# 非同步重复测量数据转换

## 一、 非结构化数据

数据结构的一个基本共同点是每一列代表一个观测指标，非结构化数据则不然，不同观测指标的观测值用同一个变量记录在同一列但不同的记录（行）中，如下表的 TestValue 变量，另有一个变量代表所观测的指标名称，如下表的 TestVAR 变量。临床电子病例中的实验室检测资料，多为非结构化数据，其基本结构如下例练习数据 nstructure.xls：

下载数据： http://www.empowerstats.com/empowerStats/exdata/nstructure.xls

nstructure

| # | TESTDATE | TESTVALUE | TESTVAR | ID | GENDER | BMI | TESTDAY |
|---|---|---|---|---|---|---|---|
| 3 | 2009/1/6 | 303 | PLT | 255486 | 1 | 21.63 | 0 |
| 4 | 2009/1/6 | 21.91 | WBC | 255486 | 1 | 21.63 | 0 |
| 5 | 2009/1/7 | 29.8 | CRP | 255486 | 1 | 21.63 | 1 |
| 6 | 2009/1/7 | 1.4 | RC | 255486 | 1 | 21.63 | 1 |
| 7 | 2009/1/7 | 2.1 | LC | 255486 | 1 | 21.63 | 1 |
| 8 | 2009/1/7 | 15 | A | 255486 | 1 | 21.63 | 1 |
| 9 | 2009/1/7 | 0.14 | LCA | 255486 | 1 | 21.63 | 1 |
| 10 | 2009/1/7 | 0.09 | RCA | 255486 | 1 | 21.63 | 1 |
| 11 | 2009/1/10 | 2.3 | NEUTn | 255486 | 1 | 21.63 | 4 |
| 12 | 2009/1/10 | 553 | PLT | 255486 | 1 | 21.63 | 4 |
| 13 | 2009/1/10 | 8.6 | WBC | 255486 | 1 | 21.63 | 4 |
| 14 | 2009/1/17 | 2.9 | LC | 255486 | 1 | 21.63 | 11 |
| 15 | 2009/1/17 | 15 | A | 255486 | 1 | 21.63 | 11 |
| 16 | 2009/1/17 | 1.2 | RC | 255486 | 1 | 21.63 | 11 |
| 17 | 2009/1/17 | 0.08 | RCA | 255486 | 1 | 21.63 | 11 |
| 18 | 2009/1/17 | 0.19 | LCA | 255486 | 1 | 21.63 | 11 |
| 19 | 2009/1/6 | 18.55 | NEUTn | 255509 | 1 | 20.76 | 0 |
| 20 | 2009/1/6 | 28.01 | WBC | 255509 | 1 | 20.76 | 0 |
| 21 | 2009/1/6 | 27 | ESR | 255509 | 1 | 20.76 | 0 |
| 22 | 2009/1/6 | 158 | PLT | 255509 | 1 | 20.76 | 0 |

上表中 ID=255486 的患者，在 2009 年 1 月 6 日检测了 PLT 这个指标，检测结果为 303；随后在 2009 年 1 月 10 日复查 PLT，检测值为 553。该患者其它检测指标有 WBC、CRP,…。TESTDATE 转换成随访天数为 TESTDAY，此时 TESTDATE 可以丢弃。数据中 GENDER 与 BMI 表示观察对象个体特征，不随时间变化，也可以像 PLT、WBC 等重复测量指标一样分别列成一行，TESTVAR 分别为 GENDER 与 BMI，TESTVALUE 分别为两变量的取值。此时 TESTDATE 与 TESTDAY 赋值 NA 以表示观察对象的个体特征变量。格式如：

http://www.empowerstats.com/empowerStats/exdata/nstructure1.xls

这两个文件内容完全一样，只是研究对象的个体特征变量 GENDER 与 BMI 的位置有所不同。部分数据截图如下：

| ID | TESTDAY | TESTVAR | TESTVALUE |
|---|---|---|---|
| 255486 | NA | GENDER | 1 |
| 255486 | NA | BMI | 21.63 |
| 255486 | 0 | NEUTn | 12.19 |
| 255486 | 0 | ESR | 8 |
| 255486 | 0 | PLT | 303 |
| 255486 | 0 | WBC | 21.91 |
| 255486 | 1 | CRP | 29.8 |
| 255486 | 1 | RC | 1.4 |
| 255486 | 1 | LC | 2.1 |
| 255486 | 1 | A | 15 |
| 255486 | 1 | LCA | 0.14 |
| 255486 | 1 | RCA | 0.09 |
| 255486 | 4 | NEUTn | 2.3 |
| 255486 | 4 | PLT | 553 |
| 255486 | 4 | WBC | 8.6 |
| 255486 | 11 | LC | 2.9 |
| 255486 | 11 | A | 15 |
| 255486 | 11 | RC | 1.2 |
| 255486 | 11 | RCA | 0.08 |
| 255486 | 11 | LCA | 0.19 |

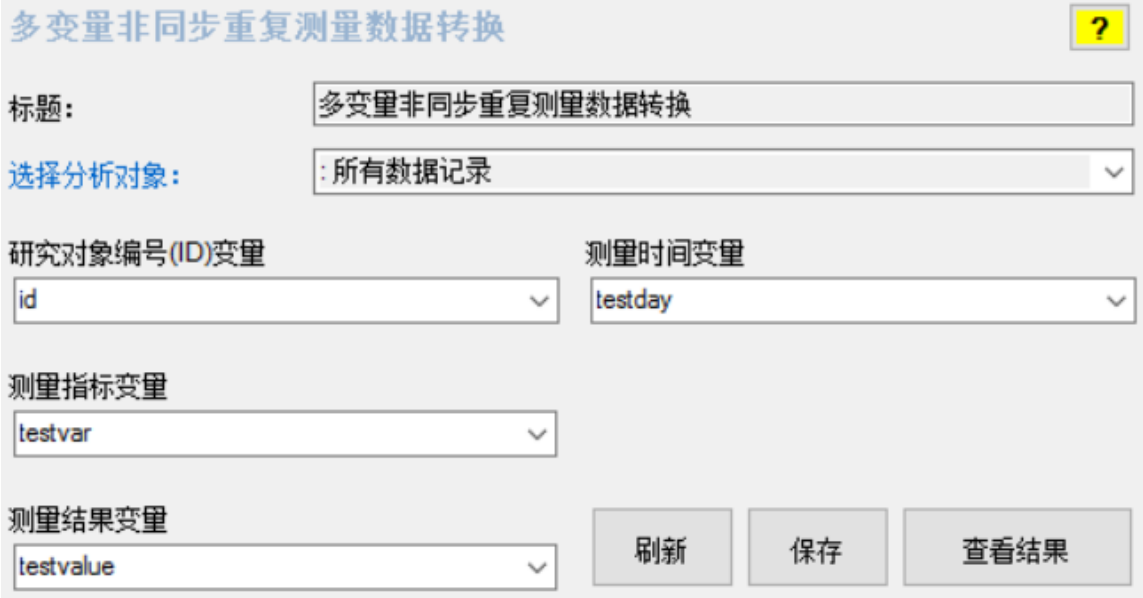
非结构化数据不要求预先固定测量指标，非常灵活，易扩展。但不能直接进入分析，需要转换成结构化的数据后才能进行分析，

本模块是将上述的数据结构转换成横向结构数据便于分析。

**二、非同步重复测量插补与数据结构转换**

要分析两指标之间的关系，需要两指标的同步测量值，如果没有同时测量的值，可以用附近的测量值取代，但临床上有些指标随时间变化明显，用临近的值取代不一定合理，这时候可以考虑根据前后的测量值对中间测量时点缺失进行插补，插补方法采用非参数曲线拟合的方法。本模块在将非结构化数据转换成结构化数据的同时，对缺失测量点数据进行插补，以便后续的数据分析。

例1. 练习数据 nstructure.xls（或 nstructure1.xls）是非结构化的重复测量数据，对测量时点缺失值进行插补并转换数据结构，易侕软件操作界面如下：




输出结果如下：

**Multivariate Asynchronous Repeated Measurement**

Data imputation and synchronization using nonparametric (Kernel) smoothing

| | |
|---|---|
| Original data | |
| # records (N): | 329 |
| # subjects (IDs): | 14 |
| measured items ( 10 ): | NEUTn ESR PLT WBC CRP RC LC A LCA RCA |
| Output data | nstructure_1_tbl.xls |
| # records (N): | 73 |
| Variables: | ID TESTDAY GENDER BMI NEUTn NEUTn.imp ESR ESR.imp PLT PLT.imp WBC WBC.imp CRP CRP.imp RC RC.imp LC LC.imp A A.imp LCA LCA.imp RCA RCA.imp<br><br>imp: imputed value for NA |

该数据有 14 个观察对象，329 条测量记录
检测指标 measured items 有：NEUTn ESR PLT WBC CRP RC LC A LCA RCA

输出数据文件 nstructure_1_tbl.xls 里有以下变量：
个体特征变量：ID TESTDAY GENDER BMI
检测指标变量、其插补值变量（_imp）：
NEUTn NEUTn.imp ESR ESR.imp PLT PLT.imp WBC WBC.imp CRP CRP.imp
RC RC.imp LC LC.imp A A.imp LCA LCA.imp RCA RCA.imp
在每个患者每个检测时点(如有观察值，_imp 变量等于其观测值，如没有实际观测值，_imp 为插补值。

以 ID=255509 为例，该患者有 5 个观测时点：TESTDAY=0、1、7、10、11，TESTDAY=1 时，NEUTn 观测值缺失，根据前后值对其插补，为 17.57065； 而 TESTDAY=11 时的 NEUTn 也缺失，但因为没有随后的实际观测值，没有足够的信息对之进行插补。TESTDAY=1、7 时 ESP 均缺失，尽管 TESTDAY =0 与 10 即前后有实际观测值，但因为实际观察点数<3, 没有足够的信息对中间缺失值进行插补。

nstructure_1_tbl.xls

| # | ID | TESTDAY | GENDER | BMI | NEUTn | NEUTn.imp | ESR | ESR.imp | PLT | PLT.imp | WBC | WBC.imp | CRP | CRP.imp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 255486 | 0 | 1 | 21.63 | 12.19 | 12.19 | 8 | 8 | 303 | 303 | 21.91 | 21.91 | NA | NA |
| 2 | 255486 | 1 | 1 | 21.63 | NA | NA | NA | NA | NA | NA | NA | NA | 29.8 | 29.8 |
| 3 | 255486 | 4 | 1 | 21.63 | 2.3 | 2.3 | NA | NA | 553 | 553 | 8.6 | 8.6 | NA | NA |
| 4 | 255486 | 11 | 1 | 21.63 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 5 | 255509 | 0 | 1 | 20.76 | 18.55 | 18.55 | 27 | 27 | 158 | 158 | 28.01 | 28.01 | NA | NA |
| 6 | 255509 | 1 | 1 | 20.76 | NA | 17.57065... | NA | NA | NA | 158 | NA | 26.8711... | 47.3 | 47.3 |
| 7 | 255509 | 7 | 1 | 20.76 | 3.08 | 3.08 | NA | NA | 440 | 440 | 8.5 | 8.5 | NA | NA |
| 8 | 255509 | 10 | 1 | 20.76 | 1.44 | 1.44 | 17 | 17 | 420 | 420 | 6.69 | 6.69 | 1.18 | 1.18 |
| 9 | 255509 | 11 | 1 | 20.76 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

将原数据文件换成 nstructure1.xls，操作同上，转换后的结果完全相同。该模块自动将测量时间变量（TESTDAY）赋值为“NA”的检测指标识别为个体特征变量。

## 三、生成边缘结构模型（MSM）分析数据

重复测量的生存分析数据，当暴露因素与混杂因素随时间变化时，时间依赖性混杂因素既可以看作暴露与结局的混杂因素，也可以看成暴露与结局之间的一个中间变量。在估计暴露的效应时，采用传统的多因素回归模型可以校正混杂因素的影响。然而，当把中间变量纳入模型时，则会产生有偏的估计。由于时间依赖性性混杂因素同时具有混杂因素和中间变量的性质，因此传统的回归模型不能很好地解决纵向数据中时依性混杂的问题。针对这个问题，Robins 在 1997 年提出了边缘结构模型(marginal structural model,MSM)这一新方法。但使用 MSM 要求每个观察对象的随访时点是相同的，当每个观察对象的随访时点（即暴露与混杂变量的测量时点）不固定时，需要通过数据拟合与数据结构转换生成固定时点的结构化数据。

例 3：下载练习数据：http://www.empowerstats.com/empowerStats/exdata/cd4tbdeath.xls

该数据为模拟数据，为非固定时点的重复测量数据。数据的部分截图如下：

每个 ID 的随访时间（FUPTIME）各异，数据记录为流水账式的非结构化的数据。

Y:  death (0/1);

X:  tuberculosis (1 = active);

Z:  cd4.sqrt

tuberculosis 与 cd4.sqrt 相互影响，同时与 death 有关。

| ID | FUPTIME | ITEM | RESULT |
|---|---|---|---|
| 120 | 5 | cd4.sqrt | 19.75 |
| 120 | 426 | cd4.sqrt | 18.11 |
| 120 | 440 | cd4.sqrt | 18.84 |
| 120 | 513 | cd4.sqrt | 18.65 |
| 120 | 626 | cd4.sqrt | 18.71 |
| 120 | 721 | cd4.sqrt | 17.2 |
| 120 | 786 | cd4.sqrt | 17.09 |
| 120 | 892 | cd4.sqrt | 17.61 |
| 120 | 995 | cd4.sqrt | 18.03 |
| 120 | 1200 | cd4.sqrt | 17.26 |
| 120 | 1293 | tuberculosis | 1 |
| 120 | 1374 | death | 0 |
| 121 | 17 | cd4.sqrt | 30.38 |
| 121 | 329 | cd4.sqrt | 27.51 |
| 121 | 348 | cd4.sqrt | 24.41 |
| 121 | 701 | cd4.sqrt | 23.26 |
| 121 | 751 | cd4.sqrt | 23.37 |
| 121 | 848 | cd4.sqrt | 22.38 |
| 121 | 1262 | cd4.sqrt | 18.95 |
| 121 | 1377 | cd4.sqrt | 17.86 |
| 121 | 1411 | cd4.sqrt | 16.97 |
| 121 | 1450 | cd4.sqrt | 16.22 |
| 121 | 1467 | cd4.sqrt | 18.22 |
| 121 | 1484 | cd4.sqrt | 16.94 |
| 121 | 1597 | cd4.sqrt | 16.09 |
| 121 | 1680 | cd4.sqrt | 15.33 |
| 121 | 1761 | death | 0 |

| ID | FUPTIME | ITEM | RESULT |
|---|---|---|---|
| 123 | 41 | cd4.sqrt | 23.41 |
| 123 | 55 | cd4.sqrt | 24 |
| 123 | 73 | cd4.sqrt | 23.71 |
| 123 | 263 | cd4.sqrt | 21.86 |
| 123 | 386 | cd4.sqrt | 20.93 |
| 123 | 436 | cd4.sqrt | 22.27 |
| 123 | 502 | cd4.sqrt | 21.49 |
| 123 | 699 | cd4.sqrt | 19.82 |
| 123 | 821 | cd4.sqrt | 19.1 |
| 123 | 974 | cd4.sqrt | 18.44 |
| 123 | 1001 | cd4.sqrt | 17.83 |
| 123 | 1066 | cd4.sqrt | 18 |
| 123 | 1259 | cd4.sqrt | 17.03 |
| 123 | 1359 | cd4.sqrt | 16.43 |
| 123 | 1584 | cd4.sqrt | 15.49 |
| 123 | 1668 | cd4.sqrt | 14.28 |
| 123 | 1779 | cd4.sqrt | 13.75 |
| 123 | 1835 | tuberculosis | 1 |
| 123 | 1840 | cd4.sqrt | 9.33 |
| 123 | 1902 | cd4.sqrt | 9.17 |
| 123 | 2011 | cd4.sqrt | 8.25 |
| 123 | 2085 | death | 1 |

调用非同步重复测量数据转换模块，输入界面如下：

输出结果（报告数据转换方法与结果）

Create structured data for Marginal structural model (MSM) analysis

Data imputation and synchronization using R (nlme package) lme()
To use MSM, follow up time need to be uniform for all IDs.
If it is not, need to impute data at time points other than the measurement times.
Use random effects model to smooth the original measurements.
with random effects for ID and for time t, fixed effect for time-varying exposure status X before the t.

| | |
|---|---|
| Original data | |
| # records (N): | 6782 |
| # subjects (IDs): | 386 |
| Variables: | ID, FUPTIME, ITEM, RESULT |
| measured items: | cd4.sqrt 6291<br>death 386<br>tuberculosis 105 |
| Output data | **cd4tb_1_tbl1.xls** |
| # records (N): | 84494 |
| # subjects (IDs): | 386 |
| Variables: | ID, START.TIME, FUP.TIME, TUBERCULOSIS, DEATH, CD4.SQRT |
| Follow up times: | Min. 1st Qu. Median Mean 3rd Qu. Max.<br>2 113 217.5 218.9 327.8 442 |

转换后的数据部分截图如下：

| ID | START.TIME | FUP.TIME | TUBERCULOSIS | DEATH | CD4.SQRT |
|---|---|---|---|---|---|
| 1 | -1 | 19 | 0 | 0 | 22.8160951... |
| 1 | 19 | 56 | 0 | 0 | 22.6547479... |
| 1 | 56 | 67 | 0 | 0 | 22.6067798... |
| 1 | 67 | 93 | 0 | 0 | 22.4934006... |
| 1 | 93 | 105 | 0 | 0 | 22.4410718... |
| 1 | 105 | 116 | 0 | 0 | 22.3931037... |
| 1 | 116 | 117 | 0 | 0 | 22.3887430... |
| 1 | 117 | 119 | 0 | 0 | 22.3800215... |
| 1 | 119 | 155 | 0 | 0 | 22.2230350... |
| 1 | 155 | 161 | 0 | 0 | 22.1968706... |

| ID | START.TIME | FUP.TIME | TUBERCULOSIS | DEATH | CD4.SQRT |
|---|---|---|---|---|---|
| 1 | 1841 | 1845 | 0 | 0 | 14.8533913... |
| 1 | 1845 | 1846 | 0 | 1 | 14.8490305... |
| 2 | -1 | 19 | 0 | 0 | 25.9840563... |
| 2 | 19 | 56 | 0 | 0 | 25.6938327... |
| 2 | 56 | 67 | 0 | 0 | 25.6075500... |
| 2 | 67 | 93 | 0 | 0 | 25.4036091... |
| 2 | 93 | 105 | 0 | 0 | 25.3094825... |
| 2 | 105 | 116 | 0 | 0 | 25.2231998... |
| 2 | 116 | 117 | 0 | 0 | 25.2153560... |
| 2 | 117 | 119 | 0 | 0 | 25.1996682... |
| 2 | 119 | 155 | 0 | 0 | 24.9172885... |
| 2 | 155 | 161 | 0 | 0 | 24.8702252... |
| 2 | 161 | 169 | 0 | 0 | 24.8074742... |
| 2 | 169 | 190 | 0 | 0 | 24.6427527... |

生成的新数据中每个 ID 的 START.TIME 与 FUP.TIME 都相同。各观察时点 CD4.SQRT 值是通过对原数据进行拟合后，对新观察时点进行预测得出来的。拟合方法采用(R nlme package)随机效应模型，研究对象与时间点为随机效应，暴露变量（tuberculosis）对混杂因素(cd4.sqrt)的影响为固定效应。（如输出页面提示：Use random effects model to smooth the original measurements.with random effects for ID and for time t, fixed effect for time-varying exposure status X before the t.）

转换后的数据可以直接使用 MSM 模块进行分析。