数字信号处理的MATLAB实现

万永革 ◎ 编著

（第二版）

科学出版社

数字信号处理的 MATLAB 实现

（第二版）

万永革 编著

科学出版社

北京

内 容 简 介

本书介绍了数字信号处理的基本概念、理论及其 MATLAB 实现，并给出具体应用实例。全书共分为 12 章。由于数字记录信号常有断记情况，需要补充完整才能进行信号处理，因此第 1 章介绍回归与插值，以便对断记的数字信号进行处理；第 2 章介绍振动与信号，为后续信号分析与处理打下基础；第 3 章介绍 Fourier 变换，以便对数字信号进行频谱分析；第 4 章介绍系统、z 变换，使读者对系统有较全面的认识；第 5 章介绍模拟滤波器设计；第 6 章和第 7 章分别介绍 IIR 滤波器和 FIR 滤波器设计；第 8 章介绍参数化建模；第 9 章介绍随机信号分析；第 10 章介绍采样率转换，以便对更长的数据进行分析处理；作为频域分析的推广，第 11 章介绍非平稳信号的时频分析；第 12 章针对数字信号处理的几个前沿问题进行简单介绍，以拓宽读者的知识面。

本书结合大量实例和程序阐述如何将数字信号处理知识用于实际问题中。本书适合于数理基础相对薄弱、着重提高动手能力的读者学习数字信号处理技术，也可作为本科生、研究生和工程技术人员学习数字信号处理的参考书。

图书在版编目(CIP)数据

数字信号处理的 MATLAB 实现 / 万永革编著. —2 版. —北京：科学出版社，2012. 5

ISBN 978-7-03-033990-4

Ⅰ. ①数… Ⅱ. ①万… Ⅲ. ①数字信号处理 - 计算机辅助计算 - 软件包 - 高等学校 - 教材 Ⅳ. ①TN911. 72

中国版本图书馆 CIP 数据核字(2012)第 061575 号

责任编辑：罗 吉 曾佳佳 / 责任校对：冯 琳

责任印制：赵德静 / 封面设计：许 瑞

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2007 年 4 月第 一 版 开本：B5(720 × 1000)

2012 年 5 月第 二 版 印张：32

2012 年 5 月第七次印刷 字数：620 000

定价：59. 00 元(含光盘)

(如有印装质量问题，我社负责调换)

第二版前言

数字信号处理是近年来发展最为迅猛的学科之一，并已在多个科学技术领域获得了极为广泛的应用。可以说，学习信号处理的理论、方法与应用已成为通信、电子、自动化、生物医学、地球物理等众多学科或专业工作人员的迫切需要。

拙作《数字信号处理的 MATLAB 实现》（第一版）于 2007 年出版，在短短 5 年里已先后印刷 6 次，被学术期刊、硕博士论文引用上百次。为适应信号处理领域的发展，我们保留本书侧重于实践操作的特点，进行了重大修改：①由于在实际工作中的信号经常有断记的情况，为采用通常的数字信号处理技术对这些信号进行处理，需要对数据进行插值，而插值又需要理解最小二乘拟合，因此增加了“回归与插值”一章；②在数字信号处理中经常需要对采样率进行转换，因此增加了“采样率转换”一章；③由于时频分析是近年来应用较为广泛的分析方法，且是前面内容描述的推广，故增加了“时频分析”一章；④第 3 章增加了 Fourier 变换谱分析误差的内容；⑤第 4 章增加了 z 变换的理论描述；⑥第 6 章增加了特殊滤波器的阐述；⑦第7章增加了线性相位条件的理论阐述和频率采样法进行 FIR 滤波器设计的论述及实例；⑧由于自适应滤波技术是最新发展的滤波技术，在最后一章增加了自适应滤波技术举例；⑨各章增加了部分例题、习题和应用举例。

在本书的编写过程中，作者采纳了很多中国地震局台站培训人员的建议和意见，这些建议对于改进本书的可读性和易懂性起到了重要作用。

感谢陈运泰院士在百忙之中审阅了本书第一版并作序。本书的出版得到了国家自然科学基金（40874022、41074072）和中国地震局教师基金（20100101）的资助。

虽然作者努力而为，但本书可能仍然存在一些不足之处，在此恳请各位学术前辈、同仁和读者批评和指正。

万永革

2012 年 3 月

第一版序

数字信号处理是20世纪60年代中期随着数字电子计算机和大规模集成电路技术的不断发展,而迅速发展起来的一门新兴学科。近年来,数字信号处理的理论、算法及实现手段获得了飞速的发展,它已广泛应用于雷达、通信、语音、图像、地震、地质勘探、航空航天、生物医学工程等各个领域,并已成为这些领域的一种重要的现代化工具。

目前关于数字信号处理已有许多优秀的著作,但是"传统"的数字信号处理著作大多专注于算法的理论及其推导,较少涉及实现方法及其应用,目前结合具体软件作数字信号处理,并以此帮助读者澄清基本概念和理解理论进而掌握处理技术的著作还很少。近年来功能强大、交互性好的MATLAB软件的引入使得通过实践掌握数字信号处理技术的学习方式变得非常容易。

《数字信号处理的MATLAB实现》是作者结合数字信号处理理论和MATLAB操作技术提供给读者的一本实践性很强的工具书。这本书在介绍数字信号处理基本原理的同时,非常重视信号处理的实现问题,对所有例题都给出了具体实现的MATLAB程序,读者通过上机实践可以形象生动地加深对理论问题的理解。把理论与仿真实验结合在一起,既突出了理论的物理概念,又使读者能在实践中掌握数字信号处理的基本概念、基本方法和基本应用,达到学以致用的目的,起到事半功倍的效果。

作者结合他在地球物理领域的研究,编制了采用数字信号处理技术处理地球物理信号的大量实例,如地震波与地形变数据的频谱分析与滤波、地球自由振荡振型的提取、地面运动的恢复与地震仪仿真、运用小波方法分析地震活动的周期性等。读者只要对这些实例做适当修改即可应用于解决其感兴趣的问题。因此,这本书是从事地球物理信号分析人员与台站观测人员的一本非常实用的参考书。

当然,由于这本书侧重于数字信号处理技术的应用,相应地淡化了数字信号处理理论。不过,对于希望更多地了解信号处理理论的读者可以通过参阅目前已出版的浩若瀚海的有关数字信号处理理论的参考书予以弥补,而对于注重信号处理技术应用的读者,这本书不失为一本很有裨益的参考书。我衷心地祝贺《数字信号处理的MATLAB实现》出版并很高兴地将它推荐给读者,希望它能成为读者学习和应用数字信号处理技术的一件得心应手的工具。

陈运泰

中国科学院院士、发展中国家科学院院士

第一版前言

数字信号处理课程涉及较深的数学功底，其内容以 Fourier 变换、Laplace 变换、z 变换、复变函数的环路积分为数学基础，这些内容对数学基础比较薄弱的读者来说掌握起来有一定困难。为了使读者既能掌握足够的数字信号处理技能，又不致陷入烦琐的数学推导之中，真正体现以“必须够用为度”的原则，作者认为以大量的数字信号处理实例训练他们数字信号处理的技能是一种可行的思路。这样就避免了“学院式”的理论推导，重点放在数字信号处理方法和技能的掌握。这本小册子采用 MATLAB 作为我们的数字信号处理实验室对信号处理技能进行训练。之所以选择 MATLAB 作为我们的数字信号处理实验室，是因为 MATLAB 是一种面向科学和工程计算的高级语言，现在已成为国际公认的最优秀的科技应用软件，在世界范围内广为流传和使用。该软件的特点是：强大的计算功能、计算结果和编程可视化一体及较高的编程效率。这是该语言无与伦比之处。MATLAB 汲取了当今信号处理的小波变换、神经网络等一系列最新研究成果，已经成为从事科学研究和工程设计不可缺少的工具软件。今天在欧美高等院校内，使用 MATLAB 已成为大学生、研究生和教师必备的基本技能，MATLAB 广泛应用于科学研究、工程计算、教学等方面。

本书力求达到的特点是：理论知识以必须够用为度，力求使读者从实验中总结信号处理的基本概念和规律、尽量避免烦琐的数学推导；知识讲授和实习演练紧密衔接；例题中给出的 MATLAB 原码粘贴到 MATLAB 命令窗口，即可运行并看到结果，因此适用于多媒体教学；另外，课后备有作业题以供读者上机实习之用。

作者一直担心，这样一本数字信号处理讲稿的付印是否有些为时过早。确实目前运用 MATLAB 处理数字信号的教学研究正方兴未艾，然而为大多数人所接受的传授知识的方式和实例的形成通常需要较长时间的“沉淀”。实际上，对已有的运用 MATLAB 进行数字信号处理的实例加以总结和系统化也是一项重要的、艰苦的、富有创造性的工作。以作者的学识水平和能力要胜任这项工作几乎是不可能的。但面对目前技术应用类数字信号处理教材的奇缺使得作者不得不冒着风险将这份讲稿呈献给读者。由于本人水平有限，加之时间仓促，“突击”出来的这份教材肯定存在很多问题，殷切希望同行和广大读者提出宝贵意见。

欲掌握数字信号处理技术，应该与实际工作结合，这正是职业技术教育的最终目标。因此本书采用了作者所熟悉的地震学和地球物理学中的实例来展示信号处

理的技巧和效果。不过这些例子很容易移植到其他学科中去。

本书的编写得到防灾科技学院防灾技术系孟晓春主任的鼓励和支持。全国地震台台长及岗位培训班的学员们给予了大力支持，本书很多实例原型来自于这些培训班的实训，武晔和韩秋莹在文字编辑上提供了帮助，谨向他们表示衷心感谢。虽然本书所述及的数字信号处理技术为较成熟的技术，然而，本书所设计的程序是作者费了一定心血编出来的，如用到这些程序，请注明引自本书。

感谢陈运泰院士在百忙之中审阅本书原稿并作序。本书的出版得到国家重点基础研究发展规划项目(2001CB711005)和国家自然科学基金(40374012)的资助。

万永革*
2006 年 6 月 25 日

* E-mail:wanyg 217217@ vip. sina. com。

目　　录

第1章　回归与插值

1.1　回归分析

人们在从事生产、科学研究的过程中积累了大量的观测数据。在这些数据资料中，隐含着许多事物本身发生、发展的规律和各种事物之间的相互关系，如果仅以直观的或所谓“看图识字”的方法来考察这些数据，往往只能给人以模糊不清的印象或似是而非的感觉，最多也只能得到定性的认识，这对我们的研究目的来说是很不够的。为了深化我们对客观事物的规律性认识，就必须做到“心中有数”，也就是要想尽办法从定性的认识上升到定量的认识。具体地说，就是要从我们所掌握的科学数据中，通过数据处理的方法寻找事物发展变化的定量的规律或事物之间的定量的关系，即把它们理论化、函数化，以便利用这些规律更好地理解更多的问题。

现在我们讨论事物之间关系的定量处理方法，也就是变量之间关系的确定。

1.1.1　统计相关

变量之间的关系可以分为两种：一种是函数关系，另一种叫做相关关系，又叫统计相关。本节只对相关关系作详细的讨论。

设两个变量，当我们对其中一个变量给定一个数值时，另一个变量不能肯定是某个数值，但总的来说，其中一个变量变化时，另一个变量也大体遵循一定的规律而变化。或者说，在两个随机变量 x 和 y 之间，如果自变量 x 变动时，因变量 y 的平均值也跟着按一定的规律而变动，但二者之间并无确定的关系，这时我们称随机变量 y 和 x 之间存在着相关关系。在日常生活和生产中，两个变量存在相关关系的情况是极其普遍的，例如暴雨量和河流洪水量、相邻两个气象站的气温、儿童的年龄和身高、农作物的施肥量和亩产量等。在地震预测研究中也普遍存在着这种相关关系。如在一次大地震发生前，一般会产生地面的倾斜，但用倾斜仪或水准测量、三角测量、GPS 等方法测量出来的数值却包含许多非地震因素引起的变化，例如固体潮、日照、温度的影响，仪器本身的不稳定性及偶然误差等。总之，两个变量之间不存在绝对的函数关系，而是存在一定紧密程度的相关关系，造成这种情况的原因是，当自变量对因变量施加影响时还有许多因素也同时在对因变量施加影响，其中也包含了偶然误差的影响。

我们还注意到，在具有相关关系的两个现象之间，有时关系比较密切，有时关系不太密切，因此，从现象之间相关关系的紧密程度来说，相关关系有三种情况：

(1)完全相关。当 x 取已确定的某个数值时,y 只能按特定的线性关系确定,即 y 严格地随 x 的变化而变化,而不受其他因素的影响,这叫完全相关。这时,x 与 y 之间就是函数关系,可以表示成 $y=f(x)$ 的形式。将 y 和 x 的数值点绘于坐标图中,能配上一条完全通过各点的曲线或直线。这是相关关系中最紧密相关的表现。从这个意义上说,函数形式是相关现象的一种特例,当自变量和因变量完全相关时,其关系就退化为函数形式。

(2)零相关。这代表两种现象的特征值 y 和 x 之间不存在任何关系,x 的变化不影响 y 的取值。这是与完全相关截然相反的另一种情况,也是相关关系中的另一个极端形式。实际上,它表示了 x 和 y 相互独立,互不影响。

(3)统计相关。现象之间既不相互独立又不完全相关的情况都属于统计相关,它是介于完全相关和零相关之间的一种形式。若把这种情况的 y 和 x 的数值点绘于坐标图中,可以发现点的分布具有某种趋势,即 y 的取值随 x 的取值而出现趋势性的变化,这种趋势性的变化可以通过一条适当的曲线或直线来拟合。

在存在相关(包括完全相关和统计相关)的情况中,大体上又可分为直线相关和曲线相关(又叫非直线相关)两种形式。在直线相关或曲线相关的某一段取值范围内,又有正相关或负相关两种情况。

另外,从相关因素的情况来说,还可以将相关关系分为三种情况:

(1)简单相关。如果某一现象 y 只与 x 存在相关关系,而与其他现象无任何关系,则称 y 和 x 的关系为简单相关,研究这种关系时只用到概率论中的二元分布。

(2)复相关。如果某一现象 y 不仅与另一现象 x_1 有关,而且还与其他一些现象 x_2,x_3,…有关,这种多个现象统计相关的情况称为复相关。解决复相关的问题需要用到多元分布的知识。复相关的情况在自然界是普遍存在的。但是,在与因变量存在相关关系的诸多因素,也就是在若干个自变量中,各个自变量与因变量的相关的紧密程度是各不相同的。从相关的紧密程度对因变量影响的大小来说,有些较大,有些较小。为了不使数学处理过于烦琐,又能达到我们的目的,经常着重分析那些影响大的或者说起主要作用的相关关系,而对影响小的、比较次要的相关关系予以忽略。

(3)偏相关。在存在复相关的情况下,为了某种研究目的,只去研究其中的一个自变量对因变量的关系,而把其他自变量看做不变,这种相关关系叫做偏相关。在研究工作中,偏相关的分析方法是经常使用的。有时,要在数据较多的情况下采用逐个分析偏相关的情况来最后得到复相关的全貌。

1.1.2　回归方程

对于两个或多个存在着统计相关的随机变量,可以根据大量有关的观测数据来确定它们之间统计的定量关系,即求出一定的数学公式来表达这些关系,这些公式叫做回归方程或经验公式。这种数学处理过程叫做拟合过程。专门研究如何得

到合适的回归方程及应用这些公式来分析处理有关问题的知识叫回归分析，在不太严格的情况下，也可以叫做相关分析。

假定变量 y 与 x 存在相关关系，并按照 $y=f(x)$ 的函数形式相关。同时，由实际观测分别得到 y 和 x 相互对应的 N 组数值，这些数值亦可在坐标系中绘出 N 个点的散点图。现在的问题是如何根据这些实测的数值和点，用最合理的数学形式来表示或拟合 y 和 x 的相关关系，以及求出 $f(x)$ 的具体数学形式。

一般来说，任何一条曲线都可以用某一个高次方的代数多项式 $y=a+bx+cx^2+dx^3+\cdots$ 来描述，其中 $a,b,\cdots$ 是常数，可以根据曲线上各点所对应的 y 和 x 的数值，通过求解方程组的方法来求得。

显然，根据代数学的知识，对于具有 N 个实测点（即 N 组实测数据）的散点图，可以配置出 $N-1$ 次方的代数多项式，它的曲线将是通过每个点的扭曲线。但是，这样做不仅计算烦琐，也没有什么意义，而且也是不合理的。因为观测点起伏不定的情况是由于其他随机因素的影响和偶然误差所致，而非 y 和 x 之间的关系真的如此复杂，故这样的变化曲线并不能反映 y 和 x 的真实的相关关系，而点的分布的趋势变化才是二者相关关系的合理解释。为了获得能合理表示这种趋势的曲线，通常采用“最小二乘法”。这种方法可以在已知 $y=f(x)$ 的形式时计算出它的参数，使具体的函数所代表的曲线与实测点的趋势最好地吻合。

回归方程的求解包括两个内容：①回归方程数学形式的确定；②回归方程中所含参数的估计。

一般情况下，对回归方程数学形式的类型（如直线型、抛物线型、双曲线型和对数型等）的确定，主要根据实测点在坐标中的分布趋势，有时也可以根据两个变量之间的物理关系来确定，而且必须事先确定好。对于参数的估计，则正是回归分析的主要任务。

在实际工作中，当取得了某一物理量的一系列观测数据以后，常常还不知道这个物理量究竟与什么因素存在相关关系。因此，第一步的工作往往是通过大量的实践来寻找这种关系，这是一件检验性比较强的工作。这一步工作做得好才能真正得到有用的结果。在寻找相关关系的过程中，除了根据经验和实测数据进行分析外，还要注意一个问题，即自然界的各种现象之间的关系是极其复杂的，有些现象之间本来并没有什么本质上的因果关系，但在时间或空间上也可能出现某种程度的数值上的巧合，这时如果硬要把它们“扯”在一起进行相关关系的分析，就有可能得到毫无意义的甚至是荒谬的结论。因此，应适当地考虑我们所研究的现象之间，它们的影响条件和性质是否类同、是否存在物理上的某种可能的联系等。虽然这方面的工作难度较大，也许要牵涉到较为深广的理论问题，但作为研究工作者来说，仍应在这些方面多做努力。不过，反过来，由于人们对事物的认识总是有限的，我们进行相关关系的研究也正是为了揭示那些尚未被人们认识的规律和本质。因此，在实际工作中，又不能被现有的知识所束缚。在这些方面，没有明确、具体的

界限，需要我们在不断积累资料、丰富知识的基础上，辩证地、实事求是地深入研究分析这些问题。

利用回归分析所得到的回归方程，一方面可以作为依据，根据一个或几个变量的变化来预测或控制另一个变量的取值范围。例如，在地震预测中，如果已发现某一种前兆手段的观测量与地震存在相关关系，且已求出观测量的异常值与地震三要素（即时间、地点和震级）之间的相关公式（即经验公式），则根据前兆异常的大小，可以按经验公式提出较为准确的预测意见；另一方面，经验公式也可以是某种干扰因素与观测值之间统计的定量关系。例如，由大量资料得到了某一地应力元件的观测值与温度的经验关系，则可以根据每次观测中的温度值按此公式来对观测值作温度干扰的校正。从目的上来说，上述两种情况是不同的，但从数学处理的方法上来说则是完全相同的。

下面主要讨论两个随机变量的回归分析。

1.1.3　一条直线的最佳配置——最小二乘法求直线回归方程

根据误差理论知道，算术平均值是最佳值，由此算出的均方误差最小，该值出现于概率最大处。假定已经给散点图配置了一条直线：$y' = a + bx$，则 y_i' 是相应于 x_i 的回归直线的理论值，而 y_i 则是相应于 x_i 的实测值。则称 $v_i = \Delta y_i = y_i - y_i'$ 为回归直线的残差，v_i 实际上就是 y_i 这一点相对于回归直线的误差。这样，把求最佳值（即算术平均值）的原理用于求解回归直线方程，就能得到最佳配置的直线。

于是，得到求解最佳回归方程的原则是：由这条直线的方程与全部测定值所计算的残差平方和为最小，即

$$\sum v_i^2 = \min$$

若有 N 对测定值 (x_i, y_i)，$i = 1, 2, \cdots, N$，则可以列出 N 个残差方程。将每一个方程的两边取平方，然后左右两边分别取总和，即得到求解最佳回归方程的条件

$$\sum (y_i - bx_i - a)^2 = \sum v_i^2 = \min \tag{1-1}$$

由于这个条件中存在平方运算，故又称为二乘运算，所以这种方法叫做最小二乘法。满足这个条件的直线就是最佳配置的回归直线。从概率分布的角度来说，直线上各点出现的概率是最大的，这些点所代表的数值即为回归的最佳值或最可信赖值。

为了满足上式的条件，根据微分学的知识知道，必须使上式左边对 a 和 b 的偏微分分别为零，即

$$\begin{cases} \dfrac{\partial}{\partial b} \sum (y_i - bx_i - a)^2 = 0 \\ \dfrac{\partial}{\partial a} \sum (y_i - bx_i - a)^2 = 0 \end{cases} \tag{1-2}$$

微分后得到

$$\begin{cases} \sum (y_i - bx_i - a)x_i = 0 \\ \sum (y_i - bx_i - a) = 0 \end{cases} \tag{1-3}$$

分解开为

$$\begin{cases} \sum x_i y_i - b\sum x_i^2 - a\sum x_i = 0 \\ \sum y_i - b\sum x_i - Na = 0 \end{cases} \tag{1-4}$$

解上面的方程组为

$$\begin{cases} b = \dfrac{N\sum x_i y_i - (\sum y_i)(\sum x_i)}{N\sum x_i^2 - (\sum x_i)^2} \\ a = \dfrac{(\sum x_i^2)(\sum y_i) - (\sum x_i y_i)(\sum x_i)}{N\sum x_i^2 - (\sum x_i)^2} \end{cases} \tag{1-5}$$

这就是用最小二乘法求直线回归方程中系数 a,b 的最后结果。将这些系数带入直线方程即得到最佳配置的回归直线。结果中所有 $\sum$ 都表示从 1 ~ N 项的数值相加。计算中要注意,$\sum x_i^2$ 是将每个 x_i 值先平方后相加;$(\sum x_i)^2$ 则是将各个 x 值相加后得到的总和再平方。$\sum y_i$ 和 $\sum x_i$ 分别为 y 值和 x 值的全部数值的总和;$\sum x_i y_i$ 则表示每一对(x_i, y_i) 值相乘以后再将这些乘积相加。

实际上在求出系数 b 后,a 值不一定根据上式给出,根据上式可以推导出

$$a = \frac{1}{N}\sum y_i - \frac{b}{N}\sum x_i = \bar{y} - b\bar{x} \tag{1-6}$$

这样就可以简单地求出 a 值。

此外,对 b 值计算公式作进一步推演,可以得到

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \tag{1-7}$$

由此得到计算 b 值的另一种计算方式。

由上式还可以得到

$$\bar{y} = a + b\bar{x} \tag{1-8}$$

它表示回归直线理论值中,当 x 接近 $\bar{x}$ 时,y 接近 $\bar{y}$,即回归直线通过点$(\bar{x}, \bar{y})$处,它是由两个观测值的平均值所表示的点。这个点在力学上就是 N 个散点的重心位置。换句话说,回归直线必须通过 N 个散点的重心。

1.1.4　相关系数

在用上述方法求解回归方程的过程中,实际上并不需要事先假定两个变量之

间具有相关关系，也就是说，可以不加任何条件作为约束就能进行这种运算，即使散点图杂乱分布也总能给它配上一条直线，即求出 a 和 b 的值以表示 x 和 y 的关系。显然，在这种情况下，所配的直线是毫无意义的。那么，究竟在怎样的情况下所配的直线才有意义呢？或者说，如果两个随机变量具有相关关系，究竟相关的紧密程度有多大？大到什么程度时回归方程才算有效？所以，首先需要有一个表示相关系数的特征数，来定量地描述两个随机变量之间相关的紧密程度；其次是制定一种指标作为鉴定回归方程是否有效的标准。而下面所讲的相关系数 r 就是表示相关关系紧密程度的特征数。

为此，我们首先看直线回归方程中的回归系数 b。b 是方程中的斜率，其大小决定 y 随 x 变化的程度。如果 y 与 x 没有相关关系，即零相关，此时 y 将不随 x 的变化而变化，b 应为零，y 就是一条水平直线。于是

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 0 \tag{1-9}$$

如果反过来把 y 看成自变量，x 作为因变量，则可得到这种情况下的回归方程：$x = a' + b'y$，很容易得到

$$b' = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (y_i - \bar{y})^2} \tag{1-10}$$

由于相关关系是“相互的”，既然 x 不随 y 变化，y 也不随 x 变化，则 $b' = 0$，此时可以得到 x 和 y 不相关时的一种情况

$$bb' = \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} = 0 \tag{1-11}$$

相反，如果 x 与 y 全相关，即所有观测值的点都通过回归直线。显然 $y = a + bx$ 与 $x = a' + b'y$ 代表着同一条直线，直线斜率 b 和 b' 的关系互为倒数，由此得到

$$bb' = \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} = 1 \tag{1-12}$$

由此看到上面两式的数学形式是一样的，它的数值代表了两个变量之间相关的紧密程度。因此把这个式子的平方根定义为相关系数 r，即

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{1-13}$$

上式还可以表示为

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[N\sum x_i^2 - (\sum x_i)^2][N\sum y_i^2 - (\sum y_i)^2]}} \tag{1-14}$$

根据前面所述内容，可以得知 r 具有如下性质：

(1)由于 r 与 b 值和 b' 值的数学式中的分子相同，而它们的分母由于是平方项，因而肯定为非零的正值，故当 $r=0$ 时，$b=0$ 且有 $b'=0$，x 和 y 毫无线性相关关系。

(2)当 $0<|r|<1$ 时，y 与 x 存在一定的线性相关关系。当 $r>0$ 时，b 和 b' 均大于零，即 y 有随 x 的增大而增大的趋势。此即正相关。当 $r<0$ 时，b 或 b' 小于零，此即负相关。$|r|$ 值越大，相关关系越紧密，散点图的分布越集中，趋势越明显，反之亦然。

(3)$|r|=1$ 时，所有的点都落在回归直线上，即完全线性相关，其中 $r=1$ 为完全正相关，$r=-1$ 为完全负相关。这时 x 与 y 为确定的函数关系。

(4)相关系数与回归系数的关系为 $r=\sqrt{bb'}$，所以 r 值介于两个回归系数之间，是它们的几何平均值。

1.1.5　回归直线的误差和因变量取值的预测

假定已经求出了 y 与 x 之间的直线回归方程

$$y'=a+bx \tag{1-15}$$

式中，y' 代表的是预测值，而 y 代表观测值。(1-15)式是代表散点图趋势变化的一条直线。在绝大多数情况下，观测值的点不在直线上，而分布在直线的两旁。直线上每一点只能看成是取定某一 x 值时所对应的 y 的平均值。换句话说，实测的 y 值将在这条回归直线的两侧按一定的分布在波动。如果这种波动的原因是由随机因素和偶然因素所引起的话，它的分布就是正态分布。因此，可以根据误差理论用均方误差(即标准误差)的概念来作为描述这种波动的特征值。于是我们定义

$$S_{y'}=\sqrt{\frac{\sum(y_i-y_i')^2}{N-2}} \tag{1-16}$$

为剩余均方误差，或叫剩余标准误差。其中 y_i 为实测值，y_i' 为回归直线的理论值，(y_i-y_i') 又称回归直线的余差。$N-2$ 是自由度，可以这样理解：两个实测点只能作出一条通过这两个点的直线，这时不存在剩余均方误差，只有三个点才能开始作出回归直线，故此时的自由度为 $3-2=1$，对于 N 个实测点，自由度为 $N-2$，当 N 很大时可以近似为 N。$S_{y'}$ 就是表征回归直线误差大小的一个数值，它的大小也就表示了所作出的回归线的精度。

令 Q 为上式根号内的分子，并称它为剩余平方和或残差平方和，即

$$Q=\sum(y_i-y_i')^2 \tag{1-17}$$

将上式进行推演

$$Q=\sum[(y_i-\bar{y})-(y_i'-\bar{y})]^2=\sum(y_i-\bar{y})^2-2\sum(y_i-\bar{y})(y_i'-\bar{y})+\sum(y_i'-\bar{y})^2$$