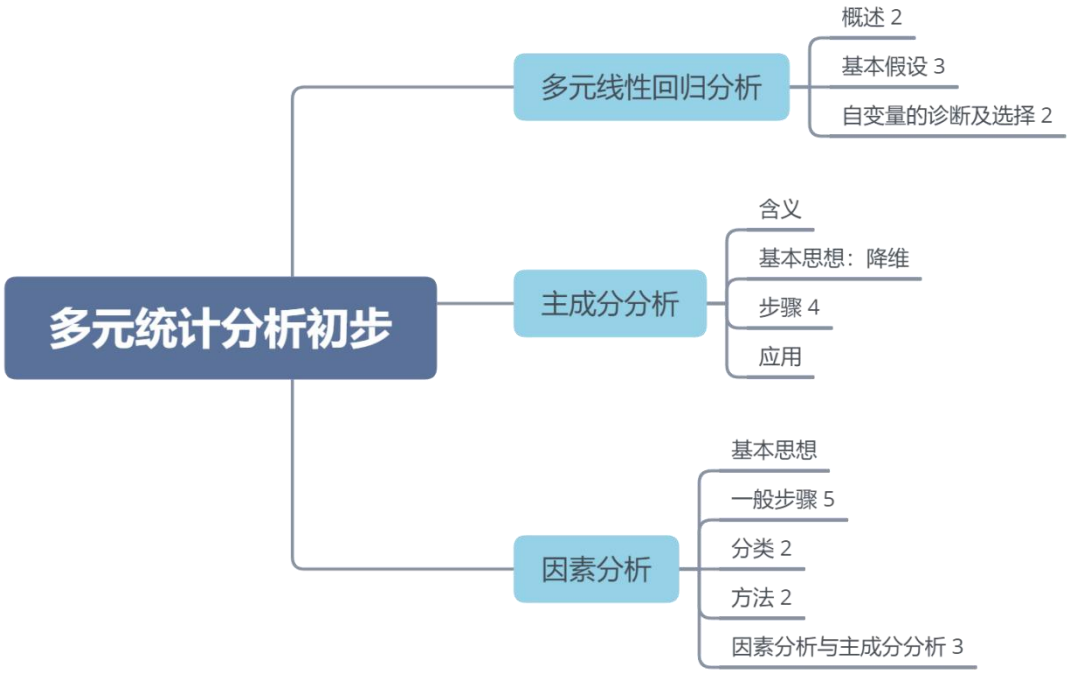
第十三章　多元统计分析初步



一、多元线性回归分析

（一）概述

1.概念

在回归分析中，若有两个或以上的自变量，就称为多元回归。在需要用多个计量资料的自变量来解释单个计量资料的因变量时，多元回归是最合适的选择；它能提供多个自变量对因变量的函数关系、提供多个备选的函数关系、提供每个关系式对实验数据的解释能力，研究者可以结合理论预期，据此做出选择。

2.回归模型

（1）多重线性回归模型

$Y=b_0+b_1X_1+b_2X_2+...+b_pX_p$

b_0:相当于一元回归方程中的常数项；

b_i：偏回归系数。当其他自变量对因变量的影响固定时，b_i 反映了第 i 个自变量 X_i 对因变量 Y 的线性影响的大小。

（2）多重线性回归模型的变型

当需要比较哪个自变量在估计 Y 时的贡献大小时，须将原始数据分别转换为标准分数，建立标准回归方程：$Z_Y=\beta_1Z_{X1}+\beta_2Z_{X2}+...+\beta_pZ_{Xp}$

Z_Y：因变量 Y 的标准分数的估计值；

Z_{Xp}：以标准分数出现的自变量；

β_1、β_2：标准偏回归系数，其中：

$$b_1=\beta_1\frac{S_Y}{S_{X1}},\quad b_2=\beta_2\frac{S_Y}{S_{X2}}$$

（二）线性回归模型的基本假设

1.自变量 X_i 是确定变量，不是随机变量；

2.自变量之间互不相关，即无多重共线性；

3.随机误差服从 0 均值、同方差的正态分布，且不存在序列相关关系；随机误差与自变量间不相关。

（三）自变量的诊断及选择

1.诊断

在进行回归分析之前，需要确定自变量是否符合基本假设，这就是诊断过程，一般需要经过异常点诊断（检测是否有个别观测点与多数观测点偏离很远，或出现过失误差）和共线性诊断（若自变量之间有较强相关关系，将很难求得理想回归方程，共线性诊断便是先对自变量间的相关性做出的判断与剔除）。

2.选择

基本上都是基于决定系数 r^2 最大原则：最优方程选择法；同时分析法（标准回归）；逐步分析法（顺向进入法——向前回归、 从无到有，反向淘汰法——向后回归、逐一剔除）；逐步回归法（先顺向进入，再反向淘汰）；阶层分析法（分层回归）；最大 r^2 增量法（先找到最大的回归方程，再增加变量）；最小 r^2 增量法等。

二、主成分分析

（一）含义

主成分分析是研究如何用少数的几个新变量来解释原来多个变量的内部结构的方法。

一般认为，两个变量相关过高，则它们是从同一个角度测量这个事物的；如果相关过低，则认为这两个变量测量的是不同事物；如果是中等程度相关，则认为两者是从不同角度来测量同一事物的。

在实际研究中，为全面系统反映事物，会用不同角度的变量来考察这个事物，但这样获得的变量往往存在较强的相关关系，即这些变量存在着较多的信息重复，直接分析现实问题，不但模型复杂，还可能因为多重共线性而引起极大的误差。

所以为了充分而有效地利用数据，要用较少的新变量代替原来较多的旧变量，同时要求这些新变量包含原变量的信息。

（二）基本思想：降维

在损失很少信息的前提下把多个指标转化为几个综合指标；通常把转化生成的综合指标称为主成分，这样在研究复杂问题时，就可以只考虑几个少数的主成分又不至于损失太多信息，从而更容易抓住主要矛盾，揭示事物内部变量之间的规律性，提高分析效率。

一般地说， 利用主成分分析得到的主成分与变量之间有如下关系:

1. 每一个主成分都是各原始变量的线性组合；

2. 主成分的数目大大少于原始变量的数目；

3.主成分保留了原始变量的绝大多数信息；

4.各主成分之间互不相关（无多重共线性）。

（三）步骤

1.对各变量数据进行标准化（消除单位等的影响）；

2.变量之间的相关性判定（若存在相关，则不能对这些变量进行主成分分析）；

3.得到主成分的表达式并确定主成分个数，选取主成分；

4.给主成分命名并结合主成分对研究问题进行深人研究。

（四）应用

主成分评价和主成分回归。

三、因素分析

（一）基本思想

因素分析在某种程度上可以看作是主成分分析的推广和扩展，也是利用降维的思想，由研究原始变量相关矩阵内部出发，把一些具有错综复杂关系的变量归结为少数几个综合因子。根据相关性把原始变量分组，使得同组内的变量之间相关性较高，不同组的变量间相关性较低每组变量代表一个基本结构，并用一个不可观测的综合变量表示，这个基本结构就称为公因子。

（二）一般步骤

1.对原始变量进行标准化并求其相关矩阵，分析变量之间的相关性；

2.求公共因子及因子负荷矩阵；

3.确定公共因子的个数，对因素负荷矩阵进行旋转；

4.根据实际需要进行因素计分；

5.对因子进行进一步的分析。

（三）分类

根据研究者对因素的确定性程度可以分为：

1.探索性因素分析

研究者事先对观察数据背后可以提取多少个因素并不确定，分析的目的在于探索因素的个数。

2.验证性因素分析

研究者根据已有的理论模型对因素的个数以及每个变量都在哪个因素上有载荷有明确的假设，分析的目的在于对假设进行验证。

对于所研究的某一具体问题，原始变量可分解为少数几个不可测的公共因子的线性丽数和与公共因子无关的特殊因子。

（四）方法

1.特征值大小（>1 时）。

2.碎石图检验又称陡坡检验（变陡时）

（五）因素分析与主成分分析

1.目的

因素分析从数据中探查能对变量起解释作用的因子及其组合系数，主成分分析则是寻找能够解释诸多变量变异的大部分彼此不相关的变量。

2.效力

在解释方面，因素分析较主成分分析更有优势。

3.其他

因素分析需要提前假设并且因素数量需要分析者假定，主成分分析不需要；

因素分析抽取因素方法有多种，主成分分析只有一种；

因素分析中因素可以旋转，主成分一般是固定的；

因素分析将变量表示成因素的线性组合，主成分分析将主成分表示成各变量的线性组合。

检验方法	t检验	卡方检验	方差分析	一元线性回归	非参数检验
应用情况	两组连续数据平均值之间是否有差异	用于计数数据	三个及三个以上平均数间差异检验	几列连续数据之间的关系，用于预测等	数据较少，使用的数据不是等距数据，而是顺序数据等
举例	初中生的男女身高有无差异	处女座与强迫症到底有没有关系	三种强度的照明条件下阅读成绩的差异	计算出智商与考试成绩的回归方程，预测在某智商下的考试成绩	10名幼儿园儿童刚入园和入园一年后两次血色素是否有明显变化