

MAT220 Lecture 18: Linear Correlation (线性相关)

Contents

1	什么是相关?	2
1.1	散点图	2
1.2	线性相关	3
1.3	相关的强度与符号	4
1.4	符号与强度	5
2	相关系数	5
2.1	样本相关系数	5
2.2	r 的公式	7
2.3	理解协方差	8
2.4	理解 r 的公式	10
2.5	由完全正相关得到 $r = 1$	11
2.6	例: 计算 r	12
3	相关的局限性	13
3.1	r 未定义的情况	13
3.2	非线性相关	14
3.3	没有因果关系的相关	14

在前两次课中, 我们同时对两个总体进行了假设检验。在本次课中, 我们将以一种完全不同的方式使用配对数据: 我们不再检验总体参数之间的差异, 而是研究配对数据的描述统计量。我们将讨论线性相关, 即判断两组数据是否呈现出类似直线的模式。我们将首先从直观的角度, 也就是从图形的角度开始讨论, 然后研究如何用数学方法描述线性关系。

1 什么是相关?

回顾第16次课，我们讨论了配对总体的概念。简单来说，在这样的总体中，一个总体中的每个数值在另一个总体中都有一个自然对应的数值：



配对数据可以通过多种方式产生：

- 对同一组对象进行连续实验，例如比较摄入咖啡因前后的反应时间；
- 对同一对象进行不同的测量，例如在同一次跑步过程中，分别使用胸带和Apple Watch 测量心率；
- 自然对称或互相匹配的对象，例如双胞胎中的第一个孩子和第二个孩子，或者同一个人的左脚和右脚。

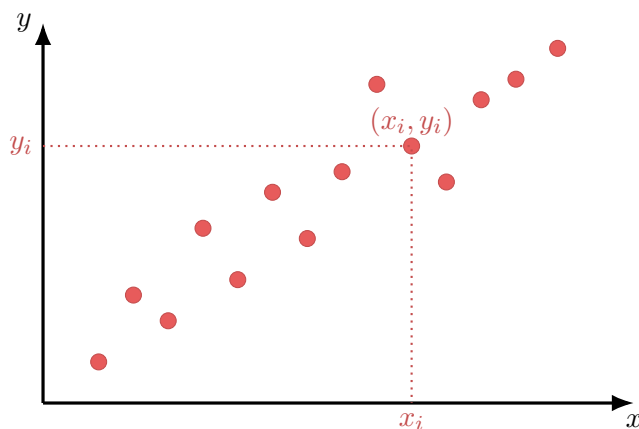
以前，我们利用配对关系构造随机变量的差，并将其写为 $D = X - Y$ 。然后，我们用它来检验 X 和 Y 的均值是否不同。今天，我们只研究配对数据的描述。我们首先介绍一种称为散点图的图形描述方法。

1.1 散点图

假设我们抽取了一个包含 n 对数据的样本。可以将其写成以下列表：

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

散点图是数据的一种图形表示，其中每一对 (x_i, y_i) 都被绘制为二维空间中的一个点。横轴表示变量 x ，纵轴表示变量 y ，因此每个点都代表一对数据：



实际上，每当我们以两种不同的方式测量同一个对象时，就会产生这种数据。通过绘制配对观测值，我们希望能够直观地检查数据中是否存在某种模式。在数学建模中，我们经常尝试使用变量 X 来解释变量 Y 。因此，我们将 x 称为解释变量，将 y 称为响应变量。¹

散点图

散点图是配对定量数据的图形，其中每一对 (x_i, y_i) 都被绘制为一个点。变量 x 称为**解释变量**，变量 y 称为**响应变量**。

这些术语表明了我们正在研究的问题类型。我们怀疑 x 的变化可能有助于解释 y 为什么发生变化。例如，假设 x 表示汽车的速度，而 y 表示其制动距离。我们可以合理地认为，速度有助于解释制动距离，并且制动距离会根据某种潜在规律响应速度的变化。

1.2 线性相关

“相关”一词表示两个变量共同呈现某种模式。原则上，两个变量之间的模式可以有許多不同的形状，也就是说，两组数据可以通过许多不同的方式相关。在本次课中，我们将研究最简单的情况：线性相关。粗略来说，当数据点的模式大致沿着一条直线分布时，数据就具有线性相关。

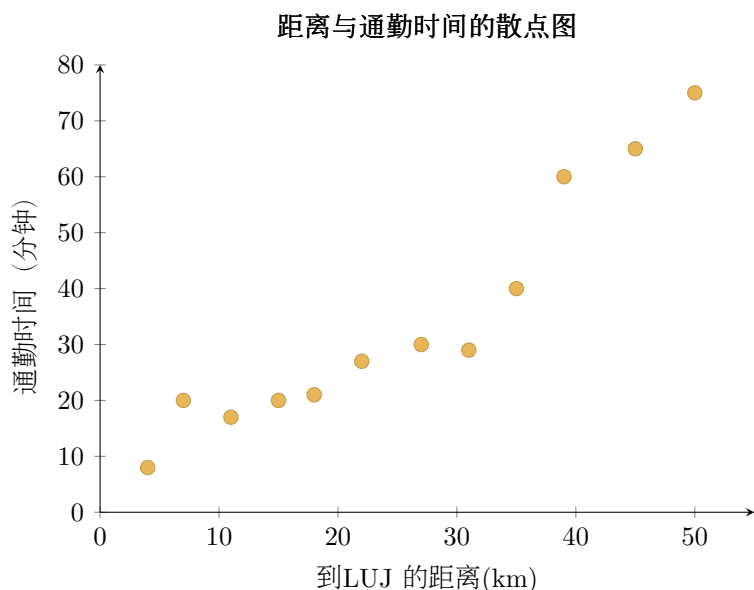
线性相关

当散点图呈现出整体上类似直线的模式时，配对数据表现出**线性相关**。

线性关系非常常见。例如，地理区域的人口与GDP、拉伸弹簧所需的力，以及耗氧量与能量消耗之间都可能存在线性关系。在第19次课中，我们将介绍线性回归：一种寻找最能近似线性模式的直线的方法。

例如，假设我们调查了12名学生，并记录他们的住所到学校的距离以及通勤到学校所需的时间。这些数据自然是配对的，因为两个数值都与同一名学生有关。数据和对应的散点图如下所示。

学生	距离 (km)	时间 (分钟)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75

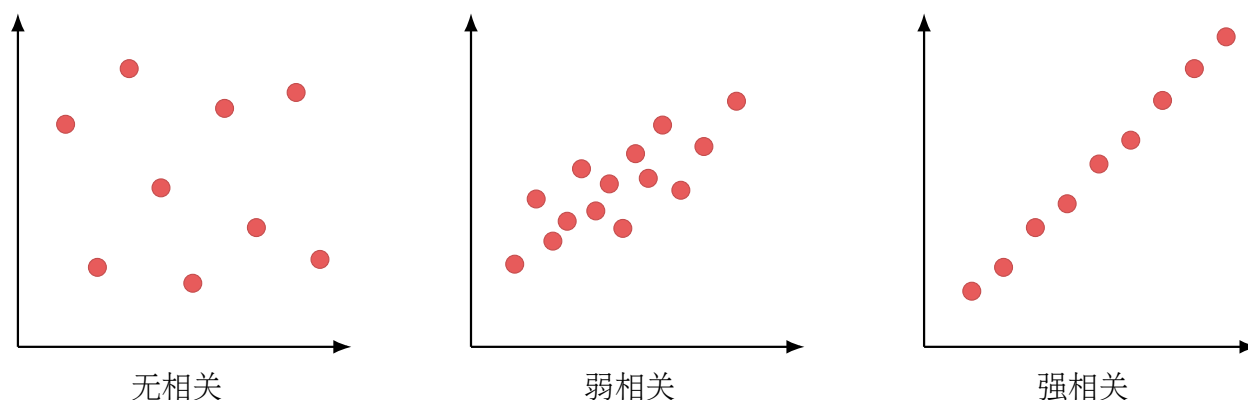


¹在文献中，这两个变量还有其他常见名称。有时， x 被称为自变量，而 y 被称为因变量。

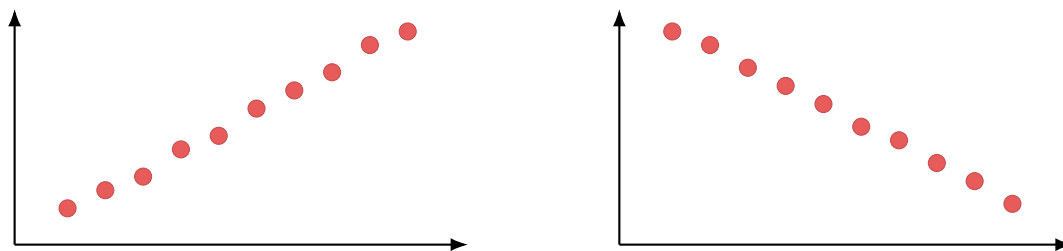
可以看到，数据点呈现出清晰的线性模式：住得离学校较近的学生通勤时间较短，而住得较远的学生通勤时间较长。当然，这些数据并没有完美地排列在一条直线上，这也是可以预料的。通勤时间会受到许多较小因素的影响，例如铁路线路、交通方式等。然而，根据这些数据，我们仍然可以看到一个总体模式。

1.3 相关的强度与符号

一般来说，有些散点图可能完全没有明显的模式，而另一些则非常接近一条直线。在日常语言中，我们使用相关的“强度”来描述这些差异。粗略来说，如果模式十分清晰，就称相关“较强”；如果仍然存在某种模式，但并不十分明显，就称相关“较弱”：



除此之外，线性相关还有另一个值得注意的重要特征。考虑下面两幅图：



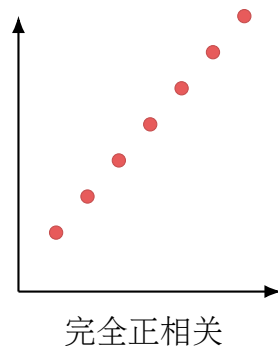
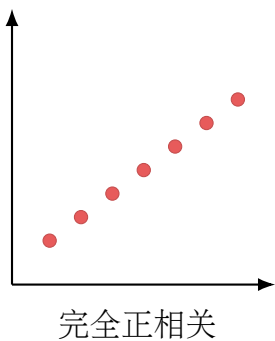
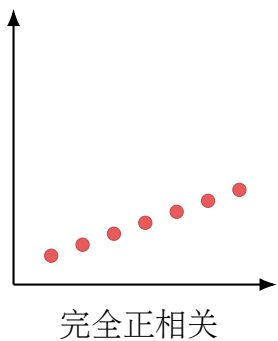
这两幅图都表现出较强的线性相关，但它们显然不同。在左图中， x 和 y 一起增加：当 x 较小时， y 也较小；当 x 较大时， y 也较大。相比之下，右图中的关系是相反的：当 x 较小时， y 较大；当 x 较大时， y 较小。换句话说，拟合这两组数据的直线具有不同符号的斜率。这一观察引出了以下术语。

相关的符号

相关的符号描述总体趋势的方向：

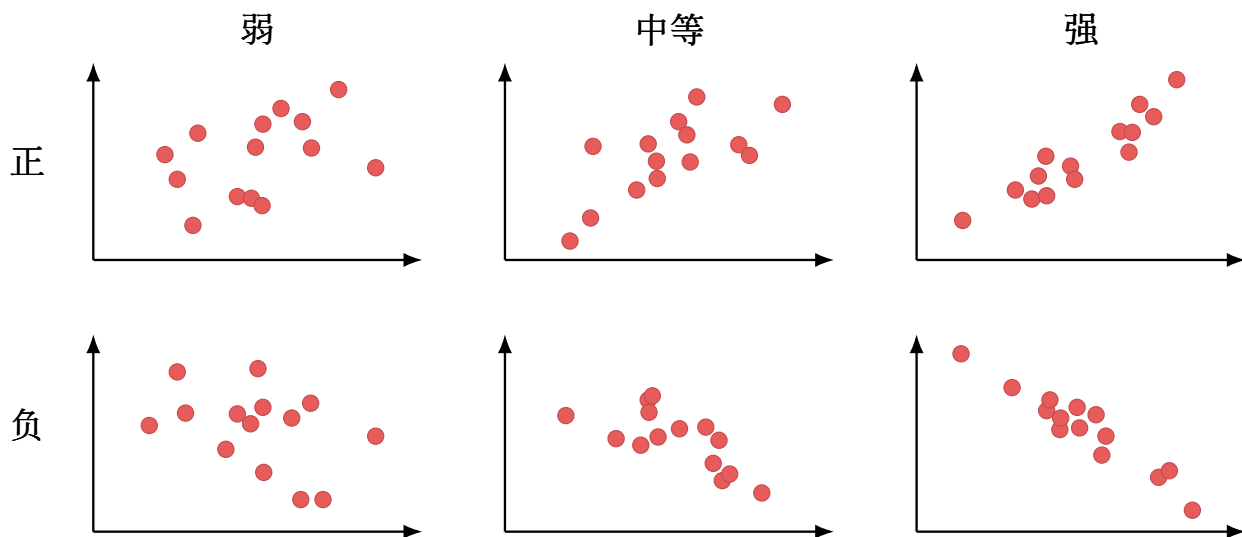
- 正相关： x 和 y 倾向于一起增加，因此总体趋势具有正斜率；
- 负相关： 当 x 增加时， y 倾向于减小，因此总体趋势具有负斜率。

请注意，这里的“正”和“负”只表示模式的符号，而不是斜率的大小。例如，下面三幅图的斜率不同，但都表现出完全的正线性相关：



1.4 符号与强度

符号和强度是线性相关的两个不同特征。如下图所示，无论符号为正还是为负，都可能出现弱、中等或强的模式。



2 相关系数

2.1 样本相关系数

尽管“弱”和“强”等自然语言表述很直观，但它们有些模糊。为了解决这一问题，我们现在引入一个数值来量化相关的强度。简单来说，我们将定义一个衡量相关强度的数：相关越强，该数的绝对值越大；相关越弱，该数的绝对值越小。这个数通常称为样本相关系数，记作 r 。根据其定义，它同时衡量变量 x 和 y 之间线性关系的强度与符号。

样本相关系数

样本相关系数 r 是一个无单位的数，并满足

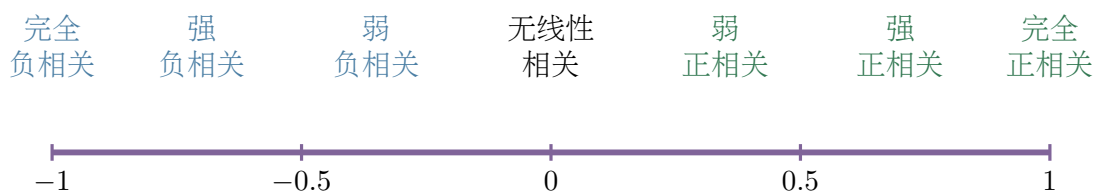
$$-1 \leq r \leq 1.$$

r 的符号表示线性相关是正还是负，而 r 的绝对值表示线性关系的强度。

“样本相关系数”这一术语可以分解为三个部分：

- “系数”表示 r 是一个数；
- “相关”表示 x 和 y 之间的模式；
- “样本”表示 r 是由样本数据 (x_i, y_i) 构造的。

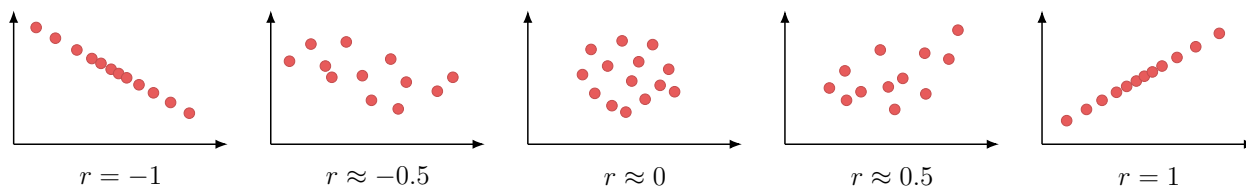
因此，“样本相关系数”实际上就是“衡量样本在多大程度上符合直线模式的数”。使用 r ，我们可以更精确地描述线性相关的强度：



用文字表示，重要的参考值如下：

- $r = 1$ 表示数据完美地排列在一条具有正斜率的直线上；
- $0 < r < 1$ 表示存在某种正线性相关；
- $r = 0$ 表示数据完全没有线性相关；
- $-1 < r < 0$ 表示数据中存在某种负线性相关；
- $r = -1$ 表示完全负线性相关。

下列图形展示了具有几个不同 r 近似值的数据：



即使 x 和 y 之间存在精确的数学规律，现实数据也很少会完美地遵循该规律。测量中包含噪声、个体差异和小误差。因此，完全线性相关最好被看作一种理想化的理论极限。

尽管本课程不会直接讨论它，但值得一提的是，也存在一个用于衡量线性相关的总体参数。与均值或标准差类似，我们使用希腊字母表示总体参数。在相关问题中，我们使用字母 ρ ：

	样本	总体
均值	$\bar{x}$	μ
标准差	s	σ
相关系数	r	ρ

样本相关系数 r 衡量给定样本中各点对之间的线性关系。相比之下，总体相关系数 ρ 描述总体中所有数据对之间的真实相关。

2.2 r 的公式

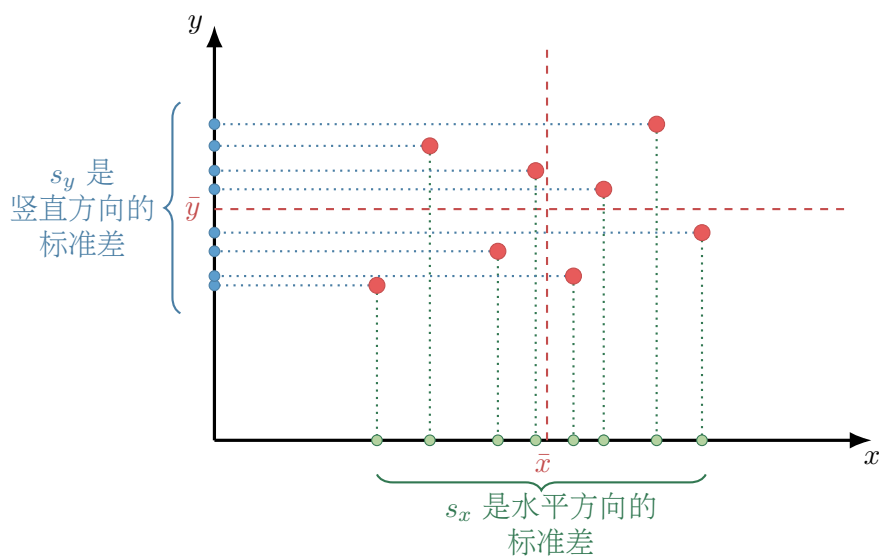
样本相关系数由以下公式定义：

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

乍看之下，这个公式似乎非常复杂。为了更好地解释其中的含义，我们将逐步分解该公式。首先请注意， r 的分母包含两个样本标准差：

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{和} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

请记住，我们正在处理配对数据，这意味着样本中的每个数据点都具有 (x_i, y_i) 的形式。每一对数据都可以分解为两个坐标，并分别加以考虑：



因此，在 r 的公式中，我们以某种方式除以了 x 和 y 各自的离散程度。 r 的分子中包含一个称为协方差的新项：

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

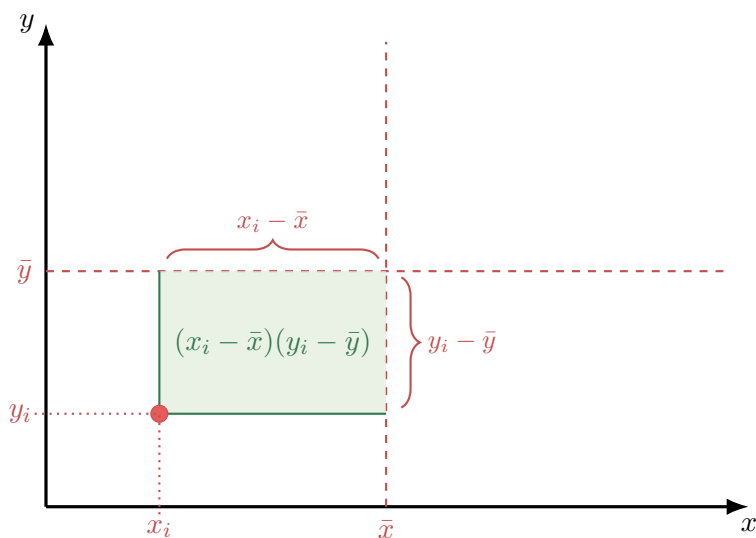
实际上，这是我们在第16次课讨论配对样本的方差时提到的协方差的样本版本。当时，我们只将它作为衡量 x 和 y 相互关系的一种量简单提及。现在，我们将详细讨论它的实际含义。

2.3 理解协方差

考察协方差的公式：

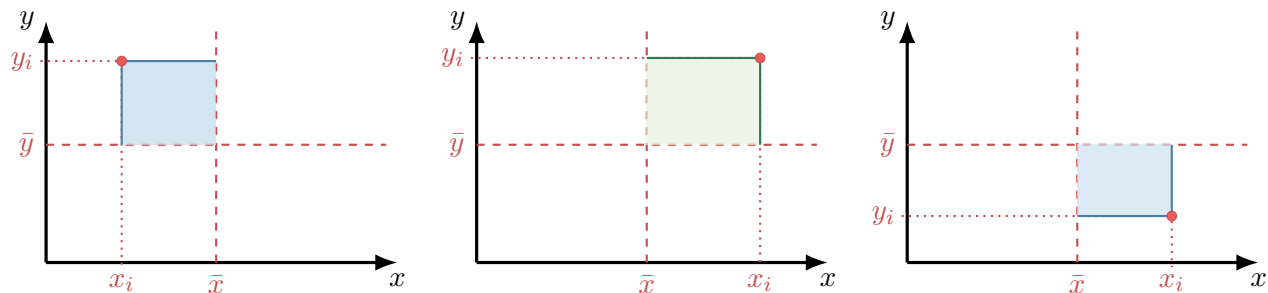
$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

这里发生了什么？首先，我们可以注意到，在公式最外层，我们对许多量求和，然后除以 $n-1$ 。公式的这一部分与样本方差的公式相似；实际上，它看起来像某种平均。那么，我们所求和的这些量究竟是什么？各个部分 $(x_i - \bar{x})$ 和 $(y_i - \bar{y})$ 实际上衡量点与均值之间的距离：它们分别衡量 x_i 距离 $\bar{x}$ 有多远，以及 y_i 距离 $\bar{y}$ 有多远。这是两个长度，将它们相乘就形成一个面积：



在上图中，我们特意将数据点 (x_i, y_i) 画在图形的左下区域。在这种情况下， x_i 位于 $\bar{x}$ 的左侧，因此 $x_i < \bar{x}$ ，从而 $x_i - \bar{x} < 0$ 。类似地， y_i 位于均值 $\bar{y}$ 的下方，因此 $y_i < \bar{y}$ ，从而 $y_i - \bar{y} < 0$ 。因此，表达式 $(x_i - \bar{x})$ 和 $(y_i - \bar{y})$ 是带有符号的距离。² 如果这两个带符号距离都是负数，那么它们的乘积就是正数。对于其他三个区域，面积具有不同的符号：

²它们通常被称为相对于均值的带符号偏差。



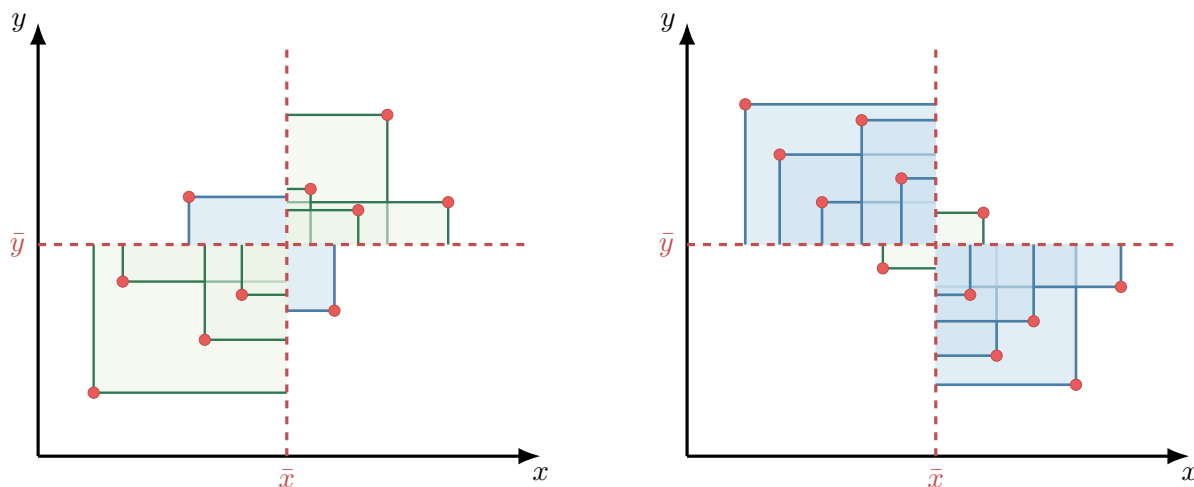
每个长方形的符号只取决于该点位于两个均值的哪一侧：

点 (x_i, y_i) 的位置	$(x_i - \bar{x})(y_i - \bar{y})$	带符号面积
$x_i < \bar{x}, y_i < \bar{y}$	$(-)(-)$	正
$x_i < \bar{x}, y_i > \bar{y}$	$(-)(+)$	负
$x_i > \bar{x}, y_i > \bar{y}$	$(+)(+)$	正
$x_i > \bar{x}, y_i < \bar{y}$	$(+)(-)$	负

如果 $x_i = \bar{x}$ 或 $y_i = \bar{y}$ ，那么带符号面积等于0。

协方差将所有这些长方形的带符号面积相加，然后除以 $n - 1$ 。换句话说，协方差平均衡量两个变量是否位于各自均值的同一侧，以及它们距离各自均值有多远。

实际上，这种“带符号面积的某种平均”使我们能够判断可能存在的相关的符号。在下图中，我们为每个数据点画出了带符号的长方形面积。我们使用绿色表示具有正号的面积，使用蓝色表示具有负号的面积。



请注意，左图表现出一定程度的正相关：大多数数据位于图的左下和右上区域。从图形上看，绿色面积多于蓝色面积，因此协方差应为正。相比之下，右图表现出负相关：大多数数据点位于图的左上或右下区域。从图形上看，带符号长方形大多为蓝色，因此协方差将为负。总结如下：

- 如果图中绿色面积多于蓝色面积，那么带符号面积之和为正，并且 $\text{cov}(x, y) > 0$ ；
- 如果图中蓝色面积多于绿色面积，那么带符号面积之和为负，并且 $\text{cov}(x, y) < 0$ ；

- 如果正面积和负面积在求和时大致互相抵消，那么协方差接近0。

回到公式，每个符号可以分解如下：

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

点与 $\bar{x}$ 之间的水平带符号距离 点与 $\bar{y}$ 之间的竖直带符号距离

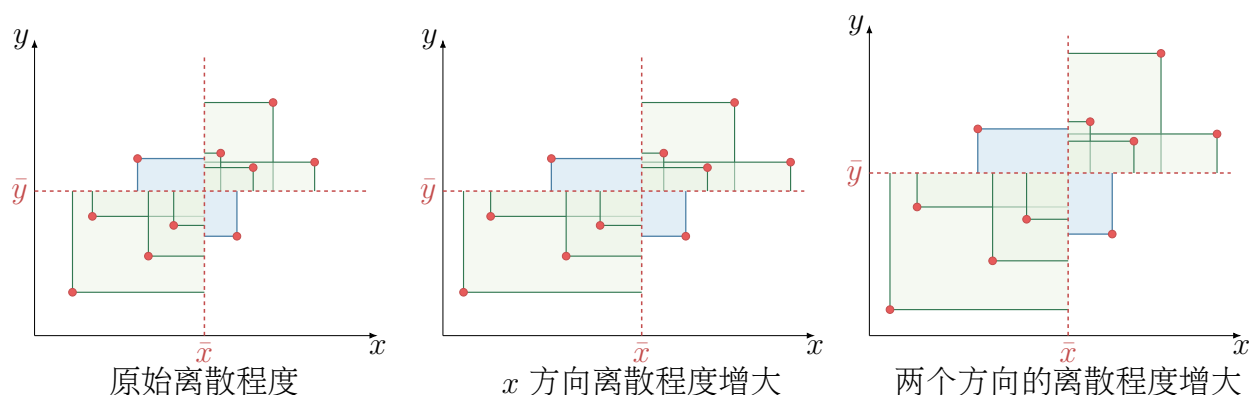
对所有面积求平均 长方形的带符号面积

协方差衡量什么

协方差衡量两个变量是否倾向于位于各自均值的同一侧。

2.4 理解 r 的公式

根据其构造，协方差公式巧妙地使用长方形的面积来记录任何潜在线性相关的符号。然而，协方差本身仍然依赖于数据的总体离散程度。为了说明这一点，我们取一组数据，并手动增加它在 x 和 y 方向上的离散程度：



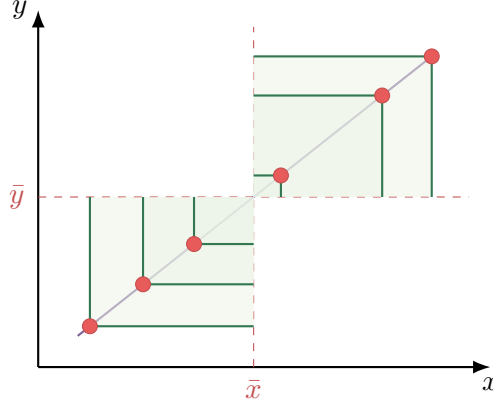
请注意，当我们增加数据的离散程度时，每个点到 $\bar{x}$ 或 $\bar{y}$ 的平均距离也会增加，因此每个长方形的平均面积也会增加。然而，一般来说，以这种方式增加离散程度并不会使绿色长方形变成蓝色长方形，反之亦然。如果我们只关心绿色和蓝色长方形之间的相互关系，而不关心它们的实际大小，那么可以除以数据的离散程度，以消除它们的绝对大小。这就是 r 的公式：

$$r = \frac{\text{cov}(x, y)}{s_x s_y}.$$

粗略来说，标准差 s_x 和 s_y 分别衡量数据在水平方向和竖直方向上的离散程度。如果我们用协方差除以这两个量，就可以消除上图的整体尺度，因此增加数据的离散程度并不会改变答案。相反，唯一重要的是由数据形成的面积之间的相对模式。

2.5 由完全正相关得到 $r = 1$

现在假设数据完美地排列在一条直线上。我们将说明 r 的公式如何给出 +1。在下图中，我们在一条具有正斜率的直线上绘制了六个点，并画出了它们各自的带符号面积。



从几何角度看，所有这些长方形都具有相同的长宽比：

$$\bar{y} - y_1 \left\{ \underbrace{\hspace{1.5cm}}_{\bar{x} - x_1} \right\} \sim \underbrace{\hspace{1.5cm}}_{\bar{x} - x_2} \bar{y} - y_2 \Rightarrow \frac{y_1 - \bar{y}}{x_1 - \bar{x}} = \frac{y_2 - \bar{y}}{x_2 - \bar{x}}$$

我们将这个长宽比记为 k 。如果所有点 (x_i, y_i) 都位于同一条直线上，那么这些比值都等于 k 。³ 重新排列后，对于所有 $i = 1, \dots, n$ ，有

$$\frac{y_i - \bar{y}}{x_i - \bar{x}} = k \Rightarrow y_i - \bar{y} = k(x_i - \bar{x}).$$

如果将其代入协方差公式，就会得到一个包含 x 的样本方差的表达式：

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})k(x_i - \bar{x}) = \frac{k}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = k s_x^2.$$

类似地，将 s_y 公式中的所有 $y_i - \bar{y}$ 替换为 $k(x_i - \bar{x})$ ，得到

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n k^2 (x_i - \bar{x})^2} = k s_x,$$

其中最后一个等式使用了 $k > 0$ 。将这些结果代入 r 的公式，得到

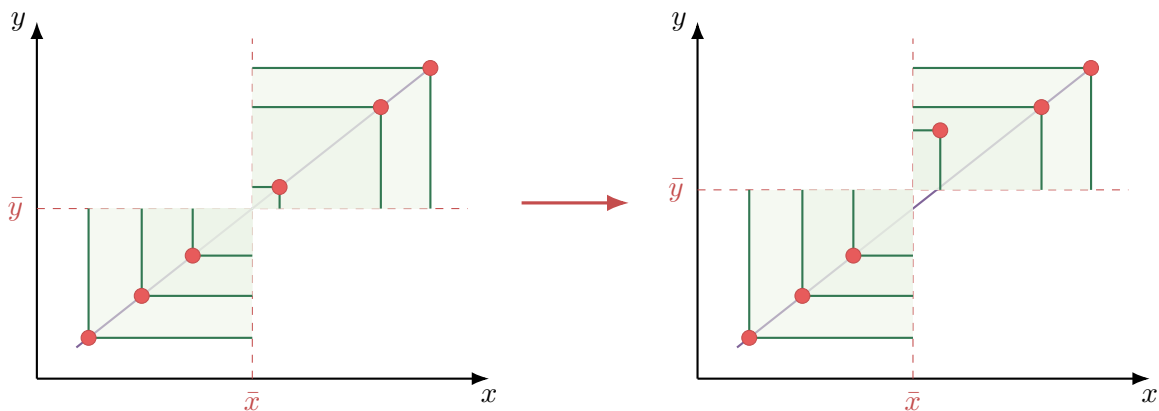
$$r = \frac{k s_x^2}{s_x (k s_x)} = 1,$$

与预期一致。

³严格来说，当点满足 $x_i = \bar{x}$ 或 $y_i = \bar{y}$ 时，这个比值无法使用。在这些情况下，“面积”会退化为一条一维线段。

2.5.1 移动一个点会发生什么？

为了更好地从几何角度理解 r 的作用，我们将一个点稍微向上移动，使数据不再完全位于一条直线上：



我们预计这一操作会使 r 从1减小到一个稍小的值。从图形上看，将一个点沿竖直方向移离直线会使总体均值 $\bar{y}$ 向上移动。因此，所有距离 $y_i - \bar{y}$ 都会发生变化。在这一过程中，这些长方形将不再具有相同的长宽比：在上图中，长方形现在具有不同的形状。

从代数角度看，将一个点沿竖直方向移动，相当于在某个固定点的 y 分量上增加一个小量。例如， $(x_4, y_4) \rightarrow (x_4, y_4 + \epsilon)$ ，其中 ϵ 是一个正数。粗略来说，对一个点进行这种改变会使协方差稍微增大，但标准差 s_y 的公式包含长度的平方，即 $(y_i - \bar{y})^2$ 形式的项。当我们将一个点移动 ϵ 时，由于平方的存在，这个整体上的小变化会引起相对较大的变化。因此， s_y 项比 $\text{cov}(x, y)$ 项增长得更快，从而使比值 r 减小到1以下。

2.6 例：计算 r

现在回到第1节的通勤时间数据。样本均值为

$$\bar{x} \approx 25.333 \quad \text{和} \quad \bar{y} \approx 34.333.$$

为了计算 r ，我们需要对几种不同类型的项求和。下表记录了所需的信息：

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
4	8	-21.333	-26.333	455.111	693.444	561.778
7	20	-18.333	-14.333	336.111	205.444	262.778
11	17	-14.333	-17.333	205.444	300.444	248.444
15	20	-10.333	-14.333	106.778	205.444	148.111
18	21	-7.333	-13.333	53.778	177.778	97.778
22	27	-3.333	-7.333	11.111	53.778	24.444
27	30	1.667	-4.333	2.778	18.778	-7.222
31	29	5.667	-5.333	32.111	28.444	-30.222
35	40	9.667	5.667	93.444	32.111	54.778
39	60	13.667	25.667	186.778	658.778	350.778
45	65	19.667	30.667	386.778	940.444	603.111
50	75	24.667	40.667	608.444	1653.778	1003.111
合计		0	0	2478.667	4968.667	3317.667

使用简化后的计算公式,

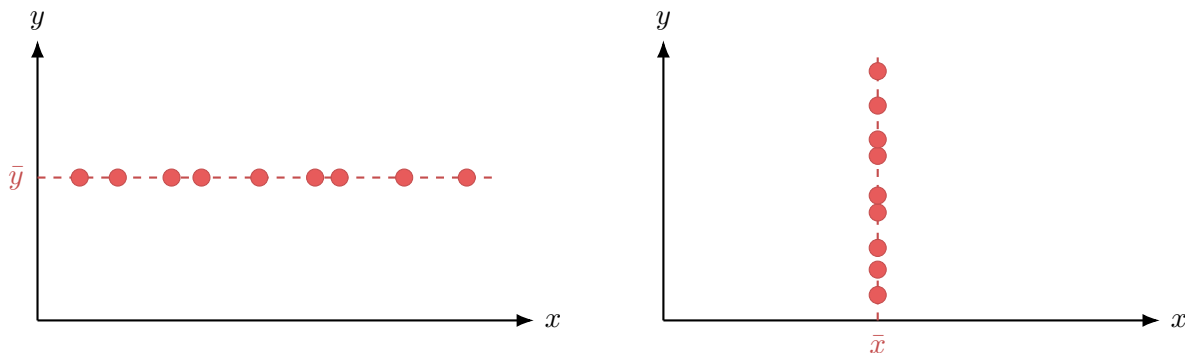
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{3317.667}{\sqrt{2478.667} \sqrt{4968.667}} \approx 0.945.$$

这是一个较强的正线性相关, 这意味着住得离LUJ 越远的学生, 其通勤时间往往越长。

3 相关的局限性

3.1 r 未定义的情况

考虑以下两组数据:



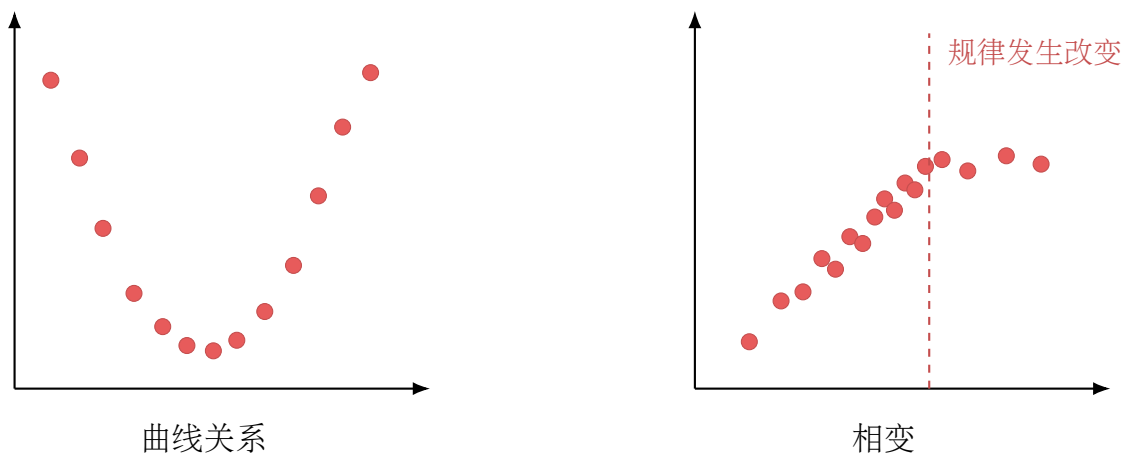
显然, 这两组数据都位于一条直线上。然而, 样本相关系数 r 实际上未定义。在每种情况下, 其中一个方向的离散程度为零, 这意味着要么每个 y_i 都等于 $\bar{y}$, 要么每个 x_i 都等于 $\bar{x}$ 。这会使协方差 $\text{cov}(x, y)$ 以及 s_y 或 s_x 中的一个等于零, 从而得到

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{0}{0},$$

这是未定义的。

3.2 非线性相关

相关系数 r 只衡量线性相关。特别地，有些数据组可能以其他方式相关，但 r 的值仍然很小。在其他情况下，变量之间的关系可能在一段范围内是线性的，但随后变成另一种关系：



在科学应用中，一种关系可能只在某个点之前近似线性，之后则由新的规律支配。在这种情况下，相关系数仍然可能给出一个很高的值，即使它在严格意义上不能代表整个系统。

3.3 没有因果关系的相关

数据中存在明显相关并不意味着两个变量之间一定存在因果联系。例如，如果 X 和 Y 相关，那么可能存在以下情况：

- X 导致 Y ；
- Y 导致 X ；
- X 和 Y 仅仅由于巧合而相关；
- 存在第三个变量，同时影响 X 和 Y 。

这里的重要结论是：

相关关系并不意味着因果关系。

例如，冰淇淋销量通常与防晒霜销量相关。这并不意味着冰淇淋销量导致防晒霜销量，也不意味着相反的因果关系。天气是同时影响两者的第三个变量。

相关关系并不意味着因果关系

r 的值可以描述线性模式的存在、符号和强度。然而，它不能为该模式提供因果解释。