第2章 回 归 线

2.1 引言

这一章是关于回归线（regression line）的. 回归线本身（对统计学家）很重要，而且它将在第4章讨论多元回归时有所帮助. 第一个例子是显示1078个父亲和他们儿子身高的散点图（图1）. 每一对父子成为图上的一个点. 父亲的身高点在 x 轴上，其儿子的身高点在 y 轴上. 左边的竖直带（在烟囱形状之内）显示了父亲身高是64英寸（四舍五入到英寸）的家庭，右边的竖直带显示了父亲身高是72英寸的家庭. 还可以画出许多其他的竖直带. 在给定父亲身高时，回归线可以对其儿子的平均身高做出近似. 这条线通过所有竖直带的中心. 回归线较SD线（虚线）更平坦，“SD”是后面要介绍的“标准差”（standard deviation）的缩写.

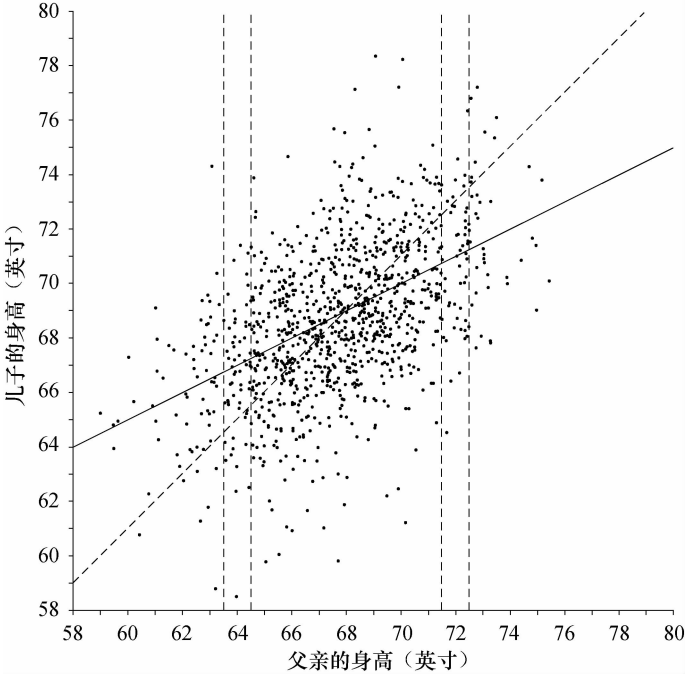


图1 父亲和儿子的身高. Pearson and Lee（1903）

2.2 回归线

假如有 n 个对象，用 $i=1,\cdots,n$ 来标记，还有两个数据变量（data variable）x 和 y. 一个数据变量对每个研究对象存储了一个值. 这样，x_i 是对象 i 的 x 值，y_i 是对象 i 的 y 值. 在图1中，一个“对象”是一个家庭：x_i 是家庭 i 的父亲身高，y_i 是家庭 i 的儿子身高. 对于Yule（1.4节），一个“对象”可能是一个都市区会，对于区会 i，$x_i=\Delta\text{Out}$，$y_i=\Delta\text{Paup}$.

回归线是由5个概括统计量计算的：（ⅰ）x的平均，（ⅱ）x的SD，（ⅲ）y的平均，（ⅳ）y的SD，（ⅴ）x和y之间的相关系数. 计算可以如下进行，其中简写var代表方差（variance），关于$\bar{y}$和var(y)的公式这里省略.

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \text{var}\,x=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{1}$$

$$x\text{的SD为}\ s_x=\sqrt{\text{var}\,x}, \tag{2}$$

$$\text{标准单位的}\ x_i\ \text{为}\ z_i=\frac{x_i-\bar{x}}{s_x}, \tag{3}$$

相关系数为

$$r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\cdot\frac{y_i-\bar{y}}{s_y}\right). \tag{4}$$

按照惯例，假定$s_x\neq 0$，$s_y\neq 0$. 必然有，$-1\leqslant r\leqslant 1$：参见后面练习B16. 在x和y之间的相关系数常常写成$r(x,y)$. 定义符号函数sign(r)：当$r>0$时，sign(r)$=+1$；当$r<0$时，sign(r)$=-1$. 下面的（5）和（6）表明，回归线比SD线要平坦些.

y在x上的回归线通过平均点$(\bar{x},\bar{y})$. 斜率为rs_y/s_x. 截距为$\bar{y}-$斜率$\cdot\bar{x}$. (5)

SD线也通过平均点. 斜率为sign(r)s_y/s_x. 截距为$\bar{y}-$斜率$\cdot\bar{x}$. (6)

y在x上的回归线，也称为由x预测y的回归线，它是平均图（graph of averages）的线性近似，平均图显示对每个x的y的平均值（图2）.

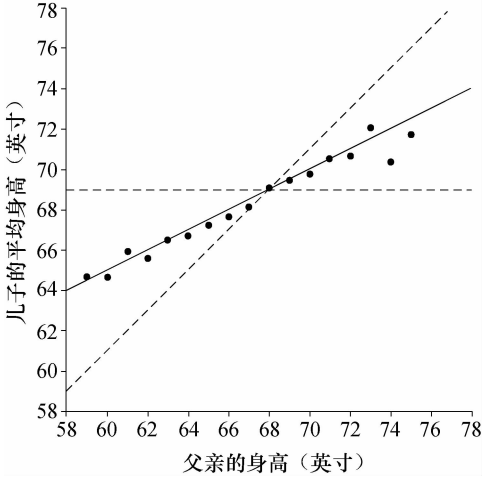


图2　平均图. 点表示对应于父亲身高的每个值，儿子身高的平均值. 回归线（实线）是沿着这些点的：它比SD线（虚线）更平坦

相关是一个关键的概念. 图3显示了三个散点图的相关系数. 所有的图都有相同数量的点（$n=50$）、同样的均值（$\bar{x}=\bar{y}=50$）和同样的SD($s_x=s_y=15$). 但这些图的形状却完全不同. 相关系数r和描述了形状.（如果变量不是成对的——成对即每个对象有两个数——那么不可能计算相关系数或者回归线.）

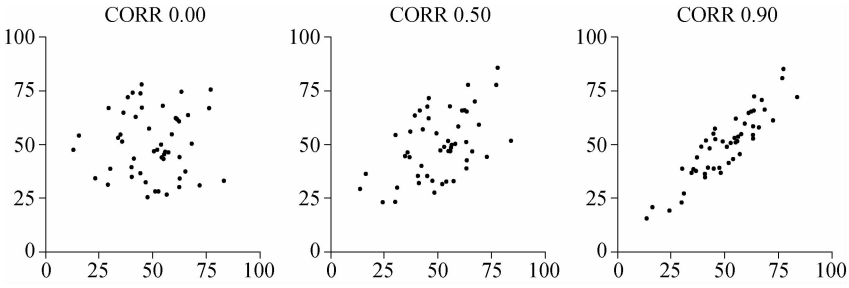


图 3　三个散点图．相关系数度量散点图在一条线周围的拥挤程度．如果符号是正的，该线向上倾斜．如果符号是负的，该线向下倾斜（这里没有显示）

如果利用直线 $y=a+bx$ 由 x 预测 y，那么关于对象 i 的误差或者*残差*（residual）为 $e_i=y_i-a-bx_i$，而 MSE 为

$$\frac{1}{n}\sum_{i=1}^{n}e_i^2.$$

RMS 误差是 MSE 的平方根．对于回归线，后面也将会看到，MSE 等于 $(1-r^2)\text{var}\ y$．MSE 是均方误差（mean square error）的缩写，RMS 是根均方（root mean square）的缩写．

一个属于 C.-F. Gauss 的定理．在所有直线中，回归线有最小的 MSE.

一个更加一般的定理将在第 3 章提供．如果对 2.1～2.2 节的内容不熟悉，可以读一下 Freedman-Pisani-Purves（2007）的第 8～12 章．

2.3　胡克定律

一个弹簧的长度在没有负载时为 a．当一个重量挂在弹簧的端点时，弹簧将被拉到一个新的长度．按照胡克定律（Hooke's law），弹簧被拉的长度和重量成比例．如果把重量 x_i 挂在弹簧上，则长度为

$$Y_i=a+bx_i+\varepsilon_i,\quad i=1,\cdots,n. \tag{7}$$

方程（7）是一个*回归模型*（regression model）．在这个方程中，a 和 b 是依赖于该弹簧的常数，称为*参数*（parameter），它们的值未知，需要根据数据来估计．ε_i 为独立同分布的，均值为 0，方差为 σ^2，称为*随机误差*（random error）或者*扰动*（disturbance）．方差 σ^2 是另一个参数．第 i 次选择重量 x_i，在该重量下，弹簧长度响应为 Y_i．你不会观测到 a，b 或 ε_i．

表 1 显示了在加州大学伯克利分校一堂物理课上关于胡克定律的实验结果．其第一行显示了载荷的重量，第二行为测量的长度．（“弹簧”是挂在一个大教室天花板上的一个长的钢琴弦．）

表 1　关于胡克定律的实验

重量（kg）	0	2	4	6	8	10
长度（cm）	439.00	439.12	439.21	439.31	439.40	439.50

利用最小二乘方法来估计参数 a 和 b．换句话说，我们拟合回归线．截距为

$$\hat{a} \doteq 439.01\text{cm}.$$

在一个参数上加帽子表示是它的一个估计：a 的估计为 439.01cm. 斜率为

$$\hat{b} \doteq 0.05\text{cm/kg}.$$

b 的估计为 0.05cm/kg. （加了点的等号“$\doteq$”意味着“约等于”，这里数量结果是四舍五入后的.）

有两个结论.（ⅰ）在弹簧上加一重物使它变长.（ⅱ）每一千克重量使它伸长约 0.05cm. 第一个是（相当明显的）定性分析，而第二个是定量的. 第 6 章将再次讨论定性和定量推断之间的区别.

练习组 A

1. 在 Pearson-Lee 数据中，父亲的平均身高为 67.7 英寸，其 SD 为 2.74 英寸. 儿子的平均身高为 68.7 英寸，其 SD 为 2.81 英寸. 相关系数为 0.501.

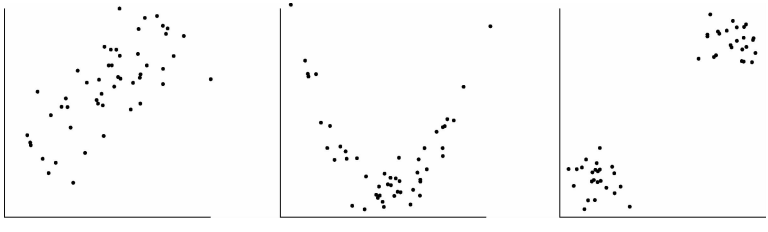
(a) 判断下面论述的对错并解释：因为儿子平均比父亲高过一英寸，如果父亲是 72 英寸高，那么有百分之五十的机会儿子高于 73 英寸.

(b) 找到儿子高度对父亲高度的回归线及其 RMS 误差.

2. 你能够通过度量没有重量时的弹簧长度来根据方程（7）确定 a 吗？度量一次呢？度量 10 次呢？简单解释.

3. 利用表 1 中的数据，为由重量预测长度的回归线求出 MSE 和 RMS 误差. 哪一个统计量更能判断数据离回归线有多远？提示：注意量纲，或者点图，或者二者.

4. 对于下面哪个图，相关系数是一个好的描述统计量？为什么？



2.4 复杂性

比较方程（7）和（8）：

$$Y_i = a + bx_i + \varepsilon_i, \tag{7}$$

$$Y_i = \hat{a} + \hat{b}x_i + e_i, \tag{8}$$

看上去一样吗？再看看. 在回归模型（7）中，我们无法观测参数 a，b 或扰动 ε_i. 在拟合的模型（8）中，估计量 $\hat{a}$，$\hat{b}$ 是可观测的，残差 e_i 也是. 对于大样本，$\hat{a} \doteq a$，$\hat{b} \doteq b$，因而 $e_i \doteq \varepsilon_i$. 但是

$$\doteq \neq =$$

(8) 中的 e_i 称为残差，而不是扰动项或随机误差项. e_i 经常被称为“误差”（error），尽管这

可能会导致混淆. 术语“残差”更清楚些.

估计量不是参数,残差不是随机误差.

在 (7) 中, 由于 ε_i 是随机变量, Y_i 也是随机的. 随机变量如何与数据相联系呢? 答案涉及观测值 (observed value) 的概念, 这将用例子说明. 均值和方差的概念如何从数据延伸到随机变量, 也将用例子通过联系这两个概念之间的某些线索来说明. 先看均值. 考虑一列表 {1, 2, 3, 4, 5, 6}, 按照公式 (1), 它有均值 3.5 及方差 35/12. 至此, 有了一个很小的数据集. 下面将看随机变量.

掷一个骰子 n 次. (一个骰子有 6 面, 每个面都等可能出现, 一面为 1 点, 还有一面为 2 点, 如此下去, 直到 6 点.) 令 U_i ($i=1,\cdots,n$) 为第 i 次掷骰子时的点数, 它为 (更确切地是在模型中设立的) 独立同分布的随机变量, 就像随机从列表 {1, 2, 3, 4, 5, 6} 选择数目一样. 每个随机变量有均值 (期望, 也称为期望值) 3.5 和方差 35/12. 这里均值和方差已经应用到随机变量上了, 即掷一个骰子所得到的点数.

样本均值和样本方差为

$$\overline{U}=\frac{1}{n}\sum_{i=1}^{n}U_i,\quad \operatorname{var}\{U_1,\cdots,U_n\}=\frac{1}{n}\sum_{i=1}^{n}(U_i-\overline{U})^2. \tag{9}$$

(9) 中的样本均值和方差本身就是随机变量. 原则上, 它们不同于作为固定数目的 $E(U_i)$ 及 $\operatorname{var}(U_i)$, 它们分别是 U_i 的期望和方差. 当 n 很大时,

$$\overline{U}\doteq E(U_i)=3.5,\quad \operatorname{var}\{U_1,\cdots,U_n\}\doteq \operatorname{var}(U_i)=35/12. \tag{10}$$

这就表明了如何从重复观测中估计一个随机变量的期望和方差.

- 随机变量有均值, 数据集也有.
- 随机变量有方差, 数据集也有.

至今讨论有些抽象. 现在, 某人实际上掷了 $n=100$ 次骰子. 这产生了一些数据. 总点数为 371. 平均每掷一次为 371/100=3.71. 这不是 $\overline{U}$, 而是 $\overline{U}$ 的观测值. 毕竟 $\overline{U}$ 有一个概率分布, 而 3.71 是其取值之一. 类似地, 胡克定律 (表 1) 中的弹簧的测量长度并不是随机变量. 按照回归模型, 它是定义在 (7) 中的随机变量 Y_i 的观测值.

在回归模型中, 数据通常是随机变量的观测值.

再重新看 (8). 如果 (7) 成立, (8) 中的 $\hat{a}$, $\hat{b}$, e_i 能够看成观测的随机变量或者观测值, 依上下文而定. 有时, 观测值称为实现 (realization). 这样, 439.01cm 是随机变量 $\hat{a}$ 的一个实现.

还有一点需要掌握. 方差常常用来度量散布. 然而, 如下面一个例子所表明的, 方差通常有错误的单位和错误的大小, 因此要取平方根来得到 SD.

例 1 美国年龄 18～24 岁的男性的平均体重为 170 磅. 这一群人的典型重量是 170 磅左右, 但将不会是刚好 170 磅. 典型的关于平均值的偏差为_____. 体重的方差为 900 平方磅: 错误的单位, 错误的大小. 不要用方差来填空. 标准差 (SD) 为 $\sqrt{\text{方差}}=30$ 磅. 关于平均体重的典型偏差大约为 30 磅.

例 2 掷 100 次骰子. 令 $S=X_1+\cdots+X_{100}$ 为总点数. 这是个随机变量，有 $E(S)=100\times 3.5=350$. 你将会得到大约 350 点，大约加或减_____. S 的方差为 $100\times 35/12\doteq 292$. (正如前面所说，35/12 是列表 {1,2,3,4,5,6} 的方差.) 不要用 292 填空. 为了利用方差，求平方根. 标准误差 (SE) 为 $\sqrt{292}\doteq 17$. 用 17 来填空. (SE 应用于随机变量，而 SD 用于数据.)

点数将在 350 左右，但将会偏离 350 约 17 左右. 点数不大可能偏离其期望值 2 倍或 3 倍的 SE. 对于随机变量，标准误差是方差的平方根. (一个随机变量的标准误差常常称为标准差，可能会引起混淆.)

2.5 比较简单回归和多元回归

一个简单 (simple) 回归方程右边有一个截距和一个解释变量及斜率系数. 一个多元 (multiple) 回归方程右边有若干解释变量，每个都有其自己的斜率系数. 为研究多元回归，需要用到矩阵代数，这将在第 3 章介绍.

练习组 B

1. 在方程 (1) 中，方差是应用于数据还是应用于随机变量? 在 (4) 中的相关系数呢?

2. 在表 1 下面，你将找到数目 439.01. 这是一个参数还是一个估计值? 0.05 呢?

3. 假定我们没有表 1 的最后一行. 基于该表的头 5 行的数据找出长度对重量的回归线.

4. 在例 1 中，900 平方磅是随机变量的方差还是数据的方差? 做简单讨论.

5. 在例 2 中，35/12 是随机变量的方差还是数据的方差? 还是两者都是? 做简单讨论.

6. 掷 180 次骰子. 找到骰子幺点的期望值及方差. 幺点的数目将大约为_____，大约加或减_____. (一个骰子有 6 面，每个面都等可能出现，有 1 点的为“幺点”.)

7. 掷 250 次骰子. 得到幺点的次数比例将约为_____，大约加或减_____.

8. 从一个装有 4 个数字“1”，“2”，“2”，“5”的盒子中放回地抽取 100 次. 抽取的结果是 17 个“1”，54 个“2”，29 个“5”. 选择填空.

(a) 关于_____，观测值比期望值高 0.8 个 SE. (提醒：SE=标准误差.)

(b) 关于_____，观测值比期望值高 1.33 个 SE.

选项 (有两个将剩下)：

“1”的数目　　“2”的数目　　“5”的数目　　所取数目之和

如果你不熟悉练习 6～8 中涉及的内容，可参看 Freedman-Pisani-Purves (2007) 的第 16～18 章或者其他教科书的类似内容.

9. 方程 (7) 为_____. 选项：

模型　　参数　　随机变量

10. 在方程 (7) 中，a 为_____. 选项 (正确的可能不止一个)：

可观测的　　不可观测的　　一个参数　　一个随机变量

对 b，ε_i，Y_i 重复上面填空.

11. 按照方程 (7)，表 1 中的 439.00 为_____. 选项：

一个参数　　一个随机变量　　一个随机变量的观测值

12. 假定 $x_1,\cdots,x_n$ 为实数. 令 $\overline{x}=(x_1+\cdots+x_n)/n$. 令 c 为一个实数.

(a) 表明 $\sum_{i=1}^{n}(x_i-\overline{x})=0$.

(b) 表明 $\sum_{i=1}^{n}(x_i-c)^2=\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]+n(\overline{x}-c)^2$.

提示：$(x_i-c)=(x_i-\overline{x})+(\overline{x}-c)$.

(c) 表明 $\sum_{i=1}^{n}(x_i-c)^2$ 作为 c 的一个函数，在 $c=\overline{x}$ 时有一个唯一的极小值.

(d) 表明 $\sum_{i=1}^{n}x_i^2=\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]+n\overline{x}^2$.

13. 一位统计学家有个样本，而且计算了到数目 q 的偏差平方和. 当 q 为_____时，偏差平方和最小. 填空（不多于 25 个词）并解释.

14. 假定 $x_1,\cdots,x_n$ 及 $y_1,\cdots,y_n$ 有均值 $\overline{x}$，$\overline{y}$，标准差为 $s_x>0$，$s_y>0$，相关系数为 r. 令

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

（“cov”是协方差（covariance）的缩写.）表明：

(a) $\mathrm{cov}(x,y)=rs_xs_y$.

(b) 由 x 预测 y 的回归线的斜率为

$$\mathrm{cov}(x,y)/\mathrm{var}(x).$$

(c) $\mathrm{var}(x)=\mathrm{cov}(x,x)$.

(d) $\mathrm{cov}(x,y)=\overline{xy}-\overline{x}\,\overline{y}$.

(e) $\mathrm{var}(x)=\overline{x^2}-\overline{x}^2$.

15. 假定 $x_1,\cdots,x_n$ 及 $y_1,\cdots,y_n$ 为实数，$s_x>0$，$s_y>0$. 令 x^*，y^* 分别为标准单位的 x，y. 表明 $r(x,y)=r(x^*,y^*)$.

16. 假定 $x_1,\cdots,x_n$ 及 $y_1,\cdots,y_n$ 为实数，$\overline{x}=\overline{y}=0$，$s_x=s_y=1$. 表明 $\frac{1}{n}\sum_{i=1}^{n}(x_i+y_i)^2=2(1+r)$ 及 $\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2=2(1-r)$，这里 $r=r(x,y)$. 表明

$$-1\leqslant r\leqslant 1.$$

17. 掷一个骰子两次. 令 X_i 为第 i 次掷时所得的点数（$i=1,2$）.

(a) 求 $P(X_1=3\mid X_1+X_2=8)$，即在已给总点数为 8 时，第一次掷得 3 的**条件概率**（conditional probability）.

(b) 求 $P(X_1+X_2=7\mid X_1=3)$.

(c) 求 $E(X_1\mid X_1+X_2=6)$，即在已给 $X_1+X_2=6$ 时，X_1 的**条件期望**（conditional expectation）.

18.（难题）假定 $x_1,\cdots,x_n$ 为实数. 假定 n 为奇数而且所有 x_i 不同. 这时，有一个唯一的**中位数**（median）μ，也就是把这些 x 按递增次序排列时位于中间的那个数. 令 c 为一个实数. 表明：

c 的函数 $f(c)=\sum_{i=1}^{n}|x_i-c|$ 在 $c=\mu$ 时达到最小值. **提示**. 你不能用微积分做这道题，因为 f 是不可微的. 但你可以表明 $f(c)$ 为：(i) 连续的，(ii) 对于 $c>\mu$，当 c 递增时是严格递增的，即 $\mu<c_1<c_2$ 意味着 $f(c_1)<f(c_2)$，(iii) 对于 $c<\mu$，当 c 递增时是严格递减的. 当 c 和所有 x 不同时，容易理解 (ii) 和 (iii). 你也可以假定 x_i 是随着 i 递增的. 如果你沿着这条思路走得足够远，你会发现 f 在这些 x 之间是线性的，在这些 x 处有角. 此外，f 是下凸的，即 $f[(x+y)/2]\leqslant[f(x)+f(y)]/2$.

评论. 如果 $-f$ 是下凸的，那么 f 称为上凸的.

2.6 札记

在 (6) 中，如果 $r=0$，你能够取 SD 线的斜率为 s_y/s_x，或者 $-s_y/s_x$. 然而，在其他应用中，sign(0) 通常定义为 0.

胡克定律 (7) 在重量相对小的时候是好的近似. 当重量较大时，可能需要一个二次项. 当接近“弹性极限”时，情况就变得更加复杂了. 实验的细节简化了. 关于数据来源，参见 Freedman-Pisani-Purves (2007) 第 A11 和 A14 页.

更多关于随机变量的材料，包括物理骰子和骰子的数学模型之间的联系，参看

http://www.stat.berkeley.edu/users/census/rv.pdf