**240327**

**一、疾病Outcome提取，生存时间提取和疾病对照label划分**

1.给定疾病信息后，按照这个代码可以清洗原始数据，制作出用于生存分析的数据

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240327\_get\_disease\_outcome/获取基线后诊断某项疾病且有蛋白组学数据的个体AS.r

需要修改代码中这两行代码，替换为想提取的疾病

target\_disease <- "AS"

disease\_ICD10 <- "I70"

1. 代码的提取原理：先选出有蛋白组学数据的个体，然后根据疾病ICD编码从住院、死亡记录、primary care里面提出该疾病的诊断及诊断时间信息，剔除自我报告该疾病、基线前住院、Primary care诊断的个体，按照基线后是否诊断分为对照组与疾病组，赋予1/0 label
2. 用代码提取出IHD和AS 的疾病及对照数据

**二、cox得出疾病风险蛋白，疾病相关蛋白**

1.代码在下面的路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240327\_get\_disease\_related\_proteins

#协变量只有age+sex，不行的，这时候还没来得及考虑有缺失值的协变量，还是用的python在提

**三、LightGBM得到疾病预测模型，顺序前向选择筛选筛选重要蛋白**

1. 代码在下面的路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240327\_LightGBM/SFS.py

**240328**

1. **Top50、Top100相关蛋白LGBM建模，shap值计算**
2. 代码在下面的路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240328\_top50&top100\_proteins/240328\_top50+top100\_proteins\_shap

1. **Top50、Top100相关蛋白每个蛋白单独建模LGBM，计算AUC**
2. 代码在下面的路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240328\_top50&top100\_proteins/240328\_top50+top100\_proteins\_AUC

#这两个代码还有没有用不好说，很可能没有用了

**240402**

**一、惩罚样条cox建模，分析cox非线性桥段，以此确定阈值**

1.那篇gut的采用的就是PS的方法，代码在这个路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240402\_ps\_nonlinear\_cox

PS\_method.R

2.RCS限制性立方样条的简单代码也在上述路径下面（RCS\_method.R），rcs和ps都是非线性样条的方式

3.复现了gut那篇胆结石和四大脂质的关系,NMR.csv是代谢组学数据中四大脂质的数据，因为这里才有LDL-c

**240403**

1. **获取全部可能用到的协变量**

1.数据放在这个路径下面，整合代码比较容易，下次现场写也完全没问题

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Data/Raw\_data

#这里指的是covariate.csv文件，但是后面重新再优化了一下，所以后续用的都是covariate\_before\_inputation.csv

2.连续变量的数据没有归一化，下次再考虑怎么归一化吧

或许也不用，做cox的时候不需要归一化

**240407-08**

一、**协变量挑选**最终选定的协变量包括以下方面，共计27种:

* 人口学相关数据：年龄、性别、种族、教育经历、Townsend指数，收入，评估中心
* 生活习惯：吸烟、喝酒、IPAQ指数、整体健康状态
* 物理测量：BMI、身高、体重、腰围、SBP、DBP
* 生化指标：总胆固醇、C反应蛋白、HDL-c、甘油三酯、HbA1c
* 病史用药状态：高血压病史、糖尿病病史、服用降脂药、服用降压药
* 家族病史：心脏病和中风

1. **cox风险蛋白火山图制作**

代码在这个路径

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240408\_cox\_volcano\_plot

**三、多重插补**

1. 插补的代码在这个路径

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240408\_cox\_get\_disease\_related\_proteins\_R/协变量处理及插补 - shell.R

2.上述协变量，多重插补5次的结果在这个路径

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Result/cox\_with\_covariate\_imputed/imp.RData

lord(“imp.RData”)后保存在imp.cov里面，可以直接调用

多重插补20次的结果也在这个路径，保存在“imp20.RData”中

**四、多重插补后，cox按照overlap得出疾病风险蛋白**

1. 代码在这个路径

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240408\_cox\_get\_disease\_related\_proteins\_R/协变量处理及插补&疾病蛋白获取（overlap法）.R

#这时候是 把五次插补得到的datasets 分别对每个蛋白和协变量建立cox模型 得到蛋白fbi-pval，排序后取交集。最大的问题是合并的p值手动求不了（不知道采用的是哪种自由度矫正方法），所以放弃这个代码

#从这时候起，所有的cox得出疾病风险蛋白过程，都是在R上实现的了（因为要使用插补数据，用R插补教程比python要全面得多）

**五、采用直接删去缺失值方法，cox获得风险蛋白**

结果在这里/home/linmiao/ZhangjunYu/Proteomics\_analysis/Result/cox\_with\_covariate

LGBM结果在这里

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Result/LightGBM\_with\_covariate/IHD

#直接删去缺失的方法不如多重插补的

**240409**

1. **调整对照组**
2. 完全健康
3. 完全健康和IHD个体做match
4. Non-CVD(没有I00~I99疾病)
5. Non-IHD（没有患过IHD疾病）

#完全健康的只有5000个左右，IHD的有4000个左右，没办法match，所以第2组去掉

1. **代码在这个路径**

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240409\_adjust\_control\_for\_IHD

**240410**

一、**多重插补后，cox按照pool得出疾病风险蛋白**

在这个路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240410\_cox\_get\_disease\_proteins\_after\_imputation

分别按照3种对照设定方法进行跑取

#多重插补，p值合并的参考代码也在这个路径，试了一下，线性回归计算的p值没有问题，但是cox计算的p值不一致，下次使用的时候要注意。但估计没有下次了，直接用pool挺方便的！

**240412**

一、**多重插补后，LGBM得到关键蛋白后，确定蛋白阈值**

1. PS方法

以排名第一的蛋白GDF15为例，离群值比较明显(取值大于2.5的仅占0.2%，100人左右)，极大地影响PS建模

#而且目前没找到能合并五次插补结果的方法

r$> summary(pro\_df)

eid GDF15

Min. :1000074 Min. :-2.24745

1st Qu.:2258547 1st Qu.:-0.31595

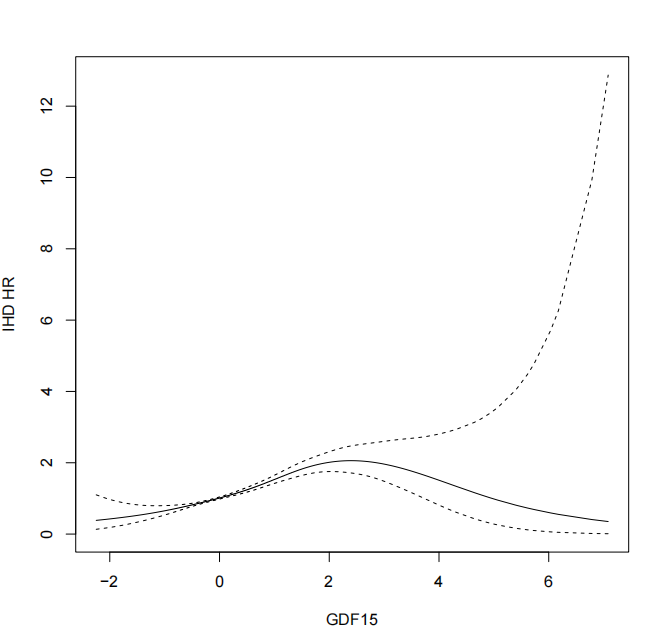
Median :3505854 Median :-0.00435

Mean :3512930 Mean : 0.07393

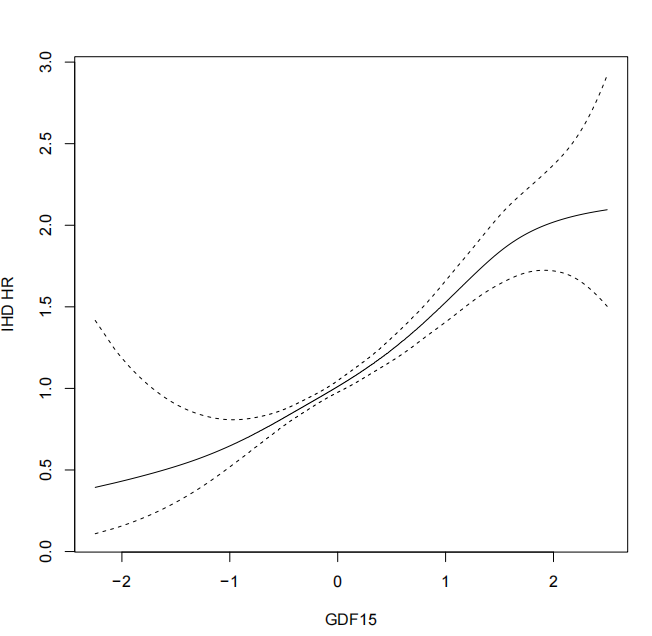
3rd Qu.:4770799 3rd Qu.: 0.35807

Max. :6024005 Max. : 7.08375

全部的GDF15用于非线性cox建模，有非线性关系，但很明显在GDF15取值较大的时候估计不准确



仅选取-2.5 < GDF > 2.5的部分，非线性又很不好（P > 0.05）



1. rcs方法

非线性不是很好，wald检验P=0.0305，但是具体到cox不太行

#而且目前没找到能合并五次插补结果的方法

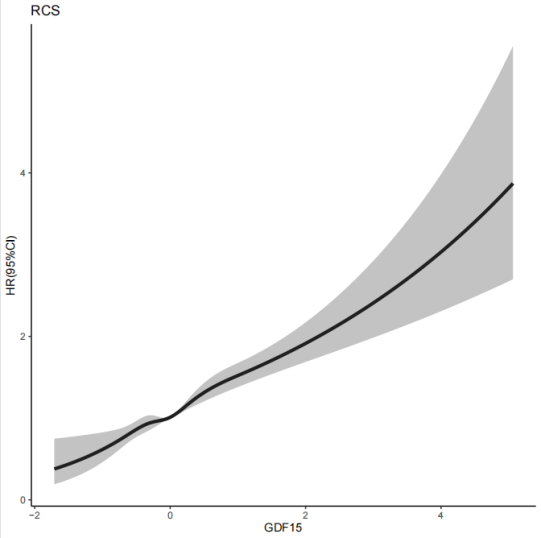
Coef S.E. Wald Z Pr(>|Z|)

target\_pro 0.6865 0.2904 2.36 0.0181

target\_pro' -2.3540 1.9196 -1.23 0.2201

target\_pro'' 10.5741 7.5038 1.41 0.1588

target\_pro''' -13.8990 8.6339 -1.61 0.1074



1. Youden index方法

代码在这个路径下面：

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240412\_determine\_cut\_off\_for\_proteins/youden\_index

得到的阈值边界以csv形式保存，在这个文件夹中

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Result/cut\_off\_determination/youden

**240413**

**一、得到阈值后，对关键蛋白进行生存分析，验证重要程度**

1. 对Youden index得到的蛋白阈值进行分析

代码在这个路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240413\_KM\_analysis/IHD\_Plot.py

High和Low代表的是蛋白浓度的高和低

1. 以pdf形式保存，代码在这个路径下面（每10个画一份pdf文件）

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240413\_KM\_analysis/IHD\_Plot\_for\_pdf.py

**240415**

**一、阴性对照的HR计算和KM分析**

1. 阴性对照包括①非缺血性心脏病的心血管疾病、②瓣膜性疾病

代码在这个路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/240415\_negative\_control

**240416**

**一、获取AS的疾病人群，处理OPSC4数据**

1. 代码在这个路径下面

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/1\_实操的代码/ASCVD/Get\_disease\_outcome/240416\_get\_disease\_outcome(operation\_included)/获取基线后诊断某项疾病且有蛋白组学数据的个体 ASCVD.r

**20240503**

重新整理了所有代码，用当前的进展制作了代码合集（模板，全套流程代码）

放在下面这个路径中了

/home/linmiao/ZhangjunYu/Proteomics\_analysis/Code/0\_代码模板汇总

**20240509**

一、蛋白表达的轨迹制作

利用性别、年龄、Townsend deprivation index进行匹配1：5

然后画拟合曲线