HZ BOOKS
华章教育

WILEY

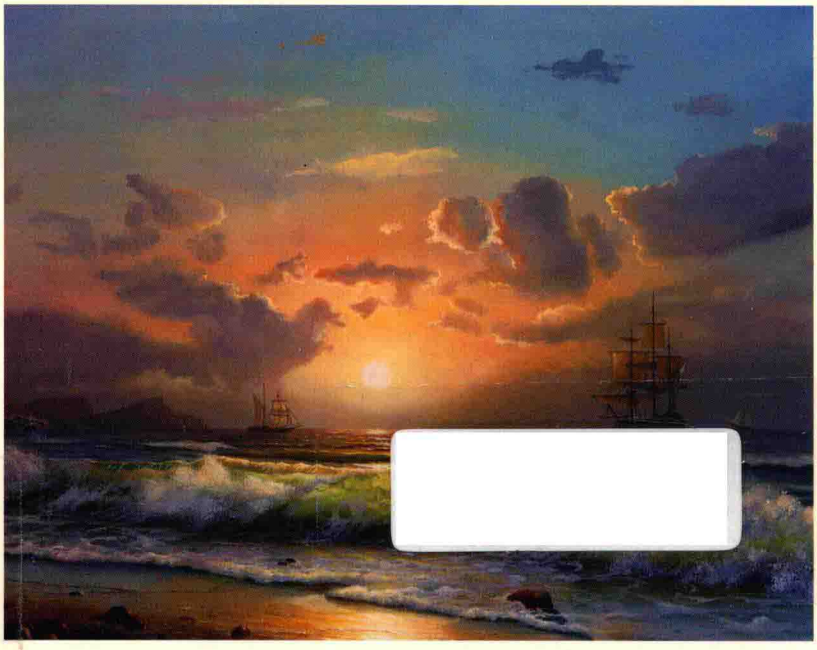
统计学精品译丛

（原书第5版）

线性回归分析导论

Introduction to Linear Regression Analysis

(Fifth Edition)



[美] 道格拉斯 C. 蒙哥马利（Douglas C.Montgomery）
伊丽莎白 A. 派克（Elizabeth A.Peck） 著
G.杰弗里·瓦伊宁（G.Geoffrey Vining）

王辰勇 译

机械工业出版社
China Machine Press

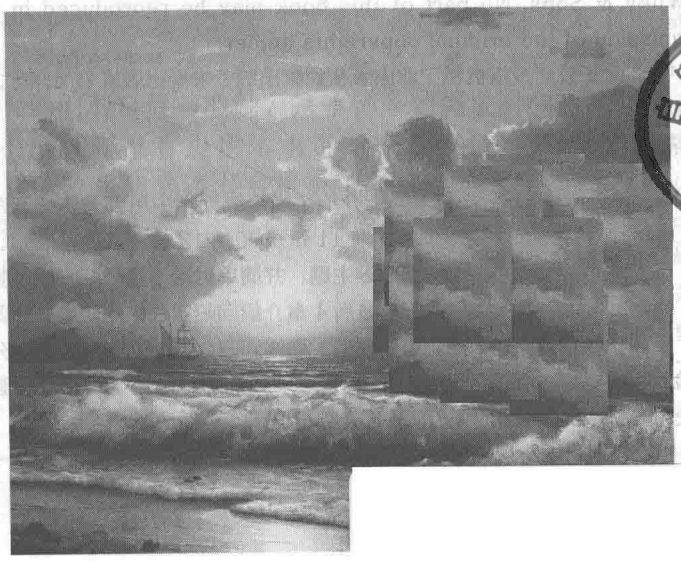
统计学精品译丛

（原书第5版）

线性回归分析导论

Introduction to Linear Regression Analysis

(Fifth Edition)



阜阳师范学院图书馆
藏书

[美] 道格拉斯 C. 蒙哥马利（Douglas C.Montgomery）
伊丽莎白 A. 派克（Elizabeth A.Peck）
G.杰弗里·瓦伊宁（G.Geoffrey Vining）
著

王辰勇 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

线性回归分析导论（原书第5版）/（美）蒙哥马利（Montgomery，D.C.）等著；王辰勇译．
—北京：机械工业出版社，2016.4
（统计学精品译丛）
书名原文：Introduction to Linear Regression Analysis, Fifth Edition

ISBN 978-7-111-53282-8

I. 线… II. ①蒙… ②王… III. 线性回归－回归分析－教材 IV. O212.1

中国版本图书馆CIP数据核字（2016）第057191号

本书版权登记号：图字：01-2013-4241

Copyright © 2012 by John Wiley & Sons,Inc.

All rights reserved. This translation published under license. Authorized translation from the English language edition, entitled *Introduction to Linear Regression Analysis, Fifth Edition*, ISBN 978-0-470-54281-1, by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Published by John Wiley & Sons. No part of this book may be reproduced in any form without the written permission of the original copyrights holder.

本书中文简体字版由约翰－威利父子公司授权机械工业出版社独家出版．未经出版者书面许可，不得以任何方式复制或抄袭本书内容．

本书封底贴有Wiley防伪标签，无标签者不得销售．

本书是世界公认的“回归分析”权威教材，不仅从理论上介绍了当今统计学中用到的传统回归方法，还补充介绍了尖端科学研究中不太常见的回归方法．本书前11章是核心内容，阐述简单回归、多元回归、诊断统计量、指示变量、有偏估计、多项式回归模型等主题，并简单讨论了用于回归模型验证的一系列方法以及如何处理强影响观测值、多重共线性问题．最后4章介绍回归实践中比较重要的各种论题，包括非线性回归、广义线性模型、时间序列数据的回归模型、稳健回归、自助回归估计值、分类回归树、神经网络以及回归试验设计等．书末还有5个附录，其中附录C简短地给出了理论性更强的某些其他论题，附录D介绍了使用SAS处理回归问题，附录E介绍了R.

本书适用于工程学、化学科学、物理科学、统计学、数学以及管理学等专业的各年级本科生与一年级研究生．

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：和 静　　责任校对：董纪丽
印　刷：北京瑞德印刷有限公司　　版　次：2016年4月第1版第1次印刷
开　本：186mm×240mm 1/16　　印　张：31
书　号：ISBN 978-7-111-53282-8　　定　价：99.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

版权所有·侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所 韩光/邹晓东

译 者 序

历经十个月，我终于翻译完成了这本 600 余页的著作. 这是我第一本译著，也是我第一次从事系统的翻译工作. 今天，我诚惶诚恐地将这本译著献给大家.

本书是一本权威的回归分析著作，被许多美国大学用作教材. 本书的第一作者蒙哥马利先生是一位著名的工程统计学家，著有十几本著作，其中包括《试验设计与分析》《统计质量控制》《响应曲面方法》《工程统计学》等权威教材. 本书秉承了蒙哥马利先生在工程统计领域的写作风格，旁征博引，大量使用各个领域的实际数据(尤其是工程数据)作为例子，这极大地增加了读者的兴趣与对回归分析方法的理解，但也为本书的翻译工作带来了不小的难度.

一本译著的完成，离不开身边支持我的人. 首先要感谢华章公司对我的信任，将这本权威的名著交给了我，并给予了我一些帮助. 还要感谢已经陪我走过了十年的小伙伴们，尤其是傅泽伟、韩旭、王尉、王潇、张旭斌；翻译本书期间恰逢我一次重大的人生转折，是他们在我最困窘与彷徨的时候，给予了我巨大的帮助与启迪. 最后，要感谢我年迈的父母，没有他们的督促与激励，这本译著的完成时间恐怕还要延后，谨将这本译著献给他们.

原书出版商没有提供本书的电子版，所以我直接将原文笔译在纸上，然后再使用 LaTeX 录入. 虽然翻译本书花费了我大量的精力，但是翻译的过程也是我个人提高的过程，令我收获颇丰. 由于经验与时间所限，译文难以做到尽善尽美，其中的不当与错误之处，请读者不吝指正.

王辰勇

2015 年 10 月

前言

回归分析是广泛用于分析多因子数据的方法之一. 回归分析使用方程来表达所感兴趣的变量(响应变量)与一系列相关预测变量之间的关系，其中所产生的概念逻辑过程使本书具有广泛的吸引力与实用性. 因为回归分析隐含着优雅的数学，同时也有完善的统计学理论，所以回归分析在理论上也是非常有趣的. 成功地使用回归分析，就要从理论与常见的实际问题两个方面，将其应用于实际数据.

本书适合作为回归分析的入门教材，包含了回归分析中的标准论题，也涉及许多新的论题. 本书理论与应用实例并重，使读者不仅理解必要的基本原理，还能将各种回归建模方法应用于具体环境中. 本书最初成书于回归分析课程的笔记，该课程面向高年级本科生与一年级研究生，学生来自不同的专业：工程学、化学与物理科学、统计学、数学，以及管理学. 本书也曾用于面向专业人士的入门培训. 本书假定读者学过统计学的入门课程，并熟悉假设检验、置信区间以及正态分布、t 分布、卡方分布与 F 分布，矩阵代数的某些知识也是必要的.

在回归分析的现代应用中，计算机扮演了重要的角色. 今天，即便是电子表格软件，也可以使用最小二乘法来拟合回归方程. 因此，本书整合了许多软件的使用方法，给出数据表格与图形输出，并总体上讨论了某些计算机软件包的功能. 本书使用了 Minitab、JMP、SAS 与 R 来处理各种问题与例子. 之所以选择这些软件包，是因为它们广泛用于实践和教学. 许多作业习题都要使用统计软件包来求解. 本书的所有数据都可以通过出版商以电子形式获得. ftp 地址为：ftp://ftp. wiley. com/public/sci_tech_med/introduciton_linear_regression，其中汇总了数据、习题解答、PowerPoint 文件，以及与本书相关的其他材料.

第 5 版的改进

本书第 5 版有很多改进，包括：重新组合了课文材料，新的例子，新的习题，关于时间序列回归分析的新的一章，以及关于回归模型试验设计的新材料. 进行修订的目的是使本书更好地用作教材与参考书，并更新对某些论题的讨论.

第 1 章从整体上介绍了回归建模，并描述了回归分析的某些典型应用. 第 2 章与第 3 章提供了简单回归与多元回归中最小二乘模型拟合的标准结果，以及基本的推断程序(假设检验、置信区间与预测区间). 第 4 章讨论了模型适用性检验的基本方法，包括残差分析，其中强调了残差图、离群点的探测与处理、PRESS 统计量，以及失拟检验. 第 5 章讨论了如何将数据变换与加权最小二乘法用于解决模型不适用这一问题，如何处理违背基本回归假设的情形. 本章也介绍了 Box-Cox(博克斯-考克斯)方法与 Box-Tidwell 方法，从分析的角度设定数据变换的形式. 第 6 章展示了诊断统计量，并简单讨论了如何处理强影响观测值. 第 7 章讨论了多项式回归模型及其各种变形. 本章的论题包括多项式拟合与推断的基本程序，以多项式、分层多项式与分段多项式为中心的讨论，同时拥有多项式与三角函数项的模型，正交多项式，响应曲面方法概述，以及非参数回归方法与光滑回归方法的介绍. 第 8 章介绍了指示变量，同时将回归模型与方差分析模型进行了联系. 第 9 章关注

试读结束：需要全本请在线购买： www.ertongbook.com

多重共线性问题，包括对多重共线性来源的讨论，多重共线性的危害、诊断量与各种诊断性度量．本章介绍了有偏估计，包括岭回归及其某些变种，以及主成分回归．第 10 章研究了变量选择与模型构建方法，包括逐步回归程序与所有可能回归．本章也讨论与解释了评估子集模型的某些准则．第 11 章展示了用于回归模型验证的一系列方法．

前 11 章是本书的核心，这 11 章贯穿着许多概念与例子．其余四章讨论回归实践中比较重要的各种论题，可以独立阅读．第 12 章介绍了非线性回归，而第 13 章简单讨论了广义线性模型．虽然这两章可能不是线性回归教材的标准论题，但是不介绍这两章，对工程学与自然科学的学生与教授将是非常不负责任的．第 14 章讨论时间序列数据的回归模型．第 15 章概述了几个重要论题，包括稳健回归、回归变量中测量误差的影响、逆估计即校准问题、自助回归估计值、分类回归树、神经网络，以及回归试验设计．

除了正文的内容外，附录 C 简短地给出了理论性更强的某些其他论题．回归分析的专家与利用本书讲授高级课程的教师会对其中某些论题更感兴趣．计算在许多回归课程中都扮演着重要角色，这些课程广泛使用 Minitab、JMP、SAS 与 R．本教材提供了这些统计软件包的输出．附录 D 介绍了使用 SAS 处理回归问题．附录 E 介绍了 R．

本书作为教材如何使用

本书覆盖了广泛的论题，有很大的灵活性．对于回归分析的入门课程，推荐详细讲授第 1 至 10 章，然后选出学生特别感兴趣的论题．举例来说，作者之一(D. C. M.)定期讲授一门面向工程学学生的回归课程，论题包括非线性回归(因为工程学中经常出现的机械模型几乎永远是非线性模型)、神经网络以及回归模型验证，其他的推荐论题有多重共线性(因为学生经常会遇到多重共线性问题)、广义线性模型导论——主要关注逻辑斯蒂回归．G. G. V. 讲授过一门面向统计学研究生的回归分析课程，大量使用了附录 C 中的材料．

我们认为，应当将计算机直接整合进课程中．近年来，在大多数课堂上都采用笔记本电脑与计算机投影设备，像在讲座中那样解释回归方法．我们发现，这样可以极大地促进学生对回归方法的理解．我们也要求学生使用回归软件来解题．在大多数情况下，习题都使用了实际数据，或是来自现实世界的议题，以表示回归分析的一般性应用．

教师手册包含了所有习题的答案、所有电子版数据集，以及可能适合于考试的习题．

致谢

感谢在准备本书的过程中提供了反馈与帮助的人．Scott M. Kowalski、Ronald G. Askin、Mary Sue Younger、Russell G. Heikes、John A. Cornell、André I. Khuti、George C. Runger、Marie Gaudard、James W. Wisnowski、Ray Hill 与 James R. Simpson 博士给出了许多建议，他们的建议极大地改良了本书的前几版与第 5 版．我们特别感激为本书提供反馈的许多研究生与实践专家，他们洞察出问题所在，丰富或拓展了本书的材料．我们也要感谢约翰-威利父子公司、美国统计学会以及生物统计学委员会，他们大度地允许我们使用其版权材料．

Douglas C. Montgomery
Elizabeth A. Peck
G. Geoffrey Vining

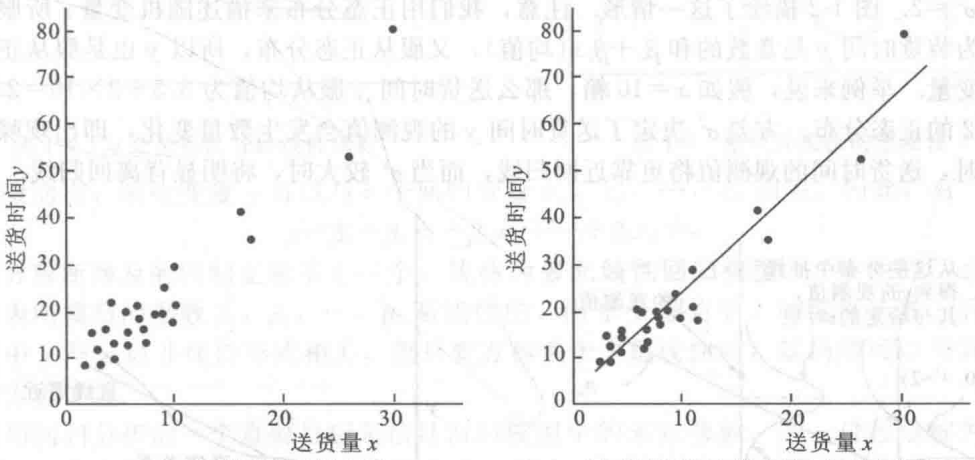
目　录

第1章 导 引

1.1 回归与建模

回归分析统计方法研究变量之间的关系并对其构建模型. 回归的应用领域广泛，几乎遍及所有学科，包括工程学、物理科学、化学科学、经济学、管理学、生命科学及社会科学等. 事实上，回归分析可能是应用得最为广泛的统计方法.

下面是一个应用回归分析解决实际问题的例子. 假设有一位工业工程师，这位工程师受聘于一家负责软饮料装瓶的公司，他正在分析自动售货机的货物运送与服务运营过程. 这位工程师估计普通的送货员装货并检修机器所需的时间(送货时间)与送货的箱数有关. 他走访了随机选取的 25 个配有自动售货机的零售网点，并依次观测了每个网点送货员的送货时间(以 min 为单位)和送货箱数(以箱为单位). 图 1-1a 画出了这 25 个观测值的**散点图**，它清楚地表明了送货时间与送货量之间的关系. 能看到所观测的数据点都大致落到一条直线上，图 1-1b 画出了这条直线，但这一直线关系并不精确.



a) 关于送货量的散点图　　b) 送货时间与送货量之间的直线关系

图 1-1

如果令 y 表示送货时间，x 表示送货量，那么这两个变量的直线方程为

$$y=\beta_0+\beta_1 x \tag{1.1}$$

式中：β_0 为截距；β_1 为斜率. 因为数据点并不是精确地落到这条直线上，所以为了解释这一现象应当对方程(1.1)进行修改. 令 y 的观测值与直线上的值$(\beta_0+\beta_1 x)$之间的差值为**误差** ε，方便起见可将 ε 理解为统计误差. 统计误差是一个随机变量，它使得模型不能精确拟合数据. 统计误差的产生可能是由于其他变量对送货时间有影响，存在测量误差，等等，因此对送货时间这一数据而言更为合理的模型为

$$y=\beta_0+\beta_1 x+\varepsilon \tag{1.2}$$

试读结束：需要全本请在线购买： www.ertongbook.com

方程(1.2)称为**线性回归模型**. 习惯上将 x 称为自变量，y 称为因变量，但由于这种命名通常会引起与统计独立性这一概念的混淆，所以又将 x 称为**预测**变量或**回归**变量，将 y 称为**响应**变量. 由于方程(1.2)仅涉及一个回归变量，故也将其称为**简单线性回归模型**.

为了深入领会线性回归模型，假设回归变量 x 的值为定值，观察响应变量 y 的相应值. x 为定值时，方程(1.2)右边的随机项 ε 将决定 y 的性质. 设 ε 的均值和方差分别为 0 和 σ^2，那么对任意定值的回归变量 x，其响应变量 y 的均值为

$$E(y|x)=\mu_{y|x}=E(\beta_0+\beta_1 x+\varepsilon)=\beta_0+\beta_1 x$$

注意，这与前文考察过的，通过图 1-1a 中的散点图得出的关系式方程是相同的. 同理对任意定值 x，其 y 的方差为

$$\mathrm{Var}(y|x)=\sigma^2_{y|x}=\mathrm{Var}(\beta_0+\beta_1 x+\varepsilon)=\sigma^2$$

因此，实际的回归模型 $\mu_{y|x}=\beta_0+\beta_1 x$ 是一条直线，y 的各均值都在这条直线上. 也就是说，取任意定值的 x 所对应的回归直线的高度，恰好是那个 x 所对应的 y 的期望值. 斜率 β_1 可以解释为 x 变化一个单位时 y 均值的变化量. 此外，对一个特定取值的 x，其 y 的随机变化情况由模型误差项的方差 σ^2 决定. 这意味着对于每一个 x，都存在 y 的一个分布，每个 x 所对应的这一 y 的分布，其方差都是相同的.

举例来说，假设送货时间与送货量之间线性关系的实际回归模型为 $\mu_{y|x}=3.5+2x$，并假设方差 $\sigma^2=2$. 图 1-2 描绘了这一情形. 注意，我们用正态分布来描述随机变量 ε 所服从的分布. 因为装货时间 y 是常数的和 $\beta_0+\beta_1 x$(均值)，又服从正态分布，所以 y 也是服从正态分布的随机变量. 举例来说，假如 $x=10$ 箱，那么送货时间 y 服从均值为 $3.5+2\times10=23.5$min、方差为 2 的正态分布. 方差 σ^2 决定了送货时间 y 的观测值会发生数量变化，即出现噪声. 当 σ^2 较小时，送货时间的观测值将更靠近回归线，而当 σ^2 较大时，将明显背离回归线.

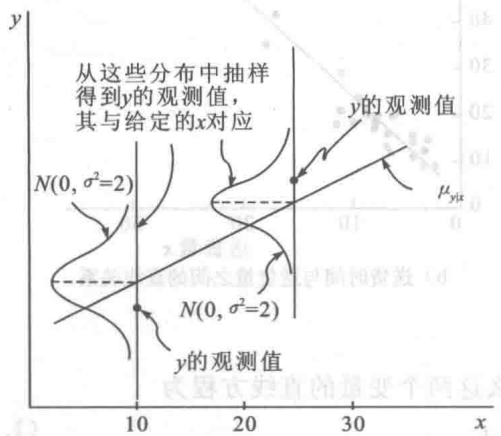


图 1-2 线性回归的观测值是如何生成的

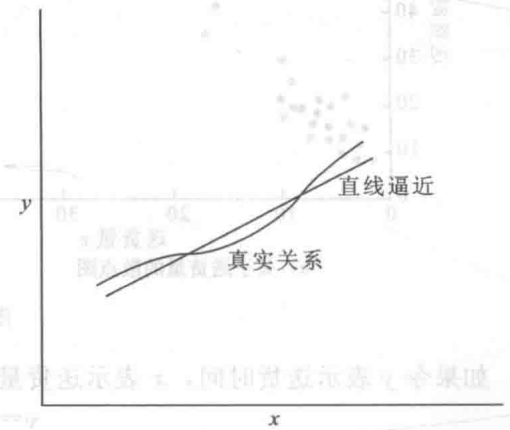


图 1-3 复杂关系的线性回归逼近

在几乎所有的回归应用领域中，对我们感兴趣的变量而言，回归方程只是实际函数关系的逼近. 实际的函数关系通常基于物理学理论、化学理论，或其他工程学理论与科学理论而产生，也就是说，实际函数关系是以人们对理论中潜在机理的了解为基础产生的. 因此，通常将这类模型称为**机理模型**. 而回归分析却不同，人们将其视为**经验模型**. 图 1-3 描绘了这样一种情形：y 与 x 之间的实际数量关系相对复杂，但这一复杂的数量关系可以

用一个线性回归方程很好地逼近. 而潜在机理有时更为复杂，这就需要用更为复杂的逼近函数来逼近 y 与 x 之间的实际关系. 如图 1-4 所示，这里的“分段线性”回归模型用来逼近 y 与 x 之间的实际关系.

通常情况下，回归方程的有效性仅限于包含观测数据在内的回归变量的区域上. 比如考虑图 1-5 中的例子. 假设关于 y 和 x 的数据位于区间 $x_1 \leqslant x \leqslant x_2$ 上，在此区间上的线性回归方程是对数据实际关系的良好逼近. 但是，假如使用该方程预测 y 的值时，其对应的回归变量的值位于区域 $x_2 \leqslant x \leqslant x_3$ 内，那么显然，对在这一区域上的 x，由于模型和方程错误，该线性回归模型将失效.

3

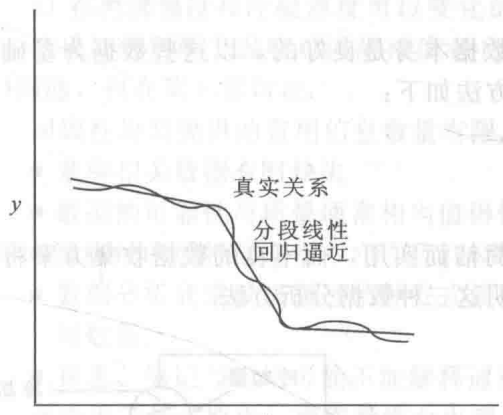


图 1-4 复杂关系的分段线性逼近

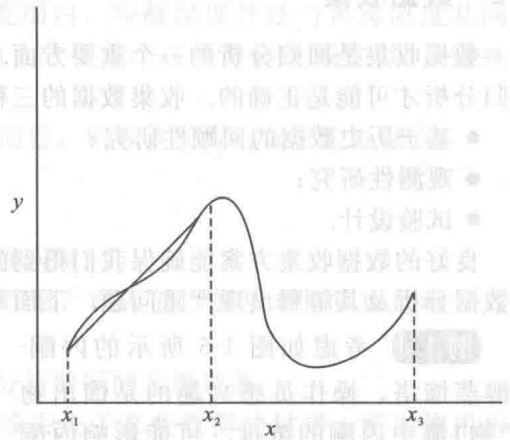


图 1-5 使用外推法的危险性

一般而言，响应变量 y 可以与 k 个回归变量 x_1，x_2，…，x_k 相关，因此，有

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\varepsilon \tag{1.3}$$

由于该方程所涉及的回归变量不止一个，故称为**多元线性回归模型**. “线性的”这一形容词用来表明模型的参数 β_1，β_2，…，β_k 是线性的，而非 y 是关于 x 的线性函数. 后文的许多模型中 y 与 x 以非线性形式相关，但只要方程关于 β 是线性的，就仍然可以将其当做线性回归方程处理.

使用回归分析的一个重要目标是估计回归模型中的未知参数，这一过程也称为用模型拟合数据. 本书将研究若干种参数估计方法，其中一种就是最小二乘法(在第 2 章介绍). 举例来说，用最小二乘法拟合送货时间数据将得到

$$\hat{y}=3.321+2.1762x$$

式中：$\hat{y}$是送货时间的拟合值，也称为估计值，其对应着送货量 x 的箱数. 该拟合方程已在图 1-1b 中画出.

回归分析的下一阶段称为**模型适用性检验**. 模型适用性检验研究模型的适当程度，确定拟合质量的高低. 模型适用性检验分析将决定回归模型的实用性. 模型适用性检验有两种可能的结果，要么表明模型是合理的，要么必须修正原来的拟合方案. 因此，回归分析是一个反复的过程，在这一过程中，数据导出了模型，而模型也拟合了数据. 研究数据拟合的质量后，要么修正模型或拟合方案，要么采用这一模型. 后续章节将多次解释这一过程.

4

回归模型并非意味着变量间存在因果联系. 即使两个或更多变量间可能存在牢固的实

证关系，也不能认为这就证明了回归变量与响应变量间存在因果联系. 确立因果关系，要求回归变量与响应变量必须存在一种基础性的、与样本数据无关的关系，比如理论分析中所暗含的关系. 回归分析有助于因果关系的确认，但不能成为判断因果关系是否存在的唯一基础.

最后一定记住，回归分析只是众多用于解决问题的数据分析方法的一种，也就是说，回归方程本身可能并非研究的主要目的. 就整个数据处理过程而言，洞察力与理解能力通常更为重要.

1.2 数据收集

数据收集是回归分析的一个重要方面. 只有数据本身是良好的，以这些数据为基础的回归分析才可能是正确的. 收集数据的三种基本方法如下：

- 基于历史数据的回顾性研究；
- 观测性研究；
- 试验设计.

良好的数据收集方案能确保我们得到的模型简洁而实用，而不良的数据收集方案将使得数据分析及其解释出现严重问题. 下面举例说明这三种数据分析方法.

例 1.1 考虑如图 1-6 所示的丙酮-丁醇蒸馏塔，操作员感兴趣的是馏出物(产物)流中丙酮的浓度. 可能影响丙酮浓度的因素有再沸温度、冷凝温度和回流率. 对该蒸馏过程而言，操作员一直记录以下数据：

- 测试样本中，产物流中丙酮的浓度，每 4 小时记录一次；
- 再沸温度控制器的日志，这是一个再沸温度图；
- 冷凝温度控制器的日志；
- 名义回流率，每小时记录一次.

在蒸馏过程中名义回流率应为常数，生产过程中该速率极少变化. 现在讨论前文所述的三种数据收集基本方法如何应用到这一蒸馏过程中.

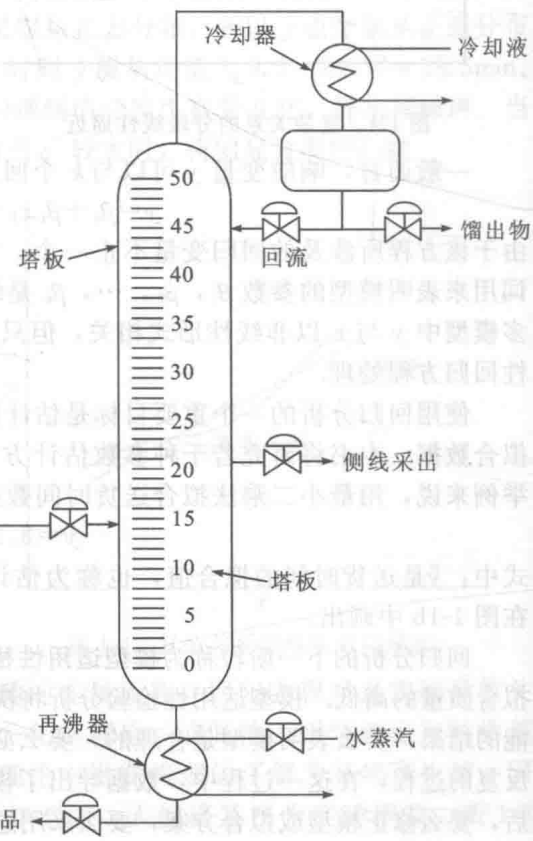


图 1-6 丙酮-丁醇蒸馏塔

回顾性研究 我们可以进行回顾性研究. 即使用一定时期内全部历史数据或其样本，来决定再沸温度、冷凝温度和回流率与产物流中丙酮浓度的关系. 回顾性研究利用以前收集的数据，并将研究成本最小化. 但是，这一过程存在

以下问题.

1) 我们实际上不知道回流率对丙酮浓度的影响，因为我们必须假设这一影响在那段历史时期内变化不大.

2) 因为再沸温度和冷凝温度与丙酮浓度的数据不直接对应，因此，构造一个逼近的对应关系需要付出极大努力才能实现.

3) 通过使用自动控制装置，在生产过程中能控制温度，使其尽可能达到特定的目标值. 因为再沸温度和冷凝温度几乎不随时间变化而变化，因此，观察其对浓度的真实影响存在极大困难.

4) 在再沸温度和冷凝温度可以变化的极小范围内，冷凝温度往往与再沸温度共同上升. 结果，辨别两个温度对丙酮浓度各自的影响十分困难. 这将导致**共线性**(即**多重共线性**)问题，将在第 9 章讨论. 6

回顾性研究提供的有用信息数量有限. 一般而言，回顾性研究有如下主要缺点：

- 某些相关数据有时缺失.
- 数据的可靠性与质量通常相当值得怀疑.
- 数据的本质特征通常不允许我们轻松地处理问题.
- 数据分析通常设法通过某种方式使用这些数据，而人们以前从未打算以这种方式使用数据.
- 日志、笔记与记忆可能不能解释通过数据分析验证的有趣现象.

无论出于什么原因，操作员常会出现未记录或丢失了某些数据的过错，所以使用历史数据时总会伴随着危险. 一般而言，历史数据包含关键的信息，也包含便于收集的信息. 便于收集的信息一般严谨而精确，而重要的信息通常并不严谨精确. 因此，历史数据通常受困于记录错误等问题，这些错误使得历史数据易于出现**离群点**，即与大多数数据差异很大的观测值. 回归分析只有在基于可靠的数据时，才是正确的.

有时候，数据便于收集并不意味着其特别有用，通常情况下，常规的过程监视装置认为不重要的数据和不便于收集的数据反而确实对过程有显著影响. 因为这些数据信息从未被收集，所以历史数据不能提供这些信息. 举例来说，外界温度可能影响蒸馏塔的热量流失，天冷时蒸馏塔流失了比天热时更多的热量. 但丙酮-丁醇蒸馏塔的生产日志没有记录外界温度，结果即使外界温度的影响比较重要，对历史数据的分析也无法包括这一因子.

在某些情况下，我们尝试使用替代数据. 收集替代数据，是为了替代真正需要去收集的数据. 只有在替代数据在很大程度上真正反映了其所代表的数据时，分析结果才能提供正确的信息. 举例来说，进料口丙酮与丁醇的混合物的性状，能显著影响蒸馏塔的性能. 蒸馏塔的设计是将(达到混合物的沸点的)饱和液体作为进料. 生产日志记录进料温度，但不记录进料流中丙酮与丁醇的特定浓度，因为这些浓度值在通常条件下很难获得. 在这种情况下，进料口的温度将替代进料口混合物的性状. 在恰当的温度进料，以及使用过冷液体或气液混合物从进料口进料，是完全可行的.

在某些情况下，数据收集过于随意，因此数据的品质、精确性和可靠性极低，这将对响应变量产生极大影响. 数据对响应变量的影响可能是真实的，也可能是不精确的“人造

7 数据”对响应变量产生了虚假的影响. 很多数据分析得出的结论是无效的，因为它过于依赖其所使用的数据，而这些数据从未打算直接用于数据分析.

最后，许多数据分析的最初目的，是隐藏暗含的有趣现象的根本起因. 使用历史数据时，这些我们感兴趣的现象可能已经发生于数月前或数年前. 日志与笔记通常不对根本起因提供显著的深入的观察数据，而记忆也会随着时间流逝而明显消退. 大多的情况下，基于历史数据的数据分析会发现未经解释的、不精确的有趣的现象.

观测性研究 对这一问题我们可以使用观测性研究来收集收据. 正如这一名字所表明的，观测性研究就是对过程或总体进行观测. 只有获得的相关数据足够多，我们对这一过程的影响和干预才足够强. 使用恰当的试验计划，观测性研究能确保数据的精确性、可靠性与完整性. 但在另一方面，观测性研究能提供的数据间特定关系的信息极为有限.

在本例中，我们建立了数据收集表. 生产人员可以使用数据收集表，记录特定时间上的再沸温度、冷凝温度与实际回流率，并观测对应的产物流中丙酮的浓度. 数据收集表应提供添加评注的功能，以便记录可能发生的任何特殊的现象. 这一处理将确保数据收集的精确性与可靠性，也将解决上文提到的问题 1)和问题 2). 离群值与数据中某些误差相关，这一方法也将观测到离群值的可能性最小化. 不幸的是，观测性研究不能解决问题 3)和问题 4)，所以观测性研究易于产生共线性问题.

试验设计 这一问题最佳的数据收集策略是试验设计，试验中我们根据一种明确定义的策略，称为试验设计，控制再沸温度、冷凝温度和回流率等因子. 这一策略能确保我们分离了与之相关的每个因子对丙酮浓度的影响. 这一过程消除了所有共线性问题. 试验中因子特定的值叫做水平，一般来说，设定一个较小的数字代表因子的水平，比如 2 或 3. 对于蒸馏塔这个例子，假设对每个因子设定高水平(+1)和低水平(−1)两种水平. 那么，我们将使用三个因子，每个因子存在两个水平. 处理组合是对每个因子水平的特定组合，使用一次处理组合，就是进行一次试验. 试验设计或试验计划包含一系列试验.

对于蒸馏塔的例子，最为合理的试验策略是，使用所有可能的处理组合，用八次不同的蒸馏试验来形成一个基本试验. 表 1-1 所示的是这些组合的高低水平.

图 1-7 表明，这一试验设计形成了以高低水平为单位的立方体. 在不同条件的蒸馏试验过程中，依次让蒸馏塔达到平衡，抽取产物流样本，确定丙酮的浓度，然后对这些因子
8 进行特定的推断. 试验设计这一方法允许我们主动地研究总体或过程.

表 1-1 蒸馏塔试验设计

再沸温度	冷凝温度	回流率
−1	−1	−1
+1	−1	−1
−1	+1	−1
+1	+1	−1
−1	−1	+1
+1	−1	+1
−1	+1	+1
+1	+1	+1

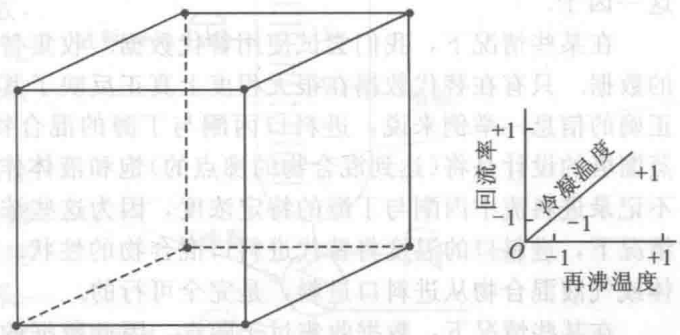


图 1-7 蒸馏塔试验设计

1.3 回归的用途

使用回归模型有以下目的：

1）描述数据；

2）参数估计；

3）预测与评估；

4）控制.

理工科资料中，往往使用方程汇总或描述数据集来建立模型，而回归分析有助于得到此类方程. 比如，通过大量收集送货时间和送货量的数据得到的回归模型，相比数据表甚至数据图形，都更加方便和实用.

回归方法有时可以解决参数估计问题. 举例来说，化学工程中使用米氏方程 $y=\beta_1 x/(x+\beta_2)+\varepsilon$ 来描述反应速率 y 与浓度 x 的关系. 在这一模型中，β_1 是反应的最终速率，即随着浓度的增大速率能达到的最大值. 如果得到了由不同浓度下速率的观测值组成的样本，那么化学工程设计中就能通过回归分析来得到能拟合数据的模型，从而得到最大速率的估计值. 第 12 章说明了此类回归模型如何拟合数据.

回归的很多应用领域都涉及对响应变量的预测. 比如，我们可能希望预测装入一定数量的软饮料瓶的包装，箱数所需的送货时间，这一预测可能有助于规划送货活动，比如，设计路线、安排行程，也可能有助于评估送货作业的效率. 由于存在前文(见图 1-5)讨论的模型或方程错误，所以利用回归模型进行预测时，使用外推法是危险的. 但是即使模型的方程形式是正确的，不良的模型参数估计量仍可能导致不良的预测效果.

使用回归模型的目的还可以是进行控制. 比如，化学工程中使用回归分析来得出有关纸张抗张强度与木浆中硬木浆浓度的模型，然后利用这一方程，通过改变硬木浆的水平，控制抗张强度使其达到合适的值. 以控制为目的使用回归方程时，重要的是变量之间要存在因果关系. 注意，如果仅使用方程进行预测，因果关系可能并不是必要的，而必要的是，用于构建回归方程的原始数据中存在的因果关系. 举例来说，在美国，根据佐治亚州亚特兰大市八月的日耗电量，可能可以较好地预测该市八月最高温度日的温度，但是通过削减电力消耗来降低最高温度的尝试显然注定会失败.

1.4 计算机的角色

回归建模过程是一个反复的过程，如图 1-8 所示. 开始，使用有关研究过程的所有理论知识和可获得的数据，来指定最初的回归模型. 数据图通常非常有助于指定最初的回归模型. 然后是估计模型的参数，一般使用最小二乘法或极大似然法，这两种方法都会在本教材中展开讨论. 然后评估模型的适用性，这包括找出可能存在的对模型方程形式的错误假设，例如，未将重要变量纳入模型，将不必要的变量纳入模型，或者使用了特殊数据或不合适的数据. 如果模型不适用，就必须重新创建模型，并重新估计参数，这一过程可能会反复进行几次，直到得到适用的模型. 最后，完成模型验证，确保在最终应用时模型能产生可接受的结果.

试读结束：需要全本请在线购买： www.ertongbook.com