

## 广义相加模型

当因变量和自变量不呈线性关系时，可用广义相加模型（GAM）。GAM 可对部分或全部的自变量采用平滑函数的方法建立模型，函数可以是非参数的形式，适用于多种分布类型、多种复杂非线性关系的分析。广义相加模型中因变量的分布类型、联系函数和广义线性模型相同。根据 Y 的分布，f(Y) 不同，常见的有：

Y 的分布	联系函数名称	f(Y)
正态分布 (normal)	Identity	Y
二项分布 (binomial)	Logit	Logit (Y)
Poisson 分布	Log	Log (Y)
$\gamma$ 分布 (gamma)	inverse	1/ (Y <sup>-1</sup> )
负二项分布 (negative binomial)	Log	Log (Y)

软件自动检测应变量的类型，如果是连续性变量，自动默认采用正态分布和 identity 作为联系函数。如是两分类的，自动用 logit 做联系函数。

### 平滑拟合自由度

GAM 用  $s(X)$  替换  $\beta * X$ 。 $\beta * X$  使用的自由度为 1。 $s(X)$  的自由度取决于平滑程度，越平滑自由度越小。最极端的情况是用一个参数估计的一条直线是最平滑的；另一极端是连接每个实际的数据点是最不平滑的，这种情况下我们用尽了所有的自由度。广义相加模型可以指定平滑拟合自由度。自由度越小，平滑程度越高，但拟合程度降低；自由度越大，平滑程度越低，拟合度越高。默认值是用最低 GCV 或 GACV 值（广义交叉验证）的方法找到适当的平滑拟合自由度。

本模块不仅输出模型，而且输出每个观察记录的预测值及其标准误。不仅可以建模，还可用于预测。输入数据文件中，应变量缺失的记录，只要模型中的自变量齐全，都可得出预测值及其标准误。右击输出文件可以看到 XX\_PRED.XLS 文件，它含原数据文件（自变量完整的记录）加预测值及其标准误两个变量。

例，DEMO 数据曲线拟合 AGE、BMI 与 SBP 的关系，同时调整 SMOKE、ALH、EDU、OCCU，按性别分层拟合。输入界面如下：

## 广义相加模型

?

标题:

广义相加模型

选择分析对象:

所有数据记录

应变量

变量名	分布类型	联系函数
Systolic BP, mmhg	Gaussian	Identity

曲线拟合自变量

变量	自由度
Age, years	.
Body mass index, kg/m2	.

曲线拟合分层因子

sex

其它自变量

变量	选择
SMOKE	.
Alcohol	.
Occupation	.
Education	.

分层变量

自动检验与选择的自变量(S)的交互作用

刷新

保存

查看结果

## 输出结果

结局变量: Systolic BP, mmhg

变量分布: gaussian

模型: SBP ~

s (AGE, fx=FALSE, by=factor (SEX) ) +s (BMI, fx=FALSE, by=factor (SEX) ) +factor (SEX) +SMOKE+ALH +OCCU.NEW+factor (EDU.NEW)

Linear terms effect

	Estimate	Std. Error	t value	95%CI low	95%CI upp	P.value
(Intercept)	132.7901	3.5116	37.8143	125.9073	139.6729	0
factor (SEX) 2	-3.2621	2.1806	-1.4959	-7.5362	1.0119	0.1351
SMOKE	-1.4649	2.0214	-0.7247	-5.4269	2.4971	0.4689
ALH	0.079	2.153	0.0367	-4.1408	4.2988	0.9707
OCCU.NEW	-0.5967	1.5663	-0.381	-3.6666	2.4731	0.7033
factor (EDU.NEW) 2	1.0006	1.9579	0.5111	-2.8369	4.8382	0.6095
factor (EDU.NEW) 3	0.227	2.2837	0.0994	-4.249	4.7029	0.9209

Chi-square tests for linear terms

	df	F	p-value
factor(SEX)	1	2.2379	0.1351
SMOKE	1	0.5252	0.4689
ALH	1	0.0013	0.9707
OCCU.NEW	1	0.1452	0.7033
factor(EDU.NEW)	2	0.1629	0.8497

Approximate significance of smooth terms

	edf	Ref.df	F	p-value
s(AGE):factor(SEX) 1	4.5392	5.5667	17.3234	0
s(AGE):factor(SEX) 2	4.2087	5.2181	30.3755	0
s(BMI):factor(SEX) 1	1.2421	1.4499	0.134	0.8062
s(BMI):factor(SEX) 2	3.3312	4.229	4.9911	4e-04

Model statistics

N:	784
Adj. r-square:	0.2833
Deviance explained:	0.301
UBRE score (sp.criterion):	373.7632
Scale estimate:	364.075278719197
family:	gaussian
link function:	identity

