

中国人民大学统计咨询研究中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心

联合推出

数 据 分 析 系 列 教 材

统计预测： 方法与应用 (第二版)

易丹辉 编著



 中国人民大学出版社

中国人民大学统计咨询研究中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心

联合推出

数 据 分 析 系 列 教 材

统计预测： 方法与应用

(第二版)

易丹辉 编著

中国人民大学出版社
· 北京 ·

图书在版编目 (CIP) 数据

统计预测：方法与应用/易丹辉编著. — 2 版. — 北京：中国人民大学出版社，2014. 6
数据分析系列教材
ISBN 978-7-300-19369-4

I. ①统… II. ①易… III. ①统计预测-高等学校-教材 IV. ①C8

中国版本图书馆 CIP 数据核字 (2014) 第 122744 号

数据分析系列教材
统计预测：方法与应用（第二版）
易丹辉 编著
Tongji Yuce: Fangfa yu Yingyong

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）		010－62511770（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	中煤涿州制图印刷厂北京分厂	版 次	1990 年 11 月第 1 版
规 格	185 mm×260 mm 16 开本		2014 年 8 月第 2 版
印 张	19 插页 1	印 次	2014 年 8 月第 1 次印刷
字 数	425 000	定 价	38.00 元

版权所有 侵权必究 印装差错 负责调换

数 据 分 析 系 列 教 材 编 委 会

编委会主任 易丹辉

编委会委员 (按姓氏笔画排序)

王惠文 吴喜之 张 波

易丹辉 柯惠新 耿 直

黄登源 谢邦昌

总 序

随着社会经济的不断发展、科学技术的不断进步，统计方法越来越成为人们必不可少的工具和手段。在多年教学过程中，老师们也越来越感到运用统计方法解决实际问题的重要，不少人在探索如何运用统计软件介绍和学习统计方法，思考如何运用这些方法解决实际问题。

自2008年数据分析系列教材出版以来，得到了不少读者的关注，这次丛书在原来的基础上，根据几位老师教学和科研的实践经验，重新策划完成。丛书特别邀请北京航空航天大学王惠文教授加盟，担任《描述统计》一书的撰写。她多年运用描述统计方法解决实际问题，积累了丰富的经验，将为读者正确运用描述统计方法提供参考。由于几位年轻老师不辞辛苦，这套数据分析系列教材在原来的基础上作了调整，更适合读者的实际需要。

在我们看来，掌握统计方法不仅要理论上弄明白，更重要的在于能够正确有效地运用这些方法，分析说明实际问题。这套丛书正是试图利用实际数据，通过统计软件的实际操作，对所能使用的统计方法加以说明，使读者不仅能够了解相应的统计方法，而且能够通过计算机操作学会运用这些方法处理分析实际数据。希望这套丛书的出版能够为读者提供这样学习的工具。

由于水平有限，难免有不足之处。恳请读者朋友们提出宝贵意见。我们也会循着这样的思路，在教学以及和读者的交流沟通中不断积累、不断提高、不断完善，奉献给读者更多、更好的成果。

感谢为这套丛书的编写付出汗水的研究生，感谢几位认真用心的年轻老师，感谢中国人民大学出版社的大力支持，特别是陈永凤编辑和王伟娟编辑，是她们的努力，让这套丛书在很短的时间里出版。为方便读者，书中的所有例题数据，都将放在中国人民大学出版社的网站上，供读者下载并练习。谢谢读者，希望在数据采集变得越来越容易、大数据时代到来的今天，能够加强沟通和联系，为提高统计方法实际运用的能力和水平共同努力。

易丹辉

第二版前言

随着社会经济的迅速发展和计算机技术更新加快，人们越来越重视并容易收集到实际发生的各种数据。如何运用这些数据认识事物发展变化的规律并预测其未来，也愈发为人们所关注，各种预测方法应运而生并不断发展。在从事教学和实际问题研究的过程中，感觉原来的内容有些过于陈旧，希望能够奉献一些新的思路和方法给学生和其他需要应用的读者，中国人民大学出版社也盛情约请，我也就决心下点工夫完成本书的修订。由于时间和精力有限，仍然没有能够达到预期的目标，只能看以后是否还有精力充实了。为方便操作，本书需要计算机完成的运算均运用的是 EViews 6.0 软件。

很感谢我的同事、我的学生，以及同仁们，在教学科研中给予的帮助和支持，特别感谢中国人民大学出版社愿意继续出版这本书。

第一版前言

在社会经济活动中，无论从宏观的角度还是从微观的角度，都存在着许多未知的因素，影响着各级的管理决策。为了克服未知因素可能带来的消极后果，必须进行有科学根据的预测。所谓预测，是人们在观察和分析客观事物发展过程的历史及现状的基础上，通过对客观事物发展规律的认识，进而推断其未来状况的过程。为了收到预期的预测效果，对于预测对象最好提出几种不同的预测方案，在各种方案中，充分衡量预测对象变化的条件以及可能变化的幅度，相应地采取有关措施，以便保持最佳的管理过程。换句话说，预测是在制定切实可行的计划时，为了避免可能产生的缺点和失误，而对事物的未来发展预先进行的多种方案的设计和研究。

预测可以按不同的标准进行分类。预测方法基本上分为两大类，即定性分析法和定量分析法。本书比较详尽地介绍了用于预测的定量分析方法：因果回归分析法和时间序列分析法。为了将每种具体方法与我国的社会经济实际相结合，在每一方法介绍之后，都配有实例说明其应用，书中所有计算均应用电子计算机完成。为帮助读者掌握和运用各种方法，特别是无法进行手工计算的方法，书后附有 TSP 软件的使用说明，它适用于 IBM—PC 机以及与它兼容的微型机，如长城 0520。介绍方法时，涉及到的比较复杂的数学公式推导和证明，均列入各章附录中，供读者参考。

本书编写的过程中，得到中国人民大学计划统计学院计算机室刘延军、陈虹同志，计划经济学系成晓梅同志以及校信息中心的同志们的帮助与支持。书中采用的某些实例，是我系袁卫同志在硕士研究生学习期间收集的资料，他为编写此书提出了不少建议。在此一并表示衷心的感谢。

本书试图将各种预测方法与我国的实际结合运用，由于水平有限，编写时间又较仓促，一定存在不少缺点，殷切期望读者们随时给予批评指教。

目 录

第一章 一元线性回归分析法	1
第一节 模型和参数估计	1
第二节 模型的检验	5
第三节 预测精度的测定	16
第四节 预测实例	21
附 录	28
第二章 多元回归分析法	32
第一节 模型和参数估计	32
第二节 模型的检验	36
第三节 自变量的选择	41
第四节 多重共线性	46
第五节 预测实例	51
第六节 滞后变量模型	53
附 录	59
第三章 非线性回归分析法	63
第一节 非线性回归模型	63
第二节 模型参数的估计	65
第三节 模型分析与评价	68
第四节 含虚拟变量的回归模型	73
第五节 预测实例	83
附 录	92

第四章 时间序列平滑法	94
第一节 概 述	94
第二节 移动平均法	95
第三节 指数平滑法	99
第四节 方法的比较	111
附 录	113
第五章 趋势模型	117
第一节 趋势模型类型	118
第二节 模型选择	122
第三节 参数估计	125
第四节 模型分析与评价	132
附 录	138
第六章 季节模型	141
第一节 季节性水平模型	142
第二节 季节性交乘趋向模型	147
第三节 季节性迭加趋向模型	155
第七章 马尔可夫法	161
第一节 基本概念	161
第二节 马尔可夫预测法	164
第三节 马氏链的稳定状态及其应用	173
第八章 ARMA 模型	177
第一节 概 述	177
第二节 时序特性的分析	182
第三节 ARMA 模型及其改进	190
第四节 随机时序模型的建立	197
第五节 时序模型预测	211
附 录	216
第九章 ARCH 类模型	222
第一节 单位根过程	222
第二节 ARCH 模型	237
第三节 广义 ARCH 模型	245
第四节 拓展的 ARCH 模型	249



第五节 多元 ARCH 模型	257
附 录	267
附表 1 t 分布表	271
附表 2 F 分布表	273
附表 3 D. W. 检验表	282
附表 4 χ^2 分布表	285
附表 5 DF 检验 t 统计量经验概率分布表	287
附表 6 Engle-Granger 检验表	288
参考文献	289

客观事物之间常存在着某种因果关系，如工业产品成本的降低常导致利润的上升；某种消费品价格的提高往往造成销售量的下降，等等。这种因果关系往往无法用精确的数学表达式描述，只有通过大量观察数据的统计处理，才能找到它们之间的关系和规律。回归分析就是通过对观察数据的统计分析和处理，研究与确定事物间相关关系和联系形式的方法。运用回归分析法寻找预测对象与影响因素之间的因果关系，建立回归模型进行预测的方法，称为因果回归分析法。其特点是，将影响预测对象的因素分解，在考察各个因素的变动中，估计预测对象未来的数量状况。它建立的是预测对象与影响因素之间的单一方程，因此也称为单方程模型分析。按方程中影响预测对象因素的多少，分为简单回归分析法和多重回归分析法。

回归分析法在预测中主要用以解决以下问题：

- (1) 分析所获得的统计数据，确定几个特定变量之间的数学关系形式，即建立回归模型。
- (2) 对回归模型的参数进行估计和统计检验，分析影响因素对预测对象的影响程度，确定预测模型。
- (3) 利用确定的回归模型和自变量的未来可能值，估计预测对象的未来可能值，并分析研究预测结果的误差范围及精度。

第一节 模型和参数估计

如果影响预测对象的主要因素只有一个，并且它们之间呈线性关系，那么可以采用简单线性回归分析法预测。由于这种方法只涉及一个自变量，因此也称为一元线性回归分析法。



一、模型形式

（一）理论回归模型

将预测对象作为因变量 Y ，主要影响因素为自变量 X ，它们之间的线性关系从理论上说，能够表述为式 (1.1) 的形式。

$$Y = \alpha + \beta X + \epsilon \quad (1.1)$$

式中， α 和 β 是固定但未知的参数，它们反映了变量 X 与 Y 之间应该有的一种线性关系； α 是常数项； β 是理论回归系数； ϵ 是那些除 X 以外，被忽略和（或）无法考虑到的因素，称为随机项。对于每一组可以观察到的因变量 Y_i 、自变量 X_i ，式 (1.1) 可以写成

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad (1.2)$$

式中， ϵ_i 服从正态分布并满足

$$\begin{aligned} E(\epsilon_i) &= 0 \\ \text{Cov}(\epsilon_i, \epsilon_j) &= \begin{cases} \sigma^2, & i=j \\ 0, & i \neq j \end{cases} \end{aligned} \quad (1.3)$$

这表明，模型的随机干扰项要求满足正态性、同方差性和独立性。

（二）实际回归模型

要得到式 (1.1) 中参数 α 和 β 的精确值几乎不可能，因为通常只有有限的样本数据和情报。利用有限的资料，只能得到参数 α 和 β 的估计值 a 和 b 。实际上，因变量 Y 和自变量 X 之间的简单线性关系能够表述为：

$$Y = a + bX + e \quad (1.4)$$

式中， a ， b 不是像 α ， β 那样取固定的数值，而是能够取多个数值的统计估计值； e 是残差项，也称为回归余项或回归残差，它是由用 $a + bX$ 估计因变量 Y 所造成的误差，是估计值与实际数值之间的离差。

相对于式 (1.2)，实际回归模型也可以写成

$$Y_i = a + bX_i + e_i \quad (1.5a)$$

式中， e_i 是 $a + bX_i$ 的估计值 $\hat{Y}_i$ 与实际观察值 Y_i 的离差，即

$$e_i = Y_i - \hat{Y}_i$$

$$\text{或} \quad Y_i = a + bX_i + e_i \quad (1.5b)$$

式中， $e_i = Y_i - \hat{Y}_i$ 。

（三）预测模型

实际预测时，残差项 e_i 是无法预测的，我们的目的是借助 $a + bX$ 得到预测对象 Y 的估计值，所以预测模型为：

$$\hat{Y} = a + bX \quad (1.6)$$

式中， a 为回归常数，是回归直线的截距。其实际含义为：不考虑自变量时，因变量

所能达到的数值。 b 为回归系数, 是回归直线的斜率。其实际含义为: 当自变量 X 每变动一个单位时因变量 Y 的平均变动量。

可以看出, 式 (1.6) $\hat{Y}=a+bX$ 实际上是式 (1.4) $Y=a+bX+e$ 的主体部分。相对于式 (1.5a), 式 (1.6) 可以写成

$$\hat{Y}_i=a+bX_i \quad (1.7a)$$

相对于式 (1.5b), 式 (1.6) 可以写成

$$\hat{Y}_t=a+bX_t \quad (1.7b)$$

注: 后面若无特殊说明, 对 i 成立的关系, 对 t 同样成立。

二、参数估计

1. 参数的最小二乘估计

对式 (1.1) 中的 α, β 进行估计, 依照不同的准则, 采用不同的统计方法, 可以得到不同的数值, 因而式 (1.4) 中的 a, b 不是唯一确定的。预测中, 通常采用最小平方方法 (Least Squares), 亦称最小二乘法。其准则是, 选择参数的估计值 a, b 使得因变量 Y 的观测值 Y_i 与估计值 $\hat{Y}_i$ 之间的离差平方和最小, 即 $\sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2 = \min$ 。^①

采用最小二乘法得到 a, b 的计算公式为:

$$\begin{cases} b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \\ a = \bar{y} - b\bar{x} \end{cases} \quad (1.8)$$

式中, x_i 为自变量 X 的第 i 个观测值; y_i 为因变量 Y 的第 i 个观测值; n 为观测值的个数, 亦称样本数据个数; $\bar{x}$ 为 n 个自变量观测值的平均数; $\bar{y}$ 为 n 个因变量观测值的平均数。

式 (1.8) 还可以写成

$$\begin{cases} b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases} \quad (1.9)$$



例 1—1

根据表 1—1 中的数据, 分析可能建立的预测国内生产总值中居民消费额的模型。

^① 有关参数 a, b 的确定参见本章附录 1—A。

表 1—1

我国国内生产总值和居民消费额

单位：亿元

年份	国内生产总值	居民消费额
1993	36 938.1	16 412.1
1994	50 217.4	21 844.2
1995	63 216.9	28 369.7
1996	74 163.6	33 955.9
1997	81 658.5	36 921.5
1998	86 531.6	39 229.3
1999	91 125	41 920.4
2000	98 749	45 854.6
2001	109 028	49 435.86
2002	120 475.6	53 056.57
2003	136 613.4	57 649.81
2004	160 956.6	65 218.48
2005	187 423.4	72 958.71
2006	222 712.5	82 575.45
2007	266 599.2	96 332.5
2008	315 974.6	111 670.4
2009	348 775.1	123 584.6
2010	402 816.5	140 758.6
2011	472 619.2	168 956.6
2012	529 238.4	190 423.8

说明：表中为现价支出法国内生产总值及居民消费额。
资料来源：中经网统计数据库。

分析：根据表 1—1 中的数据，绘制散点图，如图 1—1 所示。由图可以看出，两个变量基本呈线性关系，计算简单相关系数为 0.998 3，相关程度很高，可以考虑建立一元线性回归模型。运用式（1.9），得到

$b=0.335\ 5, \quad a=9\ 184.257$

我国国内生产总值中居民消费额的预测模型为：

$\hat{Y}=9\ 184.257+0.335\ 5X$

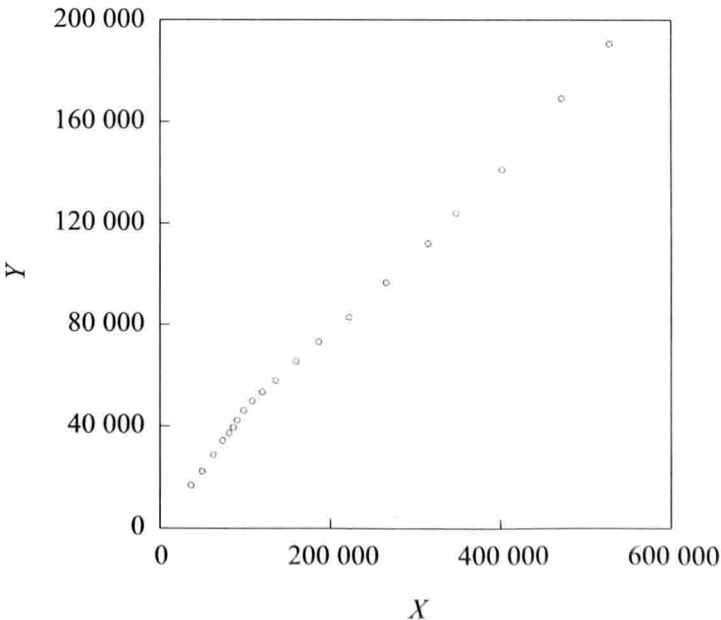


图 1—1 我国国内生产总值和居民消费额散点图

2. 参数的意义

最小二乘估计的参数,是根据求导运算得到的,因此回归系数 b 反映因变量与自变量之间的变动关系。回归系数 b 的大小表明自变量每变动一个单位所引起的因变量的平均变动量。若为正,表明随自变量增大因变量也增大,即同向变动;若为负,表明自变量与因变量反向变动。上面模型结果表明,在1993—2012年20年间,我国国内生产总值每增加1元,平均有近0.34元用于居民消费。这就是回归系数 b 所提供的经济意义。

模型中的参数,对不同的预测对象有不同的含义。参数估计值的符号和大小,要符合其实际意义。例1—1中当国内生产总值增加时,居民消费额一般也增加,因此 b 必须大于0,若得到的估计值 b 小于0,则模型不能被接受用于预测。 b 的变动范围是否适当,主要根据预测人员的经验确定。参数估计值的符号和大小不符合其实际含义,其主要原因可能是:所选用的模型不能代表变量之间的关系;统计数据不足或口径不一致;违反了最小二乘法的某些假定。鉴于模型参数的实际意义,参数估计后应先对参数的合理性进行判定。

预测模型中的回归系数 b ,反映了因变量 Y 和自变量 X 之间的一种变动结构关系。这种变动结构对未来是否合适,决定着模型能否用于预测。这一点是预测时应该予以考虑的。一般来说,回归模型适用于短期预测,因为短期内变量之间的结构关系可能会相对稳定。如果能够确认变量之间的结构关系能够延伸到未来较长时期,则回归模型也可以用于较长期的预测。

第二节 模型的检验

数理统计理论证明,采用最小二乘法得到的估计值 a, b 是 α, β 的最小方差无偏估计^①,它们是较为理想和实用的估计值。

在这一过程中,实际上是承认了几点假设:

- (1) 变量 X 是非随机的,即 X 与随机干扰项 ϵ 无关;
- (2) 变量 X 与 Y 之间为线性关系;
- (3) 随机干扰项服从期望为0,方差为 σ^2 的正态分布并相互独立,即满足正态性、独立性和同方差性。

利用变量的样本数据(实际观测值)建立预测模型后,需要判断所做的各种假设的合理性以及模型的优劣。模型检验就是利用各种统计检验来判别模型的适用性。

一、回归系数的显著性检验

对于预测模型 $\hat{Y}=a+bX$,变量 X, Y 之间的线性假设是否合理,可以通过回归

^① 有关证明参看参考文献[10]。



系数的显著性检验得到判别。回归系数的显著性检验由于要构建样本的 t 统计量，因而也称为参数的 t 检验。

检验假设

$$H_0: \beta=0^{①}$$

$$H_1: \beta \neq 0$$

计算样本的 t 统计量值：

$$t_b = \frac{b}{S_b} \sim t(n-2) \quad (1.10)$$

式中， S_b 是参数 β 估计时的标准差， $S_b = S_y / \sqrt{\sum (x - \bar{x})^2}$ ②，式 (1.10) 可以写成式 (1.11)。

$$t_b = b \cdot \sqrt{\sum (x - \bar{x})^2} / S_y \quad (1.11)$$

式中， S_y 为回归标准差， $S_y^2 = \sum (Y - \hat{Y}_i)^2 / (n-2)$ 。其中， n 是样本数据个数；2 是待估计参数个数。

$t_b = b/S_b$ 服从 t 分布，可以通过 t 分布表查得显著性水平为 α ③，自由度为 $n-2$ 的临界值 $t_{\alpha/2}$ 。将 t_b 与 $t_{\alpha/2}$ 比较，可决定是否拒绝 H_0 。若 $|t_b| > t_{\alpha/2}$ ，则拒绝 H_0 ，它表明回归系数显著不为 0，参数的 t 检验通过。回归系数显著，说明变量 X 与 Y 之间的线性假设合理，这意味着所选择的自变量能比较有效地解释预测对象的变化。若 $|t_b| \leq t_{\alpha/2}$ ，则不能拒绝 H_0 ，它表明回归系数为 0 的可能性比较大，参数的 t 检验未通过。回归系数不显著，说明变量 X 与 Y 之间的线性假设不合理，意味着模型中的自变量无法较好地解释预测对象的变化，应重新考虑。也可以利用概率 p 值做出是否拒绝 H_0 的判定。 p 值是在原假设成立的条件下，利用样本数据得到的回归系数为 0 即原假设为真的概率，若 p 值很小，如 $p < 0.05$ ，则表明利用样本数据得到的 H_0 为真的概率是一个小概率，故拒绝 H_0 ；反之不能拒绝。



例 1—2

(续例 1—1) 居民消费额预测模型中参数的 t 检验。

分析：利用表 1—1 中的数据得到参数 β 的估计值为 0.335 5，估计的标准差为 0.004 6， t 统计量的值为 74.049 3，若以显著性水平 5%，自由度 $n-2=20-2=18$ 查 t 分布表，则得到 $t_{\alpha/2}$ 为 2.101，显然， $|t_b| = 74.049 3 > t_{\alpha/2} = 2.101$ ，拒绝 H_0 。这表明回归系数显著不为 0，参数的 t 检验通过。回归系数显著，说明变量国内生产总值 (X) 与居民消费额 (Y) 之间的线性假设合理，这意味着，国内生产总值能比

① 检验回归常数 α 是否为 0 的意义不大，故通常只检验参数 β 。

② 为方便，本书中求和符号 $\sum$ 若无注明均表示 $\sum_{i=1}^n$ 或 $\sum_{t=1}^T$ 。

③ 预测时，只需检验 β 是否为 0，故为双侧假设检验。查 t 分布表时，应以显著水平为 $\alpha/2$ 查找。

较有效地解释居民消费额的变化。同时,也可以利用 p 值判定, $p=0.000\ 0$, 表明 H_0 为真的概率是一个极小的概率, 参数的 t 检验通过, 说明国内生产总值对消费额有显著影响。这一检验也可以根据经验做出判断。如果 $|t_b| > 2$, 则可以拒绝 H_0 。

二、回归方程的显著性检验

参数的 t 检验, 考察的是自变量 X 与因变量 Y 之间线性假设的合理性。但预测模型 $\hat{Y}=a+bX$ 作为一个整体, 在一定程度上也反映了变量 X 与 Y 之间的线性关系, 其是否适用于预测, 仍需检验。回归方程的显著性检验, 是利用方差分析的思想, 将因变量的总变异分解为由回归模型可以解释的变异和残差 (即无法解释的剩余变异), 通过构建两部分的方差比, 即 F 统计量, 检验预测模型的总体线性关系的显著性, 也称为方程的 F 检验。

因变量 Y 实际观察值的离散程度, 可以用各观察值 Y_i 对其平均值 $\bar{Y}$ 的离差平方和表示。每一个离差 $Y_i - \bar{Y}$ 受自变量 X 和一些未被考虑的随机因素的影响。图 1—2 是离差 $Y_i - \bar{Y}$ 的分解示意图。图中斜线是回归直线 $\hat{Y}=a+bX$, 横线是因变量实际观察值的平均值线。相对于自变量 X 的某一取值 X_i , 回归线与线段 $\bar{Y}$ 的交点是 $\hat{Y}_i$ 。由图 1—2 可知, 离差 $Y_i - \bar{Y}$ 可以分为两个部分, 即

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad (1.12)$$

(总离差) (剩余离差) (回归离差)

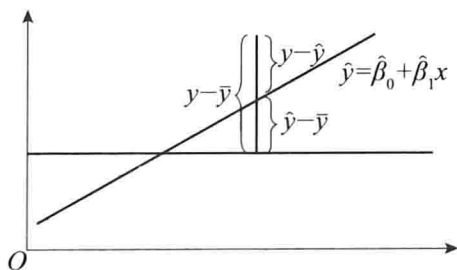


图 1—2 离差分解示意图

回归离差是由于选择自变量 X 并建立回归模型而产生的。它可以利用回归模型加以说明, 因此也称为已解释的离差。剩余离差是受那些未被考虑进模型中的随机因素的影响而产生的, 它无法由建立的回归模型说明, 因此称作未解释的离差。根据式 (1.12) 可以得到

$$\sum (Y_i - \bar{Y})^2 = \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2$$

能够证明^①式 (1.13) 成立。

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \quad (1.13)$$

(总平方和) (剩余平方和) (回归平方和)

^① 证明过程参见本章附录 1—C。