1. 目的、研究動機

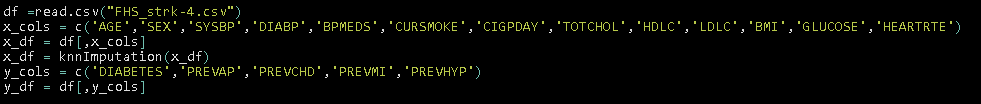
試圖以PLS手法，分析此份資料，找出對DIABETES、PREVAP、PREVCHD、PREVMI、PREVHYP等五種症狀(以下簡稱目標疾病)關鍵影響因子，若有複數，則將複數因子轉化成綜合評估指標，試圖以此指標評估得到人群是否為高風險群。

1. 手法: KNN、PLS、PCA、ANOVA
2. 資料分析
3. 資料前處理

X資料集 : AGE,SEX,SYSBP,DIABP,BPMEDS,CURSMOKE,CIGPDAY,TOTCHOL,HDLC,LDLC,BMI,GLUCOSE,HEARTRTE

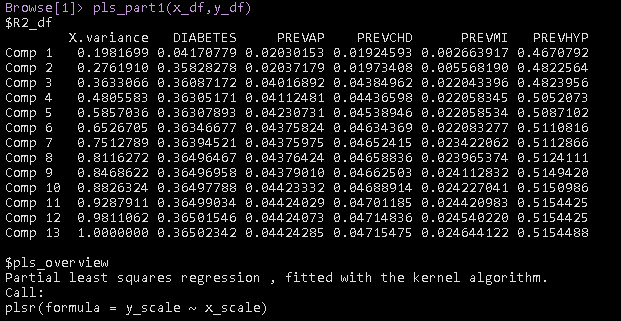
Y資料集 : 目標疾病

將X資料集缺值部分利用KNN手法進行補值

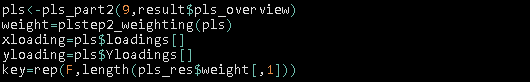


1. PLS分析

-找出X因子對Y因子解釋能力較佳的選項後，進行分析

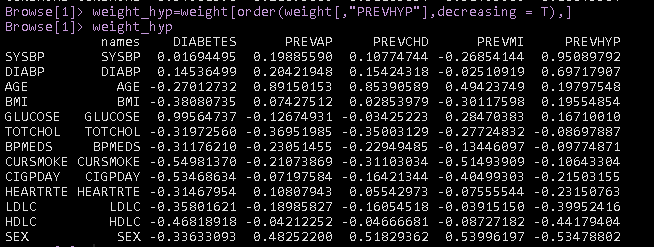


由數據可以看出，只有DIABETES、PREVHYP兩者解釋能較佳，在不遺失過多X資訊下，挑選Comp 9較佳。

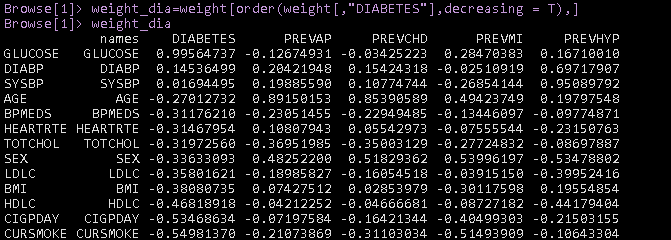


得到權重後，再分別以DIABETES、PREVHYP進行權重排序

PREHVP



DIABETES



以正值作為挑選原則，

PREHVP 關鍵因子 : SYSBP、DIABP、AGE、BMI、GLUCOSE

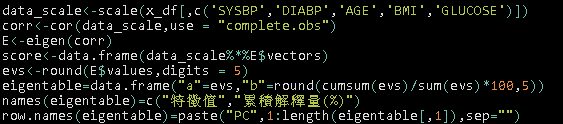
DIABETES 關鍵因子 : SYSBP、DIABP、GLUCOSE

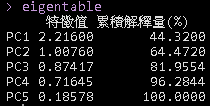
挑選PREHVP做為後續分析項目，

1. PCA資料降維、虛無假設驗證成果

將上一部驟分析所得關鍵因子，進行PCA資料降維

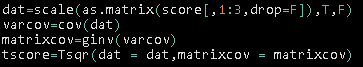
PERHBP:





依照凱莎原則，只能挑選PC1做為依據，但此部分不希望遺失過多X資訊，放寬至PC3，希望X解釋量超過80%。

依照挑選PC數將資料降成一維(綜合評估指數)



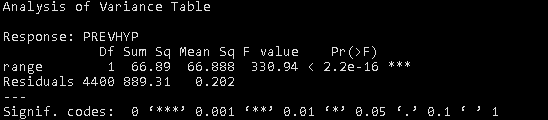
將資料綜合評估指標分成2組，高於平均(1)、低於平均(2)。

虛無假設:

H0: 1組得到高血壓人數與2組得到高血壓人數相當

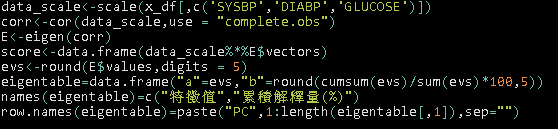
H1: 1組得到高血壓人數與2組得到高血壓人數不同

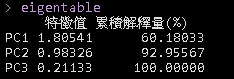




p-value <0.05，拒絕H0，1組得到高血壓人數與2組有顯著差異

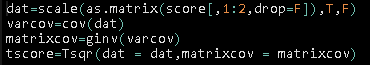
DIABETES





依照凱莎原則，無可挑選選項，但此部分不希望遺失過多X資訊，放寬至PC2，希望X解釋量超過80%。

依照挑選PC數將資料降成一維(綜合評估指數)



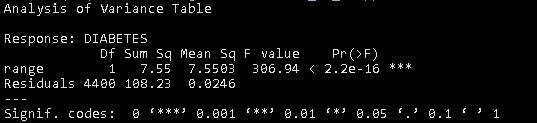
將資料綜合評估指標分成2組，高於平均(1)、低於平均(2)。

虛無假設:

H0: 1組得到糖尿病人數與2組得到糖尿病人數相當

H1: 1組得到糖尿病人數與2組得到糖尿病人數不同





p-value <0.05，拒絕H0，1組得到糖尿病人數與2組有顯著差異

四、結論

高血壓患病風險，與收縮壓、舒張壓、年齡、BMI、血糖有一定正相關性。

在高血壓確診條件中，與收縮壓、舒張壓有直接的條件定義，即休息狀態下收縮壓>140或舒張壓>90，故在此次分析中，得到高血壓關鍵因子有此

因子，是在預期中的。而年齡、BMI、血糖此三條件，也有一定的相關性，可以提供後續相關研究繼續分析。

糖尿病患病風險，與血糖、收縮壓、舒張壓有一定正相關性。

在糖尿病確診條件中，與血糖有直接的條件定義，即空腹8小時以上血糖  126 mg/dl，故在此次分析中，得到糖尿病關鍵因子有此因子，是在預期中的。而收縮壓、舒張壓此三條件，也有一定的相關性，可以提供後續相關研究繼續分析。