计量经济学STAT30021

第二讲：简单线性回归分析(1)

肖志国

复旦大学管理学院

2025年9月

为什么要研究因果关系

现在在第四范式下，基于数据和机器学习算法的分析已经取得了重大突破（蛋白质结构预测、医疗诊断、大语言模型），还需要研究因果关系吗？

需要！机器学习算法的本质是从历史数据中挖掘出现状大概率是什么。而因果关系的判断，是一个从是什么到为什么的过程，这是更高一个层级的智能。

Judea Pearl: 古巴比伦的天文学家在星相预测上远超古希腊同行。古巴比伦人满足于这种黑箱预测的技术最终湮没无闻，而古希腊人则不断提出各种怪诞的模型来解释所有的现象，但却由此开启了现代的哲学和科学。

因果关系的研究历史

即使是一些真实的观念，如果不是有人通过因果的证明把他它们联系在一起，也不具有多大的价值。– 柏拉图（前427-前347）

一切理智的或者在某种程度上具有理智的知识，都涉及了一些原因和原则。– 亚里士多德（前384-前322）

人类仅仅凭经验，只能认识事物之间恒定的前后相继关系，并不能认识任何因果关系。一切现实事物的关系都不具有必然性，因而以归纳法而认识到的关系只具有或然性。– 休谟（1711-1776）

把每一种科学和单纯的观念堆积区分开来，恰恰就是说这样的观念都是从它们的理由那里由此及彼的推导出来的。– 叔本华（1788-1860）

人类可以搞清楚因果关系吗？休谟的一种观点

我们无从得知因果之间的关系，只能得知某些事物总是会连结在一起，而这些事物在过去的经验里又是从不曾分开过的。我们并不能看透连结这些事物背后的理性为何，我们只能观察到这些事物的本身，并且发现这些事物总是透过一种经常的连结而被我们在想象中归类。–休谟《人性论》

概率论知识要点复习

重点概念与结论

- 基本概念：随机试验，事件，概率，条件概率，辛普森悖论
- 随机变量：分布，数值特征
- 多维随机变量：联合分布，边际分布，条件分布，条件期望，条件方差
- 随机变量序列的收敛：依概率收敛，依分布收敛
- 大数定律，中心极限定理

概率论知识要点复习

条件概率与辛普森悖论

为什么「患者总存活率」更低的医院，反而可能更值得推荐？

原创 丁香园 DXY 丁香园 2020-06-09

收录于话题
#庄时利和专栏
24个 >

本文作者：庄时利和

如果你家里有老人生病了，需要做某一项手术，你会如何选择医院？

假设你所在市内有 A、B 两家医院，这两家医院的手术都做得符合基本规范、术后存活者的生活质量相同、两家医院收费统一，甚至离你家的距离都一样近。而且它们都非常开放，愿意将所有治疗数据向社会公布。

网上搜索结果显示，A 院近期有 1000 名患者接受这项手术，术后 900 人存活（总存活率 90%）；B 院近期有 1000 名患者也做了同样的手术，术后 800 人存活（总存活率 80%）。

你怎么选？

概率论知识要点复习

条件概率与辛普森悖论

	A院	B院
轻症数	100	400
轻症存活数	30	210
轻症存活率	**30%**	**52.5%**
重症数	900	600
重症存活数	870	590
重症存活率	**96.7%**	**98.3%**
总人数	1000	1000
总存活数	900	800
总存活率	**90%**	**80%**

A、B 院患者按病情分组存活率比较

概率论知识要点复习

辛普森悖论背后的数学

考虑两组分数：

$$\frac{a_1}{b_1} > \frac{c_1}{d_1}, \quad \frac{a_2}{b_2} > \frac{c_2}{d_2}$$

请问$\frac{a_1+a_2}{b_1+b_2}$与$\frac{c_1+c_2}{d_1+d_2}$谁大？

一般情况下：

$$\frac{a_1 + a_2}{b_1 + b_2} > \frac{c_1 + c_2}{d_1 + d_2}$$

比如：

$$\frac{4}{5} > \frac{1}{2}, \frac{9}{10} > \frac{7}{8} \text{ 对应着 } \frac{4+9}{5+10} = \frac{13}{15} > \frac{8}{10} = \frac{1+7}{2+8}$$

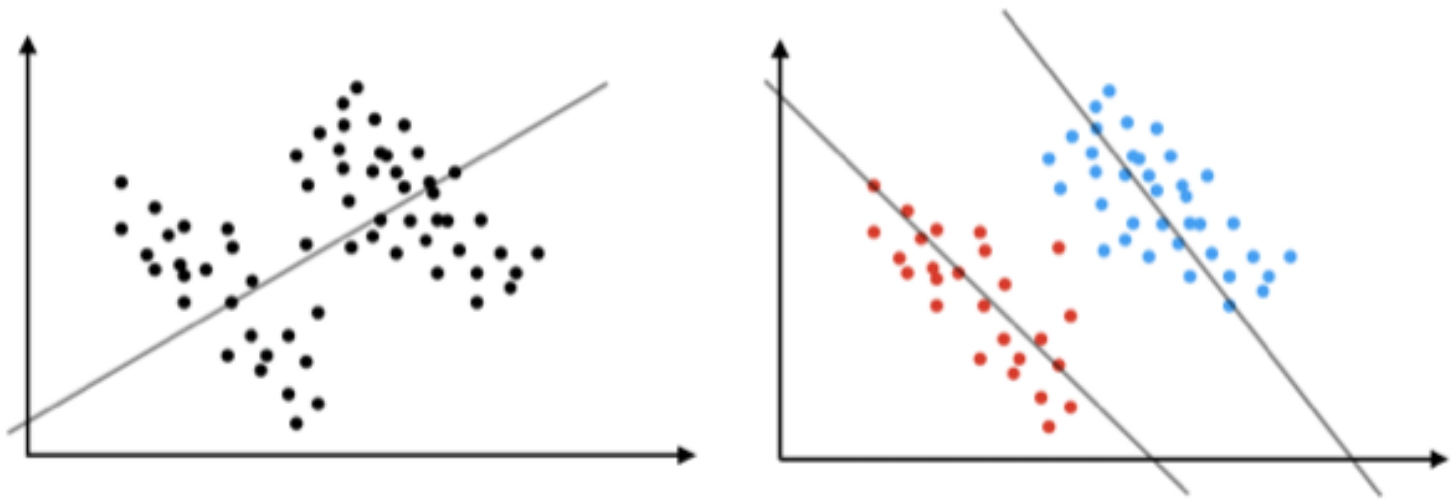
但也有反例，比如：

$$\frac{9}{10} > \frac{701}{800}, \frac{21}{100} > \frac{1}{10} \text{ 对应着 } \frac{9+21}{10+100} = \frac{30}{110} < \frac{702}{810} = \frac{701+1}{800+10}!$$

概率论知识要点复习

辛普森悖论背后的数据

出现辛普森悖论的情况，通常是因为在当时的具体情况下，存在某种因素将数据进行了分组，分组之后的数据对应着如下的图形：



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

概率论知识要点复习

条件期望

给定任何随机变量Y 和X, 设其联合密度函数为$p(y,x)$. Y 给定$X=x$ 的条件密度函数为

$$p(y|x)=\frac{p(y,x)}{p(x)}=\frac{p(y,x)}{\int p(y,x)dy}$$

Y 给定$X=x$ 的条件期望就是以上密度函数的期望:

$$E[Y|X=x]=\int yp(y|x)dy.$$

显然，$E[Y|X=x]$是x的函数，记

$$\varphi(x)=E[Y|X=x].$$

那么，我们定义Y 给定X 的条件期望为

$$E[Y|X]=\varphi(X).$$

概率论知识要点复习

条件期望的性质

1. 对任何函数h, 我们有:

$$E[h(X)Y|X] = h(X)E[Y|X].$$

2.

$$E[Y] = E(E[Y|X]).$$

3. 对任何随机变量X_1, X_2:

$$E(E[Y|X_1, X_2]|X_1) = E(E[Y|X_1]|X_1, X_2) = E[Y|X_1].$$

4. $E[Y|X]$是使得如下目标函数最小的X的函数（条件期望基本定理）:

$$E[Y|X] = \arg\min_f \ E[(Y - f(X))^2]$$

概率论知识要点复习

如何得到条件期望？

既然条件期望具有如此重要性质，那如何得到条件期望呢？

1. 方法一、通过公式推导。如果我们知道联合密度函数，则可以推导出条件期望函数的显示表达式。问题：假设条件太强。
2. 方法二、通过数据直接估算。如果我们在每一个$X = x$都有足够多的Y的观测值，那么可以直接通过这些Y的平均值作为$E[Y|X = x]$的估计值。问题：我们没有足够多的数据。事实上我们的数据量极少。
3. 方法三、通过构建模型估算。我们一般假设$E[Y|X] = f(X, \theta)$，其中f为某种已知的函数形式。然后通过数据来估算θ，从而得到$E[Y|X]$的估计值。
   - 线性回归模型：$f(X, \theta)$为线性函数
   - 机器学习模型：$f(X, \theta)$为某种形式的非线性函数

条件期望与因果关系

1. 统计学（包括机器学习）的核心内容是如何通过数据对条件期望函数进行估计。
2. 有了条件期望函数的估计值，我们是不是就能在此基础上估计出某个变量X对Y的因果效应了呢？
3. 答案是否定的。
4. 条件期望函数代表的只是一种相关关系。给定任一组变量(Y, X, Z)，都可以计算或者估算条件期望函数，但显然这一组变量未必能够用于论证和估算因果关系。
5. 只有当这一组变量(Y, X, Z)构成一个封闭的因果系统时，条件期望函数才能够用来估算因果效应（事实上我们还需要一个额外条件，那就是当Z变化时，X可以自由改变）。

统计学知识要点复习

相关系数

考虑两个随机变量X和Y，比如X = 身高，Y = 体重。我们通常关心的是研究Y如何随X变化而变化。

最直接的办法：相关系数

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}},$$

可以证明

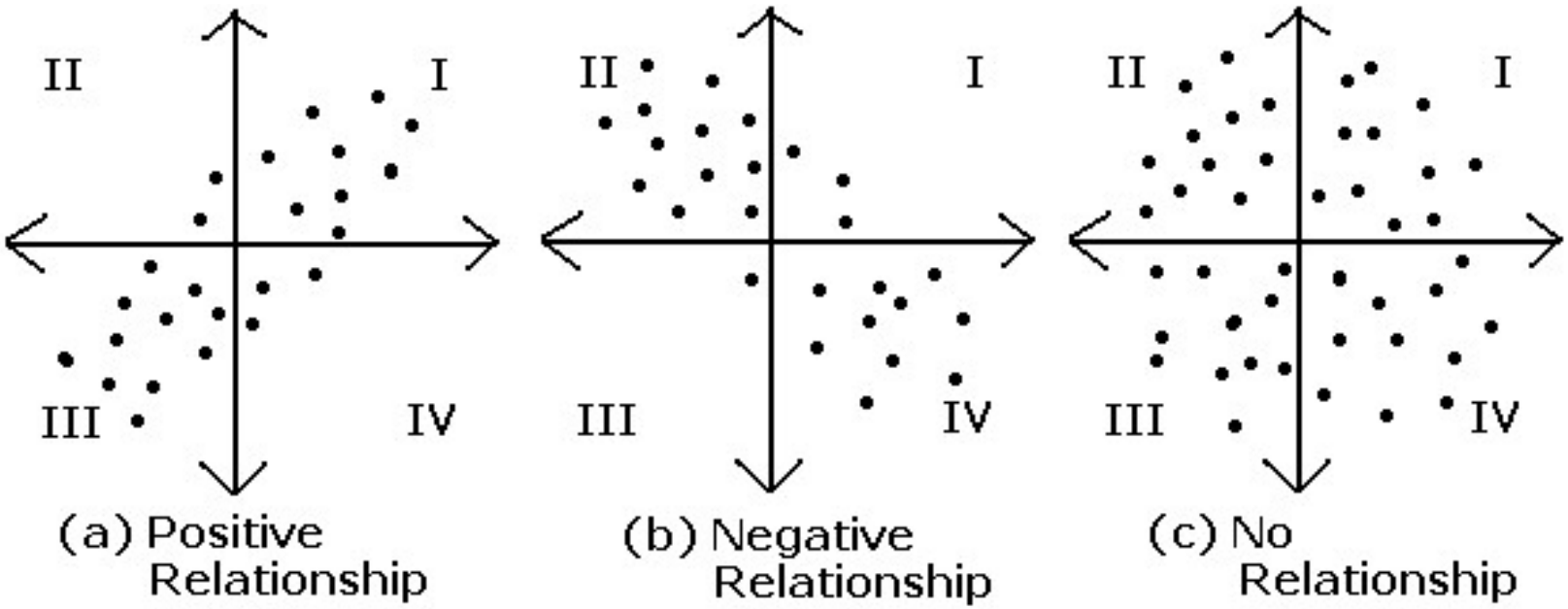
$$-1 \leq \quad \rho_{X,Y} \quad \leq 1.$$

统计学知识要点复习

相关系数

统计学知识要点复习

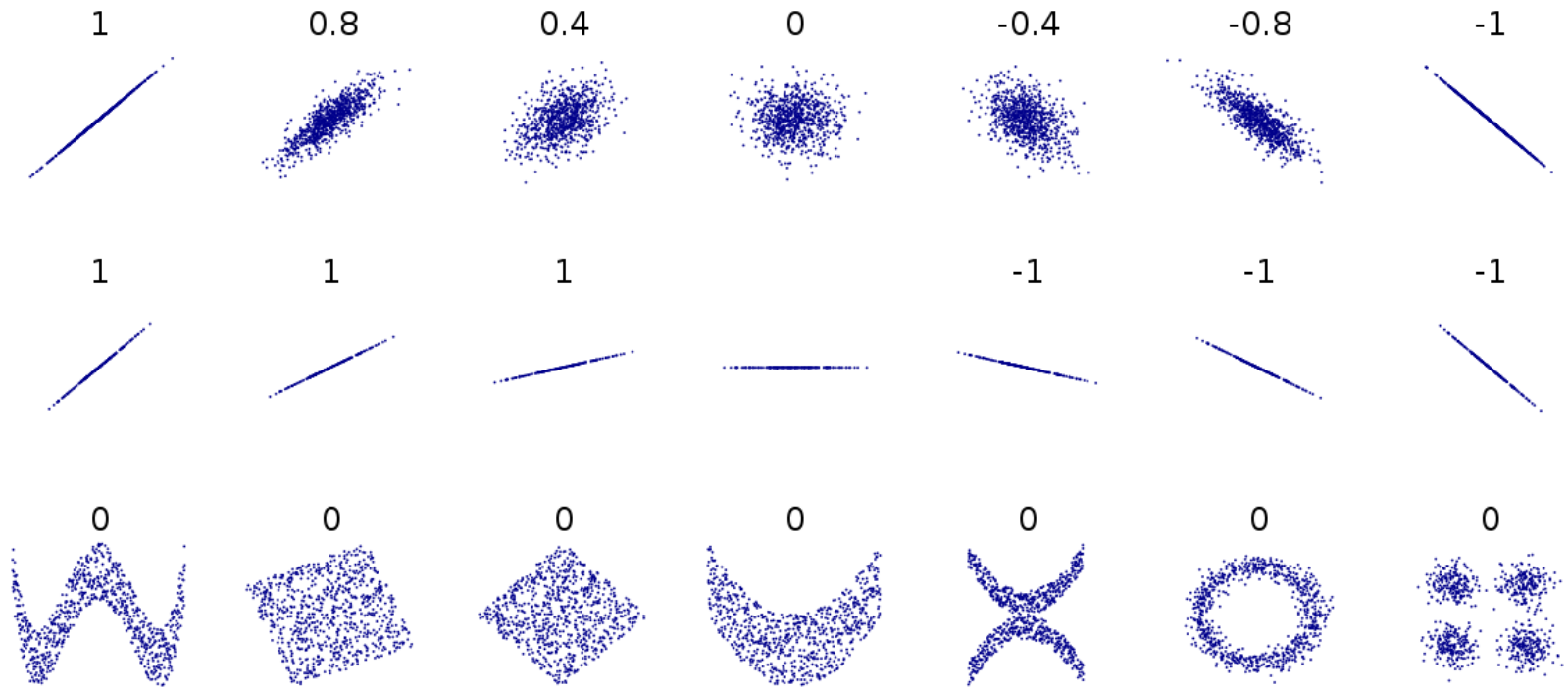
相关系数



我们可以换一种方式定义相关系数吗？

统计学知识要点复习

相关系数

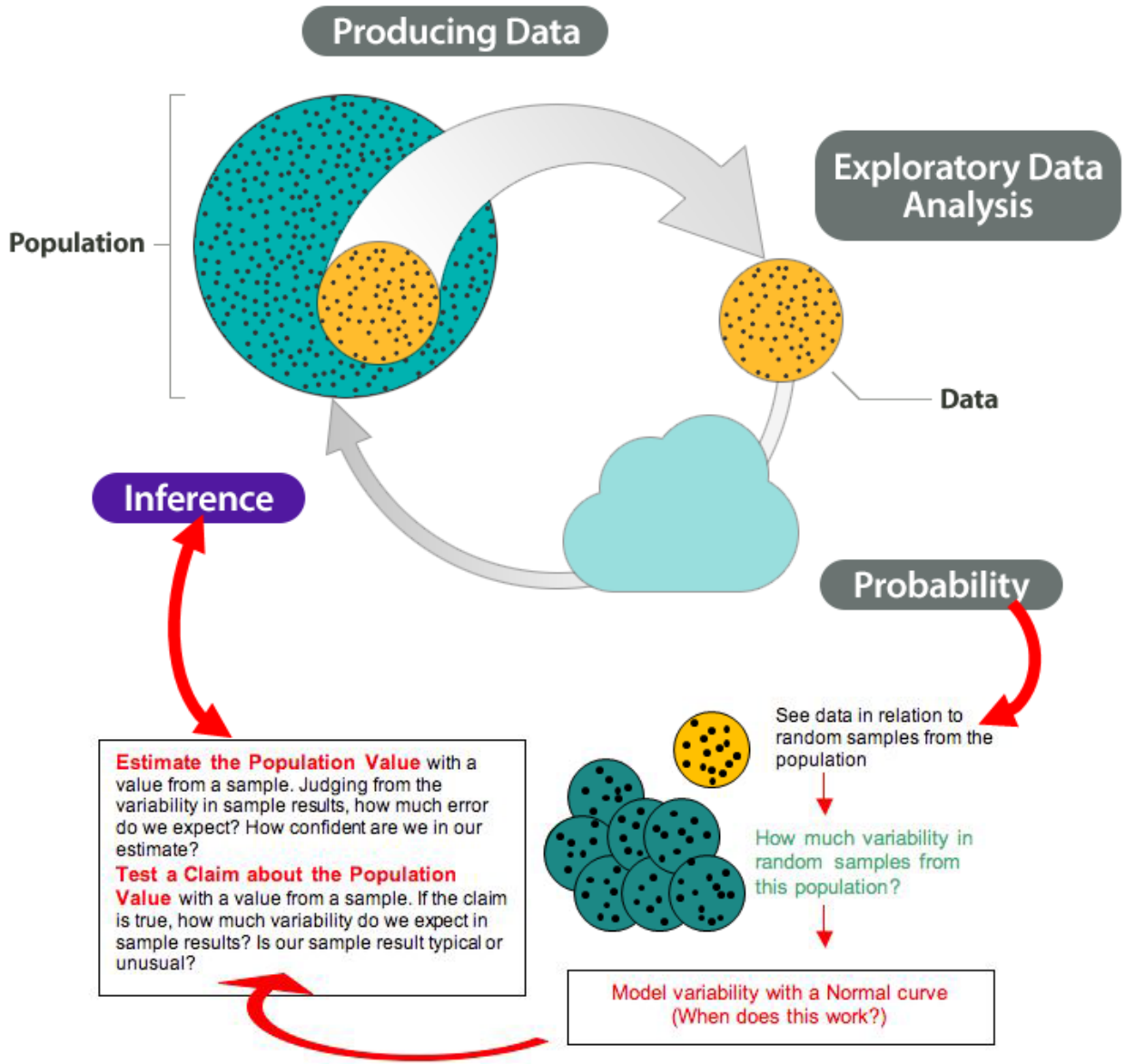


相关系数的缺陷之一：它给出的是一个含糊结果，没有给出变量之间影响关系的定量形式。

如何推广相关系数：线性回归！

统计学知识要点复习

统计建模的理论框架

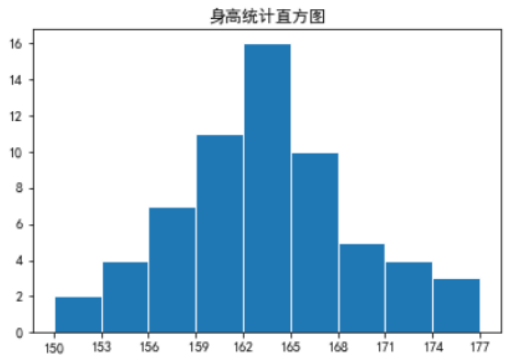


统计学知识要点复习

统计建模的出发点：数据$Z_i, i = 1, \ldots, n$

统计建模的基本思路：

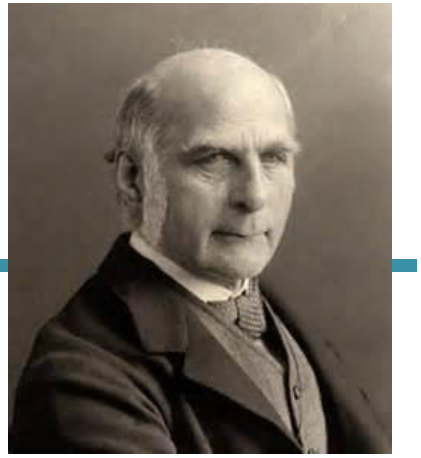
1. 我们假设我们手上的数据$Z_i, i = 1, \ldots, n$是所有数据Z的一个随机抽样。
2. 例子：我们有1000个中国人的身高数据，这个数据是所有中国人的身高数据的一个随机抽样。这1000个身高数据叫做样本，而全体中国的人身高数据称作总体。
3. 我们如何刻画全体中国人的身高数据？直方图？



4. 我们用概率分布来刻画总体。

回归分析简史

高尔顿 1886: “Regression Towards Mediocrity in Hereditary Stature”



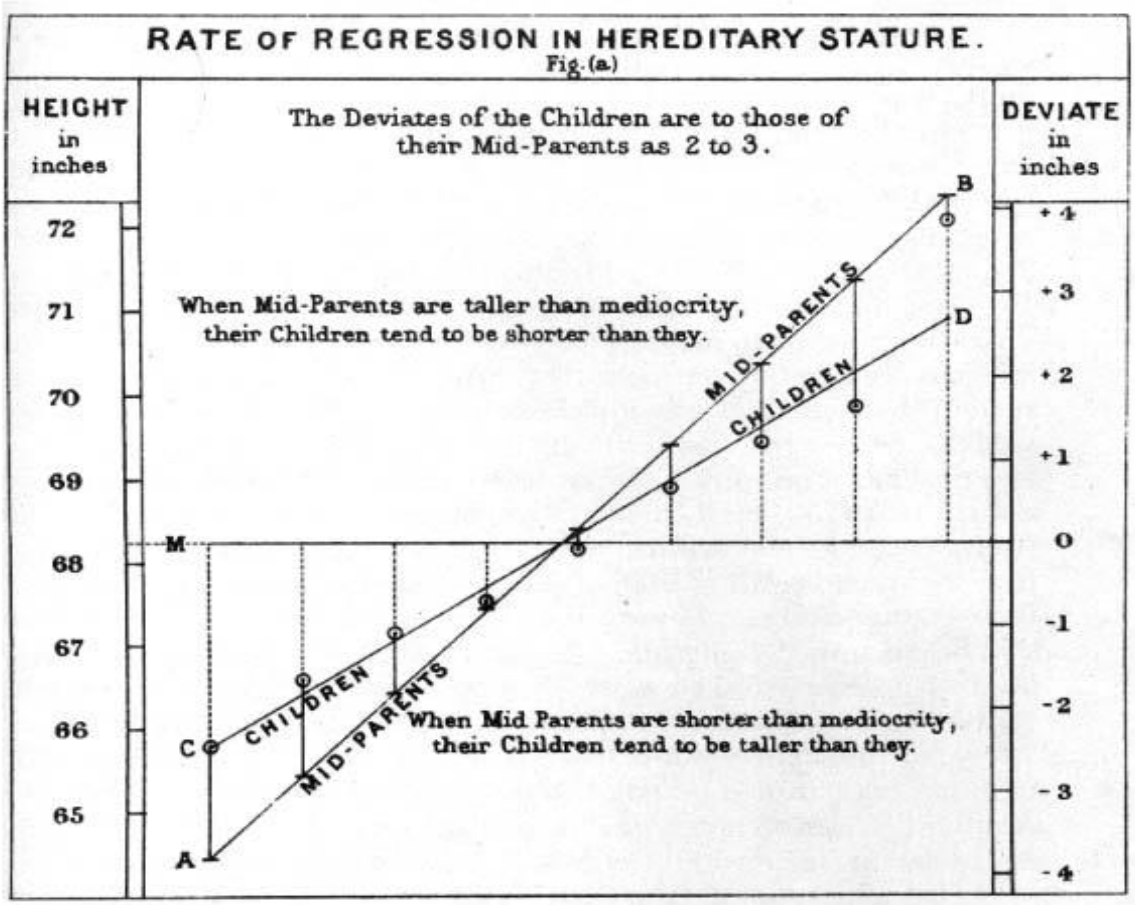
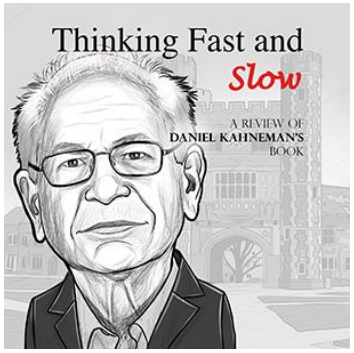


Figure 8.8. Galton's graphical illustration of regression; the circles give the average heights for groups of children whose midparental heights can be read from the line AB. The difference between the line CD (drawn by eye to approximate the circles) and AB represents regression toward mediocrity. (From Galton, 1886a.)

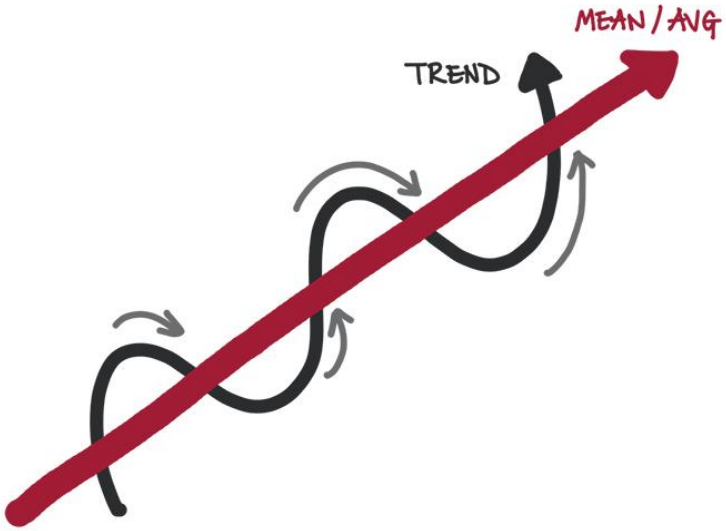
高尔顿预测公式

$$Y - X \approx -\frac{1}{3}(X - M)$$

其中
Y = 子女平均身高
X = 父母身高均值
M = 成年人身高均值



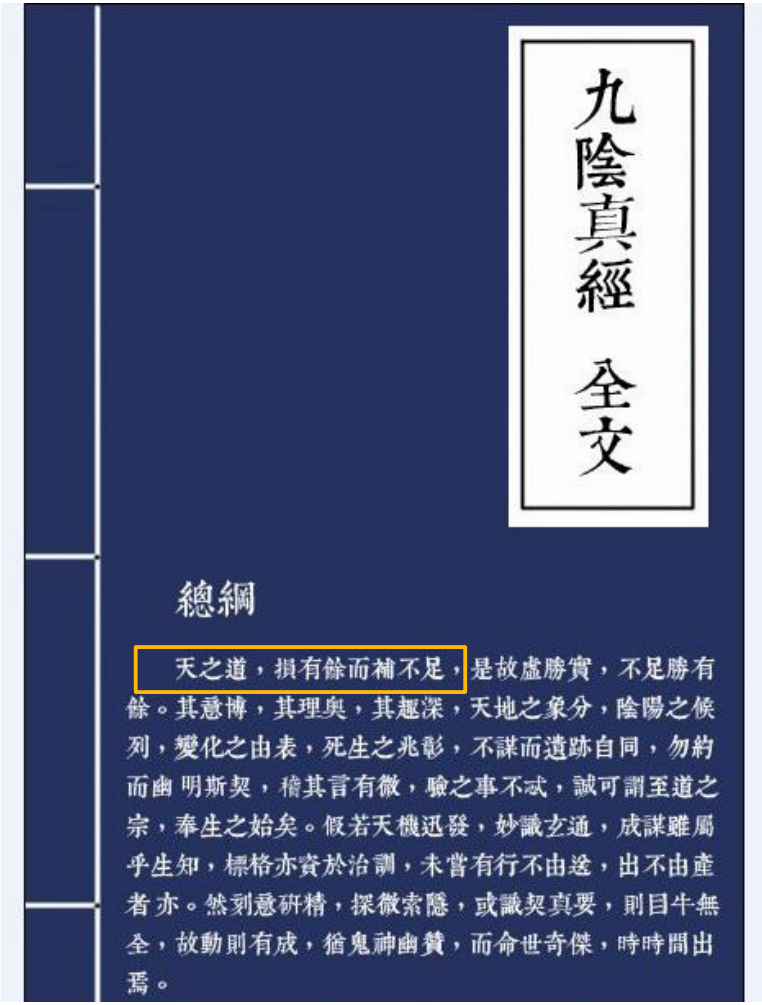
回归现象的意义不亚于发现万有引力



均值回归

万有引力

回归分析的思想缘起



天之道，其犹张弓与？高者抑下，下者举之，有余者损之，不足者补之。天之道，损有余而补不足。

人之道，则不然，损不足以奉有余。孰能有余以奉天下，唯有道者。是以圣人为而不恃，功成而不处，其不欲见贤。

回归分析的统计建模框架

- 所谓一个统计模型，就是一个关于总体分布的某种特征的假定。
- 我们考虑总体变量是多变量的情形。具体的，$Z = (Y, \mathbf{X})$，其中$\mathbf{X} = (X_1, \ldots, X_k)$ 为解释变量，Y 为目标变量。
- 我们称形如如下的模型设定为一个线性回归模型：

$$E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

其中$\beta_0, \beta_1, \ldots, \beta_k$为模型的核心未知参数。

- 由于样本$Z_i = (Y_i, \mathbf{X}_i)$与总体$Z = (Y, \mathbf{X})$具有相同的分布，以上模型设定也可以等价的表示成：

$$E[Y_i|\mathbf{X}_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}.$$

简单线性回归模型

在简单线性回归模型中，我们只有一个解释变量，不妨记做X，也就是说：

$$E[Y|X] = \beta_0 + \beta_1 X \tag{1}$$

如果我们定义

$$u = Y - \beta_0 - \beta_1 X, \tag{2}$$

那么易证$E[u|X] = 0$，因而模型(1)可以等价的写成：

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u, \\ E[u|X] &= 0. \end{aligned} \tag{3}$$

方程(3)是线性回归模型的标准表达形式，因为它看起来更容易直观理解，同时也更方便于我们后面对线性回归模型的各种假设进行推广。由于样本$Z_i = (Y_i, X_i)$与总体$Z = (Y, X)$具有相同的分布，模型(1)也可以等价的写成：

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i, \\ E[u_i|X_i] &= 0. \end{aligned} \tag{4}$$

术语

方程(3)中的Y有如下几种不同的叫法

- 因变量(Dependent variable)
- 被解释变量(Explained variable)
- 响应变量(Response variable)
- 被预测变量(Predicted variable)
- 被回归变量(Regressand)
- 结果变量(Outcome variable)

术语

对应于Y，方程(3)中的X有如下几种不同的叫法

- 自变量(Independent variable)
- 解释变量(Explanatory variable)
- 控制变量(Control variable)
- 预测变量(Predictor variable)
- 回归变量(Regressor)
- 协变量(Covariate)
- 特征变量(Feature variable)

术语

在简单线性回归模型(3)中，u表示除X以外影响Y的其它因素。u有两种叫法：

- 误差项(Error term)
- 干扰项(Disturbance)

对模型(3)来说：

- X和Y是可以观测到的变量
- u是没有观测到的变量

条件期望为零假设

模型(3)中的一个关键假设就是u关于X的条件期望为零。为了便于讨论以及后续对此假设的推广，我们不妨对此假设给一个编号：

假设1 u关于X的条件期望为零：

$$E[u|X] = 0. \tag{5}$$

说明：

- 假设1是对u关于X的条件分布的假设
- 假设1是说，u关于X的条件分布的均值不依赖于X，且恒为零。这一个数学上的表述仍然并不直观。
- 从假设1可以推出$Cov(h(u), X) = 0, \forall h$。也就是说，变量$X$与$u$的任何（线性与非线性）函数都没有相关性。

解读假设1的合理性：例子

我们以工资与受教育水平的关系为例。在这里：$X =$ 受教育水平，$Y =$ 工资水平。我们假设$u =$ 个人内在能力。

也就是说，我们假定工资水平只由两个因素决定：教育水平和个人能力。

则(5)要求：任给两个不同的受教育水平，平均的能力都是一样的。

比如：最高学历为小学毕业的人和最高学历为大学毕业的人的平均内在能力是一样的。

这个论断合理吗？

参数估计

假设$\{(X_i, Y_i), i = 1, \ldots, n\}$为来自总体$(X, Y)$的随机样本。则我们有：对于所有的$i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{6}$$

以及

$$E[u_i|X_i] = 0. \tag{7}$$

我们的首要任务就是通过一定的方法给出未知参数β_0和β_1的估计量。

方法一：矩估计法

我们有两个参数，所以需要构造两个矩条件方程。一个自然的矩条件是$E[u_i] = 0$。另外一个矩条件可以构造如下：
由$E[u_i|X_i] = 0$可得$E[X_i u_i] = 0$。所以我们总的矩条件为：

$$E\begin{bmatrix} Y_i - \beta_0 - \beta_1 X_i \\ X_i(Y_i - \beta_0 - \beta_1 X_i) \end{bmatrix} = 0$$

根据矩估计法的原理我们可以求得β_0和β_1的矩估计量为：

$$\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{X}, \quad \tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

方法二：最小二乘法(OLS)

通过最小二乘法构造β_0和β_1的估计量的思路如下。我们寻找β_0和β_1的值来最小化如下的目标函数：

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

设$\hat{\beta}_0, \hat{\beta}_1$为使得$S(\beta_0, \beta_1)$最小的值。我们称$\hat{\beta}_0, \hat{\beta}_1$分别为$\beta_0, \beta_1$的最小二乘估计量。容易求得：

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

我们发现：矩估计法与最小二乘法得到的估计量完全相同。

样本方差、样本协方差、相关系数

给定一组数据$\{(X_i, Y_i), i = 1, \ldots, n\}$。$X, Y$的样本方差分别定义为：

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, \quad s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1},$$

X, Y的样本协方差定义为：

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

X, Y的相关系数定义为：

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{s_{XY}}{s_X s_Y}$$

最小二乘估计量的解释

根据定义，

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{s_Y}{s_X} r_{XY}$$

所以：

- 如果X, Y有正相关关系，则回归系数$\hat{\beta}_1 > 0$。
- 如果X, Y有负相关关系，则回归系数$\hat{\beta}_1 < 0$。
- 如果X, Y有相同方差，则回归系数$\hat{\beta}_1 = r_{XY}$。

样本回归直线

- 我们称直线$y = \hat{\beta}_0 + \hat{\beta}_1 x$为样本回归直线。
- 称$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$为$Y_i$的拟合值。
- 称$\hat{u}_i = Y_i - \hat{Y}_i$为残差。$\hat{u}_i$是$u_i$的一个估计。

几何解释 最小二乘法实际上就是寻找一条穿过样本点区域的直线，使得其残差平方和最小。

最小二乘法的性质

性质1 所有残差之和为零：

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

性质2 回归变量X与残差的样本协方差为零：

$$\sum_{i=1}^{n} X_i \hat{u}_i = 0$$

性质3 拟合值$\hat{Y}$与残差的样本协方差为零：

$$\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = 0$$

SST, SSE和SSR

我们定义SST, SSE和SSR如下：

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
SSE &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\
SSR &= \sum_{i=1}^{n}\hat{u}_i^2
\end{aligned}
$$

由于$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}, \bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$：

- SST代表的是Y_i的总波动量: $SST = (n-1)s_Y^2$
- SSE代表的是$\hat{Y}_i$的总波动量: $SSE = (n-1)s_{\hat{Y}}^2$
- SSR代表的是$\hat{u}_i$的总波动量: $SSR = (n-1)s_{\hat{u}}^2$

波动恒等式

容易证明：SST, SSE和SSR满足如下恒等式：

$$SST = SSE + SSR \tag{8}$$

写成样本方差分解的形式，也就是

$$s_Y^2 = s_{\hat{Y}}^2 + s_{\hat{u}}^2 \tag{9}$$

等式(9)的直观解释：Y的变化可以分解成两部分：

- 一部分是回归方程所能解释的变化$s_{\hat{Y}}^2$
- 另一部分是回归方程所不能解释的变化$s_{\hat{u}}^2$

R^2

我们定义R^2为

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SSR}{SST} = \frac{s^2_{\hat{Y}}}{s^2_Y} \tag{10}$$

R^2又称为拟合优度（goodness of fit）。

- R^2表示的是回归方程所能解释的Y的变化与实际的Y的变化的比例。
- $0 \leq R^2 \leq 1$
- $R^2 = 0$当且仅当$\hat{Y}_i = \bar{Y}$对任意i：用平均值拟合任何一个值
- $R^2 = 1$当且仅当$\hat{Y}_i = Y_i$对任意i：完美拟合

R^2与相关系数

令$r_{\hat{Y}Y}$为$\hat{Y}$与Y的相关系数，也就是：

$$r_{\hat{Y}Y} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

则我们可以证明：

$$R^2 = r_{\hat{Y}Y}^2 \tag{11}$$

在只有一个解释变量的时候，我们还有：

$$R^2 = r_{\hat{Y}Y}^2 = r_{XY}^2$$

注：波动恒等式只在线性模型时才成立。也就是说，在非线性模型中，基于定义$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$计算出来的$R^2$可能会大于1。

但(11)的结果给出了一个在非线性模型中计算模型拟合程度的或许可行的指标。

更多统计学知识要点复习

假设我们对变量Z的取值感兴趣。

- 变量Z的所有可能取值的全体叫做**总体**
- 总体由一个概率分布F_Z刻画
- 我们假设θ为与这分布F_Z有关的一个关键参数
- 统计分析的核心任务是对θ的取值作出估计和推断

为完成这一任务，我们先通过抽样收集到Z的一组数据$z_1, \ldots, z_n$。

问题：如何从样本数据$z_1, \ldots, z_n$合理的对参数θ的取值作出估计和推断？

总体，样本，实现值，估计量

为回答这一问题，统计学家们建立了如下的理论框架：

- 设$Z_1, \ldots, Z_n$为来自总体F_Z的一组简单随机样本。也就是说，$Z_1, \ldots, Z_n$是互相独立且都服从F_Z分布的随机变量
- 我们观测到的数据$z_1, \ldots, z_n$是样本$Z_1, \ldots, Z_n$ 的一个**实现值**
- 我们要建立一条基于样本$Z_1, \ldots, Z_n$的决策规则，然后将这个决策规则应用于观测到的数据$z_1, \ldots, z_n$ 上
- 最常用的决策规则主要有三类：
  - 点估计：构造一个$Z_1, \ldots, Z_n$的函数$\hat{\theta}(Z_1, \ldots, Z_n)$ 来估计θ的取值，$\hat{\theta}$称作θ的**估计量**
  - 区间估计：构造一个区间$[a(Z_1, \ldots, Z_n), b(Z_1, \ldots, Z_n)]$来估计$\theta$的取值范围
  - 假设检验：判断关于θ的论断（比如$\theta \geq 0$）的正确性

计量经济学模型分类

为回答上述问题，我们需要更多的信息，或者说，我们需要一些假设条件。

- 第一类假设条件：总体分布F属于某个已知的分布族，比如正态分布族。这一类模型叫做参数模型。
- 第二类条件：总体分布F是未知的，但是我们知道它具有连续的密度函数。这一类模型叫做非参数模型。
- 第三类条件：总体分布F是未知的，但是我们知道存在某个向量值函数$\phi(\cdot)$,满足

$$E\left[\phi(Z_i, \theta)\right] = 0, \ \forall i = 1, \ldots, n \tag{12}$$

这一类模型叫做半参数模型。方程(12)称作矩条件。

参数估计方法的构造

在一个特定的模型中，有哪些基本的观念（思路或者方法）能够指导我们去找到合理的并且性质优良的估计量？

统计学给我们提供了如下的方法：

1. 矩估计方法
2. 极大似然估计方法
3. 最小二乘方法
4. 核估计方法
5. 很多其他方法，如IV, 2SLS, GMM, GEE, EL等等

参数估计方法1：矩估计

设θ为k维参数。如果存在某个k维向量值函数$\phi(\cdot)$,满足条件(12)。则我们可以通过如下方法构造一个θ的估计量：

- 令

$$\bar{\phi}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i, \theta)$$

- 我们称$\bar{\phi}(\theta)$为样本矩。
- 设$\hat{\theta}$为方程

$$\bar{\phi}(\theta) = 0$$

的解。

- 我们称$\hat{\theta}$为θ的矩估计量。

正态分布参数的矩估计

设$Z_1, \ldots, Z_n$为来自正态分布$N(\mu, \sigma^2)$的样本，其中μ, σ^2为参数。

- 根据正态分布性质，我们知道

$$E[Z_i] = \mu, \ E\left[Z_i^2\right] = \mu^2 + \sigma^2$$

- 则

$$\phi(Z_i, \mu, \sigma^2) = \begin{bmatrix} Z_i - \mu \\ Z_i^2 - \mu^2 - \sigma^2 \end{bmatrix}$$

- 从而

$$\bar{\phi}(\mu, \sigma^2) = \begin{bmatrix} \bar{Z} - \mu \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^2 - \mu^2 - \sigma^2 \end{bmatrix}$$

- 求解方程组$\bar{\phi}(\mu, \sigma^2) = 0$可得$\mu, \sigma^2$的矩估计量为：

$$\hat{\mu} = \bar{Z}, \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2$$

参数估计方法2：最大似然估计

设$Z_1, \ldots, Z_n$具有联合分布密度$p(z_1, \ldots, z_n, \theta)$.

- 令$l_z(\theta) = logp(z_1, \ldots, z_n, \theta)$。
- 称$l_z(\theta)$为对数似然函数。
- 若$\tilde{\theta}$为θ在所有可能取值中使得$l_z(\theta)$最大的取值，则我们称$\tilde{\theta}$为θ的最大似然估计量。
- 在通常情况下，$\tilde{\theta}$是如下似然方程的解：

$$\frac{\partial l_z(\theta)}{\partial \theta} = 0$$

正态分布参数的最大似然估计

设$Z_1, \ldots, Z_n$为来自正态分布$N(\mu, \sigma^2)$的样本，其中μ, σ^2为参数。

- 联合分布密为

$$p(z_1, \ldots, z_n, \mu, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} exp\{-\sum_{i=1}^{n} \frac{(Z_i - \mu)^2}{2\sigma^2}\}$$

- 对数似然函数为

$$l_z(\mu, \sigma^2) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \sum_{i=1}^{n} \frac{(Z_i - \mu)^2}{2\sigma^2}$$

- 似然方程为

$$\begin{aligned}\frac{\partial l_z(\mu,\sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_i (Z_i-\mu)=0\\ \frac{\partial l_z(\mu,\sigma^2)}{\partial \sigma^2} &= \frac{1}{2\sigma^4}\sum_i \left[(Z_i-\mu)^2-n\sigma^2\right]=0\end{aligned}$$

- 求解似然方程可得μ,σ^2的最大估计量为：

$$\tilde{\mu}=\bar{Z},\ \tilde{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)^2$$

- 矩估计原理和最大似然估计原理得到的估计量完全相同！
- 这不是一个普遍现象。但是同时也说明不同的估计原理得到的结果之间可能有密切的联系。

决策规则

- 样本数据具有随机性。
- 同一个决策规则应用到不同的样本数据上得到的决策结果很可能是不一样的。比如说，我们用样本平均值来估计总体平均值，使用两组不同的样本数据得到的平均值结果是不一样的。
- 那么我们怎么能知道基于当前样本数据$z_1, \ldots, z_n$的决策是合理的，或者说是最好的？
- 我们首先要论证决策规则本身（也就是基于随机变量$Z_1, \ldots, Z_n$）的合理性和优良性。
- 统计学的核心理论贡献就是在各种不同的情形下，找出了合适的决策规则，并论证了他们的合理性和优良性。

点估计的偏差

我们称

$$Bias(\hat{\theta}) = \hat{\theta} - \theta \tag{13}$$

为$\hat{\theta}$称的**偏差**。

- 偏差是一个随机变量，且满足$Var(Bias(\hat{\theta})) = Var(\hat{\theta})$
- 偏差是衡量一个估计量好坏（也就是其合理性和优良性）的基础
- 一个理想的估计量的偏差恒为零：但是这样的估计量不存在
- 可行的标准：我们说一个估计量越好，如果这个估计量的偏差越近似等于零
- 什么叫做一个随机变量近似等于零？均值为零，且方差很小
  - 无偏性：偏差的均值为零
  - 有效性：偏差的方差更小
  - 相合性：当n很大时，$Bias(\hat{\theta}) \approx 0$
  - 渐进正态性：当n很大时，$Bias(\hat{\theta})$近似服从$N(0, \frac{\sigma^2}{n})$ 分布

估计量的性质

设$Z_1, \ldots, Z_n$为数据(注：Z_i的维数可能大于1)，θ为我们感兴趣的k维参数。记θ所有可能取值的范围为Θ。我们称Θ为参数空间。$Z = (Z_1, \ldots, Z_n)$的任何一个函数，如果其目的是为了估计θ，称为θ 的一个(点)估计量。估计量的最关键的要求是其准确性。

1. 有限样本情形
   - 无偏性：平均起来误差为零
   - 有效性：波动更小
2. 大样本情形
   - 相合性：样本足够大时近似等于真实值
   - 渐进有效性：样本足够大时波动更小

有限样本性质

- 我们称$\hat{\theta}$是θ的一个无偏估计，若

$$E[\hat{\theta}] = \theta,\ \forall \theta \in \Theta$$

- 我们称θ的估计量$\hat{\theta}_1$比$\hat{\theta}_2$更有效，如果

$$Var[\hat{\theta}_2] - Var[\hat{\theta}_1] \geq 0,\ \forall \theta \in \Theta$$

  在上述不等式中，若θ为向量，则≥ 0表示该矩阵为非负定矩阵。
- 无偏性和有效性单独作为标准衡量一个估计量的好坏时都不有其明显缺陷。把它们同时放在一起考虑时得到一个新标准：无偏估计量中方差最小的
- 遗憾：方差最小的无偏估计量通常不存在

大样本性质

- 我们称$\hat{\theta}$是θ的一个相合估计，若$\hat{\theta}$依概率收敛于θ(记作$\hat{\theta} \xrightarrow{p} \theta$)，也就是说，$\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0, \ \forall \theta \in \Theta$$

- 设θ的两个估计量$\hat{\theta}_1$和$\hat{\theta}_2$具有如下的渐进分布：$\forall \theta \in \Theta$

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_1 - \theta\right) &\xrightarrow{d} N(0, V_1), \\ \sqrt{n}\left(\hat{\theta}_2 - \theta\right) &\xrightarrow{d} N(0, V_2). \end{aligned}$$

我们称$\hat{\theta}_1$比$\hat{\theta}_2$更渐进有效，如果$V_2 - V_1$为非负定矩阵。

渐进正态分布

设θ的估计量$\hat{\theta}$满足如下性质：

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N(0, V), \tag{14}$$

其中V为一$k \times k$维非负定矩阵，则称:

- $\hat{\theta}$服从$\sqrt{n}-$渐进正态分布
- V为$\sqrt{n}\left(\hat{\theta}-\theta\right)$的渐进方差
- V/n为$\hat{\theta}$的渐进方差

若$\hat{\theta}$满足(14)，则$\hat{\theta}$为θ的相合估计：

$$\hat{\theta}-\theta=\frac{1}{\sqrt{n}}\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{p} 0.$$

参数的区间估计

设$a(Z_1, \ldots, Z_n), b(Z_1, \ldots, Z_n)$为$Z_1, \ldots, Z_n$的两个函数。
设$\alpha \in (0, 1)$为一个给定的常数。

如果

$$P\{a(Z_1, \ldots, Z_n) \leq \theta \leq b(Z_1, \ldots, Z_n)\} = 1 - \alpha, \qquad (15)$$

则我们称区间

$$[a(Z_1, \ldots, Z_n), b(Z_1, \ldots, Z_n)]$$

为参数θ的**一个置信水平为**$1 - \alpha$**的置信区间**。

构造置信区间的基本办法

设$T(Z_1, \ldots, Z_n, \theta)$为$Z_1, \ldots, Z_n$和$\theta$的函数。我们称$T$为一个**枢轴量**，如果它的分布是一个已知的(不依赖于θ)的概率分布。

如果我们找到了这样的一个枢轴量(通常是θ的线性函数)，那么我们就可以根据T的概率分布找到两个常数t_a, t_b，使得

$$P\{t_a \leq T(Z_1, \ldots, Z_n, \theta) \leq t_b\} = 1 - \alpha, \tag{16}$$

因此我们可以由

$$t_a \leq T(Z_1, \ldots, Z_n, \theta) \leq t_b \tag{17}$$

反解出θ的范围

$$a(Z_1, \ldots, Z_n) \leq \theta \leq b(Z_1, \ldots, Z_n)$$

此即为θ的一个置信水平为$1 - \alpha$的置信区间。

构造置信区间的例子

设$X_1, \ldots, X_n$为来自总体分布为$N(\mu, \sigma^2)$的样本，其中μ, σ均为未知参数。现在我们构造 μ 的一个置信水平为$1-\alpha$的置信区间。

记$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$。令

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S}, \tag{18}$$

则我们可以证明T是一个服从自由度为$n-1$的t分布的随机变量(从而T是一个枢轴量)。

设$t_{\frac{\alpha}{2}}(n-1)$为自由度为$n-1$的$t$分布的上-$\frac{\alpha}{2}$分位数，则有

$$P\{-t_{\frac{\alpha}{2}}(n-1) \leq T \leq t_{\frac{\alpha}{2}}(n-1)\} = 1-\alpha, \tag{19}$$

由此我们反解出 μ 的一个置信水平为$1-\alpha$的置信区间

$$[\bar{X} - t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}]$$