自变量共线性和 VIF 逐步筛查

在构建多元回归方程时，如果一组自变量存在共线性，即变量之间的相关性过大，可以使得模型估计失真。

一个简单的方法来确定解释变量之间的共线性是利用方差膨胀因子（VIF）。计算某变量的 VIF：首先构建一个线性回归方程，用所有其它变量解释该变量。如一组自变量：X1、X2、X3…，构建 X1 的线性回归方程：

X1 = X2 + X3 + ……

取该方程的 R-平方值，然后计算 VIF = 1/（1-R 平方）

同理，计算 X2 的 VIF，首先构建方程：

X2 = X1 + X3 + ……

VIF 计算简单，容易理解; VIF 值越高，共线性越高。通常 VIF < 5（或 10）是可以接受的。

本模块，对一组自变量的共线性，根据 VIF 进行诊断筛查。筛查方法是：

首先计算每个变量的 VIF， 如果最大的 VIF 值>=5 （默认的筛查标准），剔除最大 VIF 的变量，重复上步，直到剩下的所有变量的 VIF 小于筛查标准。

例1. 自变量共线性导致模型估计失真。使用软件自带的 demo 数据，对 SBP 进行回归分析， 结果如下：

模型 1: SBP ~ AGE+HEIGHT+WEIGHT+BMI+OCCU.NEW+factor(EDU.NEW)+SEX
广义线性模型
结局变量: Systolic BP, mmhg
变量分布与联系函数: gaussian

	Estimate	Se	t value	95%CI.low	95%CI.upp	P.value
(Intercept)	67.9114	126.8002	0.5356	-180.617	316.4397	0.5924
AGE	0.7837	0.0588	13.3312	0.6685	0.8990	0.0000
HEIGHT	19.2738	80.1343	0.2405	-137.7895	176.3371	0.8100
WEIGHT	-0.5046	1.1682	-0.4319	-2.7943	1.7852	0.6659
BMI	2.0144	2.9115	0.6919	-3.6922	7.7209	0.4892
OCCU.NEW	-4.7628	1.4687	-3.2429	-7.6415	-1.8842	0.0012
factor(EDU.NEW)2	0.0945	1.9300	0.0489	-3.6884	3.8773	0.9610
factor(EDU.NEW)3	0.7773	2.3174	0.3354	-3.7648	5.3194	0.7374
SEX	-4.5227	2.2005	-2.0552	-8.8357	-0.2096	0.0402

AIC:	6949.8292
Log Likelihood:	-3464.9146 , df= 10
Null.deviance	402859.2183 on 787 degrees of freedom
deviance	304315.5805 on 779 degrees of freedom
residuals SD	19.6641 (pearson chi-square normality test P= <0.0001)
R-squared	0.2446
Adj R-squared	0.2369
Number of observations used:	788

模型 2：SBP ~ AGE+HEIGHT+BMI+OCCU.NEW+factor(EDU.NEW)+SEX
广义线性模型
结局变量: Systolic BP, mmhg
变量分布与联系函数: gaussian

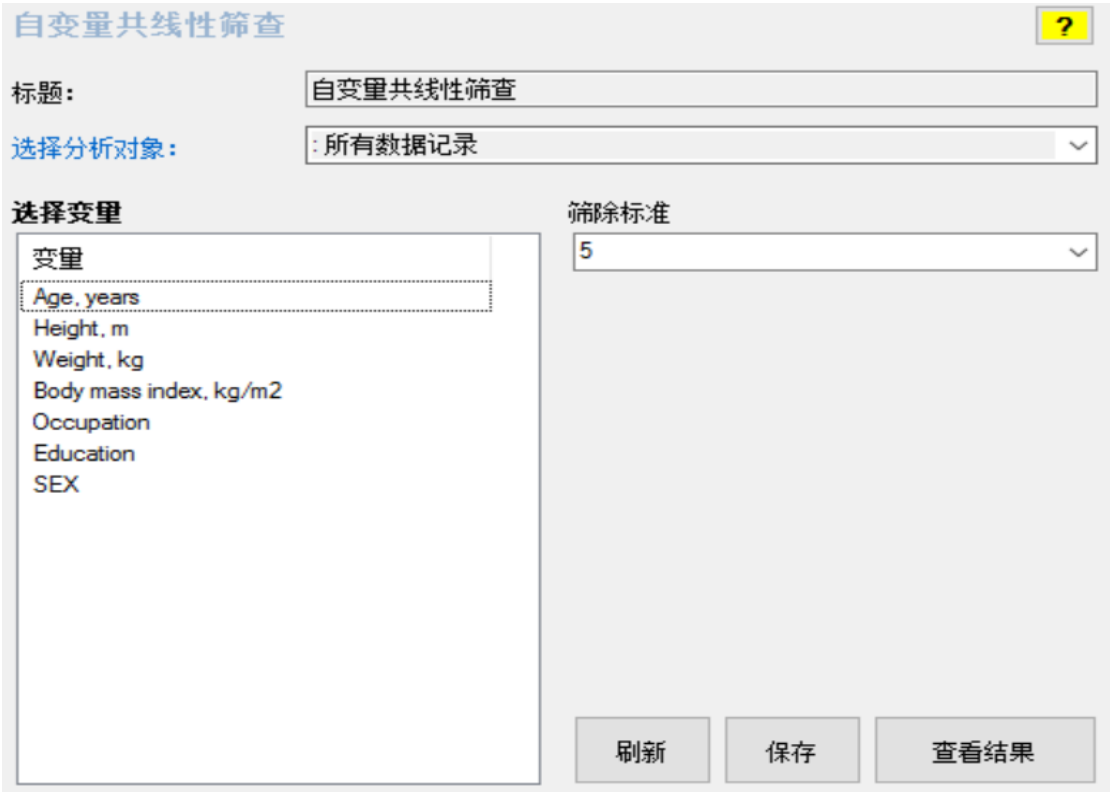
	Estimate	Se	t value	95%CI.low	95%CI.upp	P.value
(Intercept)	121.6505	24.4222	4.9811	73.783	169.518	0.0000
AGE	0.7821	0.0586	13.3389	0.6671	0.8970	0.0000
HEIGHT	-14.8607	13.2367	-1.1227	-40.8045	11.0832	0.2619
BMI	0.7641	0.3101	2.4639	0.1563	1.3719	0.0140
OCCU.NEW	-4.7318	1.4662	-3.2273	-7.6055	-1.8581	0.0013
factor(EDU.NEW)2	0.0904	1.9290	0.0469	-3.6904	3.8713	0.9626
factor(EDU.NEW)3	0.7659	2.3160	0.3307	-3.7735	5.3053	0.7410
SEX	-4.5424	2.1989	-2.0658	-8.8523	-0.2326	0.0392

AIC:	6948.0179
Log Likelihood:	-3465.0089 , df= 9
Null.deviance	402859.2183 on 787 degrees of freedom
deviance	304388.4528 on 780 degrees of freedom
residuals SD	19.6665 (pearson chi-square normality test P= <0.0001)
R-squared	0.2444
Adj R-squared	0.2376
Number of observations used:	788

比较模型 1 与模型 2:（1）模型 2 除去了 WEIGHT 变量，与模型 1 相比 AIC，R-squared， Adj-R-squared 都基本相同；（2）模型 2 的 HEIGHT、BMI 的回归系数、标准误与模型 1 相差很大。模型 1 中 BMI 不显著，模型 2 中 BMI 变得显著。

导致模型 1 的 BMI、HEIGHT 参数估计失真的原因是 HEIGHT、WEIGHT、BMI 三变量存在共线性。模型 2 中除去 WEIGHT 后，控制了自变量共线性，结果比较可靠。

例2.　对上例模型 1 中的自变量进行共线性筛查，输入界面如下：



输出结果如下：

自变量共线性逐步筛查:

	Step 1	Step 2
AGE	1.4	1.4
HEIGHT	76.9	2.1
WEIGHT	157.2	NA
BMI	90.8	1
OCCU.NEW	1.1	1.1
EDU.NEW	1.8	1.8
SEX	2.4	2.4

剔除的变量: WEIGHT

选出的变量: AGE HEIGHT BMI OCCU.NEW EDU.NEW SEX

结果解释：
第一步，计算每个变量的 VIF，发现 WEIGHT 的 VIF 最大，为 157.2；
第二步，剔除 WEIGHT 后，重新计算每个变量的 VIF，发现余下的变量里，每个变量的 VIF 均小于筛选标准 5。