数学地质丛书

实用回归分析

张启锐

地质出版社

数学地质丛书

实用回归分析

张启锐

*

责任编辑：高书平　杨友爱

地质出版社　出版发行

（北京西四）

地质出版社印刷厂印刷

（北京海淀区学院路29号）

新华书店总店科技发行所经销

*

开本：850×1168 1/32 印张：13 7/16 字数：356,000

1988年1月北京第一版·1988年1月北京第一次印刷

印数：1—2,505册　国内定价：3.80

ISBN 7-116-00089-5/P.078

统一书号：13038·新487

目 录

第一章 绪 论

§1.1 引 言

中文“回归分析”一词译自英文的“Regression Analysis”。在英文的地质文献中，“Regression”一词也经常见到，其中文的对应词是“海退”。中文“回归”和“海退”似乎风马牛不相及，但英文“Regression Analysis”中的regression和英文地质文献中使用的regression都有退回或返回的意思。

回归分析这一术语，是Galton于1886年前后提出的。他曾用回归分析法研究人类的身高变化。身材高大的父辈，其下一代的身材一般也是高大的，矮小的一代一般产生矮小的下一代。这是一般规律。但人们感兴趣的是，这两种趋势会不会向两极无限地发展下去？高大者的后代会不会越来越高大？矮小者的子孙会不会越来越矮小？Galton通过分析发现，人类的身高不会向两极发展，后代的身高最终都将逐渐“返回”或“回归”到人类的平均身高。尽管Galton的方法和结果还有可商榷之处，但这一例子对理解回归一词在回归分析中的原义是有帮助的。

回归分析是一种很实用的统计分析方法，已被广泛地应用于理论研究、生产管理和工程技术等领域。最早它主要应用于生物学领域，但很快就普及到其他科学领域。据一些西方学者估计，在计算中心的总计算时间中，约有一半是用于做回归计算。当前市场上出售的性能较好的袖珍电子计算器，都带有简单回归分析的专用程序。拥有计算机的部门，一般都配有功能齐全的回归分析程序。这些事实说明，回归分析是一种深受人们重视和被广泛使用的分析手段。

§1.2 描述、预测和控制

回归分析的基本功能是研究一事物各种特性之间的关系，或者更具体地说，是研究某一变量和其他有关变量的依赖关系。

实际问题是多种多样的，在不同场合下应用回归分析方法的目的也不尽相同。归纳起来，大致可以分为描述、预测（或称预报）和控制三个方面。三者之间密切相关。

1. 描述

这是最基本的方面。描述就是用一定形式的数学表达式去刻划实际事物的一些基本关系和规律。在数学上，可将这些关系和规律分为确定性的和随机性的。确定性的关系和规律，要用确定性的数学手段来研究；随机性的关系和规律，要用随机性的数学手段来研究。回归分析是研究随机关系和规律的一种数学方法，它描述事物的数学表达式叫做回归模型，其基本形式为：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+e \qquad (1.2.1)$$

或简写为：

$$y=\beta_0+\sum_{j=1}^{p}\beta_jx_j+e$$

其中：y习惯上称为因变量；$\beta_0,\beta_1,\cdots,\beta_p$是待定系数，称为回归系数；$x_1,x_2,\cdots,x_p$为自变量；$e$为随机误差。

Blatt和Jones（1975）曾将地表沉积岩面积按地质年代作过统计，并将统计结果点在半对数坐标纸上。他们发现沉积岩的年龄和出露面积之间有很好的线性关系，图中的点基本上落在一条直线上。1.3亿年以来形成的沉积岩，占总面积的一半；1.3到2.6亿年的占$\left(\frac{1}{2}\right)^2=\frac{1}{4}$，依次类推。反过来说，就是过去形成的沉积岩，其出露面积，经过1.3亿年后，只保留了一半，另一半转变为其它岩类或被剥蚀掉。由这个例子可见，图1.1中的回归直线对沉积岩出露面积和年龄的关系，作了很好的描述。

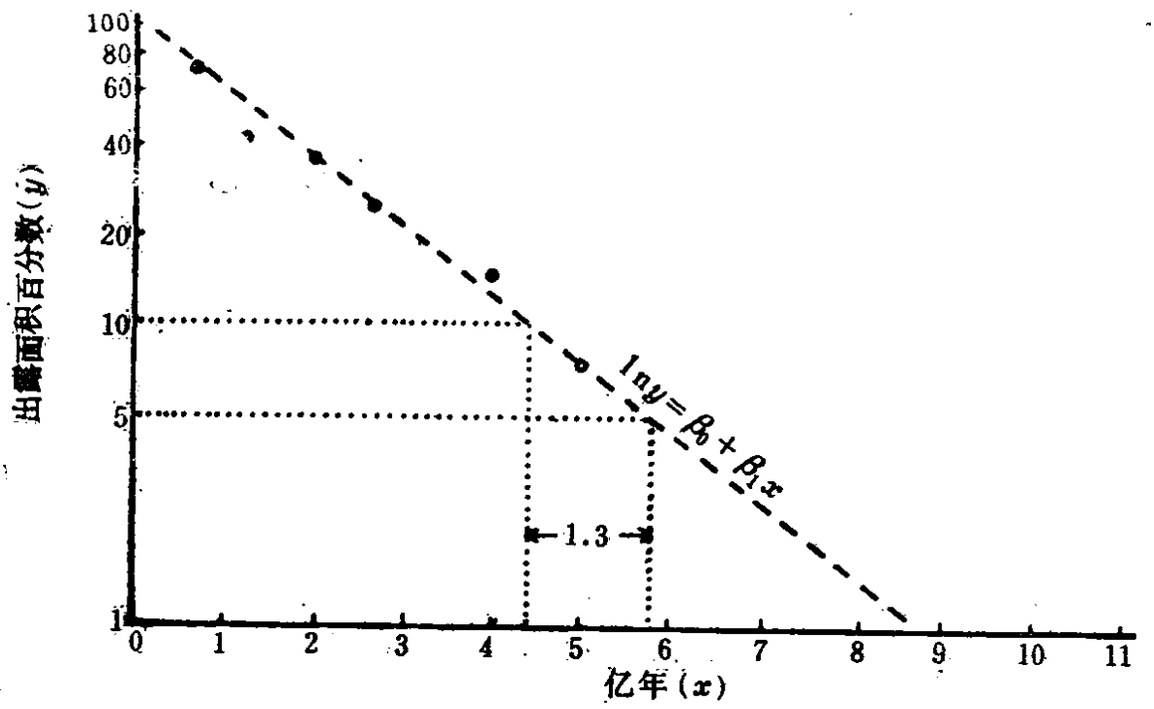


图 1.1　沉积岩地表出露面积累积百分数与形成年龄关系图
(据Blatt和Jones, 1975, 改制)

著名的统计学家 Kendall 和Stuart (1973) 对 y 和 x 的名称作过一些讨论，认为这些名称是从普通代数学中 借用过来的，用在回 归分析中不够合适。他们指出，变量x从概率的意义说，并不是自成系统，互不 相干的独立变量 (即自变量，independent variable)；另外，回归分析的目的本身正是研究 y 和 x 的相依 性的，故而把变量 x 称为自变量就令人费解了。Kendall 和Stuart的意见无疑是正确的，但目前对 y 和 x 还没有统一的名称，不同的领域、不同的学者使用着不同 的名称 (表1.1)。鉴于我们目 前的习惯，本书将仍旧称 y 为因变量，x 为自变量。

回归方程实际上就是数学模型。回归方程或回归模型在正确地估定之后，事物的一些主要关系和变化规律，就被回归方程所描述。回归方 程看起来 很简单，例如 (1.2.1) 式，但其所能描述的内容是十分丰富的，对理论研究和生产建设都很有意义。

2. 预测

预测 (预报) 是根据已知的规律展望、预言未来的可能状况。如地震预报就是根据历史经验、观测和规律性的认识，预言未来的震情。在地 质工作 中一般不 使用“预报”一词，而习 惯于说

表 1.1 回归分析中变量的各种名称比较

y	x
响应变量 (response v.)	预报变量 (predicted v.)
目标变量 (regressand v.)	回归变量 (regressor v.)
内成变量 (endogenous v.)	外成变量 (exogenous v.)
	先成变量 (predetermined v.)
因变量 (dependent v.)	说明变量 (explanatory v.)
	自 变 量 (independent v.)

"预测"，如矿产预测、远景预测等。地质领域中的预测问题，往往不同于其他领域的预报问题。预报，一般是指对未来时刻的状况的预言，是和时间密切相联的，而地质的预测，则一般和时间关系不大。地质学所研究的对象，基本上都是早已存在于自然界的地质体。地质工作中的预测，更多地是指地域上或空间上特性的预言，或者是根据较粗略的资料和认识，对较细致的地质特征的预言。特别是后一种预言，和一般概念下的预报有着较大的区别。

地质预测问题的上述特性，多少影响着回归分析预报作用的发挥。一个回归模型预报效果的好坏，主要视其能否反映事物的基本规律以及反映的程度如何。地质回归模型预测效果的优劣，主要取决于模型所赖以建立的那部分地质体和被预测的地质体两者的基本条件和性质相同或相似的程度。这种条件在地质工作中往往难以保证，因而回归分析在地质工作中的预报效果有时不够理想，这大概是其中的一个重要原因。

预测（预报）并不限于对未来状况的预言，人们也可以根据已知的现时状况，预言（即回溯）未知的过去。这在地质工作中有特殊的意义。因为这一做法与地质工作中广泛使用的"将今论古"的方法是完全一致的。如 Wells (1963) 曾根据珊瑚化石和蓝绿藻化石上的月生长线及季节纹，估计化石生活时期一年的天

数。根据显生宙不同时代标本的研究结果，得到一条回归直线（图1.2），这说明一年里的总天数不是固定不变的，随着地球自转速度的衰减，年总天数越来越少。若假设地球自转速度的衰减率是个常数，由图1.2的直线回溯20亿年前的情况，当时一年里大约有800到900天，可分为33个月左右，每个月可能为26天（对20亿年前做这种推断，根据是不充分的，见下文）。

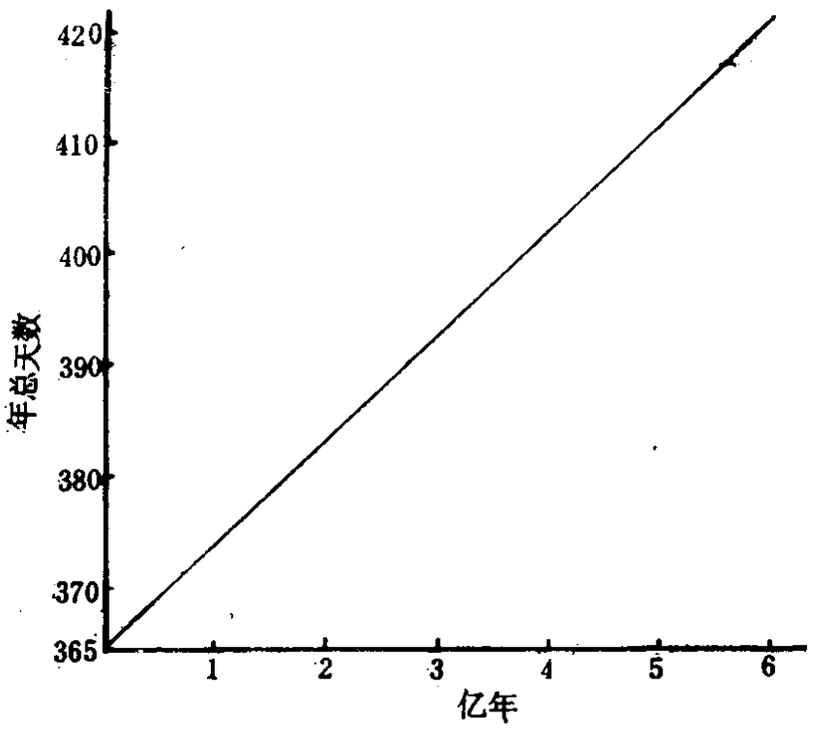


图 1.2　年总天数随地质年代的变化关系图
（据Wells, 1963改制）

3. 控制

控制是根据已知的各种因素的相互关系和规律，有意识地对其中一些因素进行调节，使系统处在所希望的状态。这对生产中如何创造优质、高产、低消耗的条件，有着指导意义。然而，在地质学中，这种应用比较有限。回归分析可以帮助分析地质现象中各种因素之间的关系，但这些因素却是不可调节的。这绝不是说回归分析的控制作用在地质工作中就完全没有意义了。我们知道地质学中的许多研究方法是属于反演性的。当条件未能充分被了解时（地质问题多是如此），反演的结果往往是多解的。不过，

当建立了地质回归模型之后，通过有关变量的调节，就可以了解各种可能的假想的地质环境，这对分析问题还是很有参考价值的。如生油岩大量生油门限的统计预测模型（张启锐，1987），其回归方程为：

$$\hat{T}=208.75/\log y+22.49\log D-141.23$$

它具体反映了生油门限的温度(T)、深度(D)和生油岩年龄(y)三要素的相互关系。若已知某盆地中心深度2500米处地层的年龄为35百万年，那么其门限温度应为97℃。若y为8百万年，D为3450米，那么其门限温度就应达到142℃。这一例子表明，虽然盆地的门限深度和生油岩年龄不是人们所能控制的。但通过上述方程可以方便地求得生油门限三要素的种种组合状况，对了解一盆地的生油条件，还是很有用的。

回归分析的描述作用还可以派生出另外两种应用：一是在已确立的回归方程的基础上，由因变量逆推相应的自变量值；二是根据回归方程中的有关参数来表示有关变量的关系，或者与理论的或经验的结果进行对比。严格地说，这两种应用所要求的变量关系，并不属于回归分析研究的范围，因为这种关系不应是随机的，而应是确定的函数关系。所谓确定的函数关系是说有关变量的关系是唯一的。如变量A的一个值只能对应于一个唯一的变量B的值。若A和B的关系为线性函数关系，那么，它们完全可以用一条直线来描述。但在实际工作中由于存在观测误差和仪器误差，实际观测值并不能正好构成一条直线，而将散布在一条直线附近。这说明有关变量观测值之间的关系包含了随机性，从而这些变量函数关系的估定就成为统计问题。我们姑且仍称这一类关系为函数关系。这一类函数关系的分析方法和回归分析的方法并没有多大区别。当变量关系不密切、拟合程度低，且有关变量都有误差时，其分析结果作第二种应用，即用于预报是比较合适的，对其他应用就不合适了。如果相关系数接近于1，这时函数关系和回归关系的区别不大，可以放心地应用于各种场合。

上述问题对于地质工作者来说比较重要。许多知名的数学地

质工作者认为：地质学中各种因素是随机性的，而不是确定性的（Krumbein,1975；Vistelius, 1975）。如果这一观点是正确的，那么回归分析在地质工作中的应用，似乎就局限于作预测这一方面了。

以上是从严格的意义讨论回归分析的应用。实际工作者往往不受上述理论的约束，总试图使回归分析的结果起更大的作用。特别是当人们对所研究的问题了解得还比较少的时候，更是如此。不过上述理论规定应当说是正确的，我们应当努力做到与规定一致。

关于预报，也有两种情况，一是内插性预报，一是外推性预报。这两种预报无论在性质上还是在方法上，都有很大区别。内插性预报是指在已知自变量的取值范围内，对因变量进行预报。这种预报一般比较有把握，因为事物的内在规律，在观测数据的范围内已得到反映。外推性预报则不同，这种预报是把观测范围内的规律，推广到观测范围以外去。当外推区段的规律和观测区段相同时，其外推结果和实际相符，若两者不一致，外推的结果将是谬误的。如前面提到的，根据生物化石的生长纹推测生物体

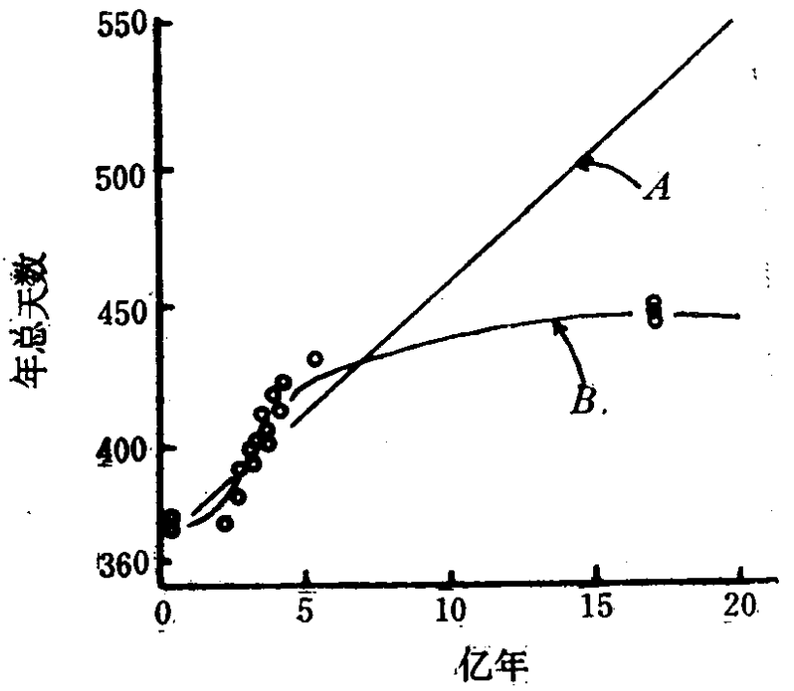


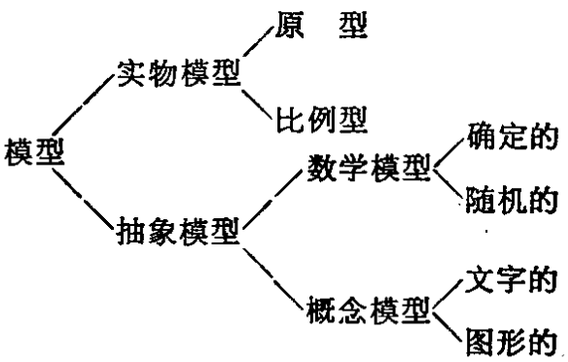
图 1.3　年总天数与年代的关系图

A—直线为理论关系；*B*—是对实际观测化石生长纹数字的拟合曲线（据Pannella, 1972）

生活年代中一年里的总天数的图1.2的直线是按每100年地球自转速度慢2毫秒计，根据显生宙的生物化石生长线测得的。若将其用于回溯前寒武纪时的情况，根据则嫌不足。Pannella（1972）曾指出，把显生宙的规律推广到前寒武纪去，是有问题的。他对20亿年前的藻化石生长纹带统计结果发现，当时的一年总天数远少于据图1.2的直线外推的结果，一年只有445天（图1.3）。

§1.3 回归模型在地质学中的位置

理论上，模型是一些具有共同性质的事物的理论抽象，其中舍弃了非本质的东西。模型有各种形式，总的说来，它可以分成实物模型（或称为物理模型、比例模型等）和抽象模型两大类。前者如李四光教授在创建地质力学中用到的泥巴模型。抽象模型又可分为概念模型和数学模型。以下给出了模型的一般分类树：



地质学中广泛应用的主要是概念模型。现在人们对数学模型的兴趣也越来越浓。

根据研究对象的性质，数学模型又分确定的和非确定的（即概率的或随机的）。这两种数学模型在地质工作中都有应用，但鉴于地质因素基本上是随机的，故而目前用得较多的还是随机模型。通过回归分析建立起来的回归模型，就是随机模型的一种。

回归模型通常包括：(1)回归函数或回归方程；(2)随机误差。如通常所说的回归模型❶

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+e \qquad (1.3.1)$$

---

❶ 更一般的形式是$y=\beta_0+\beta_1\varphi_1+\beta_2\varphi_2+\cdots+\beta_p\varphi_p+e$，其中$\varphi_i=\varphi_i(x_1,\cdots,x_q)$，$i=1,2,\cdots p$，当$q=p$，$\varphi_i=x_i$时，即为(1.3.1)式。

就是一个正态线性模型。所谓线性是对系数$\beta_0,\beta_1,\cdots,\beta_p$而言，而正态则指误差$e$遵从标准正态分布$N(0,\sigma^2)$。即

$$E(y|x)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \qquad (1.3.2)$$

是回归函数，或理论回归函数。当通过样本数据的拟合，各回归系数估定之后，得

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p \qquad (1.3.3)$$

称为经验回归函数，或简称为回归方程。

此外，回归模型还分线性的和非线性的。由于线性模型在实际应用中具有较强的适应性，故而相应的线性假设在当前仍然是许多数学理论的重要前提。它在计算上比非线性的要方便得多。尽管有不少学者对非线性模型做了大量的工作，也有不少的进展，但用起来多感不便。

实际上，自然界里许多事物之间或一事物诸特性之间的关系本质上近似于线性，其次，非线性的问题往往可以通过适当的转换，将其变为线性的。因此，只要掌握了线性回归模型的种种方法，就可以较方便地处理大量的复杂问题。

§1.4 数据类型与回归分析

在地质工作中，数据的重要性正在逐渐增加。如果要深入地认识某一地质现象，进而建立定量的回归模型，就要具备有关变量的一套观测数据。

数据是反映事物、性质的很好标志，是一定条件下的综合指标。因此，在开始一项研究工作时，最基本的工作之一，就是设法获得全面系统的数据。但是在实际地质工作中，常见的困难是搜集不到足够的、符合要求的数据。在研究程度高的地区，虽然数据资料很多，但往往不完全，或精度不完全符合要求；而在勘探程度低的新区，数据又往往很少，这些状况都给地质回归模型的建立带来不少的困难。

数据按其本义来说，是定量的。但在实际应用中，它们即可以是定量的，也可以是定性的，或者是两者的混合。

定量资料可以直接用于回归计算，而定性资料通过适当的办法，转变成数据形式后，也可以作回归分析。一般地说，数据可以根据其性质，分为四类：

1．名义型数据

这是一种纯粹的数据符号，没有量的概念。如有矿和无矿，早于、等于古生代和晚于古生代等。这些有无、是否、上下、早晚之类的概念，可以用“0”和“1”两个数构成的二态变量（或称为伪变量）来表示。若记有矿为x_A，无矿为x_B，则$x_A=1$，$x_B=0$。这时x_A和x_B之间，有且仅有以下三种关系：

$$x_A=x_A;\quad x_B=x_B;\quad x_A\neq x_B。$$

即只有等与不等的关系，此外再不能有其它的关系可言。0和1只起着“名义”的或符号的作用，其量的含义已不复存在。在线性条件下，任意两个数都可以代替0和1，其回归计算的效果完全相同、只不过0和1用起来方便罢了（参看第八章）。

2．有序型数据

有序就是指有先后次序。如地质年代就有先后顺序。震旦纪老，第三纪新。从震旦纪到第四纪就构成了一个序列。若将震旦纪年龄记为x_Z，寒武纪年龄记为$x_\in$，那么两者的新老关系，不仅可以有等与不等的概念。还可以有$x_Z>x_\in$，$x_\in<x_Z$。但这里不包含“大多少”、“小多少”的概念。若记奥陶纪年龄为x_O，那么同样有$x_Z>x_O$。三者排成序列，则有$x_Z>x_\in>x_O$。我们知道，这三个纪的年龄间隔并不相等。若以$3>2>1$代表上述序列，那么这三个数要代表的内容和它们本来量的含义就有矛盾。因此，对这类数据不能简单地直接作算术四则运算，如$x_Z-x_\in=x_O$是不成立的。实际应用中，往往不用具体的数表示有序关系。人们习惯于把它们变换成名义型数据，然后再作处理（见第八章）。序列中若有m个状态，就要用$m-1$个二态变量来表示。

3．间隔型数据

这一类数据所包含的量的含义比前者多一项内容，即不仅可以比较大小，而且还可以确定相差的量。温度是最典型的间隔型数据之一，如10℃、20℃和30℃。我们不仅可以比较温度的高低，还可以知道10℃比20℃低10℃，30℃比20℃高10℃。这一量的概念的引入，使间隔型数据和前两种类型数据有了重大的区别。它具备了通常意义下数据的性质，可以作复杂的四则运算。但这类数据仍包含了某种人为的因素。首先“间隔”的确定有一定的任意性，如温度有摄氏和华氏两种温标，两者的“间隔”并不相同，但间隔一旦确定，就成为数据比较大小的标准。如30℃比20℃多10个间隔，20℃又比10℃多同样的10个间隔。

间隔型数据的另一个重要特点是，0点的确定是任意的。如0℃就是人们参照冰点的温度确定的，它并不意味着0℃时温度等于零，即没有温度。

4．比率型数据

这种数据的突出特点是，零点具有明确的含义。如重量，它尽管可以有不同的计量单位，但重量等于零时，它的概念很明确，不会因计量单位不同而有不同的含义。而且，任何一种计量单位都可以简单地通过一个比例常数换算成另一种相应的单位。如1市斤＝0.5公斤，那么2市斤＝2×0.5公斤＝1公斤。0.5就是两者的比例常数。这样一种简单的比例关系在间隔型数据中是不存在的。

从以上的讨论，可见，四种类型数据按其叙述的先后顺序，其量的概念，后者都比前一类型多一些内容。习惯上把前两类称为定性数据，后两类称为定量数据。

在回归分析中，上述四种数据都可以得到处理，但是有序型变量不宜直接作回归分析，需变换成名义型变量，即变换为二态变量后，才能得到正确的处理。鉴于地质工作中定性资料的比重还比较大，而且定性数据的回归分析目前正受到人们的关注，故本书专辟一章，即第八章作较详尽的讨论。

关于数据，还有一点要特别引起人们的注意，即定和型数据