Concepts and Controversies 8th Edition

统计学的世界

第8版

下

[美] 戴维·穆尔（David S. Moore）
威廉·诺茨（William I. Notz） ◎著
郑磊◎译

Concepts and Controversies 8th Edition

中信出版集团

STATISTICS 统计学的世界 下

第8版 Concepts and Controversies 8th Edition

[美] 戴维·穆尔（David S. Moore） 威廉·诺茨（William I. Notz）◎著　郑磊◎译

中信出版集团 · 北京

图书在版编目（CIP）数据

统计学的世界：第8版/（美）戴维·穆尔，（美）威廉·诺茨著；郑磊译. --北京：中信出版社，2017.9

书名原文：Statistics: Concepts and Controversies 8th Edition

ISBN 978-7-5086-6672-3

I. ①统… II. ①戴… ②威… ③郑… III. ①统计学－通俗读物 IV. ①C8-49

中国版本图书馆CIP数据核字（2016）第214641号

统计学的世界（第8版）

著　　者：[美]戴维·穆尔　[美]威廉·诺茨
译　　者：郑　磊
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座　邮编　100029）
承 印 者：三河市西华印务有限公司

开　　本：787mm×1092mm　1/16　　印　　张：45　　字　　数：700千字
版　　次：2017年9月第1版　　印　　次：2017年9月第1次印刷
京权图字：01-2014-8548　　广告经营许可证：京朝工商广字第8087号
书　　号：ISBN 978-7-5086-6672-3
定　　价：148.00元（全二册）

第 15 章
描述相关关系：回归、预测与因果关系

案例分析

预测股市的走势可能让你发财，难怪有那么多人都埋头在股市信息里。

确实有些令人匪夷所思的方法。“超级碗指标”指的是每年 1 月或 2 月初举办的超级碗橄榄球赛可以预测该年股市的表现。美国国家橄榄球联盟（NFL）由原来的NFL和美国橄榄球联盟（AFL）合并而成。超级碗指标声称，若原本属于NFL的球队赢了超级碗，该年股市就会上涨；若原本属于AFL的球队赢了，股市就会下跌。从 1967 年第一届超级碗至 2011 年的 45 年间，用这个指标所做的股市预测中有 35 次是正确的。（我们把巴尔的摩乌鸦队看作老的NFL球队，因为球队在来到巴尔的摩之前是克利夫兰布朗队。我们把坦帕湾海盗队也视为原本属于NFL的球队，但它既不是一只待合并的球队，最初也不是NFL的球队，而是AFL的球队。）这个指标的预测正确率达到 75%，令人印象深刻。

“昨天（2012 年 2 月 5 日）一支NFL球队——纽约巨人队赢得了超级碗冠军，根据这个指标，今年的股票将会上涨。那么，我该投资股票吗？”

在这一章，我们将学习如何通过其他变量来预测某个变量的统计方法，而不只是数那些上上下下的点。我们还将学习变量之间的因果关系。学完这一章，你就能够对超级碗指标做出评价了。

回归直线

如果散点图显示出两个数值变量之间的线性相关关系，我们会在散点图中画一条直线，来对这个整体形态进行描述。“回归直线”（regression line）可以对两个变量间的关系进行描述，但条件是：其中一个变量可以用来解释或预测另一个变量。也就是说，回归直线描述的是一个解释变量和一个反应变量之间的相关关系。

回归直线

回归直线是一条直线，描述当解释变量x的值改变时，反应变量y的值会发生怎样的变化。我们常用回归直线来预测对于某一个给定的x值，y值是什么。

例 1　始祖鸟化石标本

始祖鸟化石的两种骨头的长度之间存在线性相关关系。图 15–1 展示了 5 件标本的两种骨头长度，图中的直线对于整体形态做了简要描述。

还有一件始祖鸟化石不完整，股骨长 50 厘米，肱骨却不见了。我们能不能猜出肱骨有多长呢？肱骨和股骨之间的线性相关关系非常强，使得我们可以放心地用股骨长度来预测肱骨长度。图 15–1 告诉我们可以这样做：从股骨长度（50 厘米）开始，在这一点的正上方找到和直线相交的点，然后查看纵轴上对应的值，我们就可以得到肱骨长度大约是 56 厘米。如果代表这件化石的这个点确实是在这条直线上，肱骨长度就应该是这个数值了。也就是说，我们的这个预测会相当准。

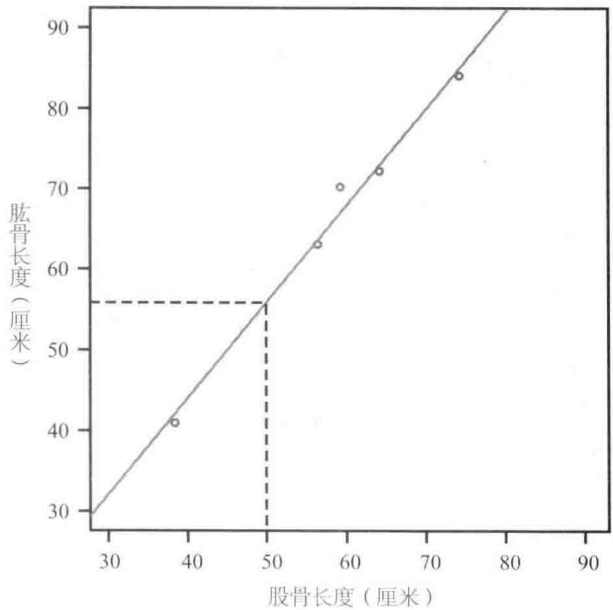

图 15–1　用回归直线来做预测

例 2　总统选举

共和党的罗纳德·里根当过两届美国总统，分别是在 1980 年和 1984 年。他的减税政策刺激了经济发展，带来税收收入的增加。图 15–2 展示了里根的竞争对手民主党候选人吉米·卡特（1980 年）和沃尔特·蒙代尔（1984 年）在各个州的支持率，并显示出正线性相关关系。我们预计会存在这种现象，因为一些州倾向于支持民主党，而另一些州倾向于支持共和党。图中只有一个异常值，即卡特的家乡佐治亚州，1980 年有 56% 的选票投给了民主党的卡特，而 1984 年只有 40% 的选票投给了民主党。

我们可以用图 15–2 上的回归直线，根据 1980 年的投票结果预测某个州 1984 年的投票情况。这个图里的点，相比图 15–1 来说，分布得离直线较远。度量线性相关程度的相关系数为 r，在图 15–1 里，r=0.994，而在图

15–2 里，r=0.704。因此可见对选举结果的预测，一般来说其准确度要比预测始姐鸟肱骨长度要差。

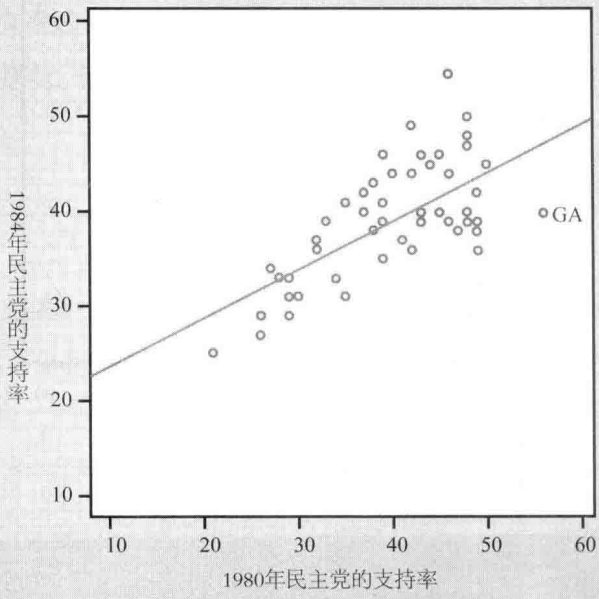

图 15–2　较弱的线性相关关系

回归方程式

当散点图显示出像图 15–1 那么强的线性相关关系时，用目测法画一条接近所有点的直线是很容易的。然而对图 15–2 来说，不同的人用目测法，可能会画出很不一样的直线。因为我们想用x来预测y，所以我们想要的直线，是在垂直方向（和y轴平行的方向）上和点尽量接近。在用目测法画直线时，很难只顾及点和直线的垂直距离。而且，用目测法只能在图上画出直线，却得不到线性方程式。我们需要找一个办法，根据数据找出垂直方向上距离那个点最近的线性方程式。有许多不同方法可以使垂直距离“越小越好”，其中最常用的就是“最小二乘法”（least-squares）。

用最小二乘法找出回归直线

用最小二乘法找到的回归直线，是使所有数据点距离直线的垂直距离的平方和最小的直线。

图 15–3 展示了最小二乘法的概念。这个图把图 15–1 的中间部分放大，聚焦在三个点上。图中画出了这三个点与回归直线之间的垂直距离。要用最小二乘法找出回归直线，就必须用到所有的垂直距离，把每一个距离值平方，然后移动直线，直到距离平方和的值达到最小。图 15–1 和图 15–2 的散点图中所画的直线，就是用最小二乘法找到的回归直线。我们无须列出计算公式，这是电脑的工作。

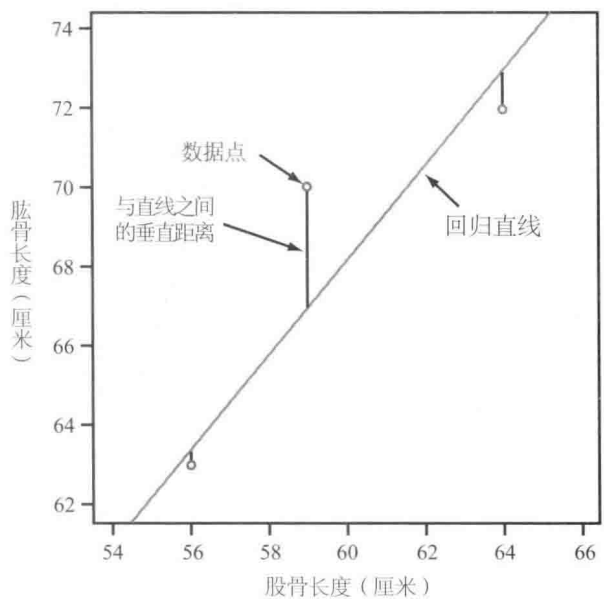

图 15–3　找到回归直线的目的是从 x 预测出 y

要写出这个线性方程式，还像以前一样，x 代表解释变量，y 代表反应变量。方程式如下：

$$y=a+bx$$

b是直线的“斜率”（slope），就是x增加一个单位时y的改变量。a是“截距”（intercept），是当$x=0$时y的值。要利用这个方程式做预测，只要把x值代入方程式中，计算出y值即可。

知识普及　向平均数回归

“回归”这个词本来的意思是往回走，为什么利用解释变量来预测反应变量的统计方法要叫作“回归”呢？最先把回归方法用在生物与心理学数据上的高尔顿爵士，检视了诸如儿童身高与其父母身高之间的相关关系等。他发现身高超过平均数的父母，通常孩子的身高也超过平均数，但是并没有父母那么高。高尔顿称这种现象为“向平均数回归”，之后这种统计方法便以此命名。

例 3　应用回归方程式

在例 1 中，我们用一种简便的方法预测了股骨为 50 厘米的化石的肱骨长度。其线性回归方程式是：

$$\text{肱骨长度} = -3.66 + (1.197 \times \text{股骨长度})$$

这条直线的斜率是$b=1.197$。这表示对于这些化石来说，股骨长度每增加 1 厘米，肱骨长度就会增加 1.197 厘米。回归直线的斜率对于理解数据来说通常很重要，斜率是变化率，即当x增加一个单位时y的改变量。

线性回归方程式的截距是$a=-3.66$，它是当$x=0$时y的值。虽然要画出直线需要知道截距，但是只有当x的值实际上有可能接近于 0 时，截距才有

统计意义。而股骨长度不可能是 0，所以截距没有统计意义。

要用方程式来做预测，只要把x值带入方程式中算出y即可。对应 50 厘米长的股骨，化石的肱骨长度预测值是：

$$肱骨长度 = -3.66 + 1.197 \times 50 = 56.2 \text{ 厘米}$$

要在散点图上画出这条直线的话，用两个不同的x值分别计算出y值，就可以得到两个点，把它们连接起来就是我们要的直线了。

练习

15.1　始祖鸟化石的肱骨长度。用线性回归方程式

$$肱骨长度 = -3.66 + 1.197 \times 股骨长度$$

预测一件股骨为 70 厘米长的始祖鸟化石的肱骨长度。

了解预测的意义

电脑使预测变得很容易而且是全自动的，即使对大量的数据而言也是一样。任何可以用全自动方式处理的事，处理时通常是不经过思考的。比如，即使数据之间存在曲线相关关系，回归软件仍然“乐于”给它们匹配（fitting）一条直线。此外，电脑也不能自行决定谁是解释变量，谁是反应变量。这一点很重要，因为如果解释变量不同，同一组数据会呈现出两条不一样的直线。

在实际应用时，我们常常用多个解释变量来预测一个反应变量。大学在处理入学申请时，可能会用学术能力评估测试的数学与阅读分数，再加上高中时期的英语、数学与科学成绩（共 5 个解释变量）来预测大一新生的表现。虽然细节很复杂，但

是所有用来预测反应变量的统计方法，都和线性回归方程式有一些共同的基本性质。

• 预测根据的是为数据匹配的某个“模型”（model）。在图 15–1 和图 15–2 里，模型就是穿过散点图中的点的一条直线。其他的预测方法会使用较复杂的模型。

• 模型离数据点越近，预测结果越好。比较图 15–1 和图 15–2，前者中的点距离直线很近，而后者则不是这样，所以图 15–1 的预测比较可靠。当变量多的时候，形态不容易看出来，而且只要数据没有呈现出很明显的整体形态，预测可能就会很不准。

• 超出现有数据范围的预测是靠不住的。假设你手上有 3~8 岁孩童的生长资料，你发现年龄x和身高y之间有很强的线性相关关系。如果你为这些数据匹配一条回归直线，然后用它来预测这些孩子 25 岁时的身高，你的预测结果将是，这个孩子 25 岁时会有 8 英尺高。人到了某个年龄阶段，长高的速度会慢下来，最后会完全停止长高，所以把直线一直延长到成人的年龄是很可笑的做法，没有人在预测身高时会犯这种错。但是，几乎所有的经济预测都在试图告诉我们下一季度或下一年会发生什么事，难怪经济预测常常出错。在可得的数据范围之外做预测，这种方法被称作“外推法”（extrapolation）。使用外推法要小心！

知识普及　计算选票的人有没有作弊?

在 1993 年宾夕法尼亚州的选举中，根据投票机的计数，共和党的布鲁斯·马克斯领先民主党的威廉·斯廷森。但是，在控制选举委员会的民主党人计算了缺席投票者的选票后，结果又变成了斯廷森领先。事情闹上了法庭。法庭传唤了一位统计学家，他用过去的选举数据绘制出回归直线，再根据投票机结果，预测缺席选票的计数。根据马克斯在投票机计数部分领先的 564 票，可以预测他应该比斯廷森多得 133 张缺席选票。而选举委员会计算出来的是斯廷森比马克斯多得了 1 025 张缺席选票。计算选票的人有没有作弊?

例 4　预测财政预算

美国国会预算办公室每年必须发布报告，预测未来 5 年的联邦预算及其盈余或赤字。这些预测和未来的经济趋势（未知）有关，也和国会对税收和开支的决定（也是未知的）有关。即便目前政策都不变，要预测预算状况也会非常不准确。比如，2004 年对 2008 年做的联邦预算预测，少算了近 1 770 亿美元。2005 年所预测的 2009 年联邦预算居然比实际少了 11 930 亿美元！正如参议员埃弗里特·德克森曾说的那样，“这里差 10 亿，那里差 10 亿，便谬以千里了”。1999 年，预算办公室预测接下来的 10 年会有 9 960 亿美元的财政盈余（不考虑社会保险）。政客们已经在讨论怎么用这笔钱了，但其他人并不相信这个预测。

相关系数与回归直线

相关系数度量线性相关关系的方向和强度，回归直线可以描述这种相关关系。相关系数和回归直线是密切相关的，即使回归直线需要选择解释变量而相关系数不需要。

相关系数和回归直线都会受异常值的严重影响。如果你的散点图有明显的异常值，你就要小心了。图 15–4 展示的是美国各州的年度最高降水量纪录和单日最高降水量纪录。夏威夷是位于图的高处的异常值，记录表明 1982 年夏威夷的年度降水量达到 704.83 英寸。

图 15–4 里所有 50 个州的相关系数是 0.510，如果把夏威夷去掉，相关系数会降为 0.248。图里面的实线，是 50 个州的回归直线。如果不计入夏威夷，回归直线就会往下落到虚线的位置。这条虚线差不多接近于水平，也就是说，一旦我们决定去除夏威夷的异常值，年度最高降水量纪录和单日最高降水量纪录之间就

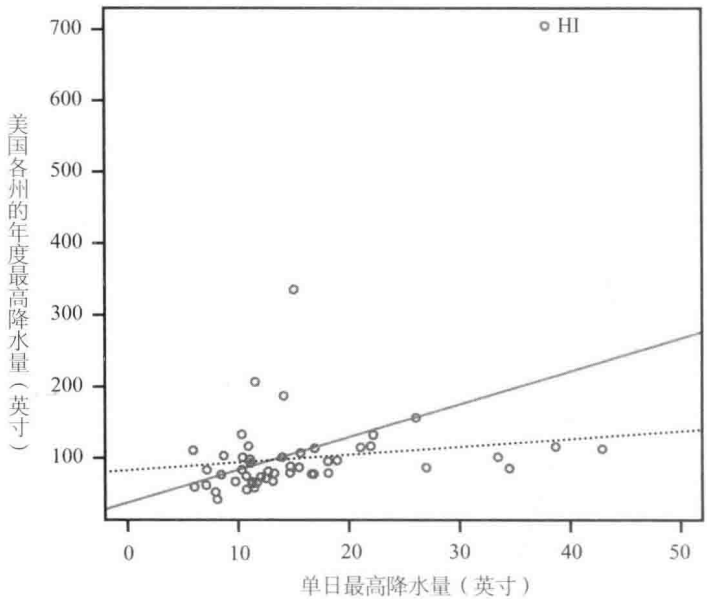

图 15–4　回归直线受异常值的严重影响。实线是根据全部 50 个数据点画的，虚线则去除了夏威夷的异常值

没有多大关系了。

回归直线的预测功能，视相关关系的强度而定。也就是说，一条回归直线有多大用处，和变量之间的相关系数密切相关。事实上，这个关系就是用相关系数的平方来度量的。

相关系数的平方 r^2

相关系数的平方 r^2，是 y 的变异值当中，可以用 y 对 x 的线性回归方程式来解释的那一部分所占的比例。

也就是说，当 y 和 x 存在线性相关关系时，y 的变异值中的一部分可以解释为，当 x 改变时 y 也随着一起改变。

例 5　r^2 的用法

再看一下图 15–1。这 5 件化石的肱骨长度的变异性很大，最短的是 41 厘米，最长的是 84 厘米。从散点图上可以看出，我们只要看看股骨长度和回归直线，就几乎可以解释所有的变异值了。当股骨长度增加时，肱骨长度也会随之增加。除此之外，肱骨长度的变异值就没剩几个了。剩下的这些变异值，从图上看，就是与直线还有些距离的点。因为这组数据的 r=0.994，所以 r^2=0.994^2=0.988，也就是说，由于股骨长度增加而使肱骨长度也随之增加，可以解释肱骨长度 98.8% 的变异值。散布在直线两侧的点只是剩下的 1.2% 的变异值，这说明预测得很准。

再看一下图 15–2。1980 年和 1984 年的民主党支持率之间虽然存在线性相关关系，但是点在直线两侧的位置比较分散。这组数据的 r=0.704，r^2=0.496，我们观察到的 1984 年民主党支持率的变异值，大约只有一半可以用回归直线来解释。把 1980 年民主党支持率为 45% 的州和支持率为 30% 的州做比较，你还是会预测前者在 1984 年的民主党支持率更高。但是，在 1980 年民主党支持率相同的各州，1984 年的支持率有不小的变异。造成这部分变异的是其他原因，诸如两次选举的主要议题不同，以及里根的两位民主党竞争对手来自不同地区等。

通常在报告回归直线时，也会同时提到 r^2 的值，它被当作回归直线预测反应变量有多成功的一个指标。当你看到一个相关系数的时候，把它平方，你会更清楚相关性的强度。完全相关系数（r= –1 或 r=1）代表所有的点都落在一条直线上，此时 r^2=1，表明一个变量的所有变异值，都可以用它和另一个变量的线性相关关系来说明。若 r= –0.7 或 r=0.7，则 r^2=0.49，表明只有差不多一半的变异值可以用线性相关关系来解释。以 r^2 的值为标准的话，相关系数 ±0.7 差不多在 0 和 ±1 的中间。

练习

15.2 棒球场。表 14–2 给出了大联盟棒球赛各个场地一瓶 16 盎司汽水的价格和一个热狗的价格。它们之间的相关系数r=0.45。热狗价格中有多大比例的变异值可以由热狗价格与 16 盎司汽水价格的线性回归方程式来解释?

因果关系

抽烟和肺癌死亡率之间有很强的相关性，那么，是不是抽烟导致人们患肺癌呢？在一个国家里，容不容易取得手枪和该国枪杀事件的发生率之间也有很强的相关性，那么，容易取得手枪是否导致发生更多谋杀案？香烟包装上已明白写着吸烟导致癌症，而有更多的人拥有手枪是否导致更多谋杀案却引起了热烈的辩论。为什么呢？我们已经知道统计数据中与因果关系有关的三大事实。

统计数据与因果关系

- 即使两个变量间有很强的相关性，也不一定意味着改变其中一个变量的值会引起另一个变量值的改变。
- 两个变量之间的相关性，常常受其他潜在变量的影响。
- 证明存在因果关系的最好证据，来自随机比较实验。

例 6 看电视会延长人们的预期寿命吗?

统计一下世界各国人均拥有的电视机台数x和民众的预期寿命y，你会发现两者之间存在很强的正相关关系：人均拥有电视机数量多的国家，其民众的预期寿命也比较长。

因果关系的基本意义是，只要改变x的值，就可以使y的值改变。我们能不能运一堆电视机到博茨瓦纳，以延长那里的民众预期寿命呢？当然不行。富国的电视机数量比穷国多，而富国民众的预期寿命之所以长，是因为他们有较好的营养条件、干净的水以及较好的医疗资源。电视机数量与预期寿命之间没有因果关系。

例 6 说明了三大事实的头两项。这类相关被叫作“胡说相关”：相关是事实，胡说的部分是“改变其中一个变量的值会导致另一个变量值的改变”的结论。像例 6 中的国家财富这种潜在变量会同时影响x和y的值，形成x和y之间的强相关关系，即使x和y之间其实并没有什么直接的关系。我们称其为“共同反应”（common response），即解释变量和反应变量都会对某个潜在变量产生反应。

“依照第三世界的新脱贫计划，援助组织今天开始送出 100 000 台电视机。”