

9.2 ANOVA 模型拟合

虽然ANOVA和回归方法都是独立发展而来，但是从函数形式上看，它们都是广义线性模型的特例。用第7章讨论回归时用到的lm()函数也能分析ANOVA模型。不过，本章我们基本都使用aov()函数。两个函数的结果是等价的，但ANOVA的方法学习者更熟悉aov()函数展示结果的格式。为保证完整性，在本章最后我们将提供一个使用lm()的例子。

9.2.1 aov()函数

aov()函数的语法为aov(formula, data = dataframe)，表9-4列举了表达式中可以使用的特殊符号。表9-4中的y是因变量，字母A、B、C代表因子。

表9-4 R表达式中的特殊符号

符 号	用 法
~	分隔符号，左边为响应变量，右边为解释变量。例如，用A、B和C预测y，代码为y~ A + B + C
+	分隔解释变量
:	表示变量的交互项。例如，用A、B和A与B的交互项来预测y，代码为y~ A + B + A:B
*	表示所有可能交互项。代码y~ A * B * C可展开为y ~ A + B + C + A:B + A:C + B:C + A:B:C
^	表示交互项达到某个次数。代码y ~ (A + B + C)^2可展开为y ~ A + B + C + A:B + A:C + B:C
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量y、A、B和C，代码y ~ .可展开为y ~ A + B + C

表9-5列举了一些常见的研究设计表达式。在表9-5中，小写字母表示定量变量，大写字母表示组别因子，Subject是对被试者独有的标识变量。

表9-5 常见研究设计的表达式

设 计	表 达 式
单因素ANOVA	y ~ A
含单个协变量的单因素ANCOVA	y ~ x + A
双因素ANOVA	y ~ A * B
含两个协变量的双因素ANCOVA	y ~ x1 + x2 + A*B
随机化区组	y ~ B + A (B是区组因子)
单因素组内ANOVA	y ~ A + Error(Subject/A)
含单个组内因子(W)和单个组间因子(B)的重复测量ANOVA	y ~ B * W + Error(Subject/W)

本章后面将深入探讨几个实验设计的例子。

9.2.2 表达式中各项的顺序

表达式中效应的顺序在两种情况下会造成影响：(a)因子不止一个，并且是非平衡设计；(b)存在协变量。出现任意一种情况时，等式右边的变量都与其他每个变量相关。此时，我们无法清晰地划分它们对因变量的影响。

例如,对于双因素方差分析,若不同处理方式中的观测数不同,那么模型 $Y \sim A*B$ 与模型 $Y \sim B*A$ 的结果不同。

R默认类型I(序贯型)方法计算ANOVA效应(参考补充内容“顺序很重要!”)。第一个模型可以这样写: $Y \sim A + B + A:B$ 。R中的ANOVA表的结果将评价:

- A对Y的影响;
- 控制A时, B对Y的影响;
- 控制A和B的主效应时, A与B的交互效应。

顺序很重要!

当自变量与其他自变量或者协变量相关时,没有明确的方法可以评价自变量对因变量的贡献。例如,含因子A、B和因变量Y的双因素不平衡因子设计,有三种效应:A和B的主效应,A和B的交互效应。假设你正使用如下表达式对数据进行建模:

$$Y \sim A + B + A:B$$

有三种类型的方法可以分解等式右边各效应对y所解释的方差。

类型I(序贯型)

效应根据表达式中先出现的效应做调整。A不做调整, B根据A调整, A:B交互项根据A和B调整。

类型II(分层型)

效应根据同水平或低水平的效应做调整。A根据B调整, B依据A调整, A:B交互项同时根据A和B调整。

类型III(边界型)

每个效应根据模型其他各效应做相应调整。A根据B和A:B做调整, A:B交互项根据A和B调整。

R默认调用类型I方法, 其他软件(比如SAS和SPSS)默认调用类型III方法。

样本大小越不平衡,效应项的顺序对结果的影响越大。一般来说,越基础性的效应越需要放在表达式前面。具体来讲,首先是协变量,然后是主效应,接着是双因素的交互项,再接着是三因素的交互项,以此类推。对于主效应,越基础性的变量越应放在表达式前面,因此性别要放在处理方式之前。有一个基本的准则:若研究设计不是正交的(也就是说,因子和/或协变量相关),一定要谨慎设置效应的顺序。

在讲解具体的例子前,请注意car包中的Anova()函数(不要与标准anova()函数混淆)提供了使用类型II或类型III方法的选项,而aov()函数使用的是类型I方法。若想使结果与其他软件(如SAS和SPSS)提供的结果保持一致,可以使用Anova()函数,细节可参考help(Anova, package = "car")。

9.3 单因素方差分析

单因素方差分析中,你感兴趣的是比较分类因子定义的两个或多个组别中的因变量均值。以