成年人高血压关联因素探索

吴沛豪

2023-05-16

目录

1 环境准备

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_China.utf8
## [2] LC_CTYPE=Chinese (Simplified)_China.utf8
## [3] LC_MONETARY=Chinese (Simplified)_China.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_China.utf8
##
## attached base packages:
## [1] stats4    grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] ggpubr_0.6.0         runway_0.0.0.9000    pROC_1.18.0
##  [4] tensorflow_2.11.0    keras_2.11.1         party_1.3-13
##  [7] strucchange_1.5-3    sandwich_3.0-2       zoo_1.8-11
## [10] modeltools_0.2-23    mvtnorm_1.1-3        RWeka_0.4-46
## [13] rattle_5.5.1         bitops_1.0-7         xgboost_1.7.5.1
## [16] glmnet_4.1-7         Matrix_1.5-4         e1071_1.7-13
## [19] MASS_7.3-58.3        caret_6.0-94         lattice_0.20-45
## [22] randomForest_4.7-1.1 rpart.plot_3.1.1     rpart_4.1.19
## [25] mice_3.15.0          yardstick_1.1.0      workflowsets_1.0.1
## [28] workflows_1.1.3      tune_1.1.0           rsample_1.1.1
## [31] recipes_1.0.5        parsnip_1.0.4        modeldata_1.1.0
## [34] infer_1.0.4          dials_1.2.0          scales_1.2.1
## [37] broom_1.0.4          tidymodels_1.0.0     VIM_6.2.2
## [40] colorspace_2.1-0     patchwork_1.1.2      qgraph_1.9.4
## [43] reshape2_1.4.4       lubridate_1.9.2      forcats_1.0.0
## [46] stringr_1.5.0        dplyr_1.1.1          purrr_1.0.1
## [49] readr_2.1.4          tidyr_1.3.0          tibble_3.2.1
## [52] ggplot2_3.4.2        tidyverse_2.0.0      igraph_1.4.1
##
```

```
## loaded via a namespace (and not attached):
##   [1] backports_1.4.1      Hmisc_5.0-1          plyr_1.8.8
##   [4] sp_1.6-0             splines_4.2.2        listenv_0.9.0
##   [7] tfruns_1.5.1         TH.data_1.1-1        digest_0.6.31
##  [10] foreach_1.5.2        htmltools_0.5.4      fansi_1.0.4
##  [13] magrittr_2.0.3       checkmate_2.1.0      cluster_2.1.4
##  [16] tzdb_0.3.0           globals_0.16.2       gower_1.0.1
##  [19] matrixStats_0.63.0   hardhat_1.3.0        timechange_0.2.0
##  [22] jpeg_0.1-10          xfun_0.37            libcoin_1.0-9
##  [25] jsonlite_1.8.4       zeallot_0.1.0        survival_3.5-5
##  [28] iterators_1.0.14     glue_1.6.2           gtable_0.3.3
##  [31] ipred_0.9-14         car_3.1-2            shape_1.4.6
##  [34] future.apply_1.10.0  DEoptimR_1.0-12      abind_1.4-5
##  [37] rstatix_0.7.2        Rcpp_1.0.10          laeken_0.5.2
##  [40] htmlTable_2.4.1      reticulate_1.28      GPfit_1.0-8
##  [43] foreign_0.8-84       proxy_0.4-27         Formula_1.2-5
##  [46] lava_1.7.2.1         prodlim_2023.03.31   vcd_1.4-11
##  [49] htmlwidgets_1.6.2    lavaan_0.6-15        rJava_1.0-6
##  [52] pkgconfig_2.0.3      nnet_7.3-18          utf8_1.2.3
##  [55] tidyselect_1.2.0     rlang_1.1.0          DiceDesign_1.9
##  [58] munsell_0.5.0        tools_4.2.2          cli_3.6.0
##  [61] generics_0.1.3       ranger_0.15.1        fdrtool_1.2.17
##  [64] evaluate_0.20        fastmap_1.1.1        yaml_2.3.7
##  [67] rticles_0.24         ModelMetrics_1.2.2.2 knitr_1.42
##  [70] robustbase_0.95-1    coin_1.4-2           glasso_1.11
##  [73] pbapply_1.7-0        future_1.32.0        nlme_3.1-162
##  [76] whisker_0.4.1        compiler_4.2.2       rstudioapi_0.14
##  [79] png_0.1-8            ggsignif_0.6.4       lhs_1.1.6
##  [82] pbivnorm_0.6.0       stringi_1.7.12       psych_2.3.3
##  [85] RWekajars_3.9.3-2    vctrs_0.6.1          pillar_1.9.0
##  [88] lifecycle_1.0.3      furrr_0.3.1          lmtest_0.9-40
##  [91] data.table_1.14.8    corpcor_1.6.10       R6_2.5.1
##  [94] gridExtra_2.3        parallelly_1.35.0    codetools_0.2-19
##  [97] boot_1.3-28.1        gtools_3.9.4         withr_2.5.0
## [100] mnormt_2.1.1         multcomp_1.4-23      parallel_4.2.2
## [103] hms_1.1.3            quadprog_1.5-8       timeDate_4022.108
## [106] class_7.3-21         rmarkdown_2.21       carData_3.0-5
## [109] base64enc_0.1-3
```

```
if (length(tf$config$list_physical_devices("GPU"))) {
  message("TensorFlow **IS** using the GPU")
} else {
```

```
  message("TensorFlow **IS NOT** using the GPU")
}
```

2 数据准备

2.1 运行 DataClean.Rmd 脚本以获取清洗后数据；

2.2 运行 DataClean2.Rmd 脚本重编码变量：

2.3 运行 DataClean3.Rmd 脚本将有序分类变量及无序二分类变量转换成 numeric

2.4 运行 data_use2.R 脚本合并数据

```
load(file = paste0(getwd(), "/data_use/data_use_4_old.RData"))
```

2.5 检查数据

```
str(YData)
```

```
## 'data.frame':    4335 obs. of  4 variables:
##  $ SEQN    : num  93705 93706 93708 93711 93712 ...
##  $ BPXSY   : num  200 111 142 101 113 ...
##  $ BPXDI   : num  68 73.3 76 66.7 70 ...
##  $ response: num  1 0 1 0 0 0 1 0 0 0 ...
```

```
str(XData)
```

```
## 'data.frame':    4335 obs. of  140 variables:
##  $ SEQN    : num  93705 93706 93708 93711 93712 ...
##  $ RIAGENDR: num  0 1 0 1 1 1 0 1 1 1 ...
##  $ RIDRETH3: Factor w/ 6 levels "Mexican American",..: 4 5 5 5 1 3 4 6 5 3 ...
##  $ DMDBORN4: num  1 1 0 0 0 1 1 1 0 1 ...
##  $ DMDCITZN: num  1 1 1 1 0 1 1 1 1 1 ...
##  $ DMDHHSIZ: num  1 5 2 3 4 1 3 5 3 2 ...
##  $ DMDFMSIZ: num  1 5 2 3 4 1 3 5 3 1 ...
##  $ DMDHHSZA: num  0 0 0 0 0 0 0 1 0 0 ...
##  $ DMDHHSZB: num  0 0 0 0 2 0 1 2 0 0 ...
##  $ DMDHHSZE: num  1 1 2 0 0 1 0 1 1 0 ...
##  $ DMDHRGND: num  0 1 1 1 0 1 1 1 1 1 ...
##  $ DMDHRAGZ: num  4 4 4 3 3 4 3 4 4 2 ...
##  $ DMDHREDZ: num  1 3 1 3 1 2 2 2 3 1 ...
##  $ DMDHRMAZ: Factor w/ 3 levels "Married/Living with partner",..: 2 1 1 1 2 2 1 1 1 3 ...
##  $ INDHHIN2: num  1 2 2 3 1 2 2 2 3 1 ...
##  $ INDFMIN2: num  1 2 2 3 1 2 2 2 3 1 ...
```

```
##  $ AGE     : num  66 18 66 56 18 67 54 71 61 22 ...
##  $ DMDEDUC : num  1 2 1 3 1 2 3 2 3 2 ...
##  $ WTDRD1  : num  7186 6464 10826 9098 60947 ...
##  $ WTDR2D  : num  5640 6464 22482 8230 89066 ...
##  $ DBQ095Z : num  1 3 1 3 3 3 3 3 3 3 ...
##  $ DRQSPREP: num  3 3 4 2 3 1 4 4 3 3 ...
##  $ DRQSDIET: num  0 0 0 1 0 0 0 0 0 0 ...
##  $ DR1TNUMF: int  17 8 14 27 12 17 16 9 18 18 ...
##  $ DR1TKCAL: int  1202 1987 1251 2840 2045 2040 2493 1287 2917 3151 ...
##  $ DR1TPROT: num  20 94.2 51 101.3 99.7 ...
##  $ DR1TCARB: num  157.4 89.8 123.7 339.6 268.2 ...
##  $ DR1TSUGR: num  91.5 14.7 49.8 148.2 125 ...
##  $ DR1TFIBE: num  8.4 7.1 16.6 44.5 22.3 14.6 11.1 2.4 31.4 18 ...
##  $ DR1TTFAT: num  57 137.4 65.5 124.2 63.9 ...
##  $ DR1TSFAT: num  16.4 35.2 17.4 41.3 15.9 ...
##  $ DR1TMFAT: num  16.4 45.8 29 39.6 24.2 ...
##  $ DR1TPFAT: num  19.8 49.9 14.8 31.3 19 ...
##  $ DR1TCHOL: int  14 462 71 546 216 176 965 470 300 384 ...
##  $ DR1TATOC: num  5.66 10.02 6.2 14.27 7.05 ...
##  $ DR1TATOA: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TRET : int  32 198 35 691 23 212 584 280 384 472 ...
##  $ DR1TVARA: int  436 431 236 1012 46 577 608 300 1222 886 ...
##  $ DR1TACAR: int  1551 872 323 414 51 1095 9 0 3314 287 ...
##  $ DR1TBCAR: int  4096 2363 2245 3639 171 3736 265 181 7461 4736 ...
##  $ DR1TCRYP: int  2 2 26 31 156 200 34 65 1774 167 ...
##  $ DR1TLYCO: int  1573 4605 0 23074 618 548 25 342 0 6435 ...
##  $ DR1TLZ  : int  1645 313 2148 5629 316 1745 928 628 2054 3932 ...
##  $ DR1TVB1 : num  0.589 1.152 1.143 1.79 1.619 ...
##  $ DR1TVB2 : num  1.24 1.03 0.84 3.22 1.51 ...
##  $ DR1TNIAC: num  7.58 26.83 15.37 17.38 31.34 ...
##  $ DR1TVB6 : num  0.458 1.821 1.096 2.177 2.59 ...
##  $ DR1TFOLA: int  179 267 260 609 437 262 203 120 655 519 ...
##  $ DR1TFA  : int  32 125 74 84 76 55 33 45 328 256 ...
##  $ DR1TFF  : int  146 139 185 526 361 206 169 75 327 263 ...
##  $ DR1TFDFE: int  202 354 311 669 490 300 224 152 888 696 ...
##  $ DR1TCHL : num  95 368 176 546 373 ...
##  $ DR1TVB12: num  0.33 2.3 1.09 3.62 3.62 2.55 3.72 3.24 3.22 3.02 ...
##  $ DR1TB12A: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TVC  : num  21.4 9.7 146.4 124 182.1 ...
##  $ DR1TVD  : num  0.2 0.7 0.8 4.7 1.3 0.6 7 6.7 4.2 7.6 ...
##  $ DR1TVK  : num  156 138 137 277 49 ...
```

```
##  $ DR1TCALC: int  314 869 412 1635 391 583 981 623 972 1959 ...
##  $ DR1TPHOS: int  466 1025 635 2141 1256 950 1908 839 1638 2027 ...
##  $ DR1TMAGN: int  162 187 248 541 260 210 276 119 451 282 ...
##  $ DR1TIRON: num  8.8 8.52 11.49 17 12.07 ...
##  $ DR1TZINC: num  2.93 8.05 6.45 13.25 15 ...
##  $ DR1TCOPP: num  0.689 0.614 1.049 1.983 1.256 ...
##  $ DR1TSODI: int  3574 3657 2135 4382 3753 2456 5000 1430 4831 6470 ...
##  $ DR1TPOTA: int  1640 1247 1631 4457 3358 2488 2449 1634 4190 3089 ...
##  $ DR1TSELE: num  22.1 118.5 54.3 129.7 109.7 ...
##  $ DR1TCAFF: int  361 0 33 347 0 385 60 432 95 70 ...
##  $ DR1TTHEO: int  120 0 69 68 0 125 161 0 0 25 ...
##  $ DR1TALCO: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TMOIS: num  1774 3405 2822 4345 3217 ...
##  $ DR1TS040: num  0.156 0.263 0.07 0.88 0.033 0.543 0.982 0.252 0.491 0.936 ...
##  $ DR1TS060: num  0.077 0.203 0.044 0.594 0.027 0.368 0.624 0.198 0.376 0.706 ...
##  $ DR1TS080: num  0.058 0.14 0.027 0.459 0.02 0.253 0.457 0.136 0.267 0.457 ...
##  $ DR1TS100: num  0.122 0.377 0.091 1.022 0.07 ...
##  $ DR1TS120: num  0.145 0.459 0.097 1.601 0.07 ...
##  $ DR1TS140: num  0.447 1.816 0.499 3.742 0.633 ...
##  $ DR1TS160: num  8.95 23.15 9.44 23.52 10.38 ...
##  $ DR1TS180: num  5.98 7.75 6.13 8.38 4.24 ...
##  $ DR1TM161: num  0.118 3.387 0.446 1.027 1.085 ...
##  $ DR1TM181: num  16 41.6 28.2 38.2 22.7 ...
##  $ DR1TM201: num  0.101 0.524 0.31 0.285 0.248 0.312 0.564 0.186 0.306 0.586 ...
##  $ DR1TM221: num  0.014 0.011 0.003 0.004 0.001 0.017 0.078 0.038 0 0.131 ...
##  $ DR1TP182: num  17.8 44.1 13.9 21.7 17.1 ...
##  $ DR1TP183: num  1.943 5.074 0.804 9.337 1.522 ...
##  $ DR1TP184: num  0 0.016 0 0 0 0 0.005 0.002 0 0.008 ...
##  $ DR1TP204: num  0.014 0.308 0.038 0.254 0.161 0.102 0.49 0.231 0.137 0.249 ...
##  $ DR1TP205: num  0.001 0.021 0.001 0.004 0.003 0.006 0.007 0.007 0.015 0.014 ...
##  $ DR1TP225: num  0.001 0.044 0.004 0.02 0.03 0.009 0.033 0.023 0.015 0.039 ...
##  $ DR1TP226: num  0.001 0.021 0 0.062 0.001 0.002 0.097 0.059 0.018 0.018 ...
##  $ DR1_300 : int  2 3 2 2 3 1 1 1 2 2 ...
##  $ DR1_320Z: num  315 3042 2160 1902 1014 ...
##  $ DR1_330Z: num  315 0 720 1902 0 ...
##  $ DR1BWATZ: num  0 3042 1440 0 1014 ...
##  $ DRD360  : num  0 1 1 0 1 0 1 1 1 1 ...
##  $ BMXWT   : num  79.5 66.3 53.5 62.1 58.9 74.9 87.1 65.6 77.7 74.4 ...
##  $ BMXHT   : num  158 176 150 171 173 ...
##  $ BMXBMI  : num  31.7 21.5 23.7 21.3 19.7 23.5 39.9 22.5 30.7 24.5 ...
##  $ BMXLEG  : num  37 46.6 31.8 40.1 44.5 39.1 26 42 37.4 44 ...
```

```
##  $ BMXARML : num  36 38.8 30.6 37.2 37.2 41.4 32 39.3 32.6 41.4 ...
##   [list output truncated]
```

3 数据标准化、虚拟化

```
source("std.r")
```

4 方差选择法

选择方差较大的特征。如果一个特征的方差很小，那么它对预测结果的影响也很小

```
df_var2 <- nearZeroVar(Data_SD[-c(1, 2, 3)], saveMetrics = TRUE)
Data_SD <- Data_SD[, !df_var2$nzv]
```

```
# 重命名变量，防止作图显示问题
Data_SD <- Data_SD %>%
  dplyr::rename(RIDRETH3MA = RIDRETH3Mexican.American,
                RIDRETH3OH = RIDRETH3Other.Hispanic,
                RIDRETH3NHW = RIDRETH3Non.Hispanic.White,
                RIDRETH3NHB = RIDRETH3Non.Hispanic.Black,
                RIDRETH3NHA = RIDRETH3Non.Hispanic.Asian,
                RIDRETH3OR = RIDRETH3Other.Race...Including.Multi.Rac,
                DMDHRMAZWDS = DMDHRMAZWidowed.Divorced.Separated,
                DMDHRMAZNM = DMDHRMAZNever.Married,
                PHDSESNA = PHDSESNafternoon,
                PHDSESNE = PHDSESNevening)
```

5 相关性计算

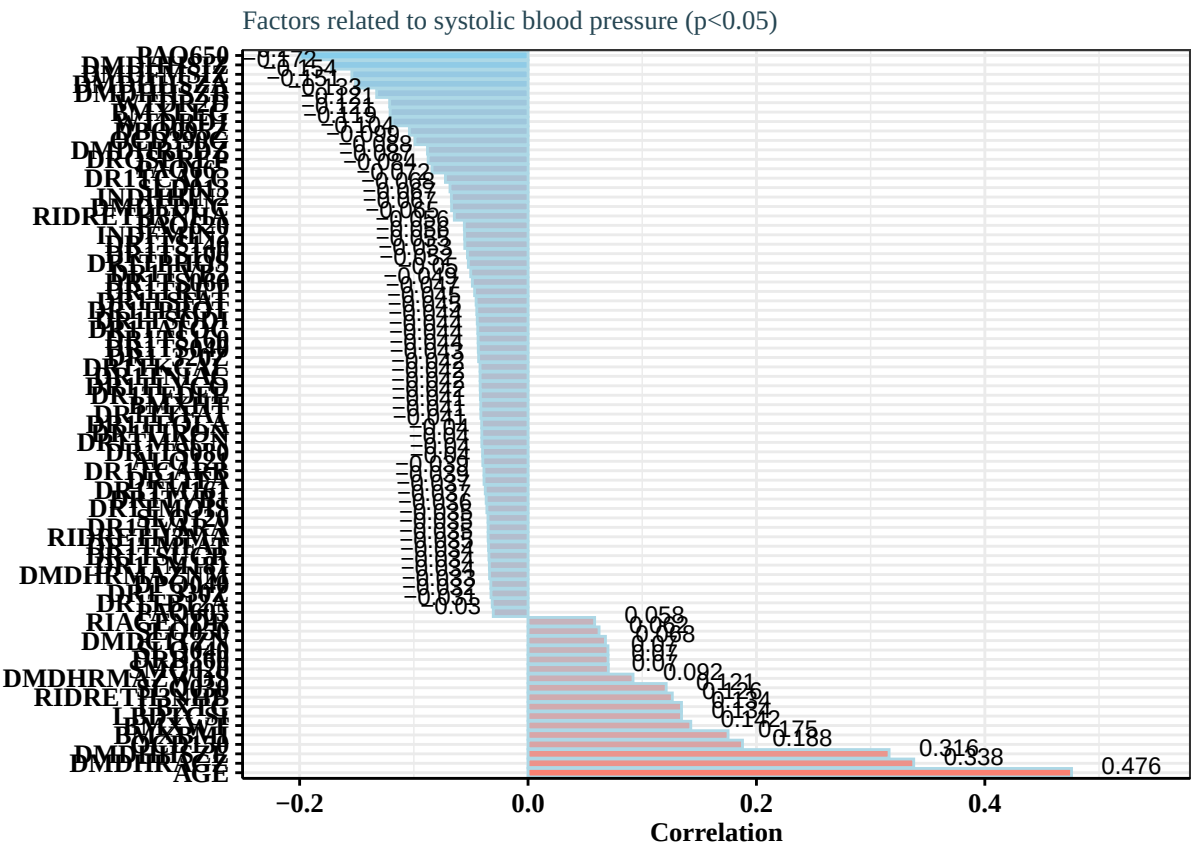
5.1 收缩压关联因素

```
cor.f <- function(m, n = 4, data = Data_SD) {
  cor <- NULL
  t_value <- NULL
  p_value <- NULL
  for (i in n:ncol(data)) {
    temp <- cor.test(data[, m], data[, i])
    cor <- cor %>% append(temp$estimate)
    t_value <- t_value %>% append(temp$statistic)
    p_value <- p_value %>% append(temp$p.value)
  }
  opt <- data.frame(
    cor = cor,
    t_value = t_value,
    p_value = p_value
  )
  rownames(opt) <- colnames(data)[n:ncol(data)]
  return(opt)
}
```

```
cor_S <- cor.f(1, data = Data_SD)
cor_S <- cor_S %>%
  dplyr::filter(p_value < 0.05) %>%
  arrange(-cor)
```

```
pA <- ggplot(data = cor_S, aes(y = cor, x = reorder(rownames(cor_S), -cor))) +
  geom_col(aes(fill = cor), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Correlation") +
  ggtitle("Factors related to systolic blood pressure (p<0.05)") +
  scale_y_continuous(expand = c(0,0),
                     limits = c(-0.25, 0.58)) +
  geom_text(aes(label = round(cor, 3)),
```

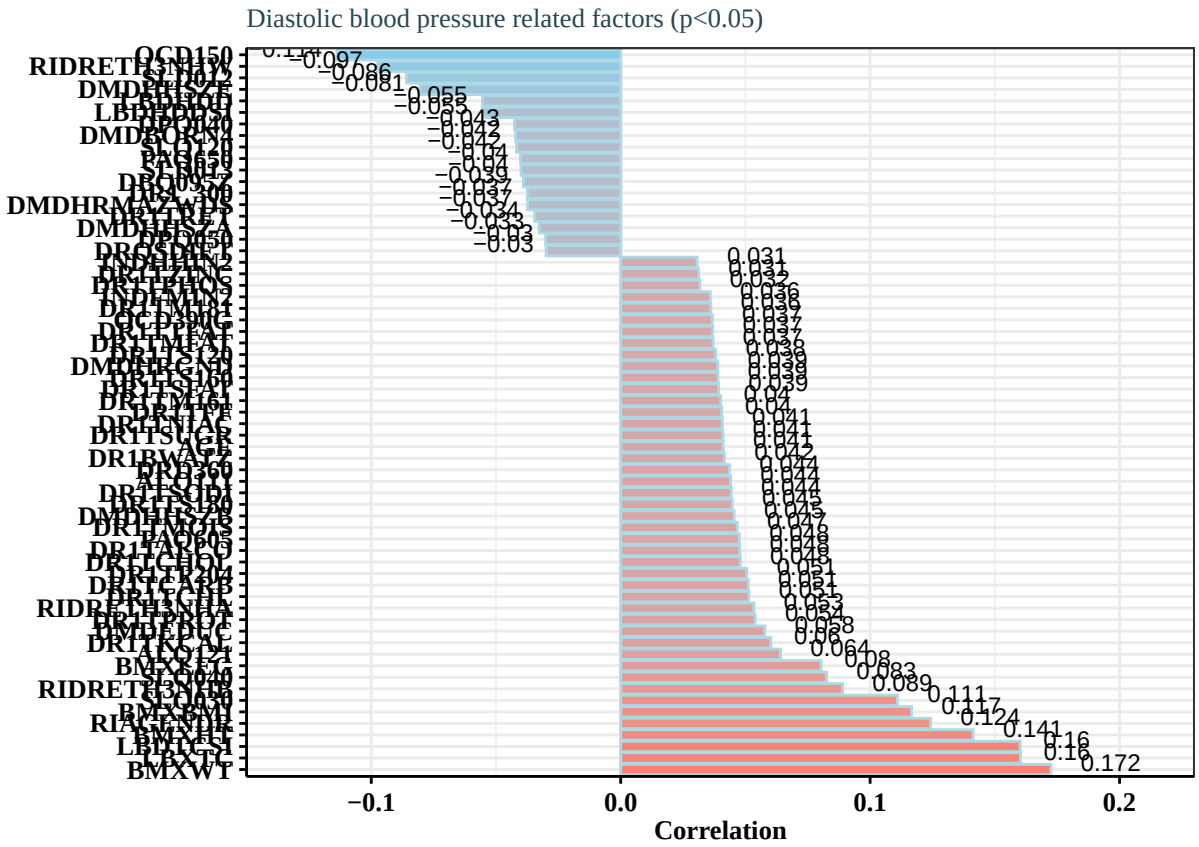
```
            vjust = 0.1, hjust = if_else(cor_S$cor >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
        axis.ticks = element_line(color = "black"),# 设置刻度线颜色
        axis.line = element_line(size = 0.5,
                                 colour = "black"),# 设置边框线颜色
        axis.title = element_text(colour = "black",
                                  size = 10,
                                  face = "bold"),# 设置标题字体
        axis.text = element_text(colour = "black",
                                 size = 10,
                                 face = "bold"),# 设置 x,y 轴标签字体
        axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
        text = element_text(size = 8,
                            color = "#264653",
                            family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pA
```



5.2 舒张压关联因素

```
cor_D <- cor.f(2, data = Data_SD)
cor_D <- cor_D %>%
  dplyr::filter(p_value < 0.05) %>%
  arrange(-cor)
```

```
pB <- ggplot(data = cor_D, aes(y = cor, x = reorder(rownames(cor_D), -cor))) +
  geom_col(aes(fill = cor), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Correlation") +
  ggtitle("Diastolic blood pressure related factors (p<0.05)") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(-0.15, 0.23)) +
  geom_text(aes(label = round(cor, 3)),
            vjust = 0.1, hjust = if_else(cor_D$cor >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pB
```

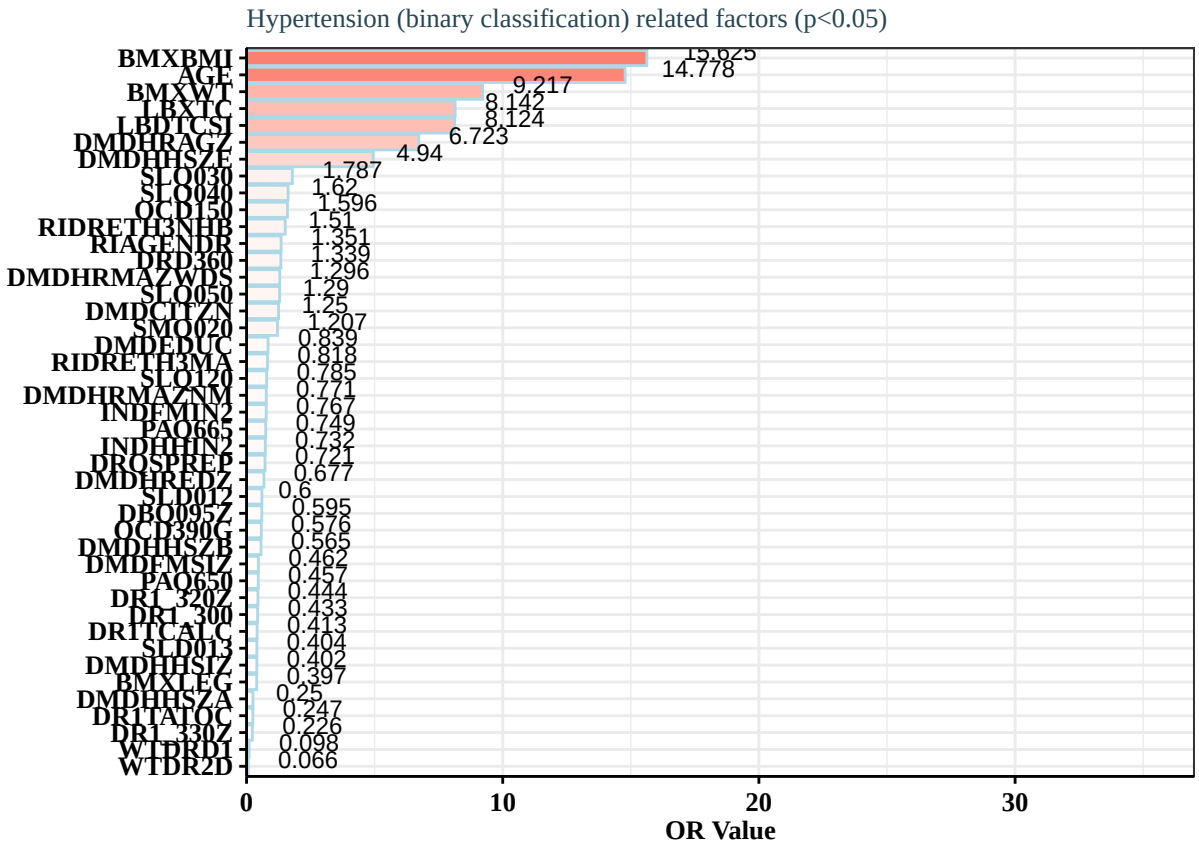


5.3 高血压关联因素（logistic）

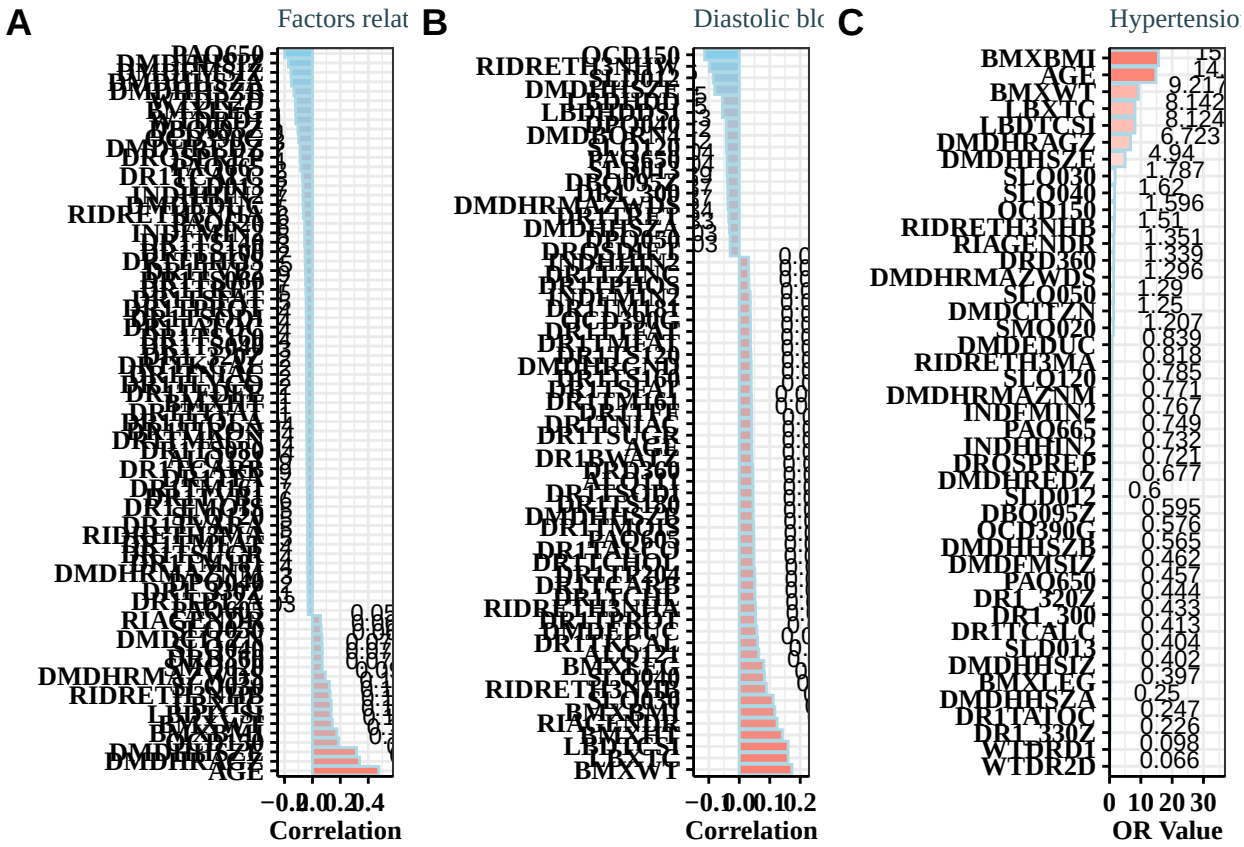
```
beta <- NULL
P_value <- NULL
for (i in 4:139) {
  temp.fit <- glm(response ~ Data_SD[, i],
    data = Data_SD,
    family = binomial(link = "logit")
  )
  temp <- summary(temp.fit)
  beta <- beta %>% append(temp$coefficients[2, 1])
  P_value <- P_value %>% append(temp$coefficients[2, 4])
}
opt <- data.frame(
  beta = beta,
  P_value = P_value
)
rownames(opt) <- colnames(Data_SD[4:139])
opt$OR_value <- exp(opt$beta)
```

```
opt$Q_value <- sprintf("%0.3f",
                       p.adjust(opt$P_value,
                                method = "BH", nrow(opt)))
opt <- opt %>%
  dplyr::filter(P_value < 0.05) %>%
  arrange(-OR_value, P_value)
```

```
pC <- ggplot(data = opt, aes(y = OR_value, x = reorder(rownames(opt), OR_value))) +
  geom_col(aes(fill = OR_value), col = "lightblue") +
  scale_fill_gradient(low = "white", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "OR Value") +
  ggtitle("Hypertension (binary classification) related factors (p<0.05)") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 37)) +
  geom_text(aes(label = round(OR_value, 3)),
            vjust = 0.1, hjust = -0.5,
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pC
```




```
ggarrange(pA, pB,pC,
          labels = c("A","B","C"), # 添加标签
          ncol = 3, nrow = 1,
          font.label = list(size = 14, face = "bold")) # 设置标签字体样式
```


A
Factors relat
Correlation
B
Diastolic blc
Correlation
C
Hypertensio
BMXBMI
AGE
BMXWT
LBXTC
LBDTCSI
DMDHRAGZ
DMDHHSZE
SLQ030
SLQ040
OCD150
RIDRETH3NHB
RIAGENDR
DRD360
DMDHRMAZWDS
SLQ050
DMDCITZN
SMQ020
DMDEDUC
RIDRETH3MA
SLQ120
DMDHRMAZNM
INDFMIN2
PAQ665
INDHHIN2
DROSPREP
DMDHREDZ
SLD012
DBQ095Z
OCD390G
DMDHHSZB
DMDFMSIZ
PAQ650
DR1_320Z
DR1_300
DR1TCALC
SLD013
DMDHHSIZ
BMXLEG
DMDHHSZA
DR1TATOC
DR1_330Z
WTDRD1
WTDR2D
OR Value

6 选择变量

6.1 特征筛选

6.1.1 合并上述相关性计算的有效变量

```
namesc <- unique(c(rownames(cor_D),
                   rownames(cor_S),
                   row.names(opt)))
namesc
```

```
##  [1] "BMXWT"       "LBXTC"       "LBDTCSI"     "BMXHT"       "RIAGENDR"
##  [6] "BMXBMI"      "SLQ030"      "RIDRETH3NHB" "SLQ040"      "BMXLEG"
## [11] "ALQ121"      "DR1TKCAL"    "DMDEDUC"     "DR1TPROT"    "RIDRETH3NHA"
## [16] "DR1TCHL"     "DR1TCARB"    "DR1TP204"    "DR1TCHOL"    "DR1TALCO"
## [21] "PAQ605"      "DR1TMOIS"    "DMDHHSZB"    "DR1TS180"    "DR1TSODI"
## [26] "ALQ111"      "DRD360"      "DR1BWATZ"    "AGE"         "DR1TSUGR"
## [31] "DR1TNIAC"    "DR1TFF"      "DR1TM161"    "DR1TSFAT"    "DR1TS160"
## [36] "DMDHRGND"    "DR1TS120"    "DR1TMFAT"    "DR1TTFAT"    "OCD390G"
## [41] "DR1TM181"    "INDFMIN2"    "DR1TPHOS"    "DR1TZINC"    "INDHHIN2"
## [46] "DRQSDIET"    "DPQ050"      "DMDHHSZA"    "DR1TRET"     "DMDHRMAZWDS"
## [51] "DR1_300"     "DBQ095Z"     "SLD013"      "PAQ650"      "SLQ120"
## [56] "DMDBORN4"    "DPQ040"      "LBDHDDSI"    "LBDHDD"      "DMDHHSZE"
## [61] "SLD012"      "RIDRETH3NHW" "OCD150"      "DMDHRAGZ"    "SMQ020"
## [66] "DMDCITZN"    "SLQ050"      "DR1TB12A"    "DR1_330Z"    "DMDHRMAZNM"
## [71] "RIDRETH3MA"  "DR1TVARA"    "DR1TVB1"     "DR1TFA"      "DR1TS080"
## [76] "DR1TMAGN"    "DR1TIRON"    "DR1TFOLA"    "DR1TFDFE"    "DR1TLYCO"
## [81] "DR1_320Z"    "DR1TS040"    "DR1TATOC"    "DR1TS060"    "DR1TVB2"
## [86] "DR1TS100"    "DR1TS140"    "PAQ620"      "DR1TCALC"    "PAQ665"
## [91] "DRQSPREP"    "DMDHREDZ"    "WTDRD1"      "WTDR2D"      "DMDFMSIZ"
## [96] "DMDHHSIZ"
```

```
Data_SD2 <- Data_SD[c("response", namesc)]
X <- Data_SD2[-1]
Y <- Data_SD2[, 1]
```

6.1.2 计算皮尔逊相关

```
library(FeatureSelection)
Featurepearson <- func_correlation(
  data = Data_SD2,
```

```
  target = "response",
  use_obs = "all.obs",
  correlation_thresh = 0.001,
  correlation_method = "pearson"
)
Featurepearson
```

```
##                 response
## AGE          0.355938709
## DMDHRAGZ     0.243454132
## DMDHHSZE     0.214914074
## BMXBMI       0.142271320
## BMXWT        0.135595445
## LBXTC        0.116209357
## LBDTCSI      0.116187276
## OCD150       0.110862022
## SLQ030       0.110840612
## RIDRETH3NHB 0.085781203
## RIAGENDR     0.074144752
## SLQ040       0.068196162
## DRD360       0.066596042
## SLQ050       0.057264346
## DMDHRMAZWDS 0.054687344
## SMQ020       0.045866003
## DMDCITZN     0.035397160
## DR1TS120     0.025810225
## DR1TCHOL     0.019210230
## DR1TCHL      0.017532844
## DR1TS080     0.017187118
## DMDHRGND     0.016879421
## DR1TP204     0.011527295
## DR1BWATZ     0.008295425
## DR1TS180     0.005334887
## BMXHT        0.005299287
## DR1TALCO     0.003353245
## DRQSDIET     0.001540367
```

6.1.3 基于集成学习选择变量

```
params_glmnet <- list(
  alpha = 1,
```

```
  family = "gaussian",
  nfolds = 3,
  parallel = TRUE
)
params_xgboost <- list(
  params = list(
    "objective" = "reg:linear",
    "bst:eta" = 0.001,
    "subsample" = 0.75,
    "max_depth" = 5,
    "colsample_bytree" = 0.75,
    "nthread" = 6
  ),
  nrounds = 1000,
  print.every.n = 250,
  maximize = FALSE
)



params_ranger <- list(
  dependent.variable.name = "y",
  probability = FALSE, num.trees = 1000,
  verbose = TRUE, mtry = 5,
  min.node.size = 10, num.threads = 6,
  classification = FALSE, importance = "permutation"
)



params_features <- list(keep_number_feat = NULL, union = TRUE)
```

```
params_barplot <- list(keep_features = 96,
                       horiz = TRUE,
                       cex.names = 1.0)
barplot_feat_select(feat, params_barplot, xgb_sort = "Cover")
```

```
feat_lasso <- cbind.data.frame(Feature = feat$all_feat$`glmnet-lasso`$Feature,
                               coef = feat$all_feat$`glmnet-lasso`$coefficients)
feat_xgboost <- cbind.data.frame(Feature = feat$all_feat$xgboost$Feature,
                               coef = feat$all_feat$xgboost$Cover)
feat_ranger <- cbind.data.frame(Feature = feat$all_feat$ranger$Feature,
                               coef = feat$all_feat$ranger$permutation)
feat_union <- cbind.data.frame(Feature = feat$union_feat$feature,
```

```
                                coef = feat$union_feat$importance)
```

```
plasso <- ggplot(data = feat_lasso, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Glmnet Lasso") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(-0.19, 0.9)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_lasso$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
```

```
feat_xgboost <- feat_xgboost %>%
  dplyr::filter(coef > 0.005)
pxgb <- ggplot(data = feat_xgboost, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "white", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
```

```
                                  vjust = 1,
                                  size = 12,
                                  hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("XGBoost") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 0.1)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_xgboost$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                  colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                  size = 10,
                                  face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
```

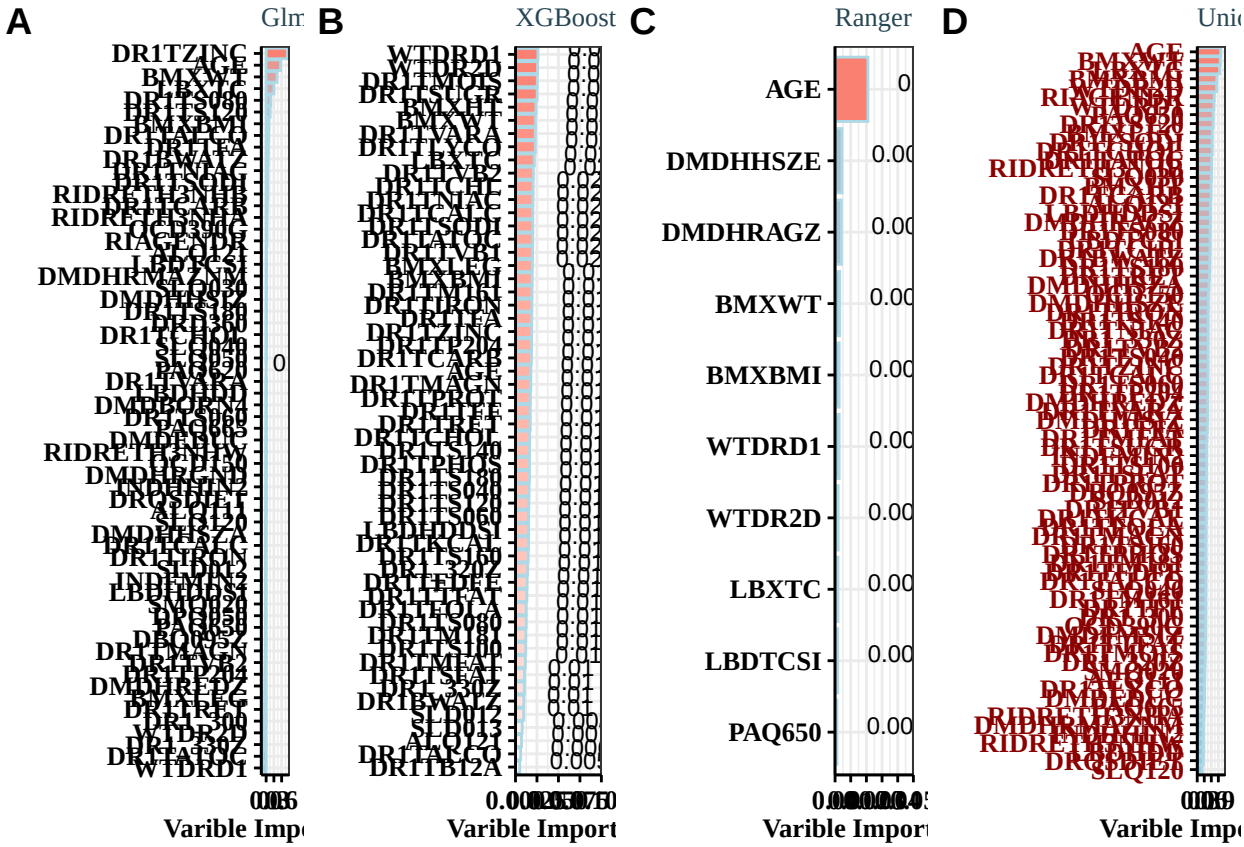
```
feat_ranger <- feat_ranger %>% dplyr::filter(coef > 0.001)
pranger <- ggplot(data = feat_ranger, aes(y = coef,
                                          x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                  vjust = 1,
                                  size = 12,
                                  hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Ranger") +
  scale_y_continuous(expand = c(0, 0),
```

```
                    limits = c(-0.0001, 0.05)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_ranger$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
```

```
feat_union <- feat_union %>%
  dplyr::filter(coef > 0.2)
punion <- ggplot(data = feat_union, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Union") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 1.2)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_union$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
```

```
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          axis.text.y = element_text(colour = 'darkred'),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
```

```
ggarrange(plasso, pxgb,pranger,punion,
          labels = c("A","B","C",'D'), # 添加标签
          ncol = 4, nrow = 1,
          font.label = list(size = 14, face = "bold")) # 设置标签字体样式
```



```
names_ts <- data.frame(
  imp = feat$union_feat$importance,
  feat = feat$union_feat$feature
)
names_fin <- names_ts %>%
  dplyr::filter(imp > 0.5)
names_final <- names_fin[, 2]
```

6.1.4 手动去除强共线性

```
match(names_final, colnames(Data_SD2))
```

```
##  [1] 30  2  3  7 95  6 94 55 38 11 26 20 84  9  8  5 18 12 59 65 76  4 17 29 25
## [26] 50
```

```
Data3 <- data.frame(
  response = Y,
  Data_SD2[, names_final]
)
```

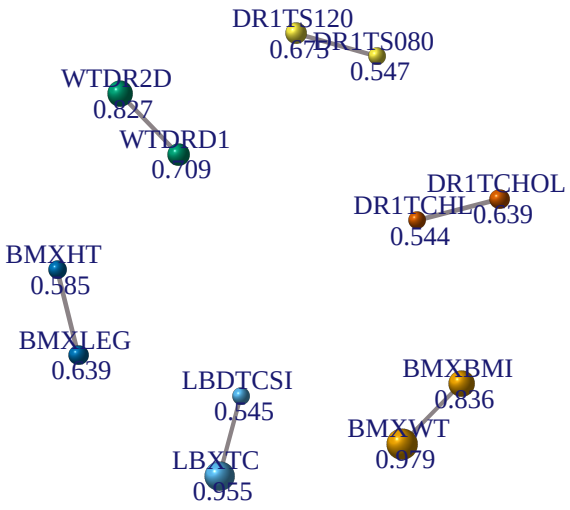
```
# Calculate the correlation matrix
cor_matrix <- cor(Data3)
```

```
# Find the pairs of variables with correlation greater than 0.75
high_cor_pairs0 <- which(cor_matrix > 0.75 & cor_matrix != 1,
                         arr.ind = TRUE)
# Create adjacency matrix of highly correlated variables
adj_matrix0 <- cor_matrix[unique(high_cor_pairs0[, 1]),
                          unique(high_cor_pairs0[, 1])]
for (i in 1:length(adj_matrix0)) {
  if (adj_matrix0[i] < 0.75 | adj_matrix0[i] == 1) {
    adj_matrix0[i] <- 0
  }
}
```

```
set.seed(1234)
g <- graph.adjacency(adj_matrix0, mode = "undirected", weighted = TRUE)
clusters <- cluster_louvain(g)
p <- plot(g, layout = layout_with_fr(g, area = nrow(adj_matrix0)^2),
     edge.color = '#8B8386', edge.width = E(g)/15+2,
```

```
    vertex.size = feat$union_feat$importance[match(colnames(adj_matrix0),
                                    feat$union_feat$feature)]*14, vertex.label = paste(
  colnames(adj_matrix0), "\n",
  feat$union_feat$importance[match(colnames(adj_matrix0),
                                    feat$union_feat$feature)] %>% round(3)
), vertex.label.cex = 0.8, vertex.label.dist = 0, vertex.label.color = "#191970",
edge.label.family = 'Times',
vertex.color = clusters$membership,
vertex.shape = 'sphere',
frame = T)
```

6.1.4.1 共线性图




```
names_ex <- c(
  'DR1TS080','WTDRD1','BMXHT','LBDTCSI','BMXBMI','DR1TCHL'
)

Data5 <- Data3[-match(names_ex, colnames(Data3))]
```

6.1.4.2 相关性大于 0.85

6.1.4.3 相关性大于 0.75

由于去除相关性 >0.85 的变量后，模型仍存在共线性。

6.2 过采样平衡结局变量

```
names_ex <- c(
  'DR1TS080','WTDRD1','BMXHT','LBDTCSI','BMXBMI','DR1TCHL'
)

Data5 <- Data3[-match(names_ex, colnames(Data3))]
set.seed(1234)
Data5 <- Data5 %>%
  mutate(response = factor(response, levels = c(0, 1)))
Data6 <- DMwR::SMOTE(response ~ ., Data5, perc.over = 100)



set.seed(1234)
# 使用 downSample 函数进行负采样
Data_ds <- caret::downSample(x = Data5[, -1], y = Data5$response, yname = "response", list = FALSE)
Data_ds <- Data_ds %>%
  dplyr::select(response,everything())
```

```
table(Data5$response)
```

```
##
##    0    1
## 2494 1841
```

```
table(Data6$response)
```

```
##
##    0    1
## 3682 3682
```

```
table(Data_ds$response)
```

```
##
##    0    1
## 1841 1841
```

6.3 交互作用

```r
glm.fit <- glm(response ~ .,
  data = Data6,
  family = binomial(link = "logit")
)
```

```r
summary(glm.fit)
```

```
##
## Call:
## glm(formula = response ~ ., family = binomial(link = "logit"),
##     data = Data6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2449  -0.9685   0.1005   0.9563   2.4939
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.340642   0.199935 -11.707  < 2e-16 ***
## AGE          2.892937   0.157600  18.356  < 2e-16 ***
## BMXWT        2.677805   0.259259  10.329  < 2e-16 ***
## LBXTC        2.163160   0.252000   8.584  < 2e-16 ***
## WTDR2D      -1.620922   0.368247  -4.402 1.07e-05 ***
## RIAGENDR     0.380387   0.073381   5.184 2.17e-07 ***
## PAQ650      -0.306036   0.065061  -4.704 2.55e-06 ***
## DR1TS120     1.688981   0.782198   2.159  0.03083 *
## BMXLEG      -1.316264   0.269451  -4.885 1.03e-06 ***
## DR1TSODI     0.202581   0.575714   0.352  0.72493
## DR1TCHOL    -0.062298   0.341545  -0.182  0.85527
## DR1TATOC    -2.959167   0.740623  -3.996 6.46e-05 ***
## RIDRETH3NHB  0.413014   0.070072   5.894 3.77e-09 ***
## SLQ030       0.006471   0.071083   0.091  0.92746
## DR1TCARB     0.989508   0.409233   2.418  0.01561 *
## ALQ121       0.262234   0.093491   2.805  0.00503 **
## LBDHDDSI    -0.385518   0.351949  -1.095  0.27335
## DMDHRAGZ    -0.197996   0.149058  -1.328  0.18407
## DR1BWATZ     0.862139   0.288748   2.986  0.00283 **
## DR1TS180     0.746903   0.405824   1.840  0.06570 .
## DR1TRET     -1.637694   0.565709  -2.895  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10208.7  on 7363  degrees of freedom
## Residual deviance:  8663.4  on 7343  degrees of freedom
## AIC: 8705.4
##
## Number of Fisher Scoring iterations: 4
```

```
# create a new model with interaction terms
glm.fit_int <- glm(response ~ . + (AGE + BMXWT+ LBXTC + WTDR2D + RIAGENDR + PAQ650 + DR1TS120 + BMXLEG + D
  data = Data6,
  family = binomial(link = "logit")
)
```

```
summary(glm.fit_int)
```

```
##
## Call:
## glm(formula = response ~ . + (AGE + BMXWT + LBXTC + WTDR2D +
##     RIAGENDR + PAQ650 + DR1TS120 + BMXLEG + DR1TATOC + RIDRETH3NHB +
##     DR1TCARB + ALQ121 + DR1BWATZ + DR1TRET) * (AGE + BMXWT +
##     LBXTC + WTDR2D + RIAGENDR + PAQ650 + DR1TS120 + BMXLEG +
##     DR1TATOC + RIDRETH3NHB + DR1TCARB + ALQ121 + DR1BWATZ + DR1TRET),
##     family = binomial(link = "logit"), data = Data6)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39311  -0.94837   0.07241   0.92599   2.65998
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.288190   0.806674  -4.076 4.58e-05 ***
## AGE                   4.979154   0.666871   7.466 8.24e-14 ***
## BMXWT                 5.137970   1.489237   3.450 0.000560 ***
## LBXTC                 3.300739   1.641900   2.010 0.044398 *
## WTDR2D               -3.792343   2.913247  -1.302 0.192999
## RIAGENDR              1.452098   0.510496   2.844 0.004448 **
## PAQ650               -1.479028   0.452374  -3.269 0.001077 **
## DR1TS120             -2.700219   6.801412  -0.397 0.691361
## BMXLEG               -1.560711   1.569548  -0.994 0.320043
## DR1TSODI              0.171927   0.618498   0.278 0.781032
## DR1TCHOL             -0.096527   0.360070  -0.268 0.788639
```

```
## DR1TATOC             -0.013985   4.685374  -0.003 0.997619
## RIDRETH3NHB           0.831887   0.502431   1.656 0.097778 .
## SLQ030                0.007617   0.074287   0.103 0.918336
## DR1TCARB             -5.630564   2.478774  -2.272 0.023116 *
## ALQ121               -0.747796   0.662394  -1.129 0.258927
## LBDHDDSI             -0.504551   0.377561  -1.336 0.181437
## DMDHRAGZ             -0.179074   0.159695  -1.121 0.262139
## DR1BWATZ              0.566107   2.024537   0.280 0.779767
## DR1TS180              1.096945   0.437461   2.508 0.012158 *
## DR1TRET              22.279256   3.939551   5.655 1.56e-08 ***
## AGE:BMXWT            -7.597307   1.086434  -6.993 2.69e-12 ***
## AGE:LBXTC            -2.136734   1.151123  -1.856 0.063423 .
## AGE:WTDR2D           -4.214178   1.822266  -2.313 0.020744 *
## AGE:RIAGENDR         -2.931103   0.309696  -9.464  < 2e-16 ***
## AGE:PAQ650            0.251250   0.279037   0.900 0.367897
## AGE:DR1TS120         -1.813720   4.062054  -0.447 0.655234
## AGE:BMXLEG            4.357222   1.154599   3.774 0.000161 ***
## AGE:DR1TATOC         -1.625637   3.069601  -0.530 0.596395
## AGE:RIDRETH3NHB      -0.834951   0.333464  -2.504 0.012284 *
## AGE:DR1TCARB          5.287795   1.583667   3.339 0.000841 ***
## AGE:ALQ121            0.379928   0.406256   0.935 0.349689
## AGE:DR1BWATZ          2.035055   1.382836   1.472 0.141115
## AGE:DR1TRET         -12.383697   2.416847  -5.124 2.99e-07 ***
## BMXWT:LBXTC           0.912470   2.541394   0.359 0.719563
## BMXWT:WTDR2D          7.879018   3.946215   1.997 0.045869 *
## BMXWT:RIAGENDR       -0.127026   0.674119  -0.188 0.850538
## BMXWT:PAQ650         -0.306742   0.675367  -0.454 0.649696
## BMXWT:DR1TS120       -3.307803   8.607116  -0.384 0.700748
## BMXWT:BMXLEG          1.884150   2.311863   0.815 0.415077
## BMXWT:DR1TATOC      -16.754269   6.695925  -2.502 0.012344 *
## BMXWT:RIDRETH3NHB     0.220348   0.635174   0.347 0.728660
## BMXWT:DR1TCARB        3.543327   3.241961   1.093 0.274412
## BMXWT:ALQ121         -2.156622   0.932066  -2.314 0.020678 *
## BMXWT:DR1BWATZ        0.494248   2.517749   0.196 0.844371
## BMXWT:DR1TRET        14.502100   6.156508   2.356 0.018494 *
## LBXTC:WTDR2D          6.734199   4.156929   1.620 0.105234
## LBXTC:RIAGENDR        1.377444   0.714800   1.927 0.053975 .
## LBXTC:PAQ650         -0.431625   0.651699  -0.662 0.507774
## LBXTC:DR1TS120       -1.464489  10.330152  -0.142 0.887263
## LBXTC:BMXLEG         -5.831493   2.648810  -2.202 0.027697 *
## LBXTC:DR1TATOC        7.719012   6.730131   1.147 0.251409
```

```
## LBXTC:RIDRETH3NHB      1.940953   0.727583   2.668 0.007638 **
## LBXTC:DR1TCARB        -1.350577   3.654210  -0.370 0.711684
## LBXTC:ALQ121           1.738998   0.907714   1.916 0.055391 .
## LBXTC:DR1BWATZ        -0.948110   2.899322  -0.327 0.743659
## LBXTC:DR1TRET         -0.584296   5.416785  -0.108 0.914101
## WTDR2D:RIAGENDR        3.639289   1.109624   3.280 0.001039 **
## WTDR2D:PAQ650         -2.909426   0.970858  -2.997 0.002729 **
## WTDR2D:DR1TS120       -3.489925  13.740443  -0.254 0.799504
## WTDR2D:BMXLEG         -4.273589   4.120792  -1.037 0.299699
## WTDR2D:DR1TATOC       12.631086   9.371235   1.348 0.177704
## WTDR2D:RIDRETH3NHB    -4.273564   2.716515  -1.573 0.115677
## WTDR2D:DR1TCARB       -9.749714   5.029413  -1.939 0.052557 .
## WTDR2D:ALQ121          2.205960   1.343048   1.643 0.100486
## WTDR2D:DR1BWATZ       10.757887   4.452933   2.416 0.015696 *
## WTDR2D:DR1TRET         5.110234   8.688990   0.588 0.556447
## RIAGENDR:PAQ650       -0.204630   0.180804  -1.132 0.257729
## RIAGENDR:DR1TS120     -7.210777   2.617320  -2.755 0.005869 **
## RIAGENDR:BMXLEG       -0.906939   0.592708  -1.530 0.125977
## RIAGENDR:DR1TATOC      1.542633   1.969480   0.783 0.433469
## RIAGENDR:RIDRETH3NHB   0.035425   0.191956   0.185 0.853586
## RIAGENDR:DR1TCARB      0.953547   1.033085   0.923 0.356002
## RIAGENDR:ALQ121       -0.058141   0.259976  -0.224 0.823037
## RIAGENDR:DR1BWATZ     -0.681414   0.854374  -0.798 0.425126
## RIAGENDR:DR1TRET       7.029967   1.738013   4.045 5.24e-05 ***
## PAQ650:DR1TS120        0.344489   2.199524   0.157 0.875545
## PAQ650:BMXLEG          2.255079   0.702347   3.211 0.001324 **
## PAQ650:DR1TATOC       -3.240801   1.705711  -1.900 0.057437 .
## PAQ650:RIDRETH3NHB    -0.374450   0.179712  -2.084 0.037196 *
## PAQ650:DR1TCARB        1.497227   0.908289   1.648 0.099270 .
## PAQ650:ALQ121          0.422591   0.246842   1.712 0.086898 .
## PAQ650:DR1BWATZ        0.713155   0.737168   0.967 0.333332
## PAQ650:DR1TRET         2.908535   1.469997   1.979 0.047861 *
## DR1TS120:BMXLEG       24.208939   9.759893   2.480 0.013122 *
## DR1TS120:DR1TATOC    -20.340827  18.659455  -1.090 0.275665
## DR1TS120:RIDRETH3NHB  -2.720973   2.656056  -1.024 0.305627
## DR1TS120:DR1TCARB     -0.026618   6.986100  -0.004 0.996960
## DR1TS120:ALQ121        1.668030   3.618515   0.461 0.644820
## DR1TS120:DR1BWATZ     -9.741829  12.481867  -0.780 0.435109
## DR1TS120:DR1TRET      10.955888  12.775453   0.858 0.391128
## BMXLEG:DR1TATOC        1.020790   7.604710   0.134 0.893220
## BMXLEG:RIDRETH3NHB    -0.499492   0.664229  -0.752 0.452059
```

```
## BMXLEG:DR1TCARB        7.710742   3.817321   2.020 0.043390 *
## BMXLEG:ALQ121          1.860763   0.980980   1.897 0.057849 .
## BMXLEG:DR1BWATZ       -2.550466   2.834764  -0.900 0.368274
## BMXLEG:DR1TRET       -51.652123   7.235108  -7.139 9.39e-13 ***
## DR1TATOC:RIDRETH3NHB   5.880033   1.940984   3.029 0.002450 **
## DR1TATOC:DR1TCARB     -1.061285   5.402641  -0.196 0.844267
## DR1TATOC:ALQ121      -10.748001   2.646718  -4.061 4.89e-05 ***
## DR1TATOC:DR1BWATZ     10.141878   8.026779   1.264 0.206408
## DR1TATOC:DR1TRET      16.668424  10.727004   1.554 0.120214
## RIDRETH3NHB:DR1TCARB  -2.473597   0.988560  -2.502 0.012342 *
## RIDRETH3NHB:ALQ121    -0.458388   0.249800  -1.835 0.066503 .
## RIDRETH3NHB:DR1BWATZ  -0.445103   0.847513  -0.525 0.599453
## RIDRETH3NHB:DR1TRET    3.831790   1.584755   2.418 0.015610 *
## DR1TCARB:ALQ121        1.279979   1.292442   0.990 0.322000
## DR1TCARB:DR1BWATZ     -6.473824   3.796777  -1.705 0.088179 .
## DR1TCARB:DR1TRET     -10.206569   5.567008  -1.833 0.066743 .
## ALQ121:DR1BWATZ        3.885199   1.141159   3.405 0.000663 ***
## ALQ121:DR1TRET         1.025916   1.932678   0.531 0.595539
## DR1BWATZ:DR1TRET      -6.735771   6.427084  -1.048 0.294625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10208.7  on 7363  degrees of freedom
## Residual deviance:  8213.1  on 7252  degrees of freedom
## AIC: 8437.1
##
## Number of Fisher Scoring iterations: 5
```

```
glm.fit_int <- glm(
  response ~ . +
    AGE:BMXWT +
    AGE:WTDR2D +
    AGE:RIAGENDR +
    AGE:BMXLEG +
    AGE:DR1TCARB +
    AGE:DR1TRET +
    BMXWT:DR1TATOC +
    BMXWT:ALQ121  +
    BMXWT:DR1TRET +
    WTDR2D:RIAGENDR +
```

```
    WTDR2D:PAQ650 +
    WTDR2D:DR1BWATZ +
    RIAGENDR:DR1TS120 +
    RIAGENDR:DR1TRET +
    PAQ650:BMXLEG +
    PAQ650:DR1TRET +
    DR1TS120:BMXLEG +
    BMXLEG:DR1TCARB +
    BMXLEG:DR1TRET +
    DR1TATOC:RIDRETH3NHB +
    DR1TATOC:ALQ121 +
    RIDRETH3NHB:DR1TCARB +
    RIDRETH3NHB:DR1TRET +
    ALQ121:DR1BWATZ,
  data = Data6,
  family = binomial(link = "logit")
)
```

```
summary(glm.fit_int)
```

```
##
## Call:
## glm(formula = response ~ . + AGE:BMXWT + AGE:WTDR2D + AGE:RIAGENDR +
##     AGE:BMXLEG + AGE:DR1TCARB + AGE:DR1TRET + BMXWT:DR1TATOC +
##     BMXWT:ALQ121 + BMXWT:DR1TRET + WTDR2D:RIAGENDR + WTDR2D:PAQ650 +
##     WTDR2D:DR1BWATZ + RIAGENDR:DR1TS120 + RIAGENDR:DR1TRET +
##     PAQ650:BMXLEG + PAQ650:DR1TRET + DR1TS120:BMXLEG + BMXLEG:DR1TCARB +
##     BMXLEG:DR1TRET + DR1TATOC:RIDRETH3NHB + DR1TATOC:ALQ121 +
##     RIDRETH3NHB:DR1TCARB + RIDRETH3NHB:DR1TRET + ALQ121:DR1BWATZ,
##     family = binomial(link = "logit"), data = Data6)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.41927  -0.95874   0.03838   0.93624   2.71602
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.874445   0.459077  -6.261 3.82e-10 ***
## AGE                   4.292591   0.521349   8.234  < 2e-16 ***
## BMXWT                 6.899319   0.735456   9.381  < 2e-16 ***
## LBXTC                 1.641931   0.262820   6.247 4.17e-10 ***
## WTDR2D               -2.068900   1.232998  -1.678 0.093358 .
```

```
## RIAGENDR               1.438217   0.188438   7.632 2.31e-14 ***
## PAQ650                -1.117892   0.273862  -4.082 4.47e-05 ***
## DR1TS120              -4.724996   3.649713  -1.295 0.195451
## BMXLEG                -3.265718   0.817512  -3.995 6.48e-05 ***
## DR1TSODI               0.336304   0.598646   0.562 0.574270
## DR1TCHOL              -0.119249   0.350043  -0.341 0.733353
## DR1TATOC               1.518096   1.860596   0.816 0.414546
## RIDRETH3NHB            0.444590   0.149245   2.979 0.002893 **
## SLQ030                -0.006754   0.072969  -0.093 0.926258
## DR1TCARB              -6.859332   1.733562  -3.957 7.60e-05 ***
## ALQ121                 0.997473   0.272760   3.657 0.000255 ***
## LBDHDDSI              -0.430176   0.366459  -1.174 0.240446
## DMDHRAGZ              -0.168315   0.155506  -1.082 0.279089
## DR1BWATZ              -0.847638   0.512470  -1.654 0.098123 .
## DR1TS180               0.992033   0.420886   2.357 0.018423 *
## DR1TRET               19.695528   3.300666   5.967 2.41e-09 ***
## AGE:BMXWT             -7.689665   0.972194  -7.910 2.58e-15 ***
## AGE:WTDR2D            -2.721259   1.605269  -1.695 0.090037 .
## AGE:RIAGENDR          -2.658604   0.265710 -10.006  < 2e-16 ***
## AGE:BMXLEG             3.927557   1.023557   3.837 0.000124 ***
## AGE:DR1TCARB           5.290942   1.326482   3.989 6.64e-05 ***
## AGE:DR1TRET          -11.418738   2.155312  -5.298 1.17e-07 ***
## BMXWT:DR1TATOC       -11.427764   5.839869  -1.957 0.050365 .
## BMXWT:ALQ121          -1.688631   0.836824  -2.018 0.043601 *
## BMXWT:DR1TRET         15.888327   5.516527   2.880 0.003975 **
## WTDR2D:RIAGENDR        3.033012   0.770120   3.938 8.20e-05 ***
## WTDR2D:PAQ650         -2.093020   0.874124  -2.394 0.016647 *
## WTDR2D:DR1BWATZ        9.580427   4.094084   2.340 0.019280 *
## RIAGENDR:DR1TS120     -6.912963   2.255976  -3.064 0.002182 **
## RIAGENDR:DR1TRET       7.015371   1.494841   4.693 2.69e-06 ***
## PAQ650:BMXLEG          1.388223   0.490036   2.833 0.004613 **
## PAQ650:DR1TRET         2.956654   1.213941   2.436 0.014868 *
## DR1TS120:BMXLEG       19.954181   8.255079   2.417 0.015640 *
## BMXLEG:DR1TCARB       10.433415   2.576426   4.050 5.13e-05 ***
## BMXLEG:DR1TRET       -48.918075   6.550224  -7.468 8.13e-14 ***
## DR1TATOC:RIDRETH3NHB   3.096261   1.552677   1.994 0.046136 *
## DR1TATOC:ALQ121       -6.605616   2.072774  -3.187 0.001438 **
## RIDRETH3NHB:DR1TCARB  -2.134542   0.855400  -2.495 0.012582 *
## RIDRETH3NHB:DR1TRET    3.478084   1.356558   2.564 0.010350 *
## ALQ121:DR1BWATZ        2.740719   1.003869   2.730 0.006330 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10208.7  on 7363  degrees of freedom
## Residual deviance:  8298.5  on 7319  degrees of freedom
## AIC: 8388.5
##
## Number of Fisher Scoring iterations: 4
```

```
# compare the two models using anova()
anova(glm.fit, glm.fit_int, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: response ~ AGE + BMXWT + LBXTC + WTDR2D + RIAGENDR + PAQ650 +
##     DR1TS120 + BMXLEG + DR1TSODI + DR1TCHOL + DR1TATOC + RIDRETH3NHB +
##     SLQ030 + DR1TCARB + ALQ121 + LBDHDDSI + DMDHRAGZ + DR1BWATZ +
##     DR1TS180 + DR1TRET
## Model 2: response ~ AGE + BMXWT + LBXTC + WTDR2D + RIAGENDR + PAQ650 +
##     DR1TS120 + BMXLEG + DR1TSODI + DR1TCHOL + DR1TATOC + RIDRETH3NHB +
##     SLQ030 + DR1TCARB + ALQ121 + LBDHDDSI + DMDHRAGZ + DR1BWATZ +
##     DR1TS180 + DR1TRET + AGE:BMXWT + AGE:WTDR2D + AGE:RIAGENDR +
##     AGE:BMXLEG + AGE:DR1TCARB + AGE:DR1TRET + BMXWT:DR1TATOC +
##     BMXWT:ALQ121 + BMXWT:DR1TRET + WTDR2D:RIAGENDR + WTDR2D:PAQ650 +
##     WTDR2D:DR1BWATZ + RIAGENDR:DR1TS120 + RIAGENDR:DR1TRET +
##     PAQ650:BMXLEG + PAQ650:DR1TRET + DR1TS120:BMXLEG + BMXLEG:DR1TCARB +
##     BMXLEG:DR1TRET + DR1TATOC:RIDRETH3NHB + DR1TATOC:ALQ121 +
##     RIDRETH3NHB:DR1TCARB + RIDRETH3NHB:DR1TRET + ALQ121:DR1BWATZ
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7343     8663.4
## 2      7319     8298.5 24   364.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.4 最终数据集

```
Data7 <- Data6 %>%
  mutate(
    AGE_BMXWT = AGE*BMXWT,
    AGE_WTDR2D = AGE*WTDR2D,
```

```
    AGE_RIAGENDR = AGE*RIAGENDR,
    AGE_BMXLEG = AGE*BMXLEG,
    AGE_DR1TCARB = AGE*DR1TCARB,
    AGE_DR1TRET = AGE*DR1TRET,
    BMXWT_DR1TATOC = BMXWT*DR1TATOC,
    BMXWT_ALQ121  = BMXWT*ALQ121,
    BMXWT_DR1TRET = BMXWT*DR1TRET,
    WTDR2D_RIAGENDR = WTDR2D*RIAGENDR,
    WTDR2D_PAQ650 = WTDR2D*PAQ650,
    WTDR2D_DR1BWATZ = WTDR2D*DR1BWATZ,
    RIAGENDR_DR1TS120 = RIAGENDR*DR1TS120,
    RIAGENDR_DR1TRET = RIAGENDR*DR1TRET,
    PAQ650_BMXLEG = PAQ650*BMXLEG,
    PAQ650_DR1TRET = PAQ650*DR1TRET,
    DR1TS120_BMXLEG = DR1TS120*BMXLEG,
    BMXLEG_DR1TCARB = BMXLEG*DR1TCARB,
    BMXLEG_DR1TRET = BMXLEG*DR1TRET,
    DR1TATOC_RIDRETH3NHB = DR1TATOC*RIDRETH3NHB,
    DR1TATOC_ALQ121 = DR1TATOC*ALQ121,
    RIDRETH3NHB_DR1TCARB = RIDRETH3NHB*DR1TCARB,
    RIDRETH3NHB_DR1TRET = RIDRETH3NHB*DR1TRET,
    ALQ121_DR1BWATZ = ALQ121*DR1BWATZ
  )
Data_ds <- Data_ds %>%
  mutate(
    AGE_BMXWT = AGE*BMXWT,
    AGE_WTDR2D = AGE*WTDR2D,
    AGE_RIAGENDR = AGE*RIAGENDR,
    AGE_BMXLEG = AGE*BMXLEG,
    AGE_DR1TCARB = AGE*DR1TCARB,
    AGE_DR1TRET = AGE*DR1TRET,
    BMXWT_DR1TATOC = BMXWT*DR1TATOC,
    BMXWT_ALQ121  = BMXWT*ALQ121,
    BMXWT_DR1TRET = BMXWT*DR1TRET,
    WTDR2D_RIAGENDR = WTDR2D*RIAGENDR,
    WTDR2D_PAQ650 = WTDR2D*PAQ650,
    WTDR2D_DR1BWATZ = WTDR2D*DR1BWATZ,
    RIAGENDR_DR1TS120 = RIAGENDR*DR1TS120,
    RIAGENDR_DR1TRET = RIAGENDR*DR1TRET,
    PAQ650_BMXLEG = PAQ650*BMXLEG,
    PAQ650_DR1TRET = PAQ650*DR1TRET,
```

```
    DR1TS120_BMXLEG = DR1TS120*BMXLEG,
    BMXLEG_DR1TCARB = BMXLEG*DR1TCARB,
    BMXLEG_DR1TRET = BMXLEG*DR1TRET,
    DR1TATOC_RIDRETH3NHB = DR1TATOC*RIDRETH3NHB,
    DR1TATOC_ALQ121 = DR1TATOC*ALQ121,
    RIDRETH3NHB_DR1TCARB = RIDRETH3NHB*DR1TCARB,
    RIDRETH3NHB_DR1TRET = RIDRETH3NHB*DR1TRET,
    ALQ121_DR1BWATZ = ALQ121*DR1BWATZ
  )
```

```
save(Data7,Data_ds,
     file = paste0(getwd(), "/data_use/featureS_old.RData"))
```