广义线性模型（GLM）

回归方程的一般表达式 为：$f(Y)= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+ \cdots\cdots + e$，式中 X（$X_1$、$X_2$、$X_3$ 等）是自变量，e 是残差，Y 是因变量，f(Y)表示 Y 的关联函数。

回归方程中因变量 Y 的类型

广义线性模型不仅适用于一般线性回归，还可以根据因变量（Y）的分布，通过联系函数 f，对 Y 进行某种函数转换。如果 Y 是连续变量，如肺功能值、收缩压值等，Y 服从正态分布，则 f(Y) = Y，这就是一般的直线回归方程，方程表达式为：

$Y= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+ \cdots\cdots + e$

如果 Y 是两分类变量，如是否患高血压、是否患慢性阻塞性肺病等，Y 呈二项分布，则 f(Y)=Logit（Y），这就是一般的 Logistic 回归方程，方程表达式为：

$\text{logit}（Y）= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+ \cdots\cdots + e$

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）
γ 分布（gamma）	inverse	$1/（Y^{-1}）$
负二项分布（negative binomial）	Log	Log（Y）

Y 的分布不同，只影响方程左边的关联函数 f(Y)，不影响方程右边的线性表达式，因此统称为广义线性模型（Generalized Linear Model）。

该模块调用 GLM，对一个或多个应变量进行回归分析。该模块自动检测应变量的类型，如果是连续性变量，则系统将自动默认采用正态分布和 identity 作为联系函数。如果两分类型，自动选择 logit 联系函数。用户可以右击联系函数，重新定义联系函数。自变量类型不限，可以引进交互作用项。同时可以进行自变量筛选，筛选方法可以有前进法、后退法、前进加后退法。如果 Y 的联系函数为 identity,还可以对自变量的各种组合模型进行比较，给出每个自变量的相当贡献大小。

本模块还可以用抽样调查(Survey)数据分析，详见易侕抽样调查数据分析 (http://www.empowerstats.com/cn/manuals/articles/surveyAnalysis.pdf)

例：DEMO 数据分析 SBP 和年龄、性别、身高、体重、吸烟、职业和教育程度的关系，筛选可能的自变量组合，输入界面如下：

输出结果：

结局变量：Systolic BP, mmhg

变量分布与联系函数：gaussian

模型：SBP ~ AGE+SEX+HEIGHT+WEIGHT+SMOKE+OCCU.NEW+factor(EDU.NEW)

#Variables	(Intercept)	AGE	SEX	HEIGHT	WEIGHT	SMOKE	OCCU.NEW	factor(EDU.NEW)2	factor(EDU.NEW)3	R-squared	Adj-Rsquared
1	101.5751	0.7581								0.2223	0.2213
1	128.4603					4.5721				0.0097	0.0085
1	131.5958								-4.9526	0.0095	0.0082
2	108.0533	0.7755					-4.8093			0.2335	0.2315
2	105.9875	0.7542	-2.8398							0.2263	0.2243
2	91.7852	0.7723			0.1732					0.2256	0.2236
3	96.9826	0.7930			0.2020		-5.0681			0.2379	0.2350

3	112.3204	0.7715	-2.7779				-4.7729			0.2373	0.2343
3	106.5878	0.7914					-4.5448		1.7371	0.2345	0.2316
4	115.7997	0.7983	-5.1140			-3.7757	-4.4368			0.2408	0.2369
4	102.6468	0.7859	-1.8508		0.1505		-4.9779			0.2393	0.2354
4	110.2547	0.7928		-10.7901	0.2738		-5.0609			0.2387	0.2348
5	154.0545	0.7741	-4.7300	-34.6305	0.3010		-4.8146			0.2439	0.2391
5	105.4996	0.8151	-4.2236		0.1629	-3.9581	-4.6423			0.2432	0.2383
5	135.3312	0.7887	-6.2374	-11.1609		-3.6088	-4.3329			0.2415	0.2366
6	153.1646	0.8018	-6.7105	-32.2679	0.3021	-3.6313	-4.5178			0.2472	0.2414
6	153.8773	0.7804	-4.5488	-35.0669	0.3014		-4.7100		0.6790	0.2441	0.2382
6	154.383	0.7728	-4.7756	-34.697	0.3006		-4.8109	-0.2756		0.2440	0.2381
7	153.0462	0.8060	-6.5655	-32.5972	0.3024	-3.6013	-4.4461		0.4820	0.2473	0.2405
7	153.3664	0.8009	-6.7338	-32.3137	0.3018	-3.6231	-4.5162	-0.1676		0.2472	0.2404
7	153.7419	0.7815	-4.5137	-35.0886	0.3016		-4.7005	0.0982	0.7496	0.2441	0.2373

调整的 R 平方值为最大的模型：

	Estimate	Std. Error	t value	95%CI low	95%CI upp	P.value
(Intercept)	153.1646	24.6229	6.2204	104.9038	201.4255	0.0000
AGE	0.8018	0.0536	14.9456	0.6966	0.9069	0.0000
SEX	-6.7105	2.2961	-2.9226	-11.2109	-2.2102	0.0036
HEIGHT	-32.2679	15.8384	-2.0373	-63.3111	-1.2247	0.0420
WEIGHT	0.3021	0.1245	2.4272	0.0582	0.5460	0.0154
SMOKE	-3.6313	1.9760	-1.8377	-7.5042	0.2416	0.0665
OCCU.NEW	-4.5178	1.4320	-3.1550	-7.3245	-1.7112	0.0017

AIC:	6935.1788
Log Likelihood:	-3459.5894 , df= 8
Null.deviance	402773.6645 on 786 degrees of freedom
deviance	303213.3999 on 780 degrees of freedom
residuals SD	19.641 (pearson chi-square normality test P= <0.0001)
R-squared	0.2472

Adj R-squared	0.2414
Number of observations used:	787