病例交叉匹配分析 （case-crossover)

时间序列数据通常指的是每天收集的数据，如气象数据（温度、湿度、PM2.5 等）、人群医疗健康数据（如门诊人数、因某病就诊人数、死亡人数等）。当分析这些数据之间的联系时（如分析 PM2.5 与心血管死亡的关系）时，常用到的一种方法是病例交叉设计（case-crossover design）。本模块是对时间序列数据，按时间分层的病例交叉设计进行单因素与多因素的回归分析。

病例交叉设计是将有病例的这一天（称为病例日：case day）与附近的其它天（称为对照日：control day）进行匹配比较，可附加匹配月份、周天或/和其它主要混杂因素。每天既可以是 case day 又可以是其它 case day 的 control day。这就是 crossover 的意义。匹配设计方法主要有两种：按时间分层的匹配(TS: time stratified)与双向对称匹配（SBI: symmetric bi-directional）。

假设所收集的时间序列数据（每天一条记录）是从 day1 到 dayN，共 N 天。

双向对称的病例交叉设计方法：

对每个有病例的天，取其前后 N 天作为对照，N 表示左右半径长度。设置排除天数，常用 2 天，表示该病例天左右最近的 2 天不能做为对照。在排除天之外，半径之内，所有满足其它匹配要求（如有）的天，都做为该病例天的对照天(control days)。case day 的 case = 1，control day 的 case =0，病例天的病例数为权重。数据匹配好后，采用条件 logistic 回归分析。

按时间分层的病例交叉设计方法：

首先设置时间窗(分层)长度，常用 28 天,表示每 28 天为一个层，也可按月份分层，整个时间就分成了若干个层。设置排除天数，常用 2 天，表示病例天周围最近 2 天不能做为对照。

在每个时间窗内进行交叉匹配，为每个有病例的天（case day），在同一个时间窗内在排除天之外的所有其它天，满足其它匹配要求（如有），都做为该病例天的对照天(control days)。case day 的 case = 1， control day 的 case =0，病例天的病例数为权重。数据匹配好后，采用条件 logistic 回归分析。

本模块自动根据参数设置进行病例交叉匹配，生成匹配后的数据。考虑到暴露因素对结局的作用可能有滞后效应，本模块同时生成滞后变量，命名为原变量名.lagD,其中 D=从 1 开始逐步递增至排除天数为止。如 TMPD.lag1、TMPD.lag2, 分别表示 TMPD 的过去 1 天、过去 2 天的值。输出数据中变量 time 表示病例天，wt 为病例天的结局变量。后续的条件 logistic 回归将以 time 为分层变量，wt 为权重。

下表显示按星期天匹配后的病例交叉设计数据结构：

CVD	O3MEAN	TMPD	date	dow	O3MEAN.lag1	TMPD.lag1	O3MEAN.lag2	TMPD.lag2	time	case	timex	wt	diffdays
55	-16.0073	54.5	1987/1/1	4	NA	NA	NA	NA	1	1	1	55	0
69	-18.5332	52.5	1987/1/8	4	-19.0819	53.5	-22.7846	49.75	1	0	2	55	7
52	-14.2682	51.5	1987/1/15	4	-17.6349	56.25	-21.1626	55.75	1	0	2	55	14
73	-11.6595	58.5	1987/1/2	5	-16.0073	54.5	NA	NA	2	1	1	73	0
52	-17.3117	54	1987/1/9	5	-18.5332	52.5	-19.0819	53.5	2	0	2	73	7
64	-13.8024	47	1987/1/16	5	-14.2682	51.5	-17.6349	56.25	2	0	2	73	14
64	-10.3241	55.25	1987/1/3	6	-11.6595	58.5	-16.0073	54.5	3	1	1	64	0
52	-14.175	56	1987/1/10	6	-17.3117	54	-18.5332	52.5	3	0	2	64	7
47	-16.9701	44.75	1987/1/17	6	-13.8024	47	-14.2682	51.5	3	0	2	64	14
57	-18.6471	54.75	1987/1/4	0	-10.3241	55.25	-11.6595	58.5	4	1	1	57	0
53	-14.5477	61.5	1987/1/11	0	-14.175	56	-17.3117	54	4	0	2	57	7
68	-16.2247	49.25	1987/1/18	0	-16.9701	44.75	-13.8024	47	4	0	2	57	14
56	-17.5291	54.5	1987/1/5	1	-18.6471	54.75	-10.3241	55.25	5	1	1	56	0
63	-19.0508	61.5	1987/1/12	1	-14.5477	61.5	-14.175	56	5	0	2	56	7
61	-13.4833	51	1987/1/19	1	-16.2247	49.25	-16.9701	44.75	5	0	2	56	14
65	-22.7846	49.75	1987/1/6	2	-17.5291	54.5	-18.6471	54.75	6	1	1	65	0
60	-21.1626	55.75	1987/1/13	2	-19.0508	61.5	-14.5477	61.5	6	0	2	65	7
43	-14.238	54	1987/1/20	2	-13.4833	51	-16.2247	49.25	6	0	2	65	14
43	-19.0819	53.5	1987/1/7	3	-22.7846	49.75	-17.5291	54.5	7	1	1	43	0
67	-17.6349	56.25	1987/1/14	3	-21.1626	55.75	-19.0508	61.5	7	0	2	43	7
53	-19.0301	50.75	1987/1/21	3	-14.238	54	-13.4833	51	7	0	2	43	14
69	-18.5332	52.5	1987/1/8	4	-19.0819	53.5	-22.7846	49.75	8	1	1	69	0

本模块自动匹配数据，并自动进行单因素与多因素分析，自动进行滞后效应分析。自变量可以是连续变量与分类变量。如自变量为连续性变量，可以设置曲线拟合，拟合方法为限制立方样条(restricted cubic spline)。在多因素分析时，本模块自动检验原变量与滞后变量之间的相关性，自动剔除相关性强的滞后变量（如 Day of Week 的滞后变量与原变量完全相关）。

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

该数据为 Los Angles 1987-2000 每天心血管死亡人数数据，分析 O3MEAN、TMPD 与 CVD 死亡之间的关系，匹配星期天(day of week)，输入界面如下：

对 TMPD 进行了曲线拟合（restricted cubic spline），结果输出如下：

Time-stratified Case-crossover design: conditional regression model

Outcome variable: CVD
Date variable: DATE
Date range: 1987-01-01 to 2000-12-31
Missing days: 0 (0 %)

Time-stratified case-crossover design:
Stratified time window length(days): 28
Exclusion days: 2
Match on day of week: TRUE
Match on confounder: NA

Number of records (days):	5114
Number of days with cases:	5114
Number of total cases:	230695
Matched results:	
Number of windows * Days of the window	# Windows * Days 1 18 182 28
Number of case days * Matched controls days:	# Case days * Control days 6 1 12 2 5096 3
Matched number of case days	5114
Matched number of total cases:	230695
Matched data:	casecrossover_18_tbl_wdata.xls

Regression results:

	Univariate	Multivariate 1	Multivariate 2	Multivariate 3
O3MEAN	0.9994 (0.9986, 1.0001) 0.1112	0.9994 (0.9986, 1.0001) 0.1063	0.9996 (0.9986, 1.0006) 0.4543	0.9996 (0.9986, 1.0006) 0.4845
TMPD				
rcs( )	0.9972 (0.9939, 1.0006) 0.1078	0.9971 (0.9938, 1.0005) 0.0944	0.9971 (0.9937, 1.0004) 0.0886	0.9969 (0.9935, 1.0003) 0.0708
rcs(')	1.0143 (0.9952, 1.0337) 0.1439	1.0152 (0.9961, 1.0347) 0.1195	1.0155 (0.9963, 1.0351) 0.1133	1.0162 (0.9970, 1.0358) 0.0982
rcs('')	0.9312 (0.8582, 1.0103) 0.0867	0.9300 (0.8571, 1.0091) 0.0814	0.9292 (0.8562, 1.0083) 0.0781	0.9270 (0.8541, 1.0060) 0.0694
rcs(''')	1.1906 (1.0391, 1.3642) 0.0120	1.1869 (1.0358, 1.3600) 0.0136	1.1879 (1.0366, 1.3613) 0.0132	1.1910 (1.0393, 1.3650) 0.0119
O3MEAN lag 1	0.9994 (0.9987, 1.0002) 0.1408		0.9996 (0.9986, 1.0006) 0.4449	0.9994 (0.9982, 1.0006) 0.3218
O3MEAN lag 2	0.9999 (0.9991, 1.0006) 0.7286			1.0003 (0.9993, 1.0013) 0.5243

Time-stratified case-crossover data was saved as: casecrossover_1_tbl1_wdata.xls

例 2：同上例数据，调整星期天（不匹配），分析 O3MEAN、TMPD 与 CVD 死亡的关系。

（1）首先产生 DAY.W (day of week)变量：右点击变量名“DATE”,选“函数转换”，点击新变量名，改为“DAY”，选择日期变量格式如“1999-1-21”，选“取 Y，M，D，W”,Y 表示年份，M 表示月份，D 表示日，W 表示星期几。点击“保存”，即产生了新变量 DAY.W。

（2）采用本模块，输入界面如下：

分析结果如下：

……

```
Outcome variable: CVD
Date variable: DATE
Date range: 1987-01-01 to 2000-12-31
Missing days: 0 ( 0 %)
Symmetric bi-directional(SBI) case-crossover design:
Maximum day difference of control from case: 14
Exclusion days: 2
Match on day of week: FALSE
Match on confounder: NA
```

Number of records (days):	5114
Number of days with cases:	5114
Number of total cases:	230695
Matched results:	
Number of case days * Matched controls days:	# Case days | * Control days 6 | 12 2 | 13 2 | 14 2 | 15 2 | 16 2 | 17 2 | 18 2 | 19 2 | 20 2 | 21 2 | 22 2 | 23 5086 | 24
Matched number of case days	5114
Matched number of total cases:	230695
Matched data:	casecrossover_17_tbl_wdata.xls

Regression results:

	Univariate	Multivariate 1	Multivariate 2	Multivariate 3
O3MEAN	0.9994 (0.9987, 1.0000) 0.0667	0.9995 (0.9988, 1.0002) 0.1553	0.9994 (0.9986, 1.0003) 0.2177	0.9995 (0.9986, 1.0004) 0.2609
TMPD				
rcs( )	0.9994 (0.9963, 1.0025) 0.7060	0.9995 (0.9964, 1.0026) 0.7507	0.9996 (0.9965, 1.0027) 0.8050	0.9994 (0.9963, 1.0026) 0.7171
rcs(')	1.0066 (0.9892, 1.0244) 0.4595	1.0060 (0.9885, 1.0238) 0.5023	1.0056 (0.9881, 1.0234) 0.5306	1.0064 (0.9889, 1.0242) 0.4756
rcs('')	0.9724 (0.9021, 1.0481) 0.4640	0.9760 (0.9054, 1.0521) 0.5258	0.9771 (0.9064, 1.0533) 0.5456	0.9743 (0.9038, 1.0504) 0.4977
rcs(''')	1.0887 (0.9609, 1.2335) 0.1823	1.0807 (0.9537, 1.2245) 0.2239	1.0795 (0.9526, 1.2232) 0.2304	1.0834 (0.9561, 1.2277) 0.2092
Day of week				
Sunday	Ref.			
Monday	1.0354 (1.0197, 1.0514) <0.0001	1.0326 (1.0165, 1.0490) <0.0001	1.0323 (1.0157, 1.0493) 0.0001	1.0317 (1.0150, 1.0487) 0.0002
Tuesday	1.0281 (1.0124, 1.0440) 0.0004	1.0240 (1.0078, 1.0404) 0.0035	1.0239 (1.0078, 1.0403) 0.0035	1.0206 (1.0035, 1.0379) 0.0177
Wednesday	1.0090 (0.9936, 1.0247) 0.2517	1.0054 (0.9896, 1.0214) 0.5064	1.0054 (0.9896, 1.0214) 0.5062	1.0044 (0.9885, 1.0205) 0.5906
Thursday	1.0121 (0.9967, 1.0278) 0.1245	1.0089 (0.9930, 1.0249) 0.2742	1.0084 (0.9926, 1.0245) 0.2998	1.0079 (0.9920, 1.0240) 0.3323
Friday	1.0203 (1.0047, 1.0361) 0.0107	1.0179 (1.0019, 1.0342) 0.0278	1.0177 (1.0017, 1.0340) 0.0296	1.0161 (1.0001, 1.0324) 0.0485
Saturday	1.0150 (0.9994, 1.0307) 0.0588	1.0141 (0.9985, 1.0299) 0.0772	1.0140 (0.9983, 1.0299) 0.0801	1.0128 (0.9971, 1.0289) 0.1104
O3MEAN lag 1	1.0001 (0.9994, 1.0007) 0.8416		1.0000 (0.9992, 1.0009) 0.9147	0.9997 (0.9986, 1.0008) 0.5681
O3MEAN lag 2	1.0007 (1.0001, 1.0014) 0.0279			1.0005 (0.9996, 1.0014) 0.2489

解释：变量 TMPD 与 TMPD.lag1, TMPD.lag2 相关系数>0.9, 变量 DAY.W (day of week)与 DAY.W.lag1, DAY.W.lag2 完全相关，因此 TMPD 与 DAY.W 的滞后效应变量从模型中自动剔除。