CK2

Group 10

2022-11-03

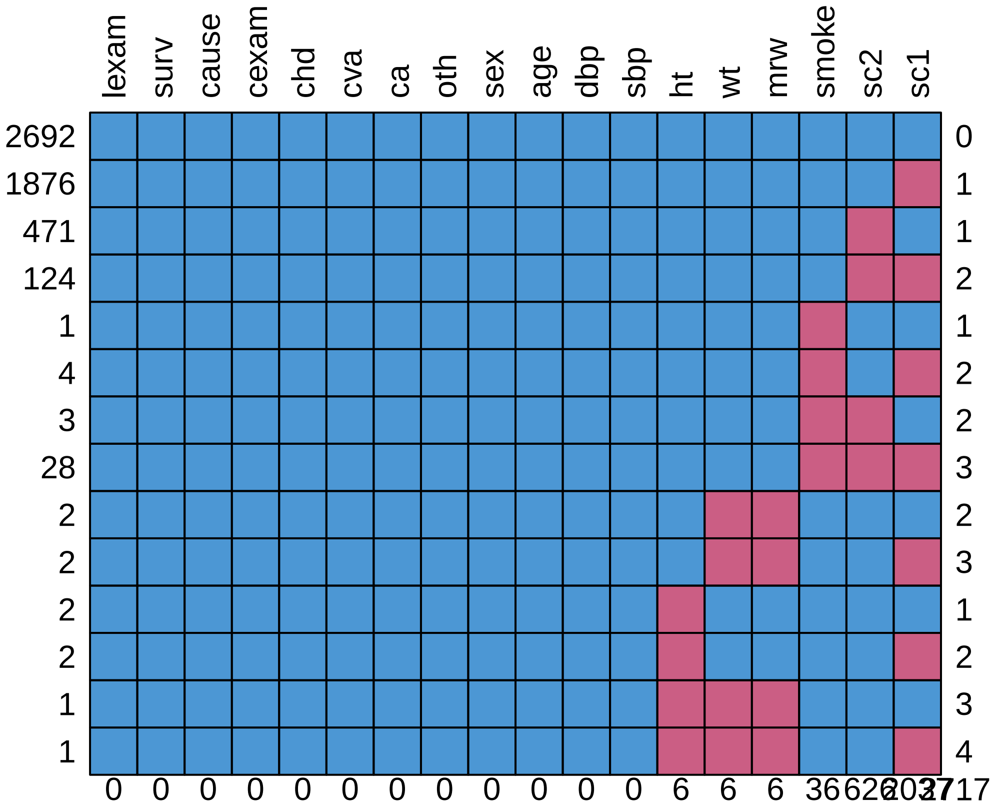
```
library(dplyr)
library(ggplot2)
library(foreign)
library(mice)
data <- read.dta("fram.dta")
```

Q4. Missing Data Analysis

Types of Missingness

From the plot below, we see 14 different patterns of missingness. Patterns with most missing values are **1. missing in sc1**, **2.missing in sc2**, and **3. missing in sc1 and sc2**.

```
md.pattern(data, rotate.names = TRUE)
```



```
##      lexam surv cause cexam chd cva ca oth sex age dbp sbp ht wt mrw smoke sc2
## 2692     1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     1   1
## 1876     1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     1   1
## 471      1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     1   0
## 124      1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     1   0
## 1        1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     0   1
## 4        1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     0   1
## 3        1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     0   0
## 28       1    1     1     1   1   1  1   1   1   1   1   1  1  1   1     0   0
## 2        1    1     1     1   1   1  1   1   1   1   1   1  1  0   0     1   1
## 2        1    1     1     1   1   1  1   1   1   1   1   1  1  0   0     1   1
## 2        1    1     1     1   1   1  1   1   1   1   1   1  0  1   1     1   1
## 2        1    1     1     1   1   1  1   1   1   1   1   1  0  1   1     1   1
## 1        1    1     1     1   1   1  1   1   1   1   1   1  0  0   0     1   1
## 1        1    1     1     1   1   1  1   1   1   1   1   1  0  0   0     1   1
##          0    0     0     0   0   0  0   0   0   0   0   0  6  6   6    36 626
##       sc1
## 2692    1    0
## 1876    0    1
## 471     1    1
## 124     0    2
## 1       1    1
## 4       0    2
## 3       1    2
## 28      0    3
## 2       1    2
## 2       0    3
## 2       1    1
## 2       0    2
## 1       1    3
## 1       0    4
##      2037 2717
```

By fitting the logistic regression of indicator `R1` with the rest of the covariate, we identify significant predictors with p-values less than 0.05. These are `cexam`, `sex`, `dbp`, `smoke`, and `sc2`. Thus we say that missingness in `sc1` is NOT MCAR.

By fitting the logistic regression of indicator `R2` with the rest of the covariate, we identify significant predictors with p-values less than 0.05. These are `Lexam`, `dbp`, and `sbp`. Thus we say that missingness in `sc2` is NOT MCAR.

By fitting the logistic regression of indicator `R3` with the rest of the covariate, we did NOT identify significant predictors with p-values less than 0.05. T= Thus we say that missingness in the (`sc1`, `sc2`) pair is MCAR.

```
### Create indicator variable R1 where R1=1 if sample experience missingness in sc1
data <- data %>% mutate(R1 = if_else(is.na(sc1), 1, 0))
data <- data %>% mutate(R2 = if_else(is.na(sc2), 1, 0))
data <- data %>% mutate(R3 = if_else(R1+R2==2, 1, 0))
### Test for MCAR by logistic regression
logit1 <- glm(R1 ~ lexam + surv + cause + cexam + cva + ca + oth + sex + age + dbp + sbp + ht + wt + mr
summary(logit1)
```

```
##
## Call:
## glm(formula = R1 ~ lexam + surv + cause + cexam + cva + ca +
##     oth + sex + age + dbp + sbp + ht + wt + mrw + smoke + sc2,
##     family = "binomial", data = data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.6299  -1.0342  -0.8277   1.2398   1.9569
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0077235  2.9288414  -0.003 0.997896
## lexam       -0.0195046  0.0144085  -1.354 0.175837
## surv         0.1975997  0.1800836   1.097 0.272525
## cause       -0.1044127  0.0744720  -1.402 0.160904
## cexam        0.0255558  0.0073209   3.491 0.000482 ***
## cva          0.0258111  0.2799328   0.092 0.926535
## ca          -0.1990322  0.1915329  -1.039 0.298733
## oth          0.3257275  0.4075533   0.799 0.424159
## sex          0.6018454  0.1051644   5.723 1.05e-08 ***
## age          0.0045063  0.0044396   1.015 0.310091
## dbp         -0.0237534  0.0041418  -5.735 9.75e-09 ***
## sbp          0.0002085  0.0023192   0.090 0.928365
## ht           0.0010009  0.0460712   0.022 0.982668
## wt           0.0160012  0.0096609   1.656 0.097665 .
## mrw         -0.0188196  0.0120381  -1.563 0.117973
## smoke       -0.0164396  0.0029867  -5.504 3.71e-08 ***
## sc2          0.0026887  0.0007274   3.696 0.000219 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6186.0  on 4567  degrees of freedom
## Residual deviance: 5993.2  on 4551  degrees of freedom
##   (641 observations deleted due to missingness)
## AIC: 6027.2
##
## Number of Fisher Scoring iterations: 4
```

```
logit2 <- glm(R2 ~ lexam + surv + cause + cexam + cva + ca + oth + sex + age + dbp + sbp + ht + wt + mr
summary(logit2)
```

```
##
## Call:
## glm(formula = R2 ~ lexam + surv + cause + cexam + cva + ca +
##     oth + sex + age + dbp + sbp + ht + wt + mrw + smoke + sc1,
##     family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1191  -0.5889  -0.5266  -0.4483   2.4133
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.488881   4.546406   0.547  0.58408
## lexam       -0.103010   0.020639  -4.991 6.01e-07 ***
## surv        -0.525694   0.298507  -1.761  0.07823 .
## cause        0.098808   0.098838   1.000  0.31745
## cexam       -0.023889   0.014509  -1.647  0.09965 .
## cva         -0.077892   0.405591  -0.192  0.84771
## ca          -0.064811   0.281683  -0.230  0.81803
## oth         -0.363254   0.572896  -0.634  0.52604
## sex          0.189586   0.173215   1.095  0.27373
## age         -0.001429   0.007283  -0.196  0.84449
## dbp         -0.016891   0.006164  -2.740  0.00614 **
## sbp          0.011860   0.003368   3.521  0.00043 ***
## ht          -0.045596   0.072156  -0.632  0.52744
## wt           0.003807   0.015291   0.249  0.80338
## mrw         -0.003140   0.018675  -0.168  0.86649
## smoke        0.007306   0.004593   1.591  0.11163
## sc1         -0.001688   0.001221  -1.382  0.16690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2662.1  on 3162  degrees of freedom
## Residual deviance: 2593.0  on 3146  degrees of freedom
##   (2046 observations deleted due to missingness)
## AIC: 2627
##
## Number of Fisher Scoring iterations: 4
```

```r
logit3 <- glm(R3 ~ lexam + surv + cause + cexam + cva + ca + oth + sex + age + dbp + sbp + ht + wt + mrw
summary(logit3)
```

```
##
## Call:
## glm(formula = R3 ~ lexam + surv + cause + cexam + cva + ca +
##     oth + sex + age + dbp + sbp + ht + wt + mrw + smoke, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6472  -0.2433  -0.2051  -0.1745   3.1445
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.254236   8.619116  -1.654   0.0982 .
## lexam        -0.029084   0.039157  -0.743   0.4576
## surv          0.508612   0.550016   0.925   0.3551
## cause        -0.091023   0.234104  -0.389   0.6974
## cexam        -0.003473   0.023085  -0.150   0.8804
## cva           0.036185   0.847738   0.043   0.9660
## ca            0.120798   0.566066   0.213   0.8310
## oth           0.023943   1.254766   0.019   0.9848
## sex           0.550062   0.325974   1.687   0.0915 .
## age          -0.009502   0.012849  -0.740   0.4596
## dbp          -0.013812   0.012401  -1.114   0.2654
## sbp          -0.005703   0.007002  -0.814   0.4154
## ht            0.168112   0.137565   1.222   0.2217
## wt           -0.033006   0.028627  -1.153   0.2489
## mrw           0.053646   0.034256   1.566   0.1173
## smoke         0.003375   0.008831   0.382   0.7024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1169.8  on 5162  degrees of freedom
## Residual deviance: 1141.0  on 5147  degrees of freedom
##   (46 observations deleted due to missingness)
## AIC: 1173
##
## Number of Fisher Scoring iterations: 7
```

To argue for whether missingness in sc1 is MAR or MNAR, we reason with context. sc1 stands for serum cholesterol exam 1, which is the serum cholesterol level of each individual from their first exam. For sc1 to be MNAR, it's missingness has to be dependent on sc1 itself. This means those with higher or lower sc1 may be less or more likely to take the examination. Since we see no logical explanation behind the former statement, we are more inclined to conclude that are data is MAR. Using the same logic, we say missingness insc2 is also MAR.

Accomodation

We first examine the total number of missing values per column. The 5% threshold is 260. We see only missing values in sc1 and sc2 with count greater than the threshold. Thus, for columns with missing value count less than the threshold, one method would be to use the Complete Case Analysis.

```
colSums(is.na(data))
```

```
## lexam  surv cause cexam   chd   cva    ca   oth   sex   age    ht    wt   sc1
##     0     0     0     0     0     0     0     0     0     0     6     6  2037
##   sc2   dbp   sbp   mrw smoke    R1    R2    R3
##   626     0     0     6    36     0     0     0
```

```
sum(is.na(data$sc1)) / nrow(data)
```

```
## [1] 0.3910539
```

```
sum(is.na(data$sc2)) / nrow(data)
```

```
## [1] 0.1201766
```

```
nrow(data) * 0.05
```

```
## [1] 260.45
```

From the pattern plot above, among the 14 patterns of missing data, only the top 4 has significant count of missing values. Hence, we usde Multivariate Imputation with 4 imputations.

To ensure our imputation did not add new information to the data, we drew correlation plots for data before (only include complete cases) and after imputation. We can see from below that the plot looks very similar.

```
library(corrplot)
```

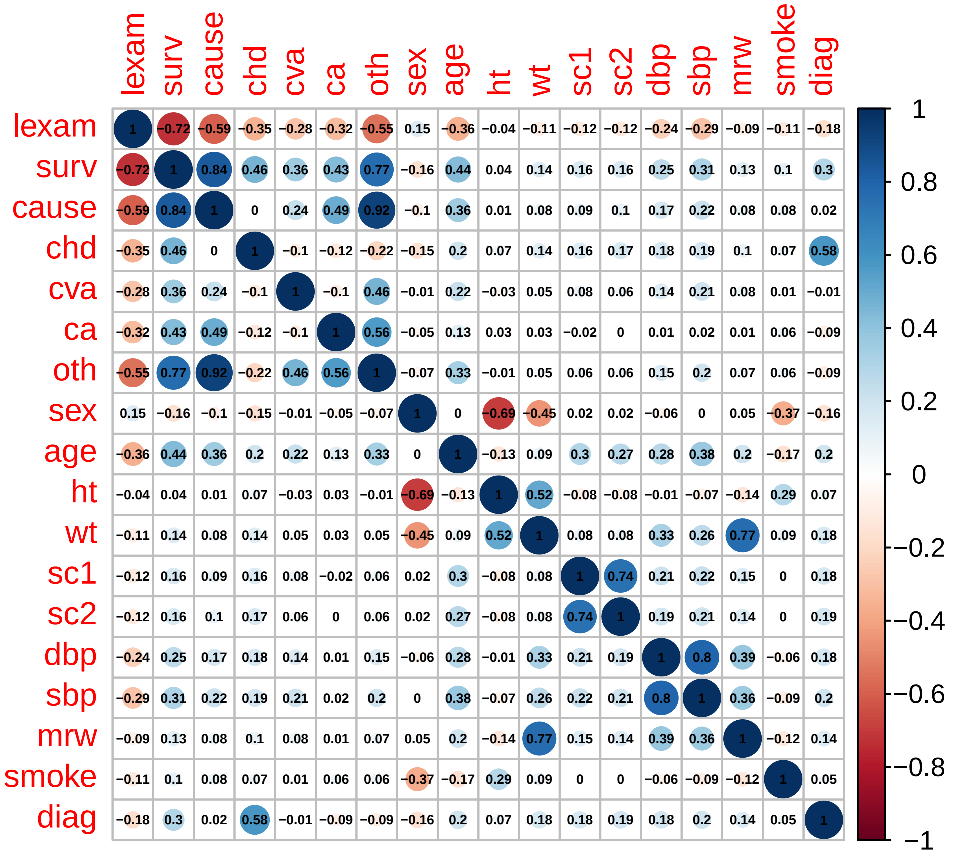
```
## corrplot 0.92 loaded
```

```
data <- data %>% mutate(diag = if_else(cexam == 0, 0, 1))
tempData = mice(data, m=4)
```

```
##
##  iter imp variable
##   1   1  ht  wt  sc1  sc2  mrw  smoke
##   1   2  ht  wt  sc1  sc2  mrw  smoke
##   1   3  ht  wt  sc1  sc2  mrw  smoke
##    1   4  ht  wt  sc1  sc2  mrw  smoke
##    2   1  ht  wt  sc1  sc2  mrw  smoke
##    2   2  ht  wt  sc1  sc2  mrw  smoke
##    2   3  ht  wt  sc1  sc2  mrw  smoke
##    2   4  ht  wt  sc1  sc2  mrw  smoke
##    3   1  ht  wt  sc1  sc2  mrw  smoke
##    3   2  ht  wt  sc1  sc2  mrw  smoke
##    3   3  ht  wt  sc1  sc2  mrw  smoke
##    3   4  ht  wt  sc1  sc2  mrw  smoke
##    4   1  ht  wt  sc1  sc2  mrw  smoke
##    4   2  ht  wt  sc1  sc2  mrw  smoke
##    4   3  ht  wt  sc1  sc2  mrw  smoke
##    4   4  ht  wt  sc1  sc2  mrw  smoke
##    5   1  ht  wt  sc1  sc2  mrw  smoke
##    5   2  ht  wt  sc1  sc2  mrw  smoke
##    5   3  ht  wt  sc1  sc2  mrw  smoke
##    5   4  ht  wt  sc1  sc2  mrw  smoke
```

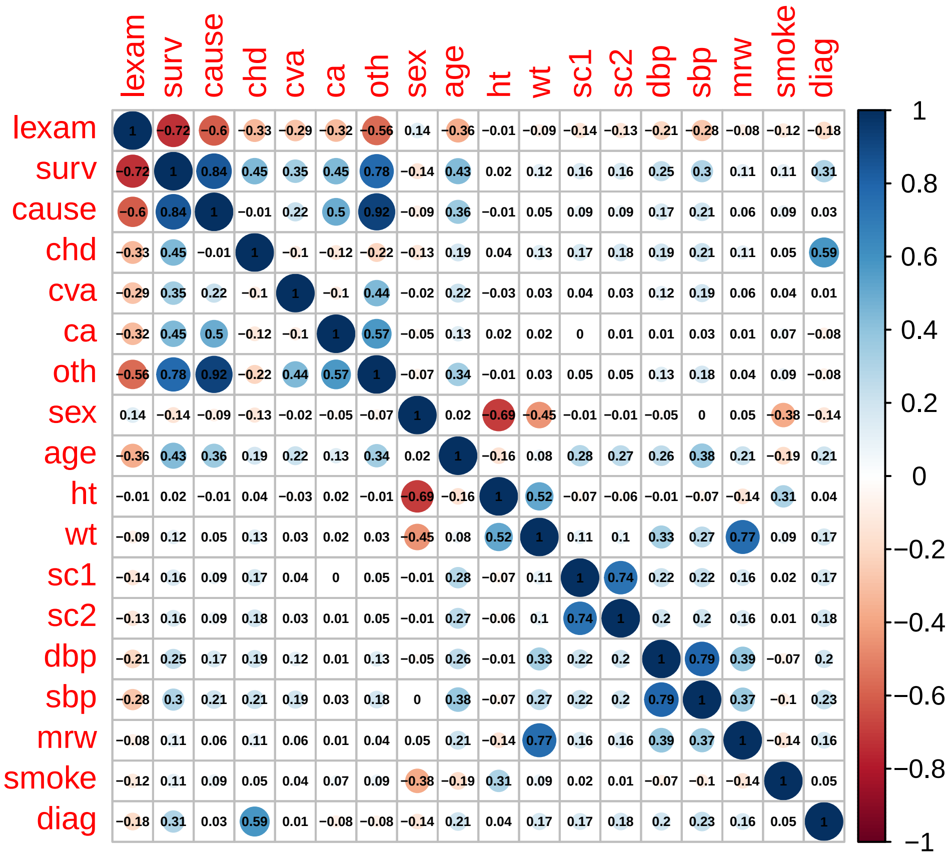
```
## Warning: Number of logged events: 120
```

```
data_imputed <- complete(tempData, action=1)
M = cor(data_imputed[, -c(4, 19, 20, 21)])
corrplot(M, addCoef.col = 'black',  number.cex= 7/(ncol(data) - 4))
```



	lexam	surv	cause	chd	cva	ca	oth	sex	age	ht	wt	sc1	sc2	dbp	sbp	mrw	smoke	diag
lexam	1	−0.72	−0.59	−0.35	−0.28	−0.32	−0.55	0.15	−0.36	−0.04	−0.11	−0.12	−0.12	−0.24	−0.29	−0.09	−0.11	−0.18
surv	−0.72	1	0.84	0.46	0.36	0.43	0.77	−0.16	0.44	0.04	0.14	0.16	0.16	0.25	0.31	0.13	0.1	0.3
cause	−0.59	0.84	1	0	0.24	0.49	0.92	−0.1	0.36	0.01	0.08	0.09	0.1	0.17	0.22	0.08	0.08	0.02
chd	−0.35	0.46	0	1	−0.1	−0.12	−0.22	−0.15	0.2	0.07	0.14	0.16	0.17	0.18	0.19	0.1	0.07	0.58
cva	−0.28	0.36	0.24	−0.1	1	−0.1	0.46	−0.01	0.22	−0.03	0.05	0.08	0.06	0.14	0.21	0.08	0.01	−0.01
ca	−0.32	0.43	0.49	−0.12	−0.1	1	0.56	−0.05	0.13	0.03	0.03	−0.02	0	0.01	0.02	0.01	0.06	−0.09
oth	−0.55	0.77	0.92	−0.22	0.46	0.56	1	−0.07	0.33	−0.01	0.05	0.06	0.06	0.15	0.2	0.07	0.06	−0.09
sex	0.15	−0.16	−0.1	−0.15	−0.01	−0.05	−0.07	1	0	−0.69	−0.45	0.02	0.02	−0.06	0	0.05	−0.37	−0.16
age	−0.36	0.44	0.36	0.2	0.22	0.13	0.33	0	1	−0.13	0.09	0.3	0.27	0.28	0.38	0.2	−0.17	0.2
ht	−0.04	0.04	0.01	0.07	−0.03	0.03	−0.01	−0.69	−0.13	1	0.52	−0.08	−0.08	−0.01	−0.07	−0.14	0.29	0.07
wt	−0.11	0.14	0.08	0.14	0.05	0.03	0.05	−0.45	0.09	0.52	1	0.08	0.08	0.33	0.26	0.77	0.09	0.18
sc1	−0.12	0.16	0.09	0.16	0.08	−0.02	0.06	0.02	0.3	−0.08	0.08	1	0.74	0.21	0.22	0.15	0	0.18
sc2	−0.12	0.16	0.1	0.17	0.06	0	0.06	0.02	0.27	−0.08	0.08	0.74	1	0.19	0.21	0.14	0	0.19
dbp	−0.24	0.25	0.17	0.18	0.14	0.01	0.15	−0.06	0.28	−0.01	0.33	0.21	0.19	1	0.8	0.39	−0.06	0.18
sbp	−0.29	0.31	0.22	0.19	0.21	0.02	0.2	0	0.38	−0.07	0.26	0.22	0.21	0.8	1	0.36	−0.09	0.2
mrw	−0.09	0.13	0.08	0.1	0.08	0.01	0.07	0.05	0.2	−0.14	0.77	0.15	0.14	0.39	0.36	1	−0.12	0.14
smoke	−0.11	0.1	0.08	0.07	0.01	0.06	0.06	−0.37	−0.17	0.29	0.09	0	0	−0.06	−0.09	−0.12	1	0.05
diag	−0.18	0.3	0.02	0.58	−0.01	−0.09	−0.09	−0.16	0.2	0.07	0.18	0.18	0.19	0.18	0.2	0.14	0.05	1

```
N = cor(data[complete.cases(data), -c(4, 19, 20, 21)])
corrplot(N, addCoef.col = 'black',  number.cex= 7/(ncol(data) - 4))
```



	lexam	surv	cause	chd	cva	ca	oth	sex	age	ht	wt	sc1	sc2	dbp	sbp	mrw	smoke	diag
lexam	1	−0.72	−0.6	−0.33	−0.29	−0.32	−0.56	0.14	−0.36	−0.01	−0.09	−0.14	−0.13	−0.21	−0.28	−0.08	−0.12	−0.18
surv	−0.72	1	0.84	0.45	0.35	0.45	0.78	−0.14	0.43	0.02	0.12	0.16	0.16	0.25	0.3	0.11	0.11	0.31
cause	−0.6	0.84	1	−0.01	0.22	0.5	0.92	−0.09	0.36	−0.01	0.05	0.09	0.09	0.17	0.21	0.06	0.09	0.03
chd	−0.33	0.45	−0.01	1	−0.1	−0.12	−0.22	−0.13	0.19	0.04	0.13	0.17	0.18	0.19	0.21	0.11	0.05	0.59
cva	−0.29	0.35	0.22	−0.1	1	−0.1	0.44	−0.02	0.22	−0.03	0.03	0.04	0.03	0.12	0.19	0.06	0.04	0.01
ca	−0.32	0.45	0.5	−0.12	−0.1	1	0.57	−0.05	0.13	0.02	0.02	0	0.01	0.01	0.03	0.01	0.07	−0.08
oth	−0.56	0.78	0.92	−0.22	0.44	0.57	1	−0.07	0.34	−0.01	0.03	0.05	0.05	0.13	0.18	0.04	0.09	−0.08
sex	0.14	−0.14	−0.09	−0.13	−0.02	−0.05	−0.07	1	0.02	−0.69	−0.45	−0.01	−0.01	−0.05	0	0.05	−0.38	−0.14
age	−0.36	0.43	0.36	0.19	0.22	0.13	0.34	0.02	1	−0.16	0.08	0.28	0.27	0.26	0.38	0.21	−0.19	0.21
ht	−0.01	0.02	−0.01	0.04	−0.03	0.02	−0.01	−0.69	−0.16	1	0.52	−0.07	−0.06	−0.01	−0.07	−0.14	0.31	0.04
wt	−0.09	0.12	0.05	0.13	0.03	0.02	0.03	−0.45	0.08	0.52	1	0.11	0.1	0.33	0.27	0.77	0.09	0.17
sc1	−0.14	0.16	0.09	0.17	0.04	0	0.05	−0.01	0.28	−0.07	0.11	1	0.74	0.22	0.22	0.16	0.02	0.17
sc2	−0.13	0.16	0.09	0.18	0.03	0.01	0.05	−0.01	0.27	−0.06	0.1	0.74	1	0.2	0.2	0.16	0.01	0.18
dbp	−0.21	0.25	0.17	0.19	0.12	0.01	0.13	−0.05	0.26	−0.01	0.33	0.22	0.2	1	0.79	0.39	−0.07	0.2
sbp	−0.28	0.3	0.21	0.21	0.19	0.03	0.18	0	0.38	−0.07	0.27	0.22	0.2	0.79	1	0.37	−0.1	0.23
mrw	−0.08	0.11	0.06	0.11	0.06	0.01	0.04	0.05	0.21	−0.14	0.77	0.16	0.16	0.39	0.37	1	−0.14	0.16
smoke	−0.12	0.11	0.09	0.05	0.04	0.07	0.09	−0.38	−0.19	0.31	0.09	0.02	0.01	−0.07	−0.1	−0.14	1	0.05
diag	−0.18	0.31	0.03	0.59	0.01	−0.08	−0.08	−0.14	0.21	0.04	0.17	0.17	0.18	0.2	0.23	0.16	0.05	1