9.3 一元线性回归分析

研究变量之间的关系，无论在理论上还是实际应用中都是重要的课题.客观世界中变量与变量之间大多存在这样或那样的联系.一般说来,变量间的关系可分为确定性的和非确定性的两种. 确定性关系是指可以用函数来描述的关系. 非确定性的关系是指统计意义上的相依关系.例如人的身高和体重存在着关系；人的血压与年龄之间存在着关系；温度与湿度存在着关系.但这些关系都没有密切到确定性的关系.回归分析是研究这种非确定性关系的一种最常用统计方法和技术,目的是寻找变量之间近似的函数关系,其中线性回归模型是一类应用最为广泛的回归模型.

在统计学中，研究变量间的相依关系时，讨论的变量可以是随机的，也可以是非随机的，但不能全部是非随机的.比如，在方差分析中，我们重点考查一个或多个因素如何影响试验指标的均值,在这里因素视为取有限个值的变量,一般是非随机的变量,而试验指标是随机变量.

9.3.1 回归模型

假设变量和变量存在相依关系.这里是随机变量,可以是随机变量也可以是非随机的变量.一般说来,如果的取值是可以事先控制的,那么看成是非随机的.为了讨论方便，我们总假定是非随机的.在实际应用中,我们本章讨论的回归技术同样适用于是随机变量的情形.

与的相依关系体现在给定条件下，的分布与有关，其分布函数记为.分布函数刻划了与的相依关系.在实际问题中,由观察数据获得的准确估计是难以做到的,也不便于应用.于是人们考虑在给定条件下的期望

.

是的函数，它在平均意义下描述了变量如何影响变量，这个函数叫做对的回归函数. 这里和分别称为自变量和因变量.这种称呼并不意味着它们之间有因果关系.如果我想利用来预测,和分别称为预测变量和响应变量.一般情况下，我们常把叫做解释变量，叫做响应变量.

记

=

则,回归模型表示为:

 (9.1.1)

这样，响应变量的值可解释为由两部分构成:一部分由解释变量的影响所致;另一部分由其他众多未加考虑的因素包括随机因素的影响所致,把它视为随机误差.

如果解释变量取定一组不完全相同的值,对应的响应变量的观察结果依次记为,那么我们的样本为.基于模型（9.1.1），用样本表示的模型为：

 (9.1.2)

寻找回归函数是回归分析的主要任务.回归分析所研究的主要问题就是如何利用样本，对回归函数进行推断，包括估计、检验等等.在寻找回归函数之前,我们需要先对回归函数作种种假设.对回归函数的不同假设，就产生种种不同的回归模型.如果只假定回归函数存在，而对其形式不作任何假定，则称为非参数回归模型.假设回归函数的数学形式已知，但其中包含有限个未知参数，这样的模型称为参数回归模型.

在参数回归模型中,我们首先需要设定的函数形式.在一些实际问题中,我们可能由专业知识知道的形式.一般情况下,我们可将数对在直角坐标系中画出它们对应的点,得到的点图称为散点图,这种散点图可以帮助我们获得的直观认识,进而指导我们设定的形式。

例9.3.1 为研究某一化学反应过程中，温度对产品得率的影响，测得数据如下

|  |  |
| --- | --- |
| 温度 | 100 110 120 130 140 150 160 170 180 190 |
| 得率 | 45 51 54 61 66 70 74 78 85 89 |

这里响应变量是随机变量，解释变量的取值是可以人为控制的，故它是非随机的.为了看出回归函数大致的形式。

散点图

由散点图可以看,这些点大致在一条直线附近,围绕一直线作上下波动.由此看出变量与有较强的线性相关关系.散点图支持了我们作出假设: .

9.3.2 线性回归系数的估计

参数回归模型中最重要的、应用最广泛的是所谓的线性回归模型.对于一元线性回归模型，我们假定回归函数为

 （9.1.3）

其中是未知参数，称为回归系数.

下面的讨论中我们均假设解释变量是非随机的，并用表示解释变量，而响应变量是随机变量，并用表示响应变量.如果解释变量取定一组不完全相同的值,对应的响应变量的观察结果依次记为,从而得到样本.

由样本表示的一元线性回归模型为

. （9.1.4）

在建立了回归模型（9.1.4）后，我们首先要做的工作就是要由样本获得回归系数的估计，进而得到回归函数的估计.为了进一步讨论估计量的性质以及作进一步推断,如假设检验,需要对回归模型作进一步的假定.通常假定随机误差相互独立,且服从方差均为的正态分布.本节我们要讨论的线性回归模型是

 (9.1.5）

在回归模型中最常用的估计方法是最小二乘估计法.下面我们用最小二乘估计法得到回归系数的估计.令

,

设的最小值点为,即有



则称分别为的最小二乘估计.

令关于的偏导数等于0得

 (9.1.6)

经整理得

 (9.1.7)

这个方程组称为正规方程组.由于不全相等，因此上述方程组有唯一解，且容易验证该唯一解是的最小值点,从而得到参数的最小二乘估计.

,

,

记

，，

,

,

参数,的最小二乘估计可表示为

, , (9.1.8)

在得到了,的估计后,便得回归函数的估计

, (9.1.9)

称为关于的经验回归函数.方程

, (9.1.10)

称为关于的经验回归方程,其图形称为经验回归直线,简称为回归直线.

方程(9.1.10)可改写为

， (9.1.11)

可见回归直线总是过中心点.

在模型(9.1.5)下,参数,的最小二乘估计也是最大似然估计.由模型(9.1.5)知,相互独立,且,因此的联合密度为

,

把上面函数看成是参数的函数,而是的观测值,便得到似然函数

,

显然,使得似然函数达到最大的正是的最大值点.最小二乘估计并不要求随机误差服从正态分布,因此说在回归模型中最小二乘估计法是适用面更宽的方法.

例9.3.2 续例9.3.1,求关于的经验回归方程.

解 计算结果见下表(P248)

由表得

,,

从而得参数估计

,,

于只得到经验回归方程

.

利用模型(9.1.5)中的假定,我们可以计算出最小二乘估计,的期望和方差.

利用,可得

,

又由模型假设知相互独立,且,于是有

,

,

,





.

进一步还可以求出与的协方差.







.

由此可见,分别是,的无偏估计.它们的方差与和随机误差的方差有关.除外，与是相关的.因此在实际应用中, 如条件允许的话，我们尽可能把自变量的值安排得分散些,以增大，从而提高,的估计精度.

进一步,由于,均为的线性组合,由正态分布的性质知,均服从正态分布,再结合以上的结果可知

,,

随机误差的方差是未知参数，也是需要估计.由于，因此很自然地，的估计可以是基于响应变量的观测值与相应的拟合值的平均平方偏差.即

, (9.1.12)

记



称为残差平方和,方差估计量可表示为

.

在模型(9.1.5)下该估计量还是最大似然估计.该估计量不具有无偏性,为得到无偏估计需求出残差平方和的期望.





 (9.1.13)

这里用到了

,

而





,



从而

.

为了得到的无偏估计，我们可取的如下估计：

, (9.1.14)

以后我们都取(9.1.14)作为的估计量.

例9.1.3 续例9.1.2,求的估计.

9.3.3 线性假设的显著性检验,或方程的显著性检验

在模型建立之前，我们也许进行过“模型识别”或某些“知识”指引着我们的选择.但到目前为止，我们建立的模型还是“尝试性的”，需要对模型作检验、诊断及改进，甚至需要对多个模型进行比较和选优.

对于线性模型，首先需要对解释变量与响应变量的线性关系作检验.对一元线性模型（9.1.6），需要检验如下假设：

 对  （9.1.15）

这里有三种等价的检验方法，使用时任选其一即可.

一、检验

在模型(9.1.6)的假定下，关于残差平方和有结论：

～，且与相互独立.

由此可得

～.

这里，从而在成立时，有

～，

选取该统计量为检验统计量，易得拒绝域为.或者计算出值，或者构造置信区间进行检验.

二、检验

取检验统计量

，即,

则在成立时，有

～，

可得假设（9.1.15）的拒绝域为 .

以上关于从检验到检验的过程纯粹是一种数学的变换.还看不出统计量的统计意义.为了看出统计量的统计意义，需考查该统计量的分子和分母的统计意义.由（9.1.13）知

 （9.1.16）

该式的左端的统计意义是明显的，它刻画了响应变量的观测值的变异性.右端的第一项是残差平方和，它与回归函数没有关系，是由随机误差引起的. 右端的第二项是什么呢？



容易证明 .由此可见右端的第二项反映了回归值的变异性，它与自变量的值、回归直线的斜率以及随机误差都有关.由的期望可清楚地看出这一点



记,称它为回归平方和,再记,那么有如下平方和分解公式：

.

容易看出，当回归直线的斜率为零时，回归平方和会偏小.而残差平方和反映了回归直线与观测数据的接近程度.因此当回归平方和偏大且残差平方和偏小时，说明直线回归的效果好.可见统计量的值越大，说明线性回归的效果越好.虽然检验和检验的效率完全相同，但检验法可以很容易地推广至多元回归模型的检验.

与方差分析类似，我们也可以列出如下的回归分析的方差分析表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F比 |
| 回归 |  |  |  |  |
| 残差 |  |  |  |  |
| 总和 |  |  |  |  |

三、相关系数检验

在实际应用中，人们常用样本相关系数

,

来刻画变量与之间的线性相关性,这里.的大小反映了与之间的线性相关性的强弱.容易想到当越大，拒绝零假设的理由越充分.

样本相关系数与统计量之间有如下关系：

,

这表示是的严增函数，故可以从分布的分位数得到的分位数.因此可得假设检验问题（9.1.15）的拒绝域为，人们已对编制了专门的表.

如果经过检验，原假设不被拒绝时，意味着不管如何变化，不随的变化作线性变化，得到的一元线性回归方程是没有价值的，这时我们称回归方程不显著.此时常需要分析线性回归效果不显著的原因以图改进模型.归结起来原因可能有以下几种:

(1)与不存在任何关系.

(2) 与存在关系,,但与的关系不是线性的,而是其他关系.此时，需要考虑其他回归模型，比如多项式回归.

（3）影响的因素除了外，还有其他不可忽视的因素（或变量）.此时，需要把其他不可忽视的因素纳入回归模型中.

当原假设被拒绝时，意味着在一定程度上会随的变化作线性变化，此时称回归方程是显著的.是否就可以认为所建立的线性回归模型就“合适了”.事件没那么简单，毕竟罗马不是一天建成的，这还得从模型的假定（前面讨论的参数估计的性质和假设检验都是基于模型的假定展开的）说起. 模型（9.1.5）中假定：（1）；（2）随机误差相互独立；（3）随机误差服从同方差的正态分布.需要检查样本数据是否与上述假定相吻合.这属于模型诊断的问题.在建立了线性回归模型后，对模型的诊断主要针对（2）和（3）.诊断的工具最常用的是残差，.常用方法是画出残差图：以为横坐标，残差为纵坐标画出残差图.理想情况下，残差与值之间不存在相依关系，并且它们的图形在水平方向看起来没有规律.

例9.1.4 续例9.1.3,对回归方程作显著性检验.

9.1.4 估计与预测

回归模型建立后，我们可用来作估计和预测.这是两个不同的问题：

(1)当自变量取定时，相应的因变量为，寻求其均值的点估计和区间估计.这是估计问题.

(2)当自变量取定时，相应的因变量为，是一个随机变量，估计一个随机变量会取何值的问题一般叫做预测问题.常采用区间形式，即寻找一个区间，使得随机变量落在该区间的概率达到为.这样的区间叫做的概率为的预测区间.

一、的估计.

由于，自然地，的点估计为



易见该估计是无偏的，其方差为





.

在模型（7.2.2）的假定下，有

～

及

～，且与相互独立，故

～,

由此可得的置信水平为的置信区间为



其中

.

二、的预测区间

如果需要给出的点预测的话，那么自然地的点预测为.由于是随机变量，我们更关注的区间预测.下面的讨论是在模型（9.1.6）的假定下.

由于

～，～，

且两者独立，故

～

因此有

～.

从而的概率为的预测区间为



其中

.

该预测区间的长度与样本量，，以及有关.愈远离，预测精度愈差.当时，预测精度可能变得很差，这种情况下的预测称作外推，作外推预测要特别小心.如果建立回归模型的目的是预测，那么在收集数据或安排试验时尽量使分散.

补充

一、可化一元线性回归的例子

在许多实际的数据处理问题中,我们常需要由观测数据得到响应变量与解释变量之间的近似的函数关系,或者说希望找出一条曲线去拟合数据.

为对数据进行分析,我们首先需要判断两个变量之间可能的函数关系，并选择可能的回归函数的形式. 如果可由专业知识确定回归函数的形式,则应尽可能利用专业知识.若不能，我们可描出数据的散点图, 通过散点图看出这些点大体上呈现什么的趋势，再与一些常见函数的图形进行比较,选择一个或几个可能的函数形式.

在选定了回归函数的形式后,通常使用最小二乘法得到参数的估计.在关于是非线性函数的场合,常常得不到估计的解析式,这时可用数值方法得到的估计.在某些情况下,我们可以对数据作各种变换将其化为一元线性回归来处理. 例如，在拟合曲线时，可两边取对数变换成线性形式：.

下面以一个例子说明上述非线性回归分析的分析步骤

例 炼钢厂出钢水时用的钢包,在使用过程中由于钢水及炉渣对耐火材料的侵蚀,其容积不断增大.钢包的容积用盛满钢水时的重量(单位:kg)表示,相应的试验次数,数据见下表.要找出变量与的定量关系式.

|  |  |
| --- | --- |
| 序号 | 序号 |
|  |  |

下面我们分三步进行.

一、确定可能的函数形式

描出数据的散点图.本例的散点图如下

观察这13个点形成的散点图,可以看出它们并不在一条直线附近,变化趋势明显偏离了直线的变化趋势.用曲线拟合这些点应该更恰当.

本例中，散点图呈现一个明显的向上且上凸的趋势,可能选择的函数关系有很多,我们可以给出如下四个曲线函数:

(1) ;

(2) ;

(3) ;

(4) .

在初步选出可能的函数关系后,接下来我们需要解决两个问题: (1)如何估计所选方程中参数? (2)如何评价所选的不同方程的优劣.

二、参数估计

以上选择的几个函数都可以通过某种变换，将非线性方程化为线性方程的形式。

以方程为例，我们作如下变换

，，

则曲线函数关系化为直线关系

，

对变换后的数据可描出其散点图

观察变换后数据的散点图,可以看出这些点大致在一条直上下波动,这初步判断作这种变换达到“线性化”是可行的.

用变换后的数据的模型可表示为

， (9.3.17)

于是可用一元线性回归的方法估计出参数.计算过程及估计见下表

从而得曲线回归方程

.

类似地可以得到其他三个曲线回归方程,它们分别是

,

,

.

4.6.3 曲线回归方程的比较

(9.3.17)表示的模型更多的是为了理论上的需要.在实际问题中,如果只是想找一条曲线去描述或拟合数据, 那么我们会更多地关注拟合效果如何,这就需要使用一些衡量拟合效果的指标,或使用诸如残差图的图形技术.尤其在涉及对多个回归方程优劣的比较时,更需要明确一些比较的指标.比较指标不是绝对的,在实用中模型的选择需要考虑多方面的因素.下面给出两个常用的指标.

(1)决定系数.决定系数定义为

,

越大表明回归曲线拟合的效果越好.

(2)剩余标准差. 剩余标准差定义为

,

越小表明回归曲线拟合的效果越好,

在给定数据后,不同的曲线回归的选择不会影响的值,但会影响残差平方和,从而两种选择准则是一致的,只是从两个不同侧面作出评价.

了第一个曲线回归方程的残差平方和的计算过程,决定系数和剩余标准差分别为

,.

其他三个方程的决定系数及剩余标准差可同样计算,将它们列于下表

从如表可以看出,第一个曲线方程的决定系数最大.以决定系数为准则,则认为第一个曲线回归模型最佳.