* 11/25 起始進度 HBP dataset已初步整理好

路徑：C:\Users\hsu\Desktop\master\TCHCData\hbp\_devtvs.csv

**待作清單：**

1. 決定預測目標(Y)
2. 挑選適合重複測量的插補資料方法 先看HBP欄位NA比例
3. 挑選變數
4. 輸入不同的input 比較模型結果

* 下次：

1. 決定預測目標(Y)

用hbp\_devtvs.csv裡的**檢測值欄位NA的比例**進行挑選

結果:

總筆數裡NA 比例最低的6個 (和血壓沒有關係的) 總筆數3032

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 變數 | HR | Total\_C | Creatinine | TG | LDL | glu\_AC |
| NA比(筆數)/有值筆數: | 44.49%  (1349)  1683 | 57.16%  (1733)  1299 | 57.49 %  (1743)  1289 | 58.91%  (1786)  1246 | 60.32 %  (1829)  1203 | 60.75 %  (1842)  1190 |

|  |  |  |  |
| --- | --- | --- | --- |
| 變數 \NA比例(人數)/有值筆數 | V1 Total 1378 | V2 Total 674 | V3 Total 346 |
| HR | 18.43% (254) 1124 | 67.8% (457) 217 | 17.63% (61) 285 |
| Total\_C | 32.15%(443)935 | 66.02%(445)229 | 64.45% (223) 123 |
| Creatinine | 33.02%(455)923 | 65.73%(443)231 | 63.87%(221)125 |
| TG | 34.03%(469)909 | 69.29$(467)207 | 65.90%(228)118 |
| LDL | 37.01%(510)868 | 68.25%(460)214 | 67.63%(234)112 |
| glu\_AC | 38.32%(528)457 | 67.8%(457)217 | 67.34%(233)113 |

因為模型是預測2類變數=>連續變數切成2類變數

和血壓有關係的NA比例最低的 :



1. 挑選適合重複測量的插補資料方法 先看**HBP欄位NA**比例

結果：

|  |  |  |  |
| --- | --- | --- | --- |
| HBP(總人數) | V1(1507) | V2(832) | V3(487) |
| 完整數值人數 | 368 | 223 | 124 |

7天\*4 =28 收縮 舒張共56筆 +平均6筆

56筆中超過3天沒有數值的(NA >12\*2)

|  |  |  |  |
| --- | --- | --- | --- |
| HBP(總人數) | V1(1507) | V2(832) | V3(487) |
| 超過3天沒有數值的 | 155 | 32 | 17 |

插補方法: k-NN 隨機森林 平均 中位數 …

1. 其他變數挑選

可能依照NA比例、回歸裡的顯著性…

1. Data分訓練測試的部分，因為是重複測量資料

測試是new patient

測試是訓練資料的new observation

1. TCHC 7天HBP (引用來源)

7天取6天 預測

7天早上 預測

**12/9與鄭醫師討論結果：**

1. Y用abpm裡的指標(morning surge 早晨血壓竄升 / nocturnal dipping 夜間血壓不降)
2. 白天和夜晚的時間採Long clock-time period的假設：

Daytime : 06:00-22:00 Night-time : 22:00-06:00

1. morning surge 有4種計算方式 目前沒有固定哪一種 好像算出來之後沒有一個判斷的臨界點=>要問 =>台灣沒有標準
2. nocturnal dipping 算出來後有4個階段(Reverse dipper / Non‐dipper / Dipper/ Extreme dipper)
3. if Y = morning surge預測 有沒有 “晨峰血壓增高”=>因為沒有一個切點=>不預測

if Y = nocturnal dipping 預測 Reverse dipper / Non‐dipper / Dipper/ Extreme dipper 4類中哪一類

* 12/12

待做：Data : ABPM

1. 計算TCHC 的 nocturnal dipping rate

3. 和hbp\_devsvt合併

* 12/15跟老師咪挺：把資料弄出來一切好說

先用hbp abpm 合併一起試試看 （其他covariate之後再補就可以）

把nocturnal dipping rate算出來（有問題的四百多個case先跳過）

先整理出complete case子集 試跑 4個模型 gee glmm bimmrf rf

nocturnal dipping 四類用兩類分類器預測 看case分佈決定 都可以嘗試看看

看完結果之後再看遺失值要怎麼插補

插補的策略 1都不補 2補ｘ（hbp） 3補y(abpm) 4兩個都補

去比較四種subset 預測的結果 如果都差不多表示遺失值沒有影響研究 有一致性

插補的同時就差不多要挑變數

插補的方法 再看看

框架類似銘洋學長的做法 可以看他怎麼做一堆模型比較的

* 12/22

目前完成

1. 計算TCHC 的 nocturnal dipping rate 還有status(abpm\_Y.csv) 1670\*194
2. 和hbp合併

hbp\_dip.csv (以Y為主合併) 1670\*125

hbp\_dip\_byx.csv(以X為主合併) 3032\*125

1. 計算hbp\_dip\_byx.csv 中status在各種情況下NA情形

|  |  |  |  |
| --- | --- | --- | --- |
| Extreme dipper | 12 | 198 6.5% | 845 |
| Dipper | 186 |
| Non dipper | 439 | 647 21% |
| Reverse dipper | 208 |
| NA | 2187 | 2187 72% |  |
| Total |  | 3032 |  |

各個訪視內的情況

**V1**

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 9 | 129 8.5% |
| Dipper | 120 |
| Non dipper | 301 | 447 29.6% |
| Reverse dipper | 146 |
| NA | 931 | 931 61.7% |
| Total |  | 1507 |

**V2**

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 2 | 36 4.3% |
| Dipper | 34 |
| Non dipper | 72 | 108 12.9% |
| Reverse dipper | 35 |
| NA | 689 | 689 82.8% |
| Total |  | 832 |

**V3**

|  |  |  |
| --- | --- | --- |
| Extreme dipper |  | 24 4.9% |
| Dipper | 24 |
| Non dipper | 43 | 61 12.5% |
| Reverse dipper | 18 |
| NA | 402 | 402 82.2% |
| Total |  | 487 |

=>有Y的樣本數少 兩類樣本數差距大

Y 是NA =>用分群方式 或是樹的方式產生(會dependent on X) 或是插補ABPM 再算Y

2類樣本數差距大 => resampling (把少的類別放大到跟另一類一樣多)

目前資料(HBP + dipping status)完整樣本數：245

|  |  |  |
| --- | --- | --- |
| Extreme dipper | 3 | 52 |
| Dipper | 49 |
| Non dipper | 143 | 193 |
| Reverse dipper | 50 |

* 12/22 咪挺結果

收縮 舒張 早上 晚上 平均 =>取4個

先抓平均完的complete case 去做模型(完整case會變多) 看效果

再對X先做隨機森林插補 再抓4個平均 再跑模型 看效果

做完再去做 對ABPM插補 再算Y 再跑模型

之後再考慮要不要加其他變數