图书情报领域实用回归分析

周爱民 编著

TUSHU QINGBAO LINGYU SHIYONG HUIGUI FENXI

郑州大学出版社

图书情报领域实用回归分析

周爱民 编著

TUSHU QINGBAO LINGYU SHIYONG HUIGUI FENXI

郑州大学出版社
郑州

图书在版编目(CIP)数据

图书情报领域实用回归分析/周爱民编著. —郑州:郑州大学出版社,2015.1

ISBN 978-7-5645-1942-1

Ⅰ.①图… Ⅱ.①周… Ⅲ.①回归分析-应用-图书情报工作-研究 Ⅳ.①G250

中国版本图书馆 CIP 数据核字(2014)第 159228 号

郑州大学出版社出版发行

郑州市大学路 40 号　　邮政编码:450052

出版人:王　锋　　发行部电话:0371-66966070

全国新华书店经销

河南防伪保密印刷公司报刊印务分公司印制

开本:787 mm×1 092 mm　1/16

印张:24

字数:568 千字

版次:2015 年 1 月第 1 版　　印次:2015 年 1 月第 1 次印刷

书号:ISBN 978-7-5645-1942-1　　定价:42.00 元

作者简介

周爱民，山西省运城市人，1956 年元月生。毕业于吉林大学数学系力学专业，现为郑州大学图书馆副研究馆员，主要从事图书馆统计研究。参与编写了《文献检索课的理论与实践》（气象出版社）、《数字图书馆研究》（中国广播电视出版社）、《网络环境下知识组织研究》（中国广播电视出版社）等三部著作，发表论文 20 多篇。提出了大等级数据的洛特卡参数的估计法、布拉德福曲线的威布尔模型、小等级数据的洛特卡参数估计法、含零等级数据的洛特卡拓展模型等理论与方法。

序　言

我们知道，数学是科学之母。同样，在图书情报领域，相关数学方法和应用的研究可以为该学科注入科学元素，建立新的研究范式。回归分析是统计学中一个非常重要的分支，在社会科学领域也有着非常广泛的应用，在图书情报领域中的文献计量学、图书馆管理、信息行为研究、用户研究、信息经济等，是比较适合采用这种方法进行研究的。

周爱民同志有数学专业的学科背景，又在图书馆工作了二十余年，且勤于钻研，精于治学，具备良好的数学及图书馆学专业素养。他一直在探索用数学的方法和理论来研究解决图书馆学的相关问题，并相继在《图书情报工作》《数学的实践与认识》《情报杂志》《统计与决策》等核心期刊上发表了《含零等级数据的洛特卡拓展模型》《级数模型在文献计量中的应用》《几种布拉德福分散曲线拟合模型的实证比较》《文献老化对数正态分布模型》《图书馆基于读者满意度测量指标及回归模型研究》《三维对数线性等级完备模型在读者测量中的应用》《不确定性数据的交互分析方法》《基于偏最小二乘法的情报组合预测法》等有较高学术水平的论文。在此基础上，他对图书情报领域使用回归分析方法进行了系统、深入的研究，最终形成该书。该书既涵盖回归分析方法，又将方法与示范案例紧密相联，还结合一些实际调查数据，帮助读者对这些方法有更全面、更深入的了解。

时值该书出版之际，我表示衷心祝贺，相信该书的出版对于促进图书情报学领域的科学化探索有积极的意义。

崔波

2014 年 6 月 30 日

前言

数学是研究一切可以抽象出来的数量、形式、结构等之间相互关系的一门精密科学。科学史证明，当一个学科发展到一定阶段，若没有适当的数学方法提练，研究只能停留在经验阶段，经验阶段向理论高度发展的钥匙就是数学方法的进入，图书情报领域的研究也概莫能外。

图书馆学和情报学关心的不只是描述发生什么事情，更关心的是事情之间的联系和规律。为了反映文献信息的生成、分布、结构功能、管理利用的一般性规律，量化分析方法必不可少。

图书馆必须有固定的经费，以保证图书馆的可持续发展。但是如何分配经费、如何配置图书馆的设施、如何评价图书的质量及如何制定采编决策等，所有这些决策还在于隐藏其背后的决策有效性。寻找有效决策的过程就是数学分析过程。

图书馆学中有关文献的增长、老化、剔除及读者分布等规律性的结论就是在量化分析的基础上获得的。1944 年美国图书馆馆员弗里蒙特·赖德首先统计和研究了美国有代表性的大学图书馆的馆藏增长情况，他得出了图书馆的馆藏图书大约每 16 年增长一倍的结论。这为以后 D. 普赖斯发现科学文献的指数增长规律奠定了基础。如果没有量化分析方法，一些重要结论就无法利用定性分析而得出。可见量化分析方法能深化定性分析的结论。

图书情报领域应用的数学方法很多，本书仅讨论回归分析方法。回归分析方法也有许多，但有些方法，如正交回归、主成分回归、偏最小二乘回归、支持向量回归、核回归、半参数回归、分位数回归、稳健回归、变系数回归、方差分量模型、回归方程组等，没有包含在 IBM SPSS 软件中，尽管其中某些方法在图书情报领域有一定的应用，但无法讨论；而有些回归，如自回归（时间序列回归）、二阶段回归、有序回归、泊松回归、岭回归、最佳尺度回归，尽管包含在 IBM SPSS 软件中，但在图书情报领域应用较少，我们不对其进行讨论。本书主要讨论线性回归、约束线性回归、逐步回归、非线性回归、列联表、Logistic 回归、对数线性模型回归、Logit 回归等。

非线性回归的计算是本书的重点，文献计量学中的所有模型严格来说都是非线性回归模型，除了文献计量学模型外，还介绍了许多其他非线性模型，这样做的目的是拓宽读者的视野，利于以后的研究，从领域外的知识中汲取营养，拓

展本领域的研究是一种普遍的方法，在这方面作者略有体会，作者曾把研究可靠性的威布尔模型引进到文献分布的研究中，取得了较好的结果。今天领域外的模型说不定就是明天领域内的模型，出于此种考虑，本书也介绍了不少领域外的非线性模型。

利用软件进行非线性回归有三个困难：第一个困难是针对实际问题利用什么模型；第二个困难是如何将模型用 SPSS 语言表述；第三个困难是初值的选取。针对第一个困难，本书给出常见的适应散点图，由于对模型可能呈现的各种图形形态认识有限，我们提供的适应散点图有限。针对第二个困难，本书给出模型的 SPSS 语句。针对第三个困难，本书给出模型的初始值，但初始值只能适应模型的一般情形，对于特殊情形仍需读者给出初始值。

作者认为给出例子，让读者从数据输入开始学习，比练习学习软件的效果要好，读者从头到尾参加无间断的学习，可以消除对学习的恐惧，有利于培养学习的信心。所以，每个例子都给出图书情报方面具体的所有数值（几个少数例子除外，实在找不到相关例子），这也是编写此书的困难所在。回归在图书情报方面的应用不少，但数据都不完整，可作为例子价值的数据非常少，有时根本找不到，不得不借用其他领域的例子，因此，碰到其他领域的例子时还请读者原谅。

IBM SPSS 软件随处可下载，而且可汉语化，利用起来非常方便，它不需要高深的统计基础，非常实用，本书以图示方法导引读者使用软件解决图书情报领域的数量问题，减少读者学习困难和利用软件解决图书情报领域的数量问题的困难。本书的例子除了需要数据模拟外，都是利用 IBM SPSS 软件计算出来的。

感谢郑州大学图书馆领导多年来对作者撰写本书的关心和支持，感谢郑州大学图书馆同事们在本书撰写过程中给予的便利和帮助。

由于作者水平有限，加上时间仓促，错误难免，欢迎批评指正。

周爱民

2014 年 5 月

目 录

第一章 回归分析概述 …… 1

第一节 回归分析分类 …… 1
第二节 回归分析的步骤 …… 2

第二章 一元线性回归分析 …… 6

第一节 一元线性回归基本概念 …… 6
第二节 一元线性回归模型的参数估计 …… 7
第三节 检验指标 …… 8
第四节 异方差检验与一元线性回归模型的加权回归 …… 9
第五节 用 SPSS 软件进行一元线性回归分析 …… 11
第六节 曲线估计的 SPSS 软件拟合步骤 …… 38

第三章 多元线性回归 …… 49

第一节 多元线性无截距回归模型 …… 49
第二节 一般多元线性回归模型 …… 67
第三节 多元线性回归模型参数的加权最小二乘估计 …… 79
第四节 一般多元线性回归模型逐步回归的 SPSS 实现 …… 91

第四章 一元非线性回归模型 …… 96

第一节 SPSS 模型表达式的符号 …… 96
第二节 一些常见非线性回归模型、SPSS 模型表达式、初始值的确定与适应图形 …… 97
第三节 非线性模型在 SPSS 软件中的拟合步骤 …… 153
第四节 非线性模型在文献计量学中的应用 …… 160
第五节 多元非线性回归 …… 234
第六节 含约束条件的非线性回归方程在 SPSS 软件中的实现 …… 241

第五章　列联表分析 …… 246

第一节　单选题列联表分析 …… 246
第二节　多选题列联表 …… 273

第六章　分类数据的 logistic 回归和对数线性模型 …… 290

第一节　Logistic 回归概述 …… 290
第二节　二分类 Logistic 回归 …… 293
第三节　Logistic 有序多分类回归 …… 304
第四节　多项无序 Logistic 回归 …… 316
第五节　多分类无序列联表 Logit 回归 …… 337
第六节　对数线性模型分析无序列联表 …… 351

参考文献 …… 372

第一章 回归分析概述

回归分析是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法，运用十分广泛。它也是一种统计学上分析数据的方法，是建立因变量 y（或称反应变量）与自变量 x（或称独变量、解释变量）之间关系模型的重要方法。

第一节 回归分析分类

按照涉及的自变量的多少，回归分析可分为一元回归分析和多元回归分析。一元回归分析包含两类：一元线性回归分析和一元非线性回归分析。如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果在回归分析中，只包括一个自变量和一个因变量，但二者的关系可用一条曲线近似表示，这种回归分析称为一元非线性回归分析。一元线性回归的模型非常简单，但一元非线性回归分析的模型非常复杂，且形式多样。一元非线性回归包括：一元 Logistic 回归、一元 Logic 回归、平方曲线、立方曲线、威布尔模型、房室模型等。一元非线性回归的方法已成熟，一般一元非线性回归模型一旦给出，利用软件从理论而言，都可回归，但这也仅仅是理论上，这里还有模型适合问题和初值设定问题，如果这两个问题得到解决，那么利用软件就可很快得到拟合结果。多元回归分析也包含两类：多元线性回归分析和多元非线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。如果在回归分析中，包括两个或两个以上的自变量和一个因变量，但二者的关系可用一个多维空间的曲面近似表示，这种回归分析称为多元非线性回归分析。多元线性回归模型也较简单，但多元非线性回归模型非常复杂，人们对其认识非常有限，几乎没有发现有名的回归模型。

按照自变量和因变量之间的关系类型进行分类，可分为线性回归分析和非线性回归分析。

按照模型关系分类，回归可分为固定模型回归和非固定模型回归。固定模型回归指

回归模型被限定，仅需估计模型参数。固定模型包括：一元线性回归、一元非线性回归、多元线性回归。常见的非固定模型回归包括：Logistic 回归、Logit 回归、对数线性模型等。非固定模型是特殊的非线性回归模型。

回归分析目的在于了解两个或多个变量间是否相关、相关方向与强度，并建立最能够代表所有观测资料的函数（回归估计式），用此函数代表因变量和自变量之间的关系的数学模型，以便观察特定变量来预测研究者感兴趣的变量值。

第二节 回归分析的步骤

一、问题的发现

回归是从问题的发现开始的，只有发现有回归价值的问题，我们才能进行回归。

二、选择相关变量

针对实际问题，要找到问题的主要方面，用有代表性的变量来刻画问题，要根据该领域专业人士的意见选择变量集合，分清何为解释变量（自变量），何为响应变量（因变量）。

三、收集数据

选择好潜在的相关变量后，下一步是从实际中收集分析问题使用的数据，有时候，我们可以在一个可控的情况下收集数据，以使不感兴趣的因素保持不变，而更多的时候，数据是在一种非实验条件下收集的，研究者只能控制很少的因素。在每种情况下，我们收集到 n 个目标的观测数据，每个目标的观测数据都是对该目标所有潜在的相关变量的测量值，收集到的数据通常如表 1-1 进行记录。

表 1-1 回归分析中的变量符号

观测序号	响应变量 y	预测变量			
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
3	y_3	x_{31}	x_{32}	…	x_{3p}
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{n1}	x_{n2}	…	x_{np}

表 1-1 中的每一列代表一个变量，而每一行表示一个观测，对应某个目标（比如，一个馆员）的 $p+1$ 个值，其中一个为响应变量的值 y_i，其他 p 个预测变量中的每一个对应一个自变量观测值，符号 x_{ij} 指第 j 个预测变量的第 i 个观测值，即第一个下标对应观测序号，

第二个下标对应预测变量的序号。

表中的每个变量按其取值情况可以分为定量变量或定性变量，定量变量如馆员的身高、体重、年龄，定性变量如性别和为人的态度等。本书主要研究响应变量是定量变量的情况，当响应变量是二值变量时，有一种研究方法是 Logistic 回归，该方法将在第六章介绍。

如果所有预测变量都是定性变量，分析这些数据的方法称为方差分析。如果预测变量有定量变量也有定性变量，此时的回归分析称为协方差分析。方差分析和协方差分析其实是特殊的回归分析，但它们多用于工程试验设计，此书不予讨论。

四、模型设定

为了将响应变量和预测变量联系起来，通常先由该研究领域的专家根据他们的知识或主客观判断给出模型的形式：$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$，这个假设的模型或者被收集的数据证实，或者被推翻。要注意的是，此处只需给出模型的形式，它可以含有未知参数，我们需要选择函数 $f(x_1, x_2, \cdots, x_p)$ 的形式，该函数可以分为两类：线性和非线性。线性函数如

$$y = a + bx + \varepsilon$$

非线性函数如

$$y = a + e^{bx} + \varepsilon$$

注意线性和非线性的含义，线性和非线性不是指 x 和 y 之间的关系，而是指 y 关于参数是线性或非线性的。有些模型 y 和 x 是非线性的，但与参数却是线性的，如

$$y = a + bx + cx^2 + \varepsilon$$

$$y = a + b\ln x + \varepsilon$$

其中，在第一个方程中有 $X_1 = X, X_2 = X^2$，在第二个方程中有 $X_1 = lnX$，这称为变量的变换，某些非线性函数，若它可以通过变量的变换转化为线性函数，则这些非线性函数称为可线性化的。因此，线性模型类实际上比从表面看上去更广泛，因为它还包括所有可线性化的函数。但要注意，并不是所有的非线性函数都可线性化，例如 $y = a + e^{bx} + \varepsilon$，其中的非线性函数就无法线性化，有些学者将不能线性化的非线性函数称为本质上的非线性函数。

仅包含一个预测变量的回归方程称为一元回归方程，而包含多于一个预测变量的方程称为多元回归方程。

五、拟合方法

确定模型和收集数据之后，接下来是利用数据估计模型参数，也称为参数估计或模型拟合，最常用的估计方法是最小二乘法，在某些假设下（本书将不详细讨论），最小二乘估计有很多好的性质。本书中我们主要采用最小二乘法和它的一些变形方法（如加权最小二乘法）。对于非线性模型我们主要采用非线性最小二乘法，对于可线性化的模型而言，直接用非线性最小二乘法比模型先线性化，然后再用线性最小二乘法拟合的结果要好得多。在某些情况下（例如，当一个或多个假设不成立时），其他估计方法可能会优于最小二乘法，本书中我们考虑的其他估计方法有加权最小二乘法。

六、模型拟合

利用选定的估算方法(例如,线性最小二乘法或非线性最小二乘法)和收集到的数据进行回归参数估计或模型拟合,

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

中回归参数 $\beta_0, \beta_1, \cdots \beta_p$ 的估计用 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ 表示,于是,线性回归方程的估计可以写成

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

参数上方的记号“ ∧ ”表示该参数的估计,$\hat{y}$(读作:y-hat,帽子的意思)称为拟舍值。注意,模型是近似关系,不是相等关系,有误差,所以最后要加上误差项 ε。拟合后的式子中没有差项 ε,但因变量和估计出的参数要带帽,即 $\hat{y}$、$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$,但估计出的具体数值不带帽。

利用

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

可以对数据中的 n 个观测计算 n 个拟合值,例如,第 i 个拟合值 $\hat{y}_i$,是

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \qquad i = 1, 2, \cdots, n$$

注意式

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

还可以用预测变量的任意值来预测相应的响应变量的值,这种情况下获得的 y 称为预测值。拟合值和预测值的不同在于,拟合值对应的预测变量的值就是数据中的某个观测,而预测值对应的可以是预测变量的任何取值,但读者不能用过多的超出数据中预测变量取值范围的值来预测响应变量,预测是有误差的,超出数据中预测变量取值范围越远,误差越大。

七、模型评价和选择

统计模型(如回归模型)的有效性依赖于某些假设,通常是指对数据和模型的假设。对分析和结论的准确性至关重要的是这些假设条件是否满足。例如,在用

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

做任何分析之前,首先需要确定特定的假设是否成立。需要解决以下问题:

(1)需要哪些假设?

(2)对于每个假设,我们如何确定该假设是否被满足?

(3)当一个或更多假设不成立时,我们该如何处理?

本书的后续内容将详细介绍标准的回归假设以及回答上面的问题。我们强调,在分析得出任何结论之前,必须验证假设的合理性。模型与数据要有很好的吻合程度。

八、回归分析的目标

回归分析最重要的目标是准确确定回归方程，其概括了响应变量 Y 和一组预测变量 $X_1, X_2, \cdots, X_p$ 之间的关系，回归方程有很多应用，如可以用来评估单个预测变量的重要性，可以分析政策改变带来的影响，回归方程也可以用来根据给定的预测变量值预测响应变量的值。回归分析作为一类数据分析技术，在用于了解某种环境下变量之间相互关系的同时，也有助于我们利用数据尽可能多地了解变量所处的环境，所以，在确定回归方程过程中得到的认识和发现，与最后的方程一样有重要价值。

第二章 一元线性回归分析

本章主要介绍最简单的一元线性回归，也就是研究一个因变量 y 和一个自变量 x 之间的关系。

第一节 一元线性回归基本概念

一元线性回归模型为

$$y = b_0 + b_1 x + \varepsilon$$

变量 y 与 x 的关系用两部分来描述，一部分是由于主要因素自变量 x 的变化引起 y 线性变化的部分，即 $b_0 + b_1 x$；另一部分是由其他一切次要随机因素引起的，记为 ε，ε 虽小，但总存在，因 ε 的存在，使得变量 y 不能由变量 x 唯一确定。否则，二者就不是统计关系，而是确定关系。例如，图书第一年流通次数 x 是影响第二年流通次数 y 主要统计因素，但不是唯一因素，两年借同一本书的读者发生的变化、两年社会环境的变化都可能影响第二年流通次数 y。

一元线性回归模型中的 y 称为被解释变量（因变量），x 称为解释变量（自变量）。b_0 称为回归常数，b_1 称为回归系数。ε 是不可观测的随机因素的影响，通常假定它服从正态分布，即

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases}$$

这里 $E(\varepsilon)$ 表示 ε 的数学期望，$\mathrm{var}(\varepsilon)$ 表示 ε 的方差。对一元线性回归模型两端求条件期望，得

$$E(y \mid x) = b_0 + b_1 x$$

此式就是一元回归方程。

在建立回归方程时，因变量与自变量之间尽量要有逻辑因果关系，自变量的变化确实影响因变量的变化。有时候，两个变量互为因果，哪一个变量作自变量都可以，如：图书馆

的读者人数与国民经济 GDP。图书馆的读者人数多，人们从书中得到知识，知识转化为生产力，提高了国民经济 GDP，所以，图书馆的读者人数可以作为自变量，国民经济 GDP 可以作为因变量；从另一个角度看，国民经济 GDP 的提高，使人们从生存需求上升到精神追求，人们有更大的愿望去图书馆看书，所以，国民经济 GDP 可以作为自变量，图书馆的读者人数可以作为因变量。

一般情况下，对我们所研究的某个实际问题，获得的 n 组样本观测值 (x_1,y_1)，(x_2,y_2)，$\cdots$，(x_n,y_n) 来说，有

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

令

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

于是一元线性回归模型可表示为

$$\begin{cases} \boldsymbol{y} = \boldsymbol{xb} + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) = 0 \\ \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n \end{cases}$$

其中 $\boldsymbol{I}_n$ 为 n 阶单位矩阵。

第二节　一元线性回归模型的参数估计

对于每一个样本观测值 y_i 与其回归值 $b_0 + b_1 x_i$ 的离差为 $y_i - b_0 - b_1 x_i$，那么所有样本观测值与其回归值的离差的综合效果如何考虑？由于离差有正有负，直接求离差的和可能正负相互抵消，不能反映所有样本观测值与其回归值的离差的综合效果，所以直接求离差的和是不行的；可以用离差的绝对值之和来反映所有样本观测值与其回归值的离差，这就是最小一乘法。遗憾的是最小一乘法只能用计算机来算，不能手算，可以用离差的绝对值平方和来反映所有样本观测值与其回归值的离差，这就是最小二乘法，离差的绝对值平方和越小越好，即

$$Q(b_0,b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

越小越好，根据极值原理，

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\, x_i = 0 \end{cases}$$

经整理后，得正规方程组：

$$\begin{cases} nb_0 + \left(\sum_{i=1}^{n} x_i\right) b_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) b_0 + \left(\sum_{i=1}^{n} x_i^{\,2}\right) b_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

令

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \; ;$$

求解以上正规方程组，得：

$$\begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

第三节　检验指标

判断拟合结果好不好有三个基本标准：第一个标准是决定系数，第二个标准是模型显著性，第三个标准是回归系数的显著性，但在一元线性回归模型中，后两个标准是一回事。

1　决定系数的计算公式为

$$R^2 = \frac{SS_{回}}{SS_{总}} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

一般而言，决定系数越接近1，回归效果越好。

2　一元线性回归模型的显著性检验

一元线性回归模型 $y = b_0 + b_1 x + \varepsilon$ 的系数 b_1 是否显著为0？若 b_1 显著为0，那么回归模型毫无意义，所以我们总希望 b_1 显著不为0。

可以证明下式总成立：

$$\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$$