超级干货： 一文读懂回归分析

微笑如酒 2021-10-20 | 1166阅读 | 4转藏

前言

1.“回归”一词的由来

我们不必在“回归”一词上费太多脑筋。英国著名统计学家弗朗西斯·高尔顿（Francis Galton,1822—1911）是最先应用统计方法研究两个变量之间关系问题的人。“回归”一词就是由他引入的。他对父母身高与儿女身高之间的关系很感兴趣，并致力于此方面的研究。高尔顿发现，虽然有一个趋势：父母高，儿女也高；父母矮，儿女也矮，但从平均意义上说，给定父母的身高，儿女的身高却趋同于或者说回归于总人口的平均身高。换句话说，尽管父母双亲都异常高或异常矮，儿女身高并非也普遍地异常高或异常矮，而是具有回归于人口总平均高的趋势。更直观地解释，父辈高的群体，儿辈的平均身高低于父辈的身高；父辈矮的群体，儿辈的平均身高高于其父辈的身高。用高尔顿的话说，儿辈身高的“回归”到中等身高。这就是回归一词的最初由来。

回归一词的现代解释是非常简洁的：回归时研究因变量对自变量的依赖关系的一种统计分析方法，目的是通过自变量的给定值来估计或预测因变量的均值。它可用于预测、时间序列建模以及发现各种变量之间的因果关系。

使用回归分析的益处良多，具体如下：

1) 指示自变量和因变量之间的显著关系；

2) 指示多个自变量对一个因变量的影响强度。

回归分析还可以用于比较那些通过不同计量测得的变量之间的相互影响，如价格变动与促销活动数量之间的联系。这些益处有利于市场研究人员，数据分析人员以及数据科学家排除和衡量出一组最佳的变量，用以构建预测模型。

2.为什么使用回归分析

1）更好地了解

对某一现象建模，以更好地了解该现象并有可能基于对该现象的了解来影响政策的制定以及决定采取何种相应措施。基本目标是测量一个或多个变量的变化对另一变量变化的影响程度。示例：了解某些特定濒危鸟类的主要栖息地特征（例如：降水、食物源、植被、天敌），以协助通过立法来保护该物种。

微笑如酒
关注 对话

TA的最新馆藏
- GPT PDF=王炸，让你的阅读效率提升10…
- 抖音创始人张一鸣：为何毕业多年后 原…
- 男子因一碗面被送去抢救，已经第3次了…
- 孟德尔随机化--带文献复现的实操（上篇…
- ChatGPT：为什么台湾又被称为“肾病…
- 濒临死亡，多次检查却没发现问题！患…

喜欢该文的人也喜欢 更多
- 原 为啥说“天地之间八万里”？八万之数是怎么得来的？古人的智慧 阅192
- 原 男女本就不一样。 阅157
- 老话：“客厅四不挂，吉祥富贵发”，是指哪四样？ 阅285
- 原 一页纸回答9个问题，让你做好2023年规划 阅384
- 40个你不知道的人性潜规则，让你少走些弯路 阅304

热门阅读 换一换
- 【全网最全】“将军饮马”模型及其各类变形全归纳 阅20240
- 小学数学基础知识整理（一到六年级） 阅322456
- 物业小区消防应急预案 阅27733
- 员工绩效考核全套方案（含评分标准） 阅68582
- 深度解读《义务教育数学课程标准（2022年版）》 阅58311

最新原创 更多
- 原 还得是国宝，太可爱了吧
- 原 医学上逆时针复位，往哪边转？哪边…
- 原 南亚之巨：印度2000年的历史概述
- 原 VBA窍门大公开！本篇你一定不能错…
- 原 土耳其最惨小镇，因辐射致死3500人…

2）建模预测

对某种现象建模以预测其他地点或其他时间的数值。基本目标是构建一个持续、准确的预测模型。示例：如果已知人口增长情况和典型的天气状况，那么明年的用电量将会是多少？

3）探索检验假设

还可以使用回归分析来深入探索某些假设情况。假设您正在对住宅区的犯罪活动进行建模，以更好地了解犯罪活动并希望实施可能阻止犯罪活动的策略。开始分析时，您很可能有很多问题或想要检验的假设情况。

回归分析的作用主要有以下几点：

1）挑选与因变量相关的自变量；

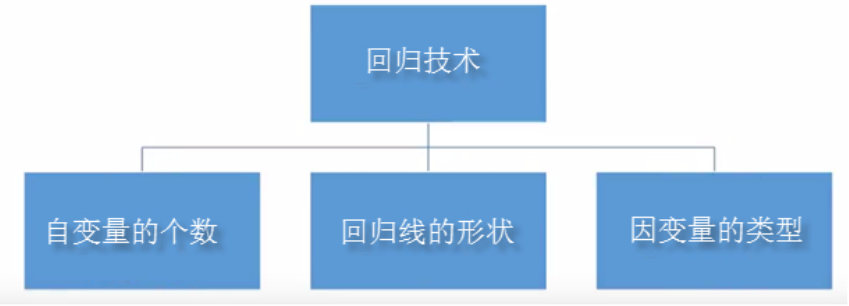
2）描述因变量与自变量之间的关系强度；

3）生成模型，通过自变量来预测因变量；

4）根据模型，通过因变量，来控制自变量。

回归分析方法

现在有各种各样的回归技术可用于预测，这些技术主要包含三个度量：自变量的个数、因变量的类型以及回归线的形状。



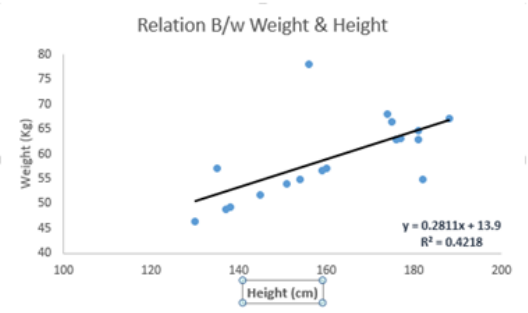
1.回归分析方法

1）线性回归

线性回归它是最为人熟知的建模技术之一。线性回归通常是人们在学习预测模型时首选的少数几种技术之一。在该技术中，因变量是连续的，自变量（单个或多个）可以是连续的也可以是离散的，回归线的性质是线性的。线性回归使用最佳的拟合直线（也就是回归线）建立因变量 (Y) 和一个或多个自变量 (X) 之间的联系。用一个等式来表示它，即：

Y=a+b*X + e

其中a 表示截距，b 表示直线的倾斜率，e 是误差项。这个等式可以根据给定的单个或多个预测变量来预测目标变量的值。



一元线性回归和多元线性回归的区别在于，多元线性回归有一个以上的自变量，而一元线性回归通常只有一个自变量。

线性回归要点：

1）自变量与因变量之间必须有线性关系；

2）多元回归存在多重共线性，自相关性和异方差性；

3）线性回归对异常值非常敏感。它会严重影响回归线，最终影响预测值；

4） 多重共线性会增加系数估计值的方差，使得估计值对于模型的轻微变化异常敏感，结果就是系数估计值不稳定；

5）在存在多个自变量的情况下，我们可以使用向前选择法，向后剔除法和逐步筛选法来选择最重要的自变量。

2）Logistic回归

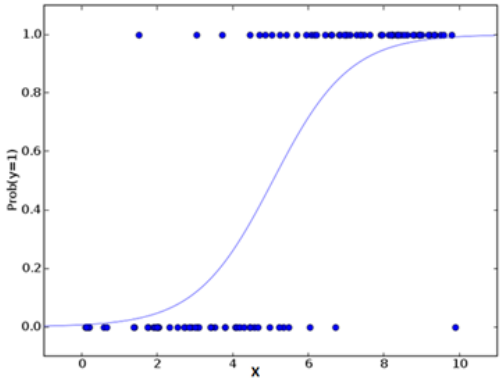
Logistic回归可用于发现 “事件=成功”和“事件=失败”的概率。当因变量的类型属于二元（1 / 0、真/假、是/否）变量时，我们就应该使用逻辑回归。这里，Y 的取值范围是从 0 到 1，它可以用下面的等式表示：

odds= p/ (1-p) = 某事件发生的概率/ 某事件不发生的概率

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) =b0+b1X1+b2X2+b3X3....+bkXk

如上，p表述具有某个特征的概率。在这里我们使用的是的二项分布（因变量），我们需要选择一个最适用于这种分布的连结函数。它就是Logit 函数。在上述等式中，通过观测样本的极大似然估计值来选择参数，而不是最小化平方和误差（如在普通回归使用的）。



Logistic要点：

1）Logistic回归广泛用于分类问题；

2）Logistic回归不要求自变量和因变量存在线性关系。它可以处理多种类型的关系，因为它对预测的相对风险指数使用了一个非线性的 log 转换；

3）为了避免过拟合和欠拟合，我们应该包括所有重要的变量。有一个很好的方法来确保这种情况，就是使用逐步筛选方法来估计Logistic回归；

4）Logistic回归需要较大的样本量，因为在样本数量较少的情况下，极大似然估计的效果比普通的最小二乘法差；

5） 自变量之间应该互不相关，即不存在多重共线性。然而，在分析和建模中，我们可以选择包含分类变量相互作用的影响；

6）如果因变量的值是定序变量，则称它为序Logistic回归；

7）如果因变量是多类的话，则称它为多元Logistic回归。

3）Cox回归

Cox回归的因变量就有些特殊，它不经考虑结果而且考虑结果出现时间的回归模型。它用一个或多个自变量预测一个事件（死亡、失败或旧病复发）发生的时间。Cox回归的主要作用发现风险因素并用于探讨风险因素的强弱。但它的因变量必须同时有2个，一个代表状态，必须是分类变量，一个代表时间，应该是连续变量。只有同时具有这两个变量，才能用Cox回归分析。Cox回归主要用于生存资料的分析，生存资料至少有两个结局变量，一是死亡状态，是活着还是死亡；二是死亡时间，如果死亡，什么时间死亡？如果活着，从开始观察到结束时有多久了？所以有了这两个变量，就可以考虑用Cox回归分析。

4）poisson回归

通常，如果能用Logistic回归，通常也可以用poission回归，poisson回归的因变量是个数，也就是观察一段时间后，发病了多少人或是死亡了多少人等等。其实跟Logistic回归差不多，因为logistic回归的结局是是否发病，是否死亡，也需要用到发病例数、死亡例数。

5）Probit回归

Probit回归意思是“概率回归”。用于因变量为分类变量数据的统计分析，与Logistic回归近似。也存在因变量为二分、多分与有序的情况。目前最常用的为二分。医学研究中常见的半数致死剂量、半数有效浓度等剂量反应关系的统计指标，现在标准做法就是调用Pribit过程进行统计分析。

6）负二项回归

所谓负二项指的是一种分布，其实跟poission回归、logistic回归有点类似，poission回归用于服从poission分布的资料，logistic回归用于服从二项分布的资料，负二项回归用于服从负二项分布的资料。如果简单点理解，二项分布可以认为就是二分类数据，poission分布就可以认为是计数资料，也就是个数，而不是像身高等可能有小数点，个数是不可能有小数点的。负二项分布，也是个数，只不过比poission分布更苛刻，如果结局是个数，而且结局可能具有聚集性，那可能就是负二项分布。简单举例，如果调查流感的影响因素，结局当然是流感的例数，如果调查的人有的在同一个家庭里，由于流感具有传染性，那么同一个家里如果一个人得流感，那其他人可能也被传染，因此也得了流感，那这就是具有聚集性，这样的数据尽管结果是个数，但由于具有聚集性，因此用poission回归不一定合适，就可以考虑用负二项回归。

7）weibull回归

中文有时音译为威布尔回归。关于生存资料的分析常用的是cox回归，这种回归几乎统治了整个生存分析。但其实夹缝中还有几个方法在顽强生存着，而且其实很有生命力。weibull回归就是其中之一。cox回归受欢迎的原因是它简单，用的时候不用考虑条件（除了等比例条件之外），大多数生存数据都可以用。而weibull回归则有条件限制，用的时候数据必须符合weibull分布。如果数据符合weibull分布，那么直接套用weibull回归自然是最理想的选择，它可以给出最合理的估计。如果数据不符合weibull分布，那如果还用weibull回归，那就套用错误，结果也就会缺乏可信度。weibull回归就像是量体裁衣，把体形看做数据，衣服看做模型，weibull回归就是根据某人实际的体形做衣服，做出来的也就合身，对其他人就不一定合身了。cox回归，就像是到商场去买衣服，衣服对很多人都合适，但是对每个人都不是正合适，只能说是大致合适。至于到底是选择麻烦的方式量体裁衣，还是选择简单到商场直接去买现成的，那就根据个人倾向，也根据具体对自己体形的了解程度，如果非常熟悉，自然选择量体裁衣更合适。如果不大了解，那就直接去商场买大众化衣服相对更方便些。

8）主成分回归

主成分回归是一种合成的方法，相当于主成分分析与线性回归的合成。主要用于解决自变量之间存在高度相关的情况。这在现实中不算少见。比如要分析的自变量中同时有血压值和血糖值，这两个指标可能有一定的相关性，如果同时放入模型，会影响模型的稳定，有时也会造成严重后果，比如结果跟实际严重不符。当然解决方法很多，最简单的就是剔除掉其中一个，但如果实在舍不得，觉得删了太可惜，那就可以考虑用主成分回归，相当于把这两个变量所包含的信息用一个变量来表示，这个变量我们称它叫主成分，所以就叫主成分回归。当然，用一个变量代替两个变量，肯定不可能完全包含他们的信息，能包含80%或90%就不错了。但有时候我们必须做出抉择，你是要100%的信息，但是变量非常多的模型？还是要90%的信息，但是只有1个或2个变量的模型？打个比方，你要诊断感冒，是不是必须把所有跟感冒有关的症状以及检查结果都做完？还是简单根据几个症状就大致判断呢？我想根据几个症状大致能确定90%是感冒了，不用非得100%的信息不是吗？模型也是一样，模型是用于实际的，不是空中楼阁。既然要用于实际，那就要做到简单。对于一种疾病，如果30个指标能够100%确诊，而3个指标可以诊断80%，我想大家会选择3个指标的模型。这就是主成分回归存在的基础，用几个简单的变量把多个指标的信息综合一下，这样几个简单的主成分可能就包含了原来很多自变量的大部分信息。这就是主成分回归的原理。

9）岭回归

当数据之间存在多重共线性（自变量高度相关）时，就需要使用岭回归分析。在存在多重共线性时，尽管最小二乘法（OLS）测得的估计值不存在偏差，它们的方差也会很大，从而使得观测值与真实值相差甚远。岭回归通过给回归估计值添加一个偏差值，来降低标准误差。

上面，我们看到了线性回归等式：

y=a+ b*x

这个等式也有一个误差项。完整的等式是：

y=a+b*x+e (误差项), [误差项是用以纠正观测值与预测值之间预测误差的值]

=> y=a+y= a+ b1x1+ b2x2+....+e, 针对包含多个自变量的情形。

在线性等式中，预测误差可以划分为 2 个分量，一个是偏差造成的，一个是方差造成的。预测误差可能会由这两者或两者中的任何一个造成。在这里，我们将讨论由方差所造成的误差。岭回归通过收缩参数 λ（lambda）解决多重共线性问题。请看下面的等式：

$$= \underset{\beta \in R^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

在这个等式中，有两个组成部分。第一个是最小二乘项，另一个是 β2（β-平方）和的 λ 倍，其中 β 是相关系数。λ 被添加到最小二乘项中用以缩小参数值，从而降低方差值。

岭回归要点：

1）除常数项以外，岭回归的假设与最小二乘回归相同；

2） 它收缩了相关系数的值，但没有达到零，这表明它不具有特征选择功能；

3）这是一个正则化方法，并且使用的是 L2 正则化。

10）偏最小二乘回归

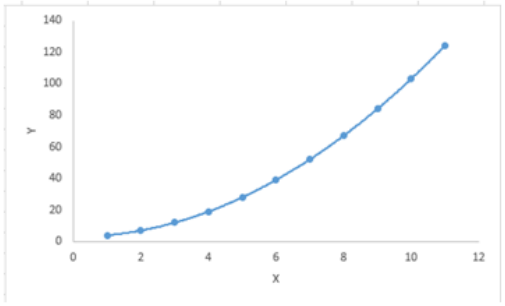
偏最小二乘回归也可以用于解决自变量之间高度相关的问题。但比主成分回归和岭回归更好的一个优点是，偏最小二乘回归可以用于例数很少的情形，甚至例数比自变量个数还少的情形。所以，如果自变量之间高度相关、例数又特别少、而自变量又很多，那就用偏最小二乘回归就可以了。它的原理其实跟主成分回归有点像，也是提取自变量的部分信息，损失一定的精度，但保证模型更符合实际。因此这种方法不是直接用因变量和自变量分析，而是用反映因变量和自变量部分信息的新的综合变量来分析，所以它不需要例数一定比自变量多。偏最小二乘回归还有一个很大的优点，那就是可以用于多个因变量的情形，普通的线性回归都是只有一个因变量，而偏最小二乘回归可用于多个因变量和多个自变量之间的分析。因为它的原理就是同时提取多个因变量和多个自变量的信息重新组成新的变量重新分析，所以多个因变量对它来说无所谓。

11）多项式回归

对于一个回归等式，如果自变量的指数大于1，那么它就是多项式回归等式。如下等式所示：

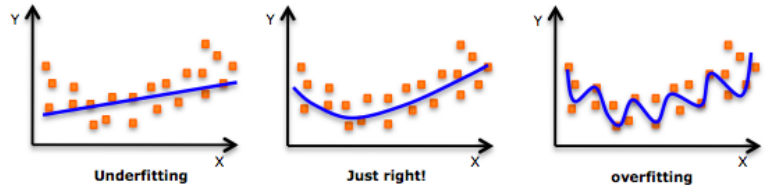
y=a+b*x^2

在这种回归技术中，最佳拟合线不是直线。而是一个用于拟合数据点的曲线。



多项式回归要点：

1）虽然存在通过高次多项式得到较低的错误的趋势，但这可能会导致过拟合。需要经常画出关系图来查看拟合情况，并确保拟合曲线正确体现了问题的本质。下面是一个图例，可以帮助理解：



2）须特别注意尾部的曲线，看看这些形状和趋势是否合理。更高次的多项式最终可能产生怪异的推断结果。

12）逐步回归

该回归方法可用于在处理存在多个自变量的情形。在该技术中，自变量的选取需要借助自动处理程序，无须人为干预。通过观察统计的值，如 R-square、t-stats和 AIC 指标，来识别重要的变量，可以实现这一需求。逐步回归通过同时添加/去除基于指定标准的协变量来拟合模型。
下面列出了一些最常用的逐步回归方法：

1）标准逐步回归法需要做两件事情，即根据需要为每个步骤添加和删除预测因子；

2）向前选择法从模型中最重要的预测因子开始，然后为每一步添加变量；

3）向后剔除法从模型中所有的预测因子开始，然后在每一步删除重要性最低的变量。

这种建模技术的目的是使用最少的预测因子变量来最大化预测能力。这也是处理高维数据集的方法之一。

13）套索回归

与岭回归类似，套索也会对回归系数的绝对值添加一个罚值。此外，它能降低偏差并提高线性回归模型的精度。看看下面的等式：

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

套索回归与岭回归有一点不同，它在惩罚部分使用的是绝对值，而不是平方值。这导致惩罚（即用以约束估计的绝对值之和）值使一些参数估计结果等于零。使用的惩罚值越大，估计值会越趋近于零。这将导致我们要从给定的n个变量之外选择变量。

套索回归要点：

1）除常数项以外，这种回归的假设与最小二乘回归类似；

2）它将收缩系数缩减至零（等于零），这确实有助于特征选择；

3）这是一个正则化方法，使用的是 L1 正则化；

4）如果一组预测因子是高度相关的，套索回归会选出其中一个因子并且将其它因子收缩为零。

14）ElasticNet 回归

ElasticNet 回归是套索回归和岭回归的组合体。它会事先使用 L1 和 L2 作为正则化矩阵进行训练。当存在多个相关的特征时，Elastic-net 会很有用。岭回归一般会随机选择其中一个特征，而 Elastic-net 则会选择其中的两个。同时包含岭回归和套索回归的一个切实的优点是，ElasticNet 回归可以在循环状态下继承岭回归的一些稳定性。

ElasticNet 回归要点：

1）在高度相关变量的情况下，它会产生群体效应；

2）选择变量的数目没有限制；

3）它可以承受双重收缩。

2.如何选择回归模型

当只了解一两种回归技术的时候，情况往往会比较简单。然而，当我们在应对问题时可供选择的方法越多，选择正确的那一个就越难。类似的情况下也发生在回归模型中。

掌握多种回归模型时，基于自变量和因变量的类型、数据的维数以及数据的其它基本特征去选择最合适的技术非常重要。以下是要选择正确的回归模型时需要考虑的主要因素：

1）数据探索是构建预测模型的不可或缺的部分。在选择合适的模型前，比如识别变量的关系和影响，应该首先执行这一步骤。

2）比较不同模型的拟合优点，我们可以分析不同的指标参数，如统计意义的参数，R-square，调整 R-square，AIC，BIC以及误差项，另一个是 Mallows’ Cp 准则。这个主要是通过将所选的模型与所有可能的子模型（或仔细挑选的一组模型）进行对比，检查可能出现的偏差。

3）交叉验证是评估预测模型最好的方法。使用该方法，需将数据集分成两份（一份用于训练，一份用于验证）。使用观测值和预测值之间的均方差即可快速衡量预测精度。

4）如果数据集中存在是多个混合变量，那就不应选择自动模型选择方法，因为我们并不愿意将所有变量同时放在同一个模型中。

5）所选择的回归技术也取决于你的目的。可能会出现这样的情况，一个不太强大的模型与具有高度统计学意义的模型相比，更易于实现。

6） 回归正则化方法（套索，岭和ElasticNet）在高维数据和数据集变量之间存在多重共线性的情况下运行良好。

诊断回归分析结果

为了理解、解释、预测某个问题，我们会进行回归分析。但事实上，选择一组优质的自变量并不是那么容易。通常我们会根据一些常识、理论基础、某些研究、专家的意见、参考文献等等选择一组自变量，来进行自变量的筛选。因此，我们需要诊断回归分析的质量——回归分析的结果诊断。

1.自变量与因变量是否具有预期的关系

每个自变量都会有一个系数，系数具有+/-号，来表示自变量与因变量的关系。从工具的得到的报告中，我们看到的系数的正负，每个自变量应该是我们期望的关系。如果有非常不符合逻辑的系数，我们就应该考虑剔除它了。

当然，有时也可能得到与常识不同的结论。举个例子，假如我们在研究森林火灾，我们通常认为降雨充沛的区域火灾的发生率会相对较低，也就是所谓的负相关，但是，这片森林火灾频发的原因可能是闪电雷击，这样降雨量这个自变量可能就不是常识中的负相关的关系了。

因此，我们除了验证自变量的系数与先验知识是否相符外，还有继续结合其他项检查继续诊断，从而得出更可靠的结论。

2.自变量对模型是否有帮助

自变量对模型有无帮助说的就是自变量是否有显著性。那如何了解这些自变量是否有显著性呢?

如果自变量的系数为零（或非常接近零），我们认为这个自变量对模型没有帮助，统计检验就用来计算系数为零的概率。如果统计检验返回一个小概率值（p值），则表示系数为零的概率很小。如果概率小于0.05，汇总报告上概率（Probability） 旁边的一个 星号（*） 表示相关自变量对模型非常重要。换句话说，其系数在95%置信度上具有统计显著性。

利用空间数据在研究区域内建模的关系存在差异是非常常见的，这些关系的特征就是不稳定。我们就需要通过 稳健概率（robust probability） 了解一个自变量是否具有统计显著性。

3.残差是否有空间聚类

残差在空间上应该是随机分布的，而不应该出现聚类。这项检查我们可以使用 空间自相关工具（Spatial Autocorrelation Tool）工具进行检查。

4.模型是否出现了倾向性

我们常说，不要戴着“有色眼镜”看人。同样，回归分析模型中，也不要带有“成见”，不能具有倾向性，否则，这不是个客观合理的模型。

我们都知道正态分布是个极好的分布模式，如果我们正确的构建了回归分析模型，那么模型的残差会符合完美的正态分布，其图形为钟形曲线。

当模型出现偏差时，可能我们看到的图形也是诡异的，这样我们就无法完全信任所预测的结果。

5.自变量中是否存在冗余

在我们建模的过程中，应尽量去选择表示各个不同方面的自变量，也就是尽量避免传达相同或相似信息的自变量。要清楚，引入了冗余变量的模型是不足以信任的。

6.评估模型的性能

最后需要做的是，评估模型的性能。 矫R2值是评估自变量对因变量建模的重要度量。

这项检查应该放到最后。一旦我们通过了前面的所有检验，接下来就可以进行评估矫正R2值。

R2值的范围介于0和1之间，以百分比形式表示。假设正在为犯罪率建模，并找到一个通过之前所有五项检查的模型，其校正 R2 值为0.65。这样就可以了解到模型中的自变量说明犯罪率是65%。在有些科学领域，能够解释复杂现象的 23% 就会让人兴奋不已。在其他领域，一个R2值可能需要更靠近80%或90%才能引起别人的注意。不管采用哪一种方式，校正R2值都会帮我们判断自己模型的性能。

另一项辅助评估模型性能的重要诊断是修正的Akaike信息准则/Akaike’ sinformation criterion (AIC)。AIC值是用于比较多个模型的一项有用度量。例如，可能希望尝试用几组不同的自变量为学生的分数建模。在一个模型中仅使用人口统计变量，而在另一个模型选择有关学校和教室的变量，如每位学生的支出和师生比。只要所有进行比较的模型的因变量（在本示例中为学生测试分数）相同，我们就可以使用来自每个模型的 AIC值确定哪一个的表现更好。模型的AIC值越小，越适合观测的数据。

回归设计常用软件

目前，用于回归设计的统计软件较多，无论是对回归方案设计，还是对试验数据处理和回归设计成果的应用分析，都有相应的软件支撑，或是自编自用的专业软件，或是具有商业性质的统计软件包，多种多样，各有特色。为了便于回归设计的更好应用，这里简要地介绍挑选或评价统计软件的基本思考以及几种回归设计常用的统计软件，以利相关人员简捷地选用。

1.统计软件的选用原则

在挑选或评价统计软件时，应从以下几个方面加以考虑：

1）可用性

一个软件如果能为用户提供良好的用户界面、灵活的处理方式和简明的语句或命令，就称这个软件可用性强。随着统计软件在可用性方面的不断进步，很多统计软件的语法规则简明、灵活、学用方便，这是人们非常欢迎的。

2）数据管理

数据录人、核查、修改、转换和选择，统称为数据管理。好的软件，如SAS(statistical analysis system)，SPSS(statistical package for thesocial science) 等的数据管理功能已近似大众化的数据库软件。统计软件与数据库软件之间建立接口，使数据管理不断深人，用起来非常方便。

3）文件管理

数据文件、程序文件、结果文件等一些文件的建立、存取、修改、合并等，统称为文件管理。它的功能越强，操作就越简单，越方便。由于操作系统本身文件管理功能较强。因此，从统计软件直接调用操作系统的命令可大大增强其文件管理功能。现在好的统计软件已设计了这类调用指令。

4）统计分析

统计分析是统计软件的核心。统计分析方法的计算机程序的数量和种类决定了数据处理的深度。有些软件，如SAS，BMDP(biomedical computer programs)等。所包括的分析过程，足够科研与管理之需。由于统计量的选择，参数估计的方法等是多种多样的，用户往往希望统计分析过程尽可能多地提供选项，这样可以提高统计分析的灵活性和深度。

5）容量

尽管处理的数据量与计算机硬件有直接关系，然而，软件的设计和程序编写技巧仍起很大作用。软件好，在一定程度上可以弥补硬件的不足，而低水平的软件会浪费很好的硬件配置。通常，统计软件应至少能同时进行不小于10个变量的上千个数据点的分析、综合、对比与预测。

2.SAS软件系统

SAS软件系统于20世纪70年代由美国SAS研究所开发。SAS软件是用于决策支援的大型集成资讯系统，但该软件系统最早的功能限于统计分析；至今，统计分析功能也仍是它的重要模组和核心功能。SAS已经遍布全世界，重要应用领域涵盖政府的经济决策与企业的决策支援应用

等，使用的单位遍及金融、医药卫生、生产、运输、通讯、科学研究、政府和教育等领域；在资料处理和统计分析领域，SAS系统被誉统计软件界的巨无霸。

SAS 是一个模块化、集成化的大型应用软件系统。它由数十个专用模块构成，功能包括数据访问、数据储存及管理、应用开发、图形处理、数据分析、报告编制、运筹学方法、计量经济学与预测等等。 SAS系统基本上可以分为四大部分：SAS数据库部分；SAS分析核心；SAS开发呈现工具；SAS对分布处理模式的支持及其数据仓库设计。 SAS系统主要完成以数据为中心的四大任务：数据访问；数据管理；数据呈现；数据分析。

SAS 是由大型机系统发展而来，其核心操作方式就是程序驱动，经过多年的发展，现在已成为一套完整的计算机语言，其用户界面也充分体现了这一特点：它采用MDI （多文档界面），用户在PGM视窗中输入程序，分析结果以文本的形式在OUTPUT视窗中输出。使用程序方式，用户可以完成所有需要做的工作，包括统计分析、预测、建模和模拟抽样等。但是，这使得初学者在使用SAS时必须要学习SAS语言，入门比较困难。

3.Excel软件

在回归设计的实践中，一些计算机软件可以解决多元回归分析的求解问题，但常常是数据的输入和软件的操作运用要经过专门训练。Excel软件为回归分析的求解给出了非常方便的操作过程，而且目前Excel软件几乎在每台计算机上都已经安装。

Excel是一个面向商业、科学和工程计算的数据分析软件，它的主要优点是具有对数据进行分析、计算、汇总的强大功能。除了众多的函数功能外，Excel的高级数据分析工具则给出了更为深入、更为有用、针对性更强的各类经营和科研分析功能。高级数据分析工具集中了Excel最精华、对数据分析最有用的部分，其分析工具集中在Excel主菜单中的“工具”子菜单内，回归分析便为其中之一。

Excel是以电子表格的方式来管理数据的，所有的输入、存取、提取、处理、统计、模型计算和图形分析都是围绕电子表格来进行的。

4.Statistica软件

Statistica是由统计软件公司（Statsoft）开发、专用于科技及工业统计的大型软件包。它除了具有常规的统计分析功能外，还包括有因素分析、质量控制、过程分析、回归设计等模块。利用其回归设计模块可以进行回归正交设计、正交旋转组合设计、正交多项式回归设计、A最优及D最优设计等。该软件包还可以进行对试验结果的统计检验、误差分析、试验水平估计和各类统计图表、曲线、曲面的分析计算工作。

5.SPSS软件

SPSS是世界上最早采用图形菜单驱动界面的统计软件，它最突出的特点就是操作界面极为友好，输出结果美观漂亮。它将几乎所有的功能都以统一、规范的界面展现出来，使用Windows的窗口方式展示各种管理和分析数据方法的功能，对话框展示出各种功能选择项。用户只要掌握一定的Windows操作技能，精通统计分析原理，就可以使用该软件为特定的科研工作服务。SPSS采用类似EXCEL表格的方式输入与管理数据，数据接口较为通用，能方便的从其他数据库中读入数据。其统计过程包括了常用的、较为成熟的统计过程，完全可以满足非统计专业人士的工作需要。输出结果十分美观，存储时则是专用的SPO格式，可以转存为HTML格式和文本格式。对于熟悉老版本编程运行方式的用户，SPSS还特别设计了语法生成窗口，用户只需在菜单中选好各个选项，然后按"粘贴"按钮就可以自动生成标准的SPSS程序。极大的方便了中、高级用户。

6.R软件

R语言是统计领域广泛使用的，诞生于1980年左右的S语言的一个分支。 R语言是S语言的一种实现。S语言是由AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言。

R是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输入，可实现分支、循环，用户可自定义功能。

与其说R是一种统计软件，还不如说R是一种数学计算的环境，因为R并不是仅仅提供若干统计程序、使用者只需指定数据库和若干参数便可进行一个统计分析。R的思想是：它可以提供一些集成的统计工具，但更大量的是它提供各种数学计算、统计计算的函数，从而使使用者能灵活机动的进行数据分析，甚至创造出符合需要的新的统计计算方法。

R是一个免费的自由软件，它有UNIX、LINUX、MacOS和WINDOWS版本，都是可以免费下载和使用的。在R主页那儿可以下载到R的安装程序、各种外挂程序和文档。在R的安装程序中只包含了8个基础模块，其他外在模块可以通过CRAN获得。

学习资料

1.书籍

1）《实用回归分析》（何晓群）

该书从数据出发，不是从假设、定理出发；从归纳出发，不是从演绎出法；强调案例分析；重统计思想的阐述，弱化数学证明的推导。

2）《应用多元统计分析》（高惠璇）

书中介绍了各种常用的多元统计分析方法的统计背景和实际意义，说明该方法的统计思想、数学原理及解题步骤，还列举了各方面的应用实例。该书将多元统计方法的介绍与在计算机上实现这些方法的统计软件（SAS系统）结合起来，不仅可以学到统计方法的理论知识，还知道如何解决实际问题。

END

作者：慕生鹏；编辑：冯夕琴；

转自：数据派THU 公众号；

原创系列文章：

来自： 微笑如酒 > 《模型》 举报/认领

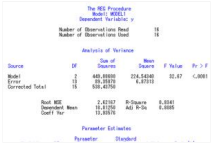
上一篇： [转] 学习笔记：一个下午的时间完成了一区文章生信分析（一）

0条评论 写评论... 发表

请遵守用户 评论公约

类似文章

更多

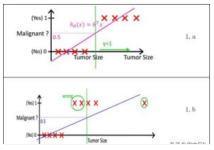


使用SAS进行变量筛选、模型诊断、多元线性回归分析

使用SAS进行变量筛选、模型诊断、多元线性回归分析。５．最小R2增量法(MINR)　　首先找到具有最小决定系数R2的单变量回归模型，然后从其...

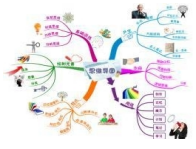
spss教程

定义变量即要定义变量名、变量类型、变量长度（小数位数）、变量标签（或值标签）和变量的格式，步骤如下：单击数据编辑窗口中的【变量视图】标签，显示如图2-5所示的变量定义视图，在出现的变量视图中...



logistic模型（logit和logistic模型的区别？）

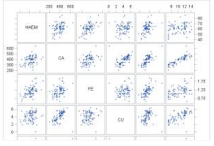
logistic模型（logit和logistic模型的区别？）关于logit和logistic模型的区别貌似是个老生常谈的问题，学习之后稍微整理一下：（1）二...



思维导图法

思维导图学习法

8555阅读



SAS系列34：多元线性回归SAS实践

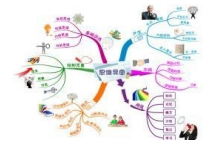
图11-1变量HAEM与变量CA、FE、ZN、P存在线性关系，但是变量CA、FE、ZN、P间也存在线性相关。图11-5至图11-7结果显示模型假设有统计学意...

数据挖掘：基于R语言的实战 | 第6章：线性模型与广义线性模型

数据挖掘：基于R语言的实战 | 第6章：线性模型与广义线性模型。本章首先在6.1节中介绍线性模型，然后在6.2节中介绍广义线性模型，在6.3...

回归分析的各种变体【一览】

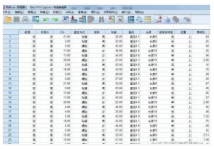
3、生成模型，通过自变量来预测因变量。4多变量回归用一个或多个自变量预测多个因变量5Logistic回归当研究的结局变量为分类变量时，采用的回归分析方法，具体包含3类，二元Logistic回归，等级Logistic...



思维导图法

思维导图学习法

8555阅读

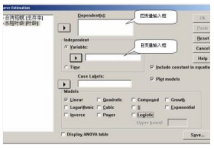


SPSS多元线性回归在医学统计分析中的应用操作及分析

SPSS多元线性回归在医学统计分析中的应用操作及分析。这里要注意的是，CA、年龄、为线性变量，可以直接作为自变量，但是吸烟、牙周炎这...

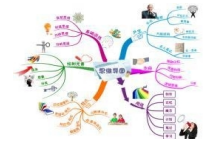
做好这些准备工作，是进行实证分析的前提！

相关性分析，也是实证分析中比较重要的部分，它一方面让我们初步把握变量之间的相关性，探讨变量之间的正负相关关系；另一方面，它可以帮助我们找到变量之间可能存在的统计分析问题，如多重共线性问题...



用SPSS进行曲线回归分析实例

用SPSS进行曲线回归分析实例曲线回归分析。(1) Linear 线性模型 (2) Quadratic 二次模型 (3) Compound 复合模型 (4) ...



思维导图法

思维导图学习法

8555阅读