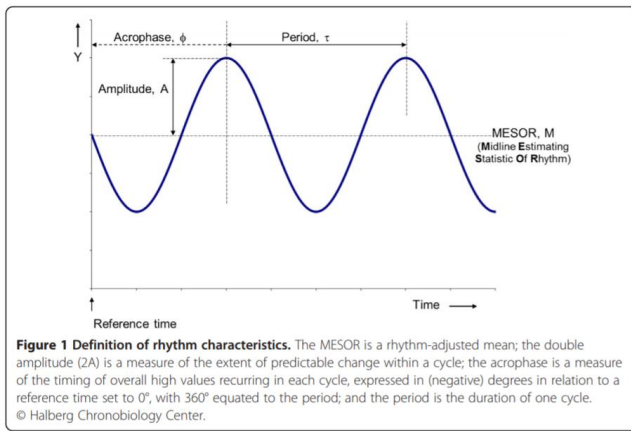


## 正弦余弦曲线回归

时间序列数据因变量和自变量都有可能随时间呈现（季节性）周期变化，本模块可用来  
 (1) 分析因变量是否有周期变化及其周期律与幅度； (2) 调整季节性周期变化分析自变量与因变量之间的联系。

正余弦曲线拟合：

即在模型中引进 sine（正弦）与 cosine（余弦）项，首先将时间变量转换成分数（f），如时间类型是每天的数据，f 即时间所对应的该年的第几天（0-365）/该年总天数。然后计算正弦值： $\sin(f) = \sin(f*2*\pi*cycles)$ ，余弦： $\cos(f) = \cos(f*2*\pi*cycles)$ ，其中 cycles 是年周期数。正余弦曲线周期图如下（摘自：Germaine Cornelissen, *Cosinor-based rhythmometry. Theoretical Biology and Medical Modelling* 2014, 11:16）：



数据类型与相应的时间变量格式：数据类型分 Daily, Weekly, Monthly, Hourly。

	Daily	Weekly	Monthly	Hourly
YYYY-MM-DD	OK	OK	OK	
YYYY-MM-DD HH:NN:SS	OK	OK	OK	OK
1-12 Month			OK	
1-53 Week		OK		
1-365 Days	OK			
0-23 Hour				OK

本模块有如下特点：

1. 可无自变量，此时只分析因变量的季节性周期变化规律，自动绘制正弦曲线周期变化图。
2. 模型中可以引进 Spline 函数对指定的自变量进行曲线拟合。平滑拟合自由度取决于平滑程度，越平滑自由度越小。本模块采用默认值，即用最低 GCV 值（广义交叉验证）的方法找到的最适当的平滑拟合自由度。
3. 如有多个自变量，自动进行单因素（一个自变量一个模型）与多因素（所有自变量同时进入模型）分析。每个模型中均调整分数的  $\cosin(f)$  与  $\sin(f)$  函数项。

4. 可以指定添加滞后效应分析。滞后效应指的是今天的暴露水平对明天的结局指标的影响。要求 (i) 时间变量时间格式为日期型变量 (YYYY-MM-DD 格式, 或 YYYY-MM-DD HH:NN:SS 格式), (ii) 1 个时间单位 1 条记录, 不能有缺失。数据类型为 Daily 最长滞后天数为 5 天, 其它数据类型自动分析滞后 1 个时间单位的效应。本模块自动构建多因素滞后效应模型。如果自变量有 X1、X2 两个, 多因素模型 1 为  $X1+X2+\cos(f)+\sin(f)$ , 模型 2 在模型 1 的基础上添加 X1.lag 与 X2.lag1, 模型 3 在模型 2 的基础上添加 X1.lag2 与 X2.lag2, 依此类推。本模块并自动检验滞后变量与原变量的相关性, 自动剔除相关系数 > 0.90 的滞后变量。例, 如 X2.lag1 与 X2 的相关系数 > 0.9, 将不分析 X2 的滞后效应。滞后变量的相关性强表示该变量短期内基本上不随时间变化。

5. 因变量的分布类型可以是正态分布、二项分布、Poisson 分布。自动检测应变量的类型, 如果是连续性变量, 默认采用正态分布和 identity 作为联系函数。如是两分类的, 默认用 logit 做联系函数。

例 1: 下载练习数据: <http://www.empowerstats.com/empowerStats/exdata/cvd.xls>

该数据为 Los Angeles 1987-2000 每天心血管死亡人数数据, 分析 O3MEAN、TMPD、Day of Week 与 CVD 死亡数之间的关系。对 TMPD 进行曲线拟合。输入界面如下:

正弦余弦曲线回归

标题: 正弦余弦曲线回归分析

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
CVD	Gaussian	Identity

时间变量名: date

时间变量格式: 1: YYYY-MM-DD

自变量

变量	曲线拟合
O3MEAN	.
TMPD	S
Day of week	.

数据类型: 1: Daily

年(或天)周期数: 1

分层变量:

添加滞后(1-5天前的暴露水平)效应

精确到小数点: 0.01

刷新 保存 查看结果

## 输出结果

Generalized additive model with sinuoidal seasonal pattern

Date variable: DATE  
Regression results:

	CVD
Sinuosid:	
Cosine ()	6.62 (6.30, 6.93) <0.0001
Sine ()	2.68 (2.37, 3.00) <0.0001

解释: 上表是基本模型的回归系数 (95%可信区间) P 值。基本模型即模型中只有正弦与余弦函数项。

Summary of sinusoidal seasonal pattern:

	CVD
n	5114
amp	7.14102355526869
peak.phase	Month = January , day = 23
low.phase	Month = July , day = 25
significant	TRUE

解释: 上表是由基本模型得出的 CVD 的季节性周期变化规律 (图形见后), amp 表示振幅, 即最高值与最低值差的一半。peak phase 即最高点对应的日期, low phase 即最低点对应的日期。significant = TRUE 表示季节周期变化显著存在。

Adjusted R square (Sample size used) for basic model:

	CVD
cos () + sin ()	0.275 (5114)

解释: 上表是由基本模型的调整的 R 平方值 (样本量)。R-square = 0.275, N=5114。

Regression results:

	Univariate	Multivariate 1	Multivariate 2	Multivariate 3	Multivariate 4	Multivariate 5	Multivariate 6
O3MEAN	-0.06 (-0.10, -0.03) 0.0001	-0.06 (-0.09, -0.02) 0.0011	-0.04 (-0.08, 0.01) 0.1208	-0.04 (-0.08, 0.01) 0.1083	-0.04 (-0.08, 0.01) 0.1282	-0.03 (-0.08, 0.01) 0.1413	-0.03 (-0.08, 0.01) 0.1424
O3MEAN lag 1	-0.04 (-0.07, -0.01) 0.0200		-0.03 (-0.08, 0.01) 0.1730	-0.03 (-0.08, 0.03) 0.3703	-0.03 (-0.09, 0.02) 0.2619	-0.03 (-0.09, 0.03) 0.2839	-0.03 (-0.09, 0.03) 0.3045
O3MEAN lag 2	-0.01 (-0.05, 0.02) 0.4130			-0.01 (-0.05, 0.04) 0.7605	0.03 (-0.03, 0.09) 0.3165	0.02 (-0.04, 0.08) 0.4594	0.02 (-0.03, 0.08) 0.4131
O3MEAN lag 3	-0.05 (-0.08, -0.01) 0.0072				-0.05 (-0.09, -0.00) 0.0374	-0.01 (-0.06, 0.05) 0.8467	-0.02 (-0.07, 0.04) 0.5819
O3MEAN lag 4	-0.07 (-0.10, -0.03) <0.0001					-0.06 (-0.10, -0.01) 0.0127	0.00 (-0.05, 0.06) 0.9321
O3MEAN lag 5	-0.08 (-0.11, -0.05) <0.0001						-0.08 (-0.13, -0.04) 0.0004
TMPD	s(edf: 5.1) <0.0001	s(edf: 5.4) <0.0001	s(edf: 5.4) <0.0001	s(edf: 5.4) <0.0001	s(edf: 5.4) <0.0001	s(edf: 5.5) <0.0001	s(edf: 5.5) <0.0001
Day of week							
Sunday	Ref.						
Monday	1.74 (0.90, 2.58) <0.0001	1.44 (0.59, 2.29) 0.0009	1.61 (0.73, 2.49) 0.0003	1.62 (0.73, 2.50) 0.0003	1.51 (0.62, 2.40) 0.0008	1.56 (0.67, 2.45) 0.0006	1.57 (0.68, 2.45) 0.0006
Tuesday	1.39 (0.55, 2.23) 0.0012	1.05 (0.20, 1.90) 0.0160	1.07 (0.22, 1.92) 0.0138	1.12 (0.21, 2.03) 0.0160	1.07 (0.16, 1.98) 0.0208	1.00 (0.08, 1.91) 0.0326	1.06 (0.15, 1.97) 0.0229
Wednesday	0.44 (-0.40, 1.28) 0.3025	0.13 (-0.72, 0.98) 0.7671	0.12 (-0.73, 0.97) 0.7757	0.14 (-0.72, 1.00) 0.7494	0.37 (-0.51, 1.25) 0.4135	0.36 (-0.52, 1.25) 0.4208	0.25 (-0.64, 1.13) 0.5821
Thursday	0.61 (-0.23, 1.45) 0.1535	0.31 (-0.54, 1.15) 0.4789	0.31 (-0.54, 1.15) 0.4805	0.31 (-0.54, 1.16) 0.4691	0.32 (-0.53, 1.17) 0.4550	0.64 (-0.24, 1.53) 0.1539	0.63 (-0.25, 1.52) 0.1604
Friday	1.04 (0.21, 1.88) 0.0146	0.79 (-0.05, 1.64) 0.0665	0.80 (-0.04, 1.65) 0.0632	0.79 (-0.07, 1.64) 0.0709	0.75 (-0.10, 1.60) 0.0838	0.81 (-0.04, 1.66) 0.0628	1.26 (0.37, 2.15) 0.0054
Saturday	0.74 (-0.10, 1.58) 0.0834	0.64 (-0.19, 1.47) 0.1334	0.57 (-0.26, 1.41) 0.1796	0.59 (-0.25, 1.43) 0.1701	0.54 (-0.30, 1.39) 0.2062	0.55 (-0.30, 1.39) 0.2045	0.62 (-0.22, 1.47) 0.1467

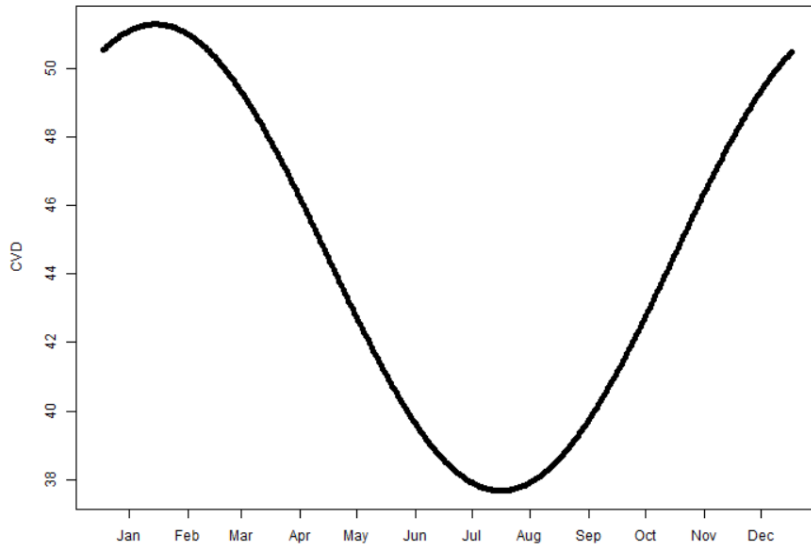
解释: 上表是单因素与多因素模型回归系数 (95%可信区间) P 值。单因素模型是在基本模型 (cosine()+sine()) 中添加 1 个自变量。单因素模型 O3MEAN 的效应 (回归系数) 是 -0.06, 滞后 1 天的效应是 -0.04、滞后 2 天是 -0.01, ..., 值得注意的是滞后 5 天的效应比较大为 -0.08。TMPD 因为是曲线拟合, 没有回归系数, 列出曲线拟合的自由度 (为 5.1) 与 P 值。Day of week 是分类变量, Sunday 为参照, 其它天与其比较。多因素模型 1 是在基本模型中同时纳入这三个变量。多因素模型 2 是在模型 1 的基础上再添加 O3MEAN.lag1 变量 (滞后 1 天的效应)。Day of Week 滞后变量完全相关, 不分析滞后效应。TMPD 滞后变量相关系数 > 0.9 (未列出) 也自动剔除。

Adjusted R square (Sample size used):

	Univariate	Multivariate 1	Multivariate 2	Multivariate 3	Multivariate 4	Multivariate 5	Multivariate 6
O3MEAN	0.277 (5114)	+	+	+	+	+	+
O3MEAN.lag1	0.276 (5113)		+	+	+	+	+
O3MEAN.lag2	0.275 (5112)			+	+	+	+
O3MEAN.lag3	0.276 (5111)				+	+	+
O3MEAN.lag4	0.277 (5110)					+	+
O3MEAN.lag5	0.278 (5109)						+
s (TMPD)	0.291 (5114)	+	+	+	+	+	+
factor (DAY.W)	0.278 (5114)	+	+	+	+	+	+
		0.294 (5114)	0.294 (5113)	0.294 (5112)	0.294 (5111)	0.295 (5110)	0.296 (5109)

解释：上表是每个模型的调整后的R平方值与样本量。单因素（univariate）模型直接列出每个模型的值，多因素模型首先列出模型中包含哪些变量，“+”表示含有该变量，最后1行为该模型的值。

季节变化周期图（该图是根据多因素分析模型中 consine() 与 sine() 两项回归系数计算出来的预测值）



例 2：下载练习数据：<http://www.empowerstats.com/empowerStats/exdata/cvdmonthly.xls>

该数据为 Los Angeles 1987-2000 每月 75 岁以上老人心血管死亡人数，分析 TMPD 与 CVD 死亡数之间的关系。

输入界面如下：

正弦余弦曲线回归

标题: 正弦余弦曲线回归分析

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
CVD	poisson	log

时间变量名: month

时间变量格式: 3: 1-12 Month

自变量

变量	曲线拟合
TMPD	

数据类型: 3: Monthly

年(或天)周期数: 1

分层变量:

添加滞后(1-5天前的暴露水平)效应

精确到小数点: 0.01

刷新 保存 查看结果

本例 CVD 死亡人数分布采用了 Poisson 分布。

## 输出结果:

Date variable: MONTH

Regression results:

	CVD
Sinuid:	
Cosine()	0.99 (0.98, 1.00) 0.0005
Sine()	0.99 (0.99, 1.00) 0.0068

解释: 上表是基本模型的回归系数 (95%可信区间) P 值。基本模型即模型中只有正弦与余弦函数项。

Summary of sinusoidal seasonal pattern:

	CVD
n	168
amp	17.9216591126738
amp.scale	(absolute scale)
peak.phase	Month = 8.3
low.phase	Month = 2.3
significant	TRUE

解释: 上表是由基本模型得出的 CVD 的季节性周期变化规律 (图形见后), amp 表示振幅, 即最高值与最低值差的一半。peak phase 即最高点对应的日期, low phase 即最低点对应的日期。significant = TRUE 表示季节周期变化显著存在。

Adjusted R square (Sample size used) for basic model:

	CVD
cos() + sin()	-0.008 (168)

解释: 上表是由基本模型的调整的 R 平方值 (样本量)。R-square = -0.008, N=168.

Regression results:

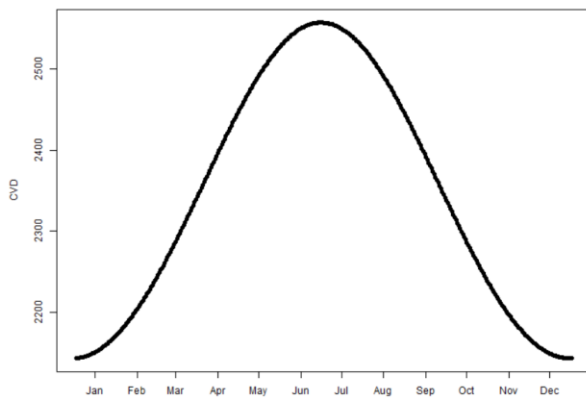
	Univariate	Multivariate
TMPD	0.98 (0.98, 0.98) <0.0001	0.98 (0.98, 0.98) <0.0001

解释: 上表是单因素模型与多因素模型 RR (95%CI) P 值, 因为只有 1 个自变量且不分析滞后效应变量, 单因素模型与多因素模型完全一样。Poisson 分布联系函数为 log, 上表中数据是 exp(beta) 即风险比及其 95% 可信区间。TMPD 每增加一个单位, CVD 死亡的风险比为 0.98, 即下降 2%

Adjusted R square (Sample size used):

	Univariate	Multivariate
TMPD	0.523 (168)	+
		0.523 (168)

解释: 同上, 单因素模型与多因素模型完全一样, 调整后的 R 平方值为 0.523, 远远高出基本模型, 表示 TMPD 有强效应。



例 3：下载练习数据：<http://www.empowerstats.com/empowerStats/exdata/indoor.xls>

该数据为卧室与客厅室内温度的 24 小时变化，分析 BEDROOM 温度的每天变化周期。  
输入界面如下：

正弦余弦曲线回归 ?

标题: 正弦余弦曲线回归分析

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
BEDROOM	Gaussian	Identity

时间变量名: datetime

时间变量格式: 2: YYYY-MM-DD HH:NN:SS

自变量

变量	曲线拟合

数据类型: 4: Hourly for daily cycle

年(或天)周期数: 1

分层变量:

添加滞后(1-5天前的暴露水平)效应

精确到小数点: 0.01

刷新 保存 查看结果

输出结果如下：

Generalized linear model with sinusoidal daily pattern

Date variable: DATETIME

Regression results:

	BEDROOM
Sinuid:	
Cosine ()	1.83 (1.63, 2.03) <0.0001
Sine ()	2.61 (2.41, 2.81) <0.0001

Summary of sinusoidal daily pattern:

	BEDROOM
n	2021
amp	3.18566344287919
peak.phase	Hour = 4.7
low.phase	Hour = 16.7
significant	TRUE

Adjusted R square (Sample size used) for basic model:

	BEDROOM
cos () + sin ()	0.326 (2021)

