9.4 多元线性回归

本节讨论因变量对多个自变量的线性回归问题，与一元回归类似，基本内容包括：回归模型，参数估计，参数及模型的检验，模型的诊断，模型的比较及选择.

9.4.1 多元线性回归模型及参数估计

对变量进行了次独立观测，得样本

，

第一个简单的元线性模型（标准模型）是

 （9.4.1）

第二个常用的元线性模型是

 （9.4.2）

记

,,,



模型(9.4.2)可表示为

 (9.4.3)

参数的LSE（即最小二估计）就是如下函数的最小化之解.



可用微分法求出的最小值点,令



即得正规方程



简单起见,设是列满秩矩阵,则正规方程有唯一解,它就是的LSE

. （9.4.4）

在模型(9.4.1)下,该估计的期望和方差(即协方差阵)分别为

.





.

可见是的无偏估计.

在线性模型中,的估计也是一个重要内容.称



为残差平方和.可以证明

,

故的无偏估计为

.

易得

，

其中.

在模型(9.4.2)下,由于是的线性函数,故

～ （9.4.5）

对于残差平方和,有结论

～,且与相互独立. （9.4.6）

下面介绍多元线性回归模型的预测问题.

给定自变量的取值，预测相应的因变量的取值.如果要用一个值预测话，自然地，预测值为



一般地，需要给出预测区间.由于

～，

～，

且两者独立，故

～

又与独立，故

～

因此的概率为的预测区间为



其中

.

8.3.2 线性假设的检验

本节的讨论都是在模型(7.3.2)的假定下.考虑检验问题

 对  （7.3.6）

其中为的已知的矩阵，且秩为.

下面几个常见检验问题都属于（7.3.6）的范畴.

(1)检验回归方程中某个系数是否为零：

 对 

取，则上述检验就化为（7.3.6）.

这个检验问题的含义是：当时，变量其实对没有影响，或者变量的作用可以由其他变量担当，因此无需把它引入回归方程.在回归系统中包含大量自变量时，人们总是希望把那些不起作用或作用不大的变量舍去，而要确定哪个变量可以舍去就需要作此检验，这实际上是“变量选取”的问题.检验问题 对 是检验回归方程是否过原点，在一些特定的回归问题中可能需要进行这样的检验.

(2)回归显著性检验



取，则该检验就化为（8.3.6）.

这个检验问题的含义是：如果原假设成立，表示所有自变量对都没有影响，这个回归方程完全失去意义.反之若原假设不成立，则说明至少有若干自变量与有关，这个回归方程不是完全没有作用.由于这一意义，这个检验问题有时被称为“回归显著性检验”.

现在来讨论检验问题（7.3.6）.在原模型下，依据LSE,可得残差平方和，它反映了实际数据与模型的偏离程度.用统计学上的习惯说法，反映了模型与数据的拟合程度.如果在原模型的基础上，把原假（7.3.6）考虑进去，那等于有了一个新模型，它是原模型的“缩小”.在新模型下，依据LSE,可得残差平方和.自然地，总有.但是，如果原假设的确是真的，那么应当会小.由此可得这样的检验想法：当较大时就拒绝原假设，不然就接受它.但是的大小要与“背景噪声”去比较才有意义，并要考虑它们的自由度，我们取检验统计量：



可以证明在原假设成立时，～，从而可得检验问题（7.3.6）的拒绝域：.

为了计算检验统计量，我们需要求在约束条件下的残差平方和，这个问题在数学上并不复杂，并且有显示表达式.这无非是个带约束条件的极值问题.即

在的约束条件下,使最小.

利用拉格朗日乘数法可得在约束条件下的参数的LSE



从而可得







即

.

在具体的检验问题中，可以使用以上公式,也可将约束条件“融入”于原模型中，得到一个无约束的线性回归模型，然后再计算.

对回归方程作显著性检验，即检验假设



将约束条件“融入”于原模型中得新模型：



该模型中，的LSE为,从而



此时是总平方和.常记为.因此该检验问题的拒绝域为



与一元回归分析一样,总平方和可分解为两项

.

可以证明.记,称它为回归平方和.并称

.

为对的样本复相关系数.它衡量作为一个整体与的线性关系的大小.称为决定系数.它度量了响应变量的变异性中能够被模型解释的比例.显然,且越接近于表明模型拟合效果越佳,因此人们常用度量模型的拟合效果.但在应用中要小心.因为选择越多的解释变量,一般会越大(至少不会变小).这样如果仅以来衡量模型的优劣就会导致在模型中引入大量的解释变量,从而出现过拟合的现象.一个模型如果拟合效果很好但预测效果很差的原因往往是过拟合造成的.

对单个参数作检验假设

 对 

这里,从而



其中是矩阵的第个主对角线元素.该检验统计量为



拒绝域为



也可等价地用检验法,检验统计量为



拒绝域为



也可通过构选的置信区间作检验.由

～

可得的置信水平为的置信区间:

7.3.3 回归诊断

本节只对回归诊断的内容作概要性介绍.

线性回归模型中，我们作了如下一些假定：

（1）是自变量的线性函数；

（2）随机误差项相互独立；

（3）随机误差项的方差相同.

在进一步讨论假设检验、区间估计以及预测区间时，还假定了服从正态分布.到目前为止，我们所有的研究都是基于以上假定展开的.这里就有一个重要问题，在实际的问题中，当有了数据后，如何考察这些假设是否合理.如果这些假设是合理的，那么前面讨论的估计、检验以及应用都是可靠的.可如果实际数据与这些假设有较大的偏离，那么前面所得的结论就不再成立，这个时候我们需要采取什么措施？这是回归诊断中所要解决第一个问题.回归诊断所要研究的另一个重要问题是探查对统计推断有较大影响的数据点，这样的点称为强影响点.

残差是诊断的重要工具，残差是因变量的实际观测值与回归值之差：

.

如果模型正确的话，我们可把残差看作误差的观测值，那么残差应该具有随机误差的一些性状.因此我们可以通过这些残差以及基于它们的统计量来考察模型假设的合理性.残差有好几种.前面所说的残差我们叫做普通残差.普通残差向量为

，

其中.

另一种残差是学生化残差：



这里用到了.

再有一种残差是预测残差.就是原模型中剔除第个数据点后拟合线性模型，然后利用该模型对作点预测得预测值为，那么预测残差为



预测残差与普通残差的关系为

.

残差图是一种回归诊断的在直观工具，这种方法简单，应用上效果很好.所谓残差图就是以某种残差为纵坐标，以任何其他量为横坐标的散点图.横坐标有多种选择，常用有如下三种：

1. 以拟合值为横坐标；
2. 以回归的自变量为横坐标；
3. 以观察序号或时间为横坐标.

诊断的项目包括：回归函数是线性函数的诊断；误差方差齐性的诊断；正态性诊断；误差独立性或不相性诊断.

可通过（或）的残差图对回归函数是线性函数作诊断.当从残差图上发现回归函数可能非线性时，就要设法改进模型，比如多项式回归，或非线性回归，或引入新的回归变量.

误差方差齐性的诊断的方法有: (1)可通过的残差图对误差方差齐性作诊断.(2)作方差齐性的假设检验.比如Hartley检验,Barlett检验.但是这种方法的实施要求每个试验点上有重复试验.

如果残差图显示了方差非齐性或经检验认为方差非齐性，则可采用两种“治疗”方案：一是对自变量或因变量作变换，使误差方差近似相等.另一种方案是改用加权最小二乘估计.

这里简单地介绍方差稳定性变换.一般根据的方差与期望的关系选择变换:

1. 若∝,则作变换;
2. 若∝,则作变换;
3. 若∝,则作变换;
4. 若∝,则作变换;
5. 若∝,则作变换.

在应用上,首先从残差图粗略地考查一下与可能存在的关系,然后求出对应的变换,对变换后的数据拟合回归模型,再考察残差图.可以同时作几种变换,再进行比较,从中选优.

正态性诊断的方法有多种.（1）通过的残差图作诊断，理想情况下大约有95%的落在[-2,2]中,大约有68%的落在[-1,1]中,且各残差不呈现任何趋势. (2)可以对残差作正态性检验.检验方法有:正态概率纸,Q-Q图,峰度偏度检验,J-B检验,W检验,D检验.当经过检验认为非正态时,可考虑对因变量作适当的变换,以使变换后的新变量更接近正态分布.

这里也简单地介绍正态性变换.假设变量取正值,Tukey于1957年提出了如下幂变换



Box和Cox于1964年提出了如下修正的幂变换



上述变换只适用变量取正值的情形.对于可以取负值的场合,可先将变量作平移,再作变换.

如果因变量的观测值是依时间顺序观测得到的,则它们构成一个时间序列.实用中常对它们建立回归模型,比如对时间的趋势拟合, 时间的周期拟合,自回归拟合.我们常用残差对拟合的模型作诊断.而其中残差的不相关性诊断是其中一项重要的内容.诊断方法有: (1)利用残差图.(2)作DW(Durbin-Watson)检验,检验统计量为

.

分位数已制成表格.

对于回归诊断要解决的第二个问题:探查对回归推断有大影响的数据,即强影响点.本节不再讨论,有兴趣的同学可查阅有关资料.

7.3.4 自变量的选择

当我们应用回归分析去处理实际问题时,碰到的第一个重要问题就是选择自变量.一般说来,根据问题本身的专业理论及有关经验,人们可罗列出很多与因变量有关的自变量,其中有一些自变量可能根本没有影响或影响很小.如果回归模型中把这样的变量都包含进来,不但计算量大,而且估计和预测的精度也会下降.在一些情况下,为获得自变量的观测数据需付出很大代价.基于这些原因,在应用回归分析时,对自变量作精心选择是十分必要.

容易想到变量的选择主要有两方面内容: (1)制定选择标准或准则.不同的标准会得到不同的“最优”.不同的应用目的会导致运用不同的标准.(2)实施选择.具体说就是根据选定的准则,设计合适的程序或算法,以选出最优模型.

下面简单介绍几种选择准则,至于算法的问题则不在本课程的内容之中.下面从不同角度给出选择准则.

1. 从拟合角度考虑.

准则1. 修正的复相关系数达到最大.这里的平方为



注意这里的不是指个自变量,而是指设计矩阵的列数.以下的都是这个意思.

准则2.平均残差平方和达到最小.平均残差平方和为



1. 从预测角度考虑

准则3.预测偏差的方差达到最小. 预测偏差的方差为



准则4.平均预测均方误差达到最小. 平均预测均方误差为



准则5.准则,统计量为



这里是从全模型导出的的估计.选择法则是:点最接近第一象限角平分线(即小),且最小.当然,在很多情况下,这两个条件不会在同一点上达到,此时需要根据具体情况以及一些附加信息来决定选哪一个模型.

准则6.预测平方和达到最小.

先说一说什么是预测平方和.考虑线性回归模型



剔除第个数据点,得到模型



此模型的LSE记为,然后利用该模型对作预测



预测偏差为



那么预测平方和为



而就是选模型的预测平方和.

1. 从极大似然估计出发

准则7.AIC准则,也叫Akaike信息准则.



选择准则是:选择使AIC达最小的模型.

变量选择问题是一个十分复杂和困难的事件.除了与选择准则有关外,还与许多其他因素有关,比如所考察的模型,参数估计的方法,是否有异常值、强影响点,自变量之间的共线性.当然,变量选取,乃至更一般的模型选择基本上不能算是一个数学问题.尽管数学方法对模型选择有一些帮助,但在处理一个具体问题时,模型的正确选择在根本上要依赖于所研究问题本身的专业知识和实践经验.这一点很重要,当应用某个准则和方法选出了“最优”模型,明显地与专业理论不一致时,首先需要重新考虑我们的统计结论,仔细审查有无异常点,有无共线性,甚至有无计算错误.

8.3.5 估计方法的改进

一、加权最小二乘估计

在线性回归模型中，如果条件不满足,那么最小二乘估计是否还具有上述BLUE等性质呢? 如果不具有,又该如何改进呢?下面就讨的这些问题.

首先指出,若不满足，但成立时,最小二乘法仍然可以使用,算得的LSE也还是的无偏估计,这是因为使用最小二乘估计的无偏性只涉及一阶矩,并不需要二阶矩具有什么性质.然而BLUE就涉及二阶矩性质,此时LSE一般不再是BLUE.

如果将条件改为,为已知的正定矩阵.因为是正定矩阵,存在非奇异对称矩阵,使.令,，则线性回归模型变为

.

且





对此线性模型，运用最小二乘估计可得参数的LSE



该估计称为的加权最小二乘估计.它是的BLUE.这个估计存在一个明显的弱点：必须知道，这在实用中不容易做到.在可以重复的试验中，可以由观测数据得到的估计.

二、多重共线性

在标准的线性模型下,参数的LSE的方差(即协方差阵)为



它涉及矩阵的逆,而且许多推断(区间估计,,假设检验,预测区间)都会涉及矩阵的逆.由线性代数的知识知,只要矩阵的行列式,那么矩阵的逆存在.可是当矩阵的行列式时,该估计的均方误差



会很大,这说明估计的精度会大大降低,稳定性将变得很差,的LSE不再是一个良好的估计.这时我们称设计矩阵呈病态.当设计矩阵呈病态时, 的列向量之间存在近似的线性相关性,因此也称回归因子之间存在多重共线性.在实际问题中,这种现象是大量存在的.比如经济问题中,许多变量往往具有某种程度的“同步”增长或减少,从而使它们存在着一定程度的多重共线性.

由于多重共线性可能会使LSE变得很差,因此就有必要改进估计方法.改进措施有:岭回归,主成分估计,特征根估计等.详细内容,我们不去介绍.

三、稳健估计

最小二乘估计是基于最小化残差平方和而得到的估计,它可能受到个别数据的较大影响.换言之,稳健性不够好.

近年来,稳健统计的思想和方法得到了很大发展和应用.下面简单地介绍稳健性概念.

理论上探讨一种统计方法的性质时，是从一定的模型出发.在模型的假定下去讨论某种统计方法有怎样的优良性.然而，在实际中，实际情况与理论模型严格符合的情况可以说是没有的.这种情况下就自然地引出一个要求：所使用的统计方法应具备一定的“抗干扰性”.就是说，当实际与理论上的假定略有背离时，该方法仍能保持良好的性能.不然的话，该方法理论上的优良性就完全是纸面上的，不仅没有实际意义，还可能把使用者引入歧途，这是统计方法稳健性意义的一个方面.

统计方法稳健性还有另一个方面的意思，这与数据有关.在统计理论上探讨一个方法的性质时，我们假定：数据确是从理论模型所假定的分布中的随机抽样.但在实际问题中，特别是数据量较多时，偶尔会发生少量的所谓“过失误差”（gross error）.例如把一个数据的小数位写错了.即使一个性质很优良的统计方法，如果数据中有受到过失误差影响的成份，其使用效果也会受到影响.稳健性另一方面的意思就在于：当数据中有少量受到过失误差影响时，使用效果不致受到太大干扰.

稳健性主要是一个相比较而存在的概念，不存在什么“最稳健的”统计方法.

在线性回归中，之所以有必要考虑稳健方法，是因为最小二乘法受到异常值的影响较大. 最小二乘法就是找，使得达到最小.因为平方函数增长太快，从而使最小二乘估计受个别异常值的影响很大.为了减轻这种影响，可改用增长较慢的函数替代平方函数，例如，取，则得到如下估计方法：找，使达到最小.把以上思想推而广之，就是下面的M估计.

设函数定义在,且满足以下条件:

(1) 在连续.

(2) ,在非增,在非降.

(3).

记



若满足



则称为的一个M估计.

可以看出, M估计不是一个确定的估计,而是指一类估计,与所选的函数有关.

