An Introduction to Statistical Methods and Data Analysis

统计学方法与数据分析引论（下）

〔美〕 R.L.奥特 /M.朗格内克 著　　张忠占 等 译

科学出版社
www.sciencep.com

统计学方法与数据分析引论

（原书第5版）

（下　册）

〔美〕R.L.奥特
M.朗格内克　著

张忠占　王建稳
王　强　杨中华　译
张忠占　校

科学出版社

北　京

内 容 简 介

本书内容分为八个部分，共20章，分上、下两册，每册10章。各章均有大量习题。本书给出了大量的实际例子，这些例子涉及众多的学科和实际领域，但又不过于专门，容易理解。在大部分章节中都使用实例来引入主题，并把统计概念和这些非常实际的问题联系在一起进行讲解，深入浅出，从而可以避免许多人对统计所抱有的粗浅的感性认识，即认为统计仅仅是另一门数学课程。作者把统计数据的收集与分析过程总结成“四步法”，并把“四步法”的讲解贯穿始终，利用实例逐步展开并阐明在设计调查研究或试验时所需要的统计技术和思路，然后讲解用直观、有效的“四步法”来收集并分析数据，非常利于初学者和实际工作人员抓住有关统计方法和模型的本质。书中提供了多种多样的图示，如正态概率图、盒形图、散点图、矩阵图和残差图等，通过这些图，读者可以一方面理解数据的特点和概括数据的方法，一方面进一步理解有关统计方法的基本思想和特点。作者很重视统计在解决实际问题中的作用，在全书中用许多篇幅讨论如何解释数据分析的结果，并专门用一章讲述了如何写数据分析报告。

本书适用于作为我国文科各专业的统计学引论教程，以及理工科各专业应用统计学课程的教材或教学参考书；也可作为有关方面实际工作人员的统计入门书。阅读本书不需要其他统计方面的基础，也不需要高等数学知识。

图字：01－2002－1678号
图书在版编目(CIP)数据

统计学方法与数据分析引论(原书第5版)/〔美〕奥特(R. Lyman Ott)、〔美〕朗格内克(Michael Longnecker)著；张忠占等译 .—北京：科学出版社，2003

ISBN 7-03-010815-9

书名原文：An Introduction to Statistical Methods and Data Analysis

Ⅰ.统… Ⅱ.①奥…②张… Ⅲ.①统计-方法 ②统计分析 Ⅳ.C8

中国版本图书馆 CIP 数据核字(2002)第079445号

责任编辑：杨 波、刘晓炜/责任校对：柏连海
责任印制：安春生/封面设计：耕者工作室

THOMSON - http://www.thomsonlearning.com

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
双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003年6月第 一 版 开本：B5(720×1000)
2003年6月第一次印刷 印张：45 3/4
印数：1—3 000 字数：884 000

定价：118.00元（上、下册）

（如有印装质量问题，我社负责调换〈路通〉）

目　录

下　册

第六部分　数据分析:回归方法和模型的建立

第七部分　试验设计与方差分析

第十一章　线性回归和相关

11.1　引言和案例

预测一个变量未来的值是重要的管理活动。财务官员必须预测未来现金流量,生产经理必须预测原材料的需求,人事经理必须预测未来职员的需求。解释过去的变化也是重要的。解释顾客数量过去的变化能够帮助经理了解对社会服务机构的服务的需求。找出可以解释某个汽车零部件规格限变异的变量能够帮助企业改进这个零部件的质量。回归分析的基本思想是利用一个定量的自变量的数据预测或解释一个定量的因变量。

我们可以区分预测(关于未来的值)和解释(关于现在或过去的值)。由于事后估计的原因,解释比预测更容易。但经常会用词"预测"来概括这两种情况,因此,这本书里我们有时会对预测和解释不加区别。

为了使预测(或解释)有意义,被预测的变量(因变量)和用来预测的变量(自变量)之间必须存在某种联系。毫无疑问,如果你试了足够多次,可能发现 28 种股价在一年中变化的公众股票完全能够被棒球联盟的 28 个主要球队在 7 月 4 日比赛的输赢比例来预测,但是,这样的预测是荒谬的,因为这二者之间毫无关联。预测需要一种**关联单元**,代表这两种变量之间存在关系的本质。对于时间序列数据,很简单关联单元就是时间。变量可以在同一时期观测,或为了做到真正地估计,自变量的观测比因变量的观测早一个时期。对于横断面数据(cross-sectional data),变量之间应该存在一个经济上或物理上的本质联系。假如我们打算预测不同的软饮料的市场份额的变化,那么就应该考虑到这些饮料的推销宣传活动,而不是各种品牌的实心面条调味汁的广告宣传。预测时需要关联单元似乎是显然的,但许多预测是在没有明显的关联单元的环境中进行的。

本章我们介绍简单线性回归分析,这种方法适用于一个自变量,并且预测因变

量 y 的方程是给定的自变量 x 的线性函数。例如，假如某县公路管理部门主任想预测用于投标的重新铺路的合同的成本，我们可以合理地预见，这个成本是将要重新铺设的公路英里数的函数。一个合理的开始是利用线性的预测函数。令 y = 这个项目总的成本(千美元)，x = 需要重新铺设的公路的英里数，$\hat{y}$ = 预测的这个项目总的成本(千美元)。预测方程 $\hat{y} = 2.0 + 3.0x$(例如)就是线性方程。常数项，比如 2.0，是**截距**，可解释为当 $x = 0$ 时 y 的预测值。在公路重新铺设的例子中，我们可以把截距解释为这个项目开始时的固定成本。x 的系数，比如 3.0，是这条直线的**斜率**，就是当 x 改变一个单位时 y 的预测变化。在公路重新铺设的例子中，假设有两个项目相差一英里，我们可以预测长项目的成本比短项目多 3(千美元)。一般地，预测方程表示为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

其中 $\hat{\beta}_0$ 是截距，而 $\hat{\beta}_1$ 是斜率，参看图 11.1。

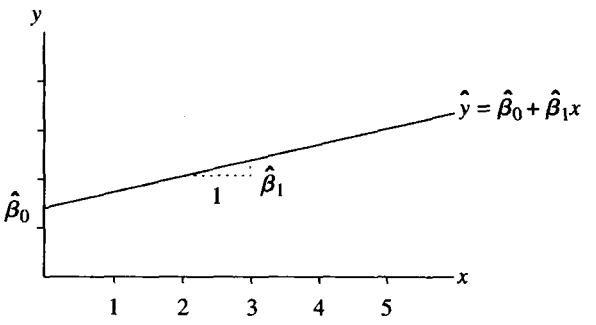


图 11.1 线性预测函数

简单线性回归的基本思想是用数据拟合与因变量 y 和单个自变量 x 相关的预测直线。简单回归的第一个假定是这两个变量的关系是线性的。按照直线性的假设，当 x 变化时方程的斜率不改变。在公路重新铺设的例子中，我们可以假定从长距离项目中不会获得(实质的)规模经济。除非直线性假定成立(至少大致)，否则利用简单线性回归就没有多大意义了。

表面上，直线性并不总是合理的假定。例如，我们打算用 x = 某汽车商的电台广告的重复次数，预测 y = 知道这家汽车商的盛夏打折销售的司机的数量，直线性的假定意味着商业广告的第一次播出不会比第一千零一次播出导致更多的司机知道此事。(你已经听过上述那样的商业广告)我们强烈怀疑这个假定在 x 的广泛范围内的有效性。我们很清楚商业广告重复的次数越多，其影响越小，因此一个直线的预测不很好。

假定存在直线性，我们将 y 表示为 x 的线性函数：$y = \beta_0 + \beta_1 x$。但是，按照这

个方程，y 是 x 的确切的线性函数；没有留下余地考虑不可避免的误差（y 的实际值与预测值的差异）。为此，对应每一个 y 我们引入一个随机误差项 ϵ_i，并假定模型为

$$y = \beta_0 + \beta_1 x + \epsilon$$

我们假定随机变量 y 是由可预测的部分（x 的线性函数）和不可预测的部分（随机误差 ϵ_i）组成。系数 β_0 和 β_1 可解释为真实的截距和斜率，而误差项 ϵ 包含所有的其他已知和未知的因素的影响。在公路重新铺设的项目中，一些不可预测的因素比如罢工、天气条件和设备故障的影响都包含在 ϵ，ϵ 还包含一些因素比如公路的陡峭或维修前的条件——那些应该在预测中考虑的但最终没有考虑到的因素。不可预测的因素和被忽略的因素的综合影响就构成了随机误差项 ϵ。

例如利用不同的新汽车本身的重量（自变量）预测这些新汽车所消耗汽油的平均里程（因变量），一个方法是每辆车安排不同的司机行驶一个月时间。预测误差可能是由哪些不可预测和忽略的因素产生呢？这个研究中不可预测（随机）因素包括司机的驾驶习惯和技术、行驶路程的类型（城市与公路）和途中所遇红灯的次数。可能被忽略的因素有发动机的规格和变速装置的类型（手动与自动）。

在回归分析的研究中，自变量的值（x_i 的值）一般看作预先确定的常数，因此随机性的惟一来源就是 ϵ_i 项。尽管绝大多数经济和商务应用中 x_i 的值都是固定的，但并不总是这样的。例如，x_i 代表某个才能测试中申请者的得分，y_i 代表申请者的生产能力，如果数据是申请者的随机样本，那么 x_i（y_i）是随机变量。在回归分析研究中把 x 看作固定的还是随机的无关紧要，如果 x_i 都是随机的，我们可以简单地把所有概率陈述视为在 x_i 取定观测值的条件下的相应的概率陈述。

当我们假定所有的 x_i 都是常数，那么关于 x_i 的模型中惟一的随机部分是随机误差项 ϵ_i。以下是模型的正式的假设条件。

定义 11.1 回归分析的正式的假定：

1. 所有误差的期望值都是零，即对所有的 i，$E(\epsilon_i)=0$。

2. 所有误差的方差都是相同的，即对所有的 i，$\mathrm{Var}(\epsilon_i)=\sigma_\epsilon^2$。

3. 误差是相互独立的。

4. 所有误差都服从正态分布，即对所有的 i，ϵ_i 都是服从正态分布。

这些假定都显示在图 11.2 中。因变量的实际值服从正态分布，其均值都落在回归线上且对自变量所有的值都具有相同的标准差。惟一未在图中显示的假定是观测值彼此的独立性。

有了这些正式的假定，就可以导出以后的显著性检验和预测方法。我们一开始可以通过观察数据的**散点图**来检验模型的这些假定。散点图就是简单地画出每对（x_i，y_i）的坐标，其中因变量作为纵轴。看看这些点是不是基本落在一条直线的周围或者是一条确定的曲线模式。除此之外，我们还观察是否有明显地远离数据

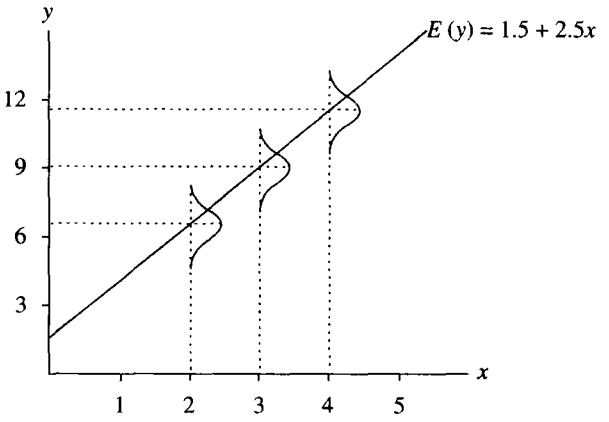


图 11.2 回归分析中 y 的理论分布

的一般模式的离群值。图 11.3(a)显示了一个散点图。

最近,**平滑法**被用来直接通过数据描绘出一条曲线而不必要假定任何特定的模型。假如这样的平滑法产生的曲线接近直线,那攻线性回归就是合理的。LOWESS(Locally weighted scatterplot smoother)就是一个这样的方法。粗略地讲,平滑法就是将数据沿 x 轴取一个非常窄小的"小段",在小段上计算出拟合数据的直线,再将这个小段沿 x 轴慢慢移动,重新计算直线,如此重复下去。最后将这些所有的小直线连接成一条平滑的曲线。小段的宽度称为*窗宽*;一般是由进行平滑的计算机软件控制的。简单的散点图(图 11.3(a))和通过这些数据的 LOWESS 曲线显示在图 11.3(b)中。散点图表明是曲线关系;LOWESS 曲线印证了这一点。

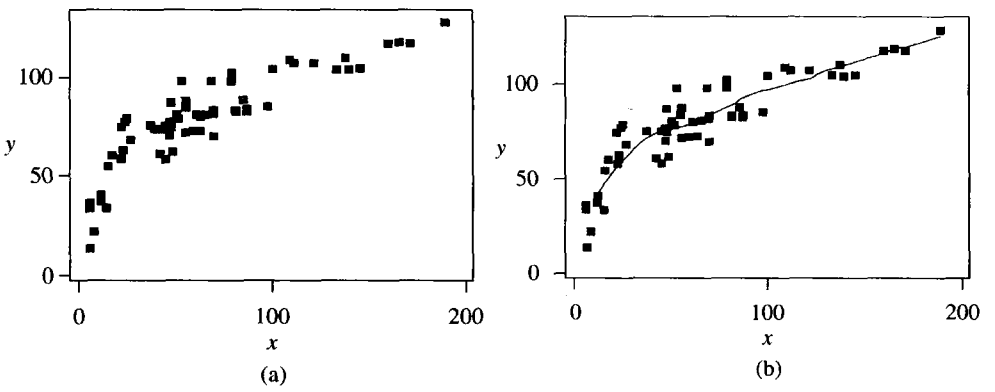


图 11.3 (a) 散点图和(b) LOWESS 曲线

另一种散点平滑是**样条拟合**。这个方法是取一个窄小的数据小段，在上面拟合出一条曲线(一般是三次方程)，再移动到下一个小段拟合出另一条曲线，如此重复下去。最后将这些曲线连接成一条连续的曲线。

许多经济关系都不是线性的。例如，任何回报递减模式都趋向于产生一种增加的关系，只是其增加率是逐渐减少的。如果散点图本身或拟合的 LOWESS 曲线没有表现出线性性，那么我们一般是通过对自变量或因变量做变换将二者的关系直线化。一个好的统计软件或表格算法都可以计算像平方根这样的函数。变换后的变量可以简单看作另一个变量。

例如，某个大城市每个春季都要派遣职员修补街道的坑洞。每天所派遣的职员人数和修补的坑洞数都被记录下来，修补的坑洞数和职员数的散点图和带有 LOWESS 曲线的散点图显示在图 11.4 中。二者的关系是非线性的，就算没有 LOWESS 曲线，斜率下降也是明显的。这并不奇怪，因为派遣职员越多，雇佣的工人工作效率越低，这些职员不得不走得更远去发现坑洞，如此下去。所有这些原因都显示会出现回报递减现象。

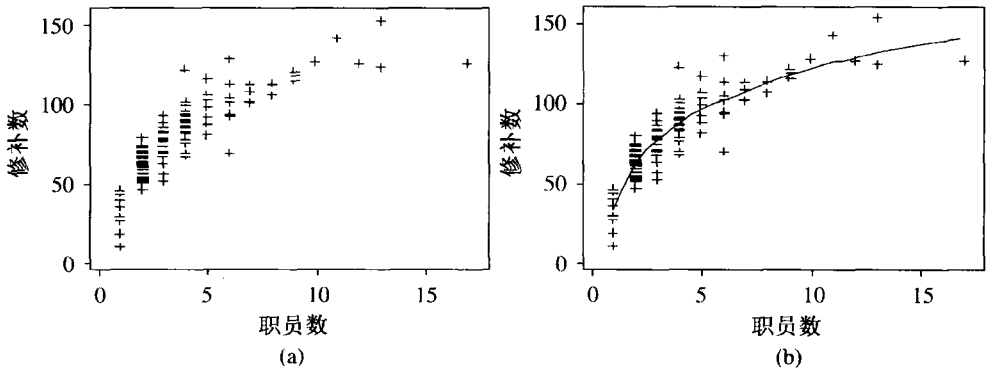


图 11.4 街道坑洞数据的散点图

我们可以通过尝试几种自变量变换的方法来寻找更接近线性的散点图。三种常用的变换是平方根变换、自然对数变换和倒数变换(1 除以变量)。将每种变换用于修补坑洞数的数据。变换后的散点图和相应的 LOWESS 曲线显示在图 11.5a—c 中。平方根变换(a)和倒数变换(c)并没有得出直线关系，而自然对数变换(b)就得出了非常好的结果。因此，我们将用 Ln(职员数)作为自变量。

寻找一个好的变换需要不断地尝试，也会犯不少错误。下面是一些帮助找出合适变换的建议。在散点图中寻找变换要注意两个方面。首先，二者的关系是非线性的吗？其次，沿着 y(纵)轴是否有变差逐渐增大的趋势？如果有这种趋势，那

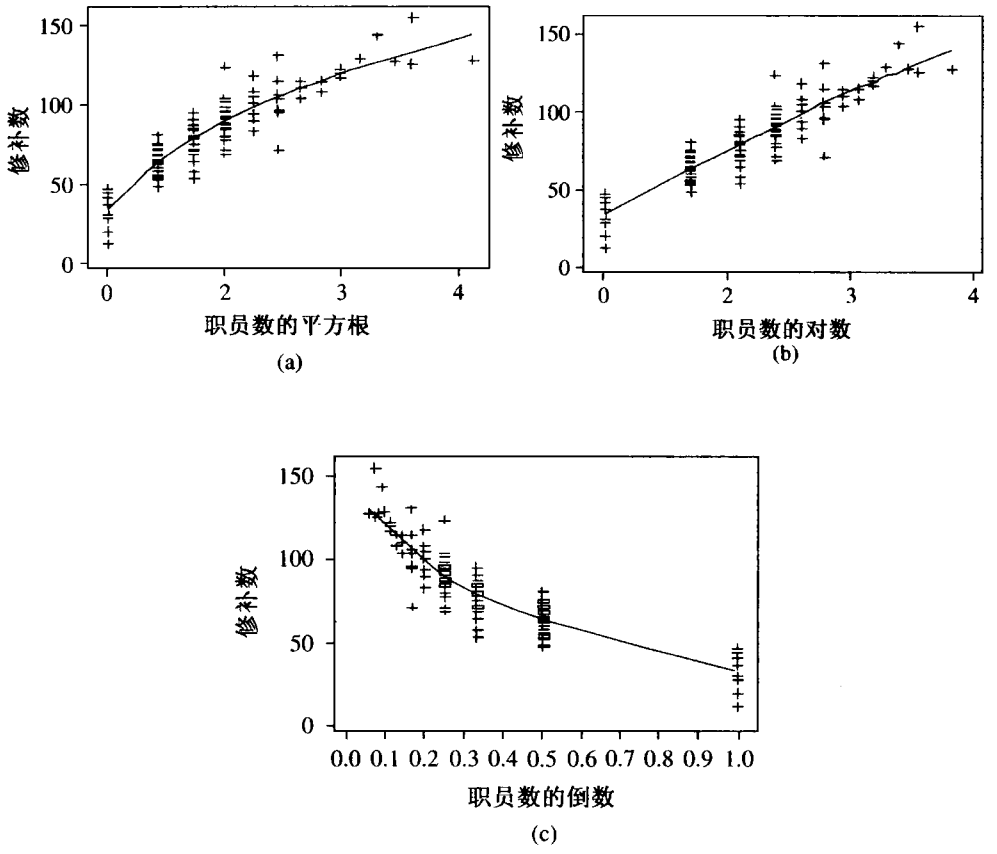


图 11.5 预测变量变换后的散点图

么常数方差的假定就有问题了。这些建议虽然不能覆盖所有的情况,但是包含了最常见的问题。

定义 11.2 选择变换的几个建议:

1.假如散点图显示因变量是随自变量增加而增加,但增加的幅度是逐渐减少的;且在曲线周围的变差大致是常数,则对自变量 x 作平方根变换、自然对数变换或倒数变换。

2.假如散点图显示因变量是随自变量增加而增加,但增加的幅度是逐渐增加的;且在曲线周围的变差大致是常数,用 x 和 x^2 作为预测变量。因为这个方法使用了两个变量,所以需要下两章所介绍的多元回归分析。

3.假如散点图显示因变量是随自变量增加到最大值后下降,且在曲线周围的变差大致是常数,用 x 和 x^2 作为预测变量。

4.假如散点图显示因变量是随自变量增加而增加,但增加的幅度是逐渐减少的;且在曲线周围的变差随着被预测变量 y 的增加而增加的,则用 y^2 作为因变量。

5.假如散点图显示因变量是随自变量增加而增加,但增加的幅度是逐渐增加的;且在曲线周围的变差随着被预测变量 y 的增加而增加的,则用 $\ln(y)$作为因变量。有时候也用 $\ln(x)$作为自变量。注意原始变量的自然对数增量十分接近原始变量相应的百分比变化,因此,变换后的变量的斜率可以用百分比的变化很好地解释。

例 11.1

某航线实行飞行常客优惠项目,该项目的参加者中得到的免费旅行大量增加。为了预测未来这些旅行数量的趋势,项目负责人收集了最近 72 个月的数据。因变量 y 为免费旅行的次数;自变量 x 为月数。图 11.6 给出了利用 Minitab 绘制的带有 LOWESS 平滑曲线的散点图。应该用什么变换呢?

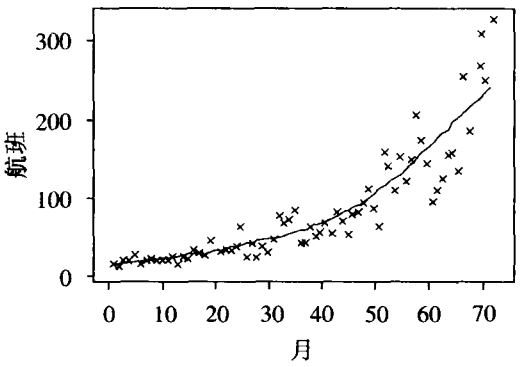


图 11.6 常客免费航班

解答 散点图的模式显示因变量是随自变量增加而增加,但增加的幅度是逐渐增加的。LOWESS 曲线是明确向上的。另外,围绕平滑曲线的变动(上下)是逐渐增加的,曲线高端周围的点(这个例中的右端)比曲线低端周围的点更分散。变差逐渐增加建议对变量 y 作变换,这时候自然对数(ln)变换常常是很好的选择。图 11.7 给出了 Minitab 对 y 取对数后重新绘制的散点图。散点图的模式非常接近直线,并且围绕直线的变差十分接近常数。

我们将在第十二章详细讨论模型假定的检查。对于一个预测变量的简单回归模型,仔细检查散点图,最好再用平滑曲线拟合数据,可以帮助我们避免严重的错误。

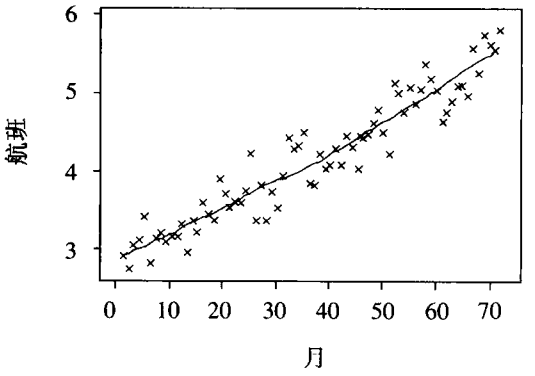


图 11.7 对数变换后的结果

一旦已经确定了任何数学变换,我们必须估计回归线的实际方程。实际上,我们惟一获得的是样本数据,总体的截距、斜率和误差的方差都必须用有限的样本数据来估计。这一节的假定使得我们能够用样本数据对总体参数做出推断。

案例:比较诊断大肠杆菌的两种方法

第七章的案例分析描述了诊断大肠杆菌(*E. coli*)的一种新的细菌方法,即 Petrifilm HEC 试验。研究人员想要评估 HEC 试验的结果和另一个复杂的在实验室中进行的试验方法 HGMF(hydrophobic grid membrane filtration)的结果的一致性。HEC 试验的接种比传统的方法更容易、更简洁,施行也更安全。但是在使用 HEC 试验之前,必须比较在同一肉体样本上所获得的 HEC 试验的读数和 HGMF 方法的读数,以确定两种方法是否产生同样的结果。如果二者的读数不同但能够获得一个接近 HEC 读数和 HGMF 读数之间关系的方程,那么研究人员可以对 HEC 试验的读数进行校准来预测用 HGMF 方法所能获得的读数。如果 HEC 试验的结果和 HGMF 方法的结果没有关系,那么 HEC 试验不能用来诊断大肠杆菌。

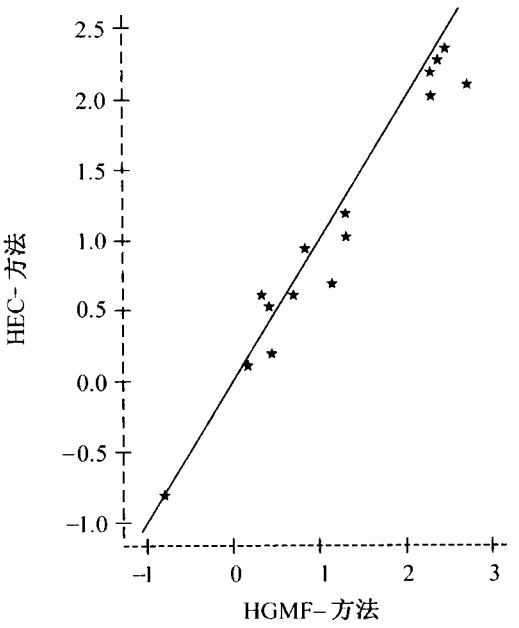
数据收集的设计 第七章描述了这个试验的第一个阶段。试验的第二阶段是将两种方法应用于人工污染的牛肉。来自于三头经检测是大肠杆菌阴性的 Holstein 母牛的牛肉块已经准备就绪。取出其中 18 块牛肉并用大肠杆菌污染。分别用 HEC 和 HGMF 方法检测这 18 个样本中的每一块,将这两种方法产生的大肠杆菌浓度进行必要的转换计量($\log_{10}$CFU/ml)。这个案例的数据是 18 对样本点,具体如下。

RUN	HEC	HGMF
1	0.50	0.42
2	0.06	0.20
3	0.20	0.42
4	0.61	0.33
5	0.20	0.42
6	0.56	0.64
7	−0.82	−0.82
8	0.67	1.06
9	1.02	1.21
10	1.20	1.25
11	0.93	0.83
12	2.27	2.37
13	2.02	2.21
14	2.32	2.44
15	2.14	2.28
16	2.09	2.69
17	2.30	2.43
18	−0.10	1.07

数据的整理 下一步,研究人员按照前面 2.5 节描述的步骤为以后的统计分析整理数据。他们仔细检查试验过程以确定每一对作为样本的肉都是基本一样的,这样使得 HEC 读数和 HGMF 读数的任何差异都来自于两种方法的不同。在这样的检查中,有关试验过程的问题都会被发现,有问题的观测值都会剔除出去。

数据的分析 研究人员感兴趣的是确定这两种方法所产生的大肠杆菌浓度的测量值是否有很强的相关性。下面是试验数据的散点图。(见下页)

散点图中的 45°直线显示了两个方法的读数的近似的一致性。如果散点落在直线上,则这两个方法所确定的达成杆菌浓度是完全一致的。其中 17 个点都靠近这条直线但有一些偏差,因此研究人员要决定一致性的程度,并获得一个表示两个方法所得读数之间关系的方程。如果利用回归方程可以表示两个方法所得读数的准确的相关性,那么已知 HEC 读数后,研究人员可以预测 HGMF 方法的读数。这使得他们可以比较用 HEC 方法和在实验室中用 HGMF 方法所测得的大肠杆菌浓度。我们将在 11.6 节对这些数据做详细的分析。



注意:2 个观测点被隐藏起来

11.2 估计模型中的参数

回归模型

$$y = \beta_0 + \beta_1 x + \varepsilon$$

的截距 β_0 和斜率 β_1 都是总体的参数,我们必须从样本数据中估计它们的值。误差的方差 σ_ε^2 是另一个必须估计的总体参数。回归分析的第一个问题就是获得斜率、截距和方差的估计值。这一节我们讨论如何获得这些估计值。

11.1 节的公路重新铺设例子是一个很合适的说明。假设以下数据是近几年的相似项目中所获得。注意我们有关联单元:特定的成本和英里数的联系是因为它们来自同一个项目。

费用 y_i(千美元):	6.0	14.0	10.0	14.0	26.0
里程 x_i(英里):	1.0	3.0	4.0	5.0	7.0

第一步是绘制数据的散点图以检查 x 和 y 的关系。记住图中每个点代表一个观测数据的坐标(x, y),如图 11.8 所示。散点图显示 x 和 y 之间存在不完全但大体上呈渐增的关系。似乎有可能存在一个直线关系,但这么有限的数据无法看出能做什么变换。

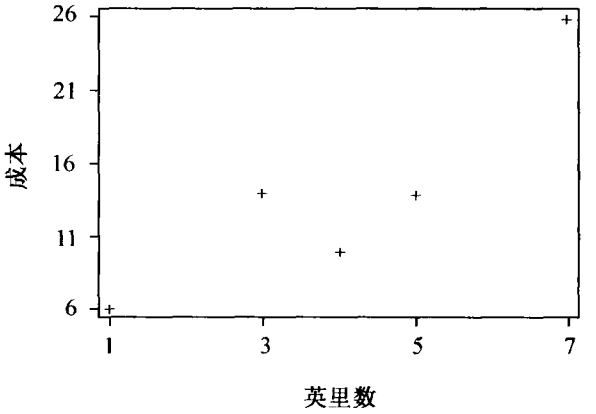


图 11.8 成本对英里数的散点图

回归分析就是找出最佳的直线预测。“最佳”的最常用的标准是根据平方预测误差。我们可以通过最小化总的平方预测误差来求出预测直线的方程—也就是，求出截距 $\hat{\beta}_0$ 和斜率 $\hat{\beta}_1$。基于这个目的的方法称为最小二乘法，因为它是通过最小化下面的量来选择 $\hat{\beta}_0$ 和斜率 $\hat{\beta}_1$ 的：

$$\sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

图 11.9 中标出了预测误差，即为与预测直线之间的垂直偏差。这些偏差就是垂直距离，这是因为我们预测的是 y，误差就应该取 y 方向的。对于这些数据，最小二乘线就是 $\hat{y} = 2.0 + 3.0x$；由此方程产生的一个偏差用较小的大括号标出。

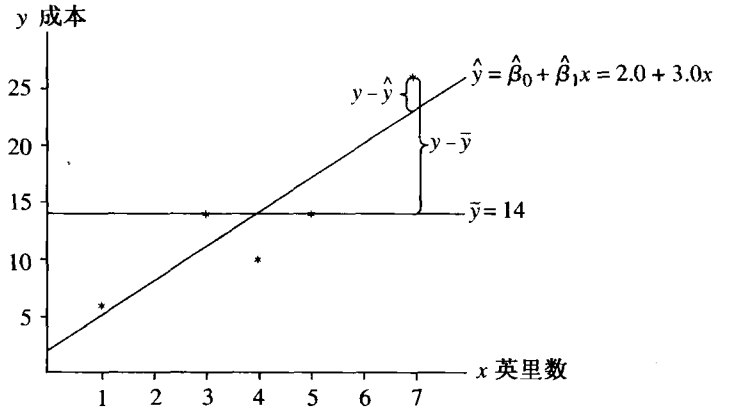


图 11.9 最小二乘线和均值的偏差