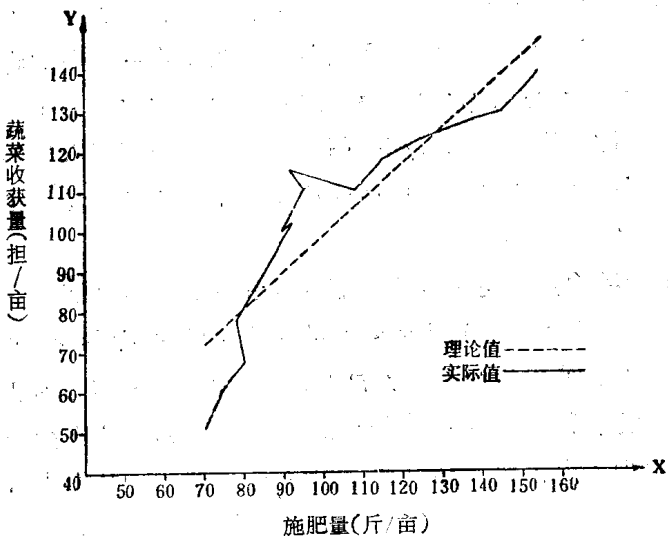

图 1-2　施肥量和蔬菜产量的回归直线

直线方程式为：

$$Y=a+bX \tag{1}$$

上式称为蔬菜产量Y对使用氮肥量X的回归直线，也称为Y对X的回归方程。回归直线的斜率b称为回归系数，它表示当X增加一个单位时Y的平均增加量，说明存在回归关系的两个变量间的数量关系。式中a为直线方程中的常数项，即当$X=0$时，$Y=a$，故a为$X=0$时直线在Y轴上的截距。根据观察资料确定了a、b之值，直线方程式也就确定了。a，b又称为参数。

下面，我们讲回归方程式的确定方法。

根据上面的散点图，我们可以作出很多条直线表示两个变量之间的关系。有些直线离散点近些，用它来表示X与Y之间的关系，与实际情况较为接近，而有些直线则可能离散点较远，用它来表示两个变量之间的关系，与实际情况相差较大。我们所要求的回归直线则是在一切直线中最接近实际资料的。也就是说，以这条直线来表示变量X与Y的关系与实际资料的误差比任何其他直线为小。

若用(X_i, Y_i)表示n组观察资料，任何一条直线的方程式为：

$$\hat{Y}_i=a+bX_i \tag{2}$$

根据上述方程式，由每一个观察资料X_i的数值即可以求得相应的Y_i的数值，这些数值我们称为理论数值(或估计值)，实际值与理论值之间存在着误差，设误差为δ_i，则

$$\delta_i=Y_i-\hat{Y}_i=Y_i-a-bX_i \tag{3}$$

而n个观察值所引起的误差的总和组成总误差，通常我们是应用最小二乘法原则使总误差的平方和为最小。

设以Q代表误差的平方总和，则

$$Q=\Sigma\delta_i^2=\Sigma(Y_i-a-bX_i)^2 \tag{4}$$

根据数学分析中求极值的原理，要使Q为最小，只需在(4)式中分别对a，b求偏导数，并令其等于零即可。

$$\frac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}(Y_i-a-bX_i)=0 \tag{5}$$

$$\frac{\partial Q}{\partial b}=-2\Sigma X_i(Y_i-a-bX_i)=0 \tag{6}$$

(5)(6)两式可以改写为：

$$\left.\begin{aligned} na+b\Sigma X&=\Sigma Y\\ a\Sigma X+b\Sigma X^2&=\Sigma XY\end{aligned}\right\}$$

以上两个方程式称为规范方程式(normal equations)。根据两个规范方程式求得a，b数值，以此代入(1)式即为所求的线性回归方程式。

方程式中a，b之值计算如下：

$$a=\frac{\Sigma Y}{n}-b\frac{\Sigma X}{n}=\overline{Y}-b\overline{X} \tag{7}$$

$$b=\frac{\Sigma XY-\frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2-\frac{(\Sigma X)^2}{n}}=\frac{\Sigma XY-n\overline{XY}}{\Sigma X^2-n\overline{X}^2}$$

$$=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2} \tag{8}$$

上式b值即直线的斜率，b的符号取决于分子$(X-\overline{X})$的数值和$(Y-\overline{Y})$的数值之乘积，当Y随X的增加而增加时，$b>0$，b为正号，表明两个变量的变化方向相同；当Y随X的增加而减少时，$b<0$，b为负号，表明两个变量的变化方向相反。

求回归方程式的具体计算步骤通常是列表进行的。

〔例 1-1〕 现以施肥量和蔬菜产量为例，列表说明如下（见表 1-1）：

表 1-1　　施肥量和蔬菜产量的回归计算表

年份	施肥量（斤/亩）X	收获量（担/亩）Y	X^2	Y^2	XY	$\hat{Y}$
1950	70	51	4,900	2,601	3,570	72.35
1952	74	60	5,476	3,600	4,440	75.92
1954	80	68	6,400	4,624	5,440	81.27
1956	78	78	6,084	6,084	6,084	79.49
1958	85	90	7,225	8,100	7,650	85.72
1960	92	102	8,464	10,404	9,384	91.96
1962	90	100	8,100	10,000	9,000	90.18
1964	95	110	9,025	12,100	10,450	94.64
1966	92	115	8,464	13,225	10,580	91.96
1968	108	110	11,664	12,100	11,880	106.22
1970	115	118	13,225	13,924	13,570	112.46
1972	123	122	15,129	14,884	15,006	119.59
1974	130	125	16,900	15,625	16,250	125.83
1976	138	128	19,044	16,384	17,664	132.96
1978	145	130	21,025	16,900	18,850	139.20
1980	154	140	23,716	19,600	21,560	147.22
合计	1,669	1,647	184,841	180,155	181,378	1,647

$$b=\frac{\Sigma XY-\frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2-\frac{(\Sigma X)^2}{n}}=\frac{181,378-\frac{1669\times 1647}{16}}{184,841-\frac{(1669)^2}{16}}=0.8913$$

$$a=\frac{\Sigma Y}{n}-b\frac{\Sigma X}{n}=\frac{1647}{16}-0.8913\times\frac{1669}{16}=9.9638$$

$$Y=a+bX=9.9638+0.8913X$$

式中 a 是常数项，是估计的固定蔬菜产量；b 表示当施肥量增加 1 斤时蔬菜平均增长量。

由于Y相应于X的估计量，可以用$\hat{Y}$表示之，上式又可以改写成：

$$\hat{Y}=a+bX=9.9638+0.8913X$$

根据上述方程式可以进行预测。例如，我们假定1982年的施肥量是每亩160斤，则预测每亩蔬菜平均收获量是：

$$\hat{Y}=9.9638+0.8913(160)=152.5718(担)$$

在进行预测时应注意以下两点：

一、假定其他条件不变；

二、如果继续大量施肥，产量达到一定高度后就会下降。这时不能用直线说明两者的关系，而要用曲线来描述了。

以上是以X为自变量，Y为因变量。有时也可以Y为自变量，X为因变量而求出另一条回归直线，可以下式表示之：

$$\hat{X}=a'+b'Y \tag{9}$$

$$a'=\overline{X}-b'\overline{Y}=\frac{\Sigma X}{n}-b'\frac{\Sigma Y}{n} \tag{10}$$

$$b'=\frac{\Sigma XY-n\overline{X}\overline{Y}}{\Sigma Y^2-n\overline{Y}^2}=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(Y-\overline{Y})^2} \tag{11}$$

为区别起见，b称为Y对X的回归系数，b'称为X对Y的回归系数。$\hat{Y}=a+bX$称为Y对X的回归直线，$\hat{X}=a'+b'Y$称为X对Y的回归直线，这个方程式是由观察值Y来估计X值的。

由上例所计算的回归方程式中，回归系数b为正数，表明蔬菜产量与施肥量的变化方向相同。

现再举例1-2说明两个变量之间的变化方向相反。

〔例1-2〕 由表1-2资料，生产量与单位产品成本之间成反方向变动：当生产量增加时，单位产品成本就降低；生产量减少时，单位产品成本就增加。这时回归方程式中的回归

表 1-2　　某工厂产量和实际成本的回归计算

月　份	产量（件）X	每件实际成本（元）Y	X^2	XY	Y^2
1	55	72	3,025	3,960	5,184
2	53	74	2,809	3,922	5,476
3	71	68	5,041	4,828	4,624
4	81	67	6,561	5,427	4,489
5	86	69	7,396	5,934	4,761
6	82	67	6,724	5,494	4,489
7	84	68	7,056	5,712	4,624
8	92	64	8,464	5,888	4,096
合　计	604	549	47,076	41,165	37,743

系数 b 为负数。由例中可知，当生产量增加 1 件时，平均单位产品成本降低 0.193 元。当产量增加至 100 件时，预计平均单位成本为 63.89 元。

$$Y=a+bX$$

$$b=\frac{n\Sigma XY-\Sigma Y\Sigma X}{n\Sigma X^2-(\Sigma X)^2}=\frac{8\times 41,165-549\times 604}{8\times 47,076-(604)^2}$$

$$=\frac{329,320-331,596}{376,608-364,816}=\frac{-2276}{11,792}=-0.1930$$

$$a=\frac{\Sigma Y}{n}-b\frac{\Sigma X}{n}=\frac{549}{8}+0.1930\times\frac{604}{8}=68.625+14.5715$$

$$=83.1965$$

$$\hat{Y}=83.1965-0.1930X$$

第二节　一元线性回归方程的简化计算

当实际资料的数字较大时，可以简化计算方法，将原来的资料加（减）一个常数或乘（除）一个常数，然后将变换后的资