UMSS

大学数学科学丛书 —3

线性模型引论

王松桂　史建红　尹素菊　吴密霞 编著

科学出版社
www.sciencep.com

大学数学科学丛书 3

线性模型引论

王松桂 史建红
尹素菊 吴密霞 编著

科学出版社
北京

内 容 简 介

本书系统阐述线性模型的基本理论、方法及其应用，其中包括理论与应用的近期发展. 全书共分九章. 第一章通过实例引进各种线性模型. 第二章讨论矩阵论方面的补充知识. 第三章讨论多元正态及有关分布. 从第四章起，系统讨论线性模型统计推断的基本理论与方法，包括：最小二乘估计、假设检验、置信区域、预测、线性回归模型、方差分析模型、协方差分析模型和线性混合效应模型.

本书可作为高等院校数学科学系、数理统计或统计系、生物统计系，计量经济系等有关学科的高年级本科生、硕士生或博士生的学位课或选修课教材，以及数学、生物、医学、工程、经济、金融等领域的教师或科技工作者的参考书.

图书在版编目（CIP）数据

线性模型引论/王松桂等编著. —北京：科学出版社，2004
（大学数学科学丛书；3）
ISBN 7-03-012772-2

Ⅰ.线… Ⅱ.王… Ⅲ.线性模型-教材 Ⅳ.O212

中国版本图书馆 CIP 数据核字（2004）第 005441 号

责任编辑：吕 虹/责任校对：钟 洋
责任印制：钱玉芬/封面设计：王 浩

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
新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2004 年 5 月第 一 版 开本：B5(720×1000)
2004 年 5 月第一次印刷 印张：19 1/4
印数：1—3 500 字数：350 000

定价：35.00 元
（如有印装质量问题，我社负责调换〈环伟〉）

第一作者简介

王松桂，北京工业大学教授、博士生导师．1965年毕业于中国科技大学并留校执教，曾任数学系副主任．1993年调入北京工业大学，曾任应用数学系主任和应用数理学院院长．长期从事线性模型和多元统计分析等方面的科学研究．

曾先后应邀赴美国、加拿大、日本、瑞典、瑞士、芬兰、波兰等国家和中国香港地区的20余所大学讲学和合作研究．曾获得第三世界科学院研究基金、瑞士国家基金和芬兰科学院研究基金．曾任中国数学会理事、中国概率统计会常务理事，现任中国工业与应用数学会常务理事、美国统计刊物“Journal of Statistical Planning and Inferences”副主编以及美国“Mathematics Review”特约评论员．曾获中国科学院重大科技成果二等奖和两项北京市科技进步二等奖，所著教材《概率论与数理统计》获教育部优秀教材二等奖．

在《中国科学》、《科学通报》、《数学学报》、《数学进展》、美国“Linear Algebra and Its Applications”、“Annals of Statistics”、“Journal of Multivariate Analysis”等国内外刊物发表论文100余篇．出版的学术专著有“Advanced Linear Models”（英文版，美国Marcel Dekker公司出版，1994）、《线性模型的理论及其应用》、《近代回归分析》、《实用多元统计分析》、《矩阵论中的不等式》、《广义逆矩阵及其应用》、《线性统计模型》、《概率论与数理统计》等9部．

《大学数学科学丛书》序

按照恩格斯的说法，数学是研究现实世界中数量关系和空间形式的科学.从恩格斯那时到现在，尽管数学的内涵已经大大拓展了，人们对现实世界中的数量关系和空间形式的认识和理解已今非昔比，数学科学已构成包括纯粹数学及应用数学内含的众多分支学科和许多新兴交叉学科的庞大的科学体系，但恩格斯的这一说法仍然是对数学的一个中肯而又相对来说易于为公众了解和接受的概括，科学地反映了数学这一学科的内涵. 正由于忽略了物质的具体型态和属性、纯粹从数量关系和空间形式的角度来研究现实世界、数学表现出高度抽象性和应用广泛性的特点，具有特殊的公共基础地位，其重要性得到普遍的认同.

整个数学的发展史是和人类物质文明和精神文明的发展史交融在一起的. 作为一种先进的文化，数学不仅在人类文明的进程中一直起着积极的推动作用，而且是人类文明的一个重要的支柱. 数学教育对于启迪心智、增进素质、提高全人类文明程度的必要性和重要性已得到空前普遍的重视. 数学教育本质是一种素质教育；学习数学，不仅要学到许多重要的数学概念、方法和结论，更要着重领会到数学的精神实质和思想方法. 在大学学习高等数学的阶段，更应该自觉地去意识并努力体现这一点.

作为面向大学本科生和研究生以及有关教师的教材，教学参考书或课外读物的系列，本丛书将努力贯彻加强基础、面向前沿、突出思想、关注应用和方便阅读的原则，力求为各专业的大学本科生或研究生（包括硕士生及博士生）走近数学科学、理解数学科学以及应用数学科学提供必要的指引和有力的帮助，并欢迎其中相当一些能被广大学校选用为教材，相信并希望在各方面的支持及帮助下，本丛书将会愈出愈好.

李大潜

2003 年 12 月 27 日

前　言

线性模型是现代统计学中理论丰富、应用广泛的一个重要分支，随着高速电子计算机的日益普及，在生物、医学、经济、管理、农业、工业、工程技术等领域的应用获得长足发展. 因此，在国内外很多高等院校已将线性模型列入数学科学系、数理统计系或统计系、生物统计系、计量经济系等高年级本科生、硕士生或博士生的学位课或选修课. 本书是为适应上述需要而编写的教材或教学参考书.

全书共分九章. 第一章通过实例引进各种线性模型，使读者对模型的丰富实际背景有一些了解，这将有助于对后面引进的统计概念和方法的理解. 第二章讨论矩阵论方面的补充知识. 第三章讨论多元正态及有关分布. 从第四章起，系统讨论线性模型统计推断的基本理论与方法. 本书的第一作者先后在中国科学技术大学、北京工业大学、复旦大学、安徽大学、云南大学等国内院校以及芬兰的坦佩雷大学和美国的科罗拉多州立大学讲授过本书的部分内容.

借本书出版之际，我们要向我们的老师陈希孺院士表示衷心的感谢，感谢他对我们多年来的研究给予的热情鼓励和指导.

本书的出版得到科学出版社和吕虹先生的支持和关心，樊亚莉小姐为本书部分章节打字，另外，本书的写作得到国家自然科学基金和北京市自然科学基金资助，编者愿借此机会向他们表示诚挚的谢意.

本书由王松桂等编著. 第一至四章由王松桂执笔，第五、六章由史建红执笔，第七、八章由尹素菊执笔，第九章由吴密霞执笔，最后由王松桂统一修改定稿. 由于编者水平所限，书中错误或不当之处在所难免，恳请国内同行及广大读者不吝赐教.

编　者

2003 年 6 月 30 日

符 号 表

$\triangleq$	“定义为”或“记为”
$A \geq 0$	A 为对称半正定方阵
$A > 0$	A 为对称正定方阵
$A \geq B$	$A \geq 0, B \geq 0$ 且 $A - B \geq 0$
A^-	矩阵 A 的广义逆
A^+	矩阵 A 的 Moore-Penrose 广义逆
$A^\perp$	满足 $A'A^\perp = 0$ 且具有最大秩的矩阵
$\mathrm{rk}(A)$	矩阵 A 的秩
$\mid A \mid$	矩阵 A 的行列式
$\parallel A \parallel$	矩阵 A 的范数
$\mathrm{tr}(A)$	方阵 A 的迹
$\lambda_i(A)$	A 的第 i 个顺序特征根
$\mathcal{M}(\mathrm{A})$	矩阵 A 的列向量张成的子空间
P_A	向 $\mathcal{M}(\mathrm{A})$ 的正交投影变换阵
$\mathbf{1}' = (1, \cdots, 1)$	分量皆为 1 的列向量 ①
$\mathrm{Vec}(A)$	将 A 的列向量依次排成的列向量
$A \otimes B$	A 与 B 的 Kronecher 乘积
$E(X)$	随机变量或向量 X 的均值
$\mathrm{Var}(X)$	随机变量 X 的方差
$\mathrm{Cov}(X, Y)$	随机变量或向量 X, Y 的协方差
$u \sim (\mu, \Sigma)$	均值为 μ, 协方差阵为 Σ 的随机向量
$u \sim N_p(\mu, \Sigma)$	均值为 μ, 协方差阵为 Σ 的 p 维正态向量
LS 估计	最小二乘估计
BLU 估计	最佳线性无偏估计
MVU 估计	最小方差无偏估计
MINQUE	最小范数二次无偏估计
RSS	回归平方和
SS_e	残差平方和
MSE	均方误差
MSEM	均方误差矩阵
GMSE	广义均方误差

① 在不致引起混淆的情况下，本书向量除分量为 1 的向量 $\mathbf{1}$ 用黑体表示外，其余均用白体英文小写字母表示，如 $a, b, \cdots$

目　录

第一章 模型概论

线性模型是一类统计模型的总称，它包括了线性回归模型、方差分析模型、协方差分析模型和线性混合效应模型 (或称方差分量模型) 等. 许多生物、医学、经济、管理、地质、气象、农业、工业、工程技术等领域的现象都可以用线性模型来近似描述. 因此线性模型成为现代统计学中应用最为广泛的模型之一. 本书将系统讨论线性模型统计推断的基本理论与方法.

本章将通过实例引进各种线性模型，使读者对模型的丰富实际背景有一些了解，这将有助于对后面引进的统计概念和方法的理解. 我们先从线性回归模型谈起.

§1.1 线性回归模型

在现实世界中，存在着大量的这样的情况：两个变量例如 X 和 Y 有一些依赖关系. 由 X 可以部分地决定 Y 的值，但这种决定往往不很确切. 常常用来说明这种依赖关系的最简单、直观的例子是体重与身高. 若用 X 表示某人的身高，用 Y 表示他的体重. 众所周知，一般来说，当 X 大时，Y 也倾向于大，但由 X 不能严格地决定 Y. 又如，城市生活用电量 Y 与气温 X 有很大的关系，在夏天气温很高或冬天气温很低时，由于空调、冰箱等家用电器的使用，用电量就高. 相反，在春秋季节气温不高也不低，用电量就相对少. 但我们不能由气温 X 准确地决定用电量 Y. 类似的例子还很多. 变量之间的这种关系称为“相关关系”，回归模型就是研究相关关系的一个有力工具.

在以上诸例中，Y 通常称为因变量或响应变量，X 称为自变量或预报变量. 我们可以设想，Y 的值由两部分组成：一部分是由 X 能够决定的部分，它是 X 的函数，记为 $f(X)$. 在许多情况下，这个函数关系或者是线性的或者是近似线性的，即

$$f(X) = \beta_0 + \beta_1 X, \tag{1.1.1}$$

这里 β_0 和 β_1 是未知参数. 而另一部分则由其它众多未加考虑的因素 (包括随机因素) 所产生的影响，它被看作随机误差，记为 e. 这里 e 作为随机误差，我们有理由要求它的均值 $E(e) = 0$, 其中 $E(\cdot)$ 表示随机变量的均值. 于是，我们得到

$$Y = \beta_0 + \beta_1 X + e. \tag{1.1.2}$$

在这个模型中，若忽略掉 e, 它就是一个通常的直线方程. 因此，我们称 (1.1.2) 为线性回归模型或线性回归方程. 关于“回归”一词的由来，我们留在后面作解释.

常数项 β_0 是直线的截距，β_1 是直线的斜率，也称为回归系数. 在实际应用中，β_0 和 β_1 皆是未知的，需要通过观测数据来估计.

假设自变量 X 分别取值为 $x_1, x_2, \cdots, x_n$ 时，因变量 Y 对应的观测值分别为 $y_1, y_2, \cdots, y_n$. 于是我们有 n 组观测值 (x_i,y_i), $i=1,\cdots,n$. 如果 Y 与 X 有回归关系 (1.1.2), 则这些 (x_i,y_i) 应该满足

$$y_i=\beta_0+\beta_1 x_i+e_i, \qquad i=1,\cdots,n, \tag{1.1.3}$$

这里 e_i 为对应的随机误差. 基于 (1.1.3), 应用适当的统计方法 (这将在第四章讨论) 可以得到 β_0 和 β_1 的估计值 $\hat{\beta}_0,\hat{\beta}_1$, 将它们代入 (1.1.2), 再略去误差项 e_i 得到

$$Y=\hat{\beta}_0+\hat{\beta}_1 X, \tag{1.1.4}$$

称之为经验回归直线，也称为经验回归方程. 这里“经验”两字表示这个回归直线是基于前面的 n 次观测数据 (x_i,y_i), $i=1,\cdots,n$ 而获得的.

例 1.1.1　肥胖是现代社会人们普遍关注的一个重要问题，那么体重多少才算是肥胖呢？这当然跟每个人的身高有关，于是许多学者应用直线回归方法研究人的体重与身高的关系. 假设 X 表示身高 (cm), Y 表示体重 (kg). 我们假设 Y 与 X 之间具有回归关系 (1.1.2). 在这里误差 e 表示除了身高 X 之外，所有影响体重 Y 的其它因素，例如遗传因素、饮食习惯、体育锻炼多少等. 为了估计其中的参数 β_0 和 β_1, 研究者测量了很多人的身高 x_i 和体重 y_i, $i=1,\cdots,n$ 得到关系 (1.1.3). 从而应用统计方法可以估计出 β_0 和 β_1. 一种研究结果是，若用 $X-150$ 作自变量，则得到 $\hat{\beta}_0=50$, $\hat{\beta}_1=0.6$, 也就是说我们有经验回归直线

$$Y=50+(X-150)\times 0.6.$$

我们可以把它改写成如下形式:

$$Y=-40+0.6X, \tag{1.1.5}$$

这个经验回归方程在一定程度上描述了体重与身高的相关关系. 给定 X 的一个具体值 x_0, 我们可以算出对应的 Y 值 $y_0=-40+0.6x_0$. 例如某甲身高 $x_0=160$(cm), 代入 (1.1.5) 可以算出对应 $y_0=56$(kg). 我们称 56kg 为身高是 160cm 的人的体重的预测. 这就是说，对于一个身高 160cm 的人，我们预测它的体重大致为 56kg, 但实际上，它的体重不可能恰为 56kg. 可能比 56kg 多，也可能比 56kg 少.

例 1.1.2　我们知道，一个公司的商品销售量与其广告费有密切关系，一般说来在其它因素 (如产品质量等) 保持不变的情况下，用在广告上的费用愈高，它的商品销售量也就会愈多. 但这也只是一种相关关系. 某公司为了进一步研究这

种关系，用 X 表示在某地区的年度广告费，Y 表示年度商品销售量. 根据过去一段时间的销售记录 (x_i, y_i), $i = 1, \cdots, n$, 采用线性回归模型 (1.1.3)，假定计算出 $\hat{\beta}_0 = 1608.5$, $\hat{\beta}_1 = 20.1$, 于是得到经验回归直线

$$Y = 1608.5 + 20.1X.$$

这个经验回归直线告诉我们，广告费 X 每增加一个单位，该公司销售收入就增加 20.1 个单位. 如果某地区人口增加很快，那么很可能人口总数也是影响销售量的一个重要因素. 若记 X_1 为年度广告费，X_2 为某地区人口总数. 我们可以考虑如下含两个自变量的线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e.$$

同样，根据记录的历史数据，应用适当统计方法可以估计出 β_i, $i = 0, 1, 2$. 假定估计出的

$$\hat{\beta}_0 = 320.3, \quad \hat{\beta}_1 = 18.4, \quad \hat{\beta}_2 = 0.2,$$

则我们得到经验回归方程

$$Y = 320.3 + 18.4X_1 + 0.2X_2.$$

从这个经验回归方程我们可以看出，当广告费 X_1 增加或人口总数 X_2 增加时，商品销售量都增加，且当人口总数保持不变时，广告费每增加 1 个单位，销售量增加 18.4 个单位. 而当广告费保持不变，该地区人口总数每增加一个单位，该公司销售量增达 0.2 个单位. 当然，在实际应用中，并不是每个经验回归方程都能描述变量之间的客观存在的真正的关系. 关于这一点，将在第五章详细讨论.

在实际问题中，影响因变量的主要因素往往很多，这就需要考虑含多个自变量的回归问题. 假设因变量 Y 和 $p-1$ 个自变量 $X_1, \cdots, X_{p-1}$ 之间有如下关系：

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + e, \tag{1.1.6}$$

这是多元线性回归模型，其中 β_0 为常数项，$\beta_1, \cdots, \beta_{p-1}$ 为回归系数，e 为随机误差.

假设我们对 $Y, X_1, \cdots, X_{p-1}$ 进行了 n 次观测，得到 n 组观测值

$$x_{i1}, \cdots, x_{i,p-1}, y_i, \qquad i = 1, \cdots, n,$$

它们满足关系式

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1} + e_i, \qquad i = 1, \cdots, n, \tag{1.1.7}$$

这里 e_i 为对应的随机误差. 引进矩阵记号

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix},$$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix},\quad e=\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

(1.1.7) 就写为如下简洁形式:

$$y=X\beta+e, \tag{1.1.8}$$

这里 y 为 $n\times 1$ 的观测向量. X 为 $n\times p$ 已知矩阵, 通常称为设计矩阵. 对于线性回归模型, 术语 "设计矩阵" 中的 "设计" 两字并不蕴含任何真正设计的含义, 只是习惯用法而已. 几年来, 有一些学者建议改用 "模型矩阵". 但就目前来讲, 沿用 "设计矩阵" 者居多. β 为未知参数向量, 其中 β_0 称为常数项, 而 $\beta_1,\cdots,\beta_{p-1}$ 为回归系数. 而 e 为 $n\times 1$ 随机误差向量, 其均值为零, 即 $E(e_i)=0$. 关于 e 最常用的假设是:

(a) 误差项具有等方差, 即

$$\mathrm{Var}(e_i)=\sigma^2,\qquad i=1,\cdots,n,$$

(b) 误差是彼此不相关的, 即

$$\mathrm{Cov}(e_i,e_j)=0,\qquad i\neq j,\quad i,j=1,\cdots,n.$$

通常称以上两条为 Gauss-Markov 假设. 我们知道, 一个随机变量的方差刻画了该随机变量取值散布程度的大小, 因此假设 (a) 要求 e_i 等方差, 也就是要求不同次的观测 y_i 在其均值附近波动程度是一样的. 这个要求有时显得严厉些. 在一些情况下, 我们不得不放松为 $\mathrm{Var}(e_i)=\sigma_i^2,\ i=1,\cdots,n$. 假设 (b) 等价于要求不同次的观测是不相关的. 在实际应用中这个假设比较容易满足.

模型 (1.1.8) 和 Gauss-Markov 假设合在一起, 可简洁地表示为

$$y=X\beta+e,\qquad E(e)=0,\qquad \mathrm{Cov}(e)=\sigma^2 I, \tag{1.1.9}$$

这里 $\mathrm{Cov}(e)$ 表示随机向量 e 的协方差阵. (1.1.9) 就是我们以后要讨论的最基本的线性回归模型.

在一些实际问题中, $\mathrm{Var}(e_i)=\sigma_i^2,\ i=1,\cdots,n$. 这里 σ_i^2 可能不全相等. 这时观测向量或误差向量的协方差阵形为

$$\mathrm{Cov}(e)=\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{pmatrix}. \tag{1.1.10}$$

在经济问题中, $y_1,y_2,\cdots,y_n$ 表示某经济指标在 n 个不同时刻的观测值, 它们往往是相关的. 这种相关性反应在误差项上, 就是误差项的自相关性. 一种最简单的自相关关系是误差为一阶自回归形式, 即

$$e_i=\varphi e_{i-1}+\varepsilon_i,\qquad |\varphi|<1,$$

其中 $\varepsilon_i,\ i=1,\cdots,n$ 是独立同分布的随机变量, $E(\varepsilon_i)=0$, $\mathrm{Var}(\varepsilon_i)=\sigma_\varepsilon^2$. 这时

$$\mathrm{Cov}(e)=\frac{\sigma_\varepsilon^2}{1-\varphi^2}\begin{pmatrix}1 & \varphi & \cdots & \varphi^{n-1}\\ \varphi & 1 & \cdots & \varphi^{n-2}\\ \vdots & \vdots & & \vdots\\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1\end{pmatrix}. \tag{1.1.11}$$

上面我们讨论的都是线性回归模型. 有一些模型虽然是非线性的, 但经过适当变换, 可以化为线性模型.

例 1.1.3 在经济学中, 著名的 Cobb-Douglas 生产函数为

$$Q_t=aL_t^bK_t^c,$$

这里 Q_t, L_t 和 K_t 分别为 t 年的产值、劳力投入量和资金投入量, a, b 和 c 为参数, 在上式两边取自然对数, 得到

$$\ln(Q_t)=\ln(a)+b\ln(L_t)+c\ln(K_t)\,.$$

若令

$$y_t=\ln(Q_t),\ x_{t1}=\ln(L_t),\ x_{t2}=\ln(K_t)\,,$$

$$\beta_0=\ln(a),\ \ \beta_1=b,\ \ \beta_2=c\,,$$

再加上误差项，便得到线性关系

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t\,,$$

因此我们把原来的非线性模型化成了线性模型.

例 1.1.4　多个自变量的多项式

我们知道，任何光滑函数都可以用足够高阶的多项式来逼近. 因此，当因变量 Y 和诸自变量之间的关系不是线性关系时，我们可以用多元多项式来近似，有时可能还要添加若干自变量的交叉积. 例如

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + e\,.$$

这样的模型往往出现在化学工程领域的研究之中，其目的是求诸自变量的一个组合，使得因变量 Y 达到最大或最小. 这类问题称为响应曲面设计.

引进新变量 $X_3 = X_1^2, X_4 = X_2^2, X_5 = X_1 X_2$，上述模型变成了一个线性模型. 从这里我们可以看出，线性模型中“线性”二字实质上是指 Y 关于未知参数 β_i 的关系是线性的.

最后，我们解释一下“回归”一词的由来. “回归”英文为“ regression ”，是由英国著名生物学家兼统计学家 Galton(高尔顿) 在研究人类遗传问题时提出的. 为了研究父代与子代身高的关系，Galton 收集了 1078 对父亲及其一子的身高数据. 用 X 表示父亲身高，Y 表示儿子身高. 单位为英寸 (1 英寸为 2.54cm). 将这 1078 对 (x_i, y_i) 标在直角坐标纸上，他发现散点图大致呈直线状. 也就是说，总的趋势是父亲的身高 X 增加时，儿子的身高 Y 也倾向于增加，这与我们的常识是一致的. 但是，Galton 对数据的深入分析，发现了一个很有趣的现象 —— 回归效应.

因为这 1078 个 x_i 值的算术平均值 $\bar{x} = 68$ 英寸，而 1078 个 y_i 值的平均值为 $\bar{y} = 69$ 英寸，这就是说，子代身高平均增加了 1 英寸. 人们自然会这样推想，若父亲身高为 x，他儿子的平均身高大致应为 $x + 1$，但 Galton 的仔细研究所得结论与此大相径庭. 他发现，当父亲身高为 72 英寸时 (请注意，比平均身高 $\bar{x} = 68$ 要高)，他们的儿子平均身高仅为 71 英寸. 不但达不到预期的 72+1=73 英寸，反而比父亲身高低了 1 英寸. 反过来，若父亲身高为 64 英寸 (请注意，比平均身高 $\bar{x} = 68$ 要矮)，他们儿子平均身高为 67 英寸，竟比预期的 64+1=65 英寸高出了 2 英寸. 这个现象不是个别的，它反映了一个一般规律：即身高超过平均值 $\bar{x} = 68$ 英寸的父亲，他们儿子的平均身高将低于父亲的平均身高. 反之，身高低于平均身高 $\bar{x} = 68$ 英寸的父亲，他们儿子的平均身高将高于父亲的平均身高. Galton 对这个一般结论的解释是：大自然具有一种约束力，使人类身高的分布在一定时期内相对稳定而不产生两极分化，这就是所谓的回归效应. 通过这个例子，Galton 引进了“回归”