Discritive-and-Correlation-Analysis-in-R

Samuel Adabla

Importing working data into R

```
library(readxl)

Prostate <- read_excel("D:/Analytics/Data/R Projects/WEEK 4/Prostate.xlsx")
View(Prostate)
```

Examining the structure and dimension of data

```
str(Prostate)
```

```
## tibble [502 x 18] (S3: tbl_df/tbl/data.frame)
##  $ patno : num [1:502] 1 2 3 4 5 6 7 8 9 10 ...
##  $ stage : num [1:502] 3 3 3 3 3 3 3 3 3 3 ...
##  $ rx    : chr [1:502] "0.2 mg estrogen" "0.2 mg estrogen" "5.0 mg
estrogen" "0.2 mg estrogen" ...
##  $ dtime : num [1:502] 72 1 40 20 65 24 46 62 61 60 ...
##  $ status: chr [1:502] "alive" "dead - other ca" "dead - cerebrovascular"
"dead - cerebrovascular" ...
##  $ age   : num [1:502] 75 54 69 75 67 71 75 73 60 78 ...
##  $ wt    : num [1:502] 76 116 102 94 99 98 100 114 110 107 ...
##  $ pf    : chr [1:502] "normal activity" "normal activity" "normal
activity" "in bed < 50% daytime" ...
##  $ hx    : num [1:502] 0 0 1 1 0 0 0 1 0 1 ...
##  $ sbp   : num [1:502] 15 13 14 14 17 19 14 17 12 13 ...
##  $ dbp   : num [1:502] 9 7 8 7 10 10 10 11 8 8 ...
##  $ ekg   : chr [1:502] "heart strain" "heart block or conduction def"
"heart strain" "benign" ...
##  $ hg    : num [1:502] 13.8 14.6 13.4 17.6 13.4 ...
##  $ sz    : num [1:502] 2 42 3 4 34 10 13 3 4 21 ...
##  $ sg    : num [1:502] 8 NA 9 8 8 11 9 9 10 6 ...
##  $ ap    : num [1:502] 0.3 0.7 0.3 0.9 0.5 ...
##  $ bm    : num [1:502] 0 0 0 0 0 0 0 0 0 0 ...
##  $ sdate : num [1:502] 2778 2820 2933 2999 3002 ...
```

```
dim(Prostate)
```

```
## [1] 502  18
```

Creating factor variables rx_f and status_f from the character variables treatment (rx) and status

```
Prostate$rx_f     <-  factor(Prostate$rx)
Prostate$status_f <- factor(Prostate$status)
```

Summarizing the categorical variables rx_f and status_f (i.e., Obtaining frequency tables)

```
table(Prostate$rx_f)

## 
## 0.2 mg estrogen 1.0 mg estrogen 5.0 mg estrogen         placebo 
##             124             126             125             127

table(Prostate$status_f)

## 
##                        alive       dead - cerebrovascular 
##                          148                           31 
##     dead - heart or vascular              dead - other ca 
##                           96                           25 
## dead - other specific non-ca          dead - prostatic ca 
##                           28                          130 
##     dead - pulmonary embolus   dead - respiratory disease 
##                           14                           16 
##         dead - unknown cause    dead - unspecified non-ca 
##                            7                            7
```

Obtaining relative frequency tables (proportions or %s) of rx_f and status_f

```
prop.table(table(Prostate$rx_f))

## 
## 0.2 mg estrogen 1.0 mg estrogen 5.0 mg estrogen         placebo 
##        0.247012        0.250996        0.249004        0.252988

prop.table(table(Prostate$status_f))

## 
##                        alive       dead - cerebrovascular 
##                   0.29482072                   0.06175299 
##     dead - heart or vascular              dead - other ca 
##                   0.19123506                   0.04980080 
## dead - other specific non-ca          dead - prostatic ca 
##                   0.05577689                   0.25896414 
##     dead - pulmonary embolus   dead - respiratory disease 
##                   0.02788845                   0.03187251
##         dead - unknown cause    dead - unspecified non-ca
##                   0.01394422                   0.01394422
```

Creating a new variable, died, from the variable status using for loop

```
for (i in (1:502))
{
  if (Prostate$status[i] == "alive")
  {
    Prostate$died[i] = "No"
  }
  else
  {
    Prostate$died[i] = "Yes"
  }
}

## Warning: Unknown or uninitialised column: `died`.
```

Converting the new character variable, dead, to a factor

```
Prostate$died_f <-  factor(Prostate$died)
```

Obtaining a cross-tab (with counts) of rx_f and died_f

```
table(Prostate$rx_f, Prostate$died_f)

##
##                   No Yes
##   0.2 mg estrogen 29  95
##   1.0 mg estrogen 55  71
##   5.0 mg estrogen 32  93
##   placebo         32  95
```

Obtaining a cross-tab (with cell %s) of rx_f and died_f

```
(prop.table(table(Prostate$rx_f, Prostate$died_f)))*100

##
##                          No       Yes
##   0.2 mg estrogen  5.776892 18.924303
##   1.0 mg estrogen 10.956175 14.143426
##   5.0 mg estrogen  6.374502 18.525896
##   placebo          6.374502 18.924303
```

Obtaining relative frequency tables (with row %s) of rx_f and died_f

```
prop.table(table(Prostate$rx_f, Prostate$died_f), 1)
```

```
##
##                           No       Yes
##   0.2 mg estrogen 0.2338710 0.7661290
##   1.0 mg estrogen 0.4365079 0.5634921
##   5.0 mg estrogen 0.2560000 0.7440000
##   placebo         0.2519685 0.7480315
```

Obtain relative frequency tables (with column %s) of rx_f and died_f.

```
prop.table(table(Prostate$rx_f, Prostate$died_f), 2)
```

```
##
##                           No       Yes
##   0.2 mg estrogen 0.1959459 0.2683616
##   1.0 mg estrogen 0.3716216 0.2005650
##   5.0 mg estrogen 0.2162162 0.2627119
##   placebo         0.2162162 0.2683616
```

Summarizing the continuous variables age, weight(wt), systolic blood pressure

(sbp), diastolic blood pressure (dbp), hg, sz and sg.
Descriptive statistics

```
summary(Prostate$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   48.00   70.00   73.00   71.46   76.00   89.00       1
```

```
summary(Prostate$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   69.00   90.00   98.00   99.03  107.00  152.00       2
```

```
summary(Prostate$sbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   13.00   14.00   14.35   16.00   30.00
```

```
summary(Prostate$dbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   7.000   8.000   8.149   9.000  18.000
```

```
summary(Prostate$hg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.899  12.299  13.699  13.446  14.699  21.199
```

```
summary(Prostate$sz)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.00    5.00   11.00   14.63   21.00   69.00       5 
```

```
summary(Prostate$sg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    5.00    9.00   10.00   10.31   11.00   15.00      11 
```

Using Hmisc packgage to obtain additonal discriptive statistics i.e., percentiles

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
## 
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
## 
##     format.pval, units
```

Obtain summary descriptive statistics on age, wt, sbp, dbp, hg, sz and sg using the describe() function.

```
describe(Prostate$age)
```

```
## Prostate$age 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##      501        1       41    0.996    71.46    7.497       56       60 
##      .25      .50      .75      .90      .95 
##       70       73       76       78       80 
## 
## lowest : 48 49 50 51 52, highest: 84 85 87 88 89
```

```
describe(Prostate$wt)
```

```
## Prostate$wt 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##      500        2       67    0.999    99.03    14.93    77.95    82.90 
##      .25      .50      .75      .90      .95 
##    90.00    98.00   107.00   116.00   123.00 
## 
## lowest :  69  71  72  73  74, highest: 136 142 145 150 152
```

```
describe(Prostate$sbp)
```

```
## Prostate$sbp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      502        0       18     0.98    14.35    2.596       11       12
##      .25      .50      .75      .90      .95
##       13       14       16       17       18
##
## lowest :  8  9 10 11 12, highest: 21 22 23 24 30
##
## Value          8     9    10    11    12    13    14    15    16    17
18
## Frequency      1     3    14    27    65    74    98    74    72    34
17
## Proportion 0.002 0.006 0.028 0.054 0.129 0.147 0.195 0.147 0.143 0.068
0.034
##
## Value         19    20    21    22    23    24    30
## Frequency     12     3     2     3     1     1     1
## Proportion 0.024 0.006 0.004 0.006 0.002 0.002 0.002
```

```
describe(Prostate$dbp)
```

```
## Prostate$dbp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      502        0       12    0.945    8.149    1.553        6        6
##      .25      .50      .75      .90      .95
##        7        8        9       10       10
##
## lowest :  4  5  6  7  8, highest: 11 12 13 14 18
##
## Value          4     5     6     7     8     9    10    11    12    13
14
## Frequency      4     5    43   107   165    94    66     9     5     2
1
## Proportion 0.008 0.010 0.086 0.213 0.329 0.187 0.131 0.018 0.010 0.004
0.002
##
## Value         18
## Frequency      1
## Proportion 0.002
```

```
describe(Prostate$hg)
```

```
## Prostate$hg
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      502        0       91        1    13.45     2.16     10.2     10.7
##      .25      .50      .75      .90      .95
##     12.3     13.7     14.7     15.8     16.4
##
```

```
## lowest :  5.899414  7.000000  7.199219  7.799805  8.199219
## highest: 17.296875 17.500000 17.597656 18.199219 21.199219
```

```
describe(Prostate$sz)
```

```
## Prostate$sz
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      497        5       55    0.998    14.63    13.05      2.0      3.0
##      .25      .50      .75      .90      .95
##      5.0     11.0     21.0     32.0     39.2
##
## lowest :  0  1  2  3  4, highest: 54 55 61 62 69
```

```
describe(Prostate$sg)
```

```
## Prostate$sg
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      491       11       11    0.959    10.31    2.245        8        8
##      .25      .50      .75      .90      .95
##        9       10       11       13       13
##
## lowest :  5  6  7  8  9, highest: 11 12 13 14 15
##
## Value          5     6     7     8     9    10    11    12    13    14
15
## Frequency      3     8     7    67   137    33   114    26    75     5
16
## Proportion 0.006 0.016 0.014 0.136 0.279 0.067 0.232 0.053 0.153 0.010
0.033
```

```
bystats(Prostate$age, 0)
```

```
##
##  Mean of Prostate$age by
##
##       N Missing     Mean
## 0   501       1 71.45709
## ALL 501       1 71.45709
```

```
bystats(Prostate$wt, 0)
```

```
##
##  Mean of Prostate$wt by
##
##       N Missing   Mean
## 0   500       2 99.026
## ALL 500       2 99.026
```

```
bystats(Prostate$sbp, 0)
```

```
##
##  Mean of Prostate$sbp by
```

```
##
##       N     Mean
## 0   502 14.35259
## ALL 502 14.35259
```

```
bystats(Prostate$dbp, 0)
```

```
##
##  Mean of Prostate$dbp by
##
##       N     Mean
## 0   502 8.149402
## ALL 502 8.149402
```

```
bystats(Prostate$hg, 0)
```

```
##
##  Mean of Prostate$hg by
##
##       N     Mean
## 0   502 13.44645
## ALL 502 13.44645
```

```
bystats(Prostate$sz, 0)
```

```
##
##  Mean of Prostate$sz by
##
##       N Missing     Mean
## 0   497       5 14.62978
## ALL 497       5 14.62978
```

```
bystats(Prostate$sg, 0)
```

```
##
##  Mean of Prostate$sg by
##
##       N Missing     Mean
## 0   491      11 10.30957
## ALL 491      11 10.30957
```

```
bystats(Prostate$age, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(18, 30, 18, 111, 30, 111, 18, 18) of Prostate$age by 0
##
##     N     Missing Mean               Median Mode      SD                0%
## 0   "501" "1"     "71.4570858283433" "73"   "numeric" "7.0812890557171"
"48"
## ALL "501" "1"     "71.4570858283433" "73"   "numeric" "7.0812890557171"
"48"
```

```
##     25%  50%  75%  100%
## 0   "70" "73" "76" "89"
## ALL "70" "73" "76" "89"
```

```
bystats(Prostate$wt, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(19, 29, 19, 110, 29, 110, 19, 19) of Prostate$wt by 0
##
##     N     Missing Mean     Median Mode      SD                 0%   25%
50%
## 0   "500" "2"     "99.026" "98"   "numeric" "13.4364578963953" "69" "90"
"98"
## ALL "500" "2"     "99.026" "98"   "numeric" "13.4364578963953" "69" "90"
"98"
##     75%   100%
## 0   "107" "152"
## ALL "107" "152"
```

```
bystats(Prostate$sbp, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(20, 30, 20, 111, 30, 111, 20, 20) of Prostate$sbp by 0
##
##     N     Mean               Median Mode      SD                 0%  25%
50%
## 0   "502" "14.3525896414343" "14"   "numeric" "2.41609359306121" "8" "13"
"14"
## ALL "502" "14.3525896414343" "14"   "numeric" "2.41609359306121" "8" "13"
"14"
##     75%  100%
## 0   "16" "30"
## ALL "16" "30"
```

```
bystats(Prostate$dbp, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(21, 30, 21, 111, 30, 111, 21, 21) of Prostate$dbp by 0
##
##     N     Mean               Median Mode      SD                0%  25%
50% 75%
## 0   "502" "8.14940239043825" "8"    "numeric" "1.4694458476704" "4" "7"
"8" "9"
## ALL "502" "8.14940239043825" "8"    "numeric" "1.4694458476704" "4" "7"
"8" "9"
##     100%
## 0   "18"
## ALL "18"
```

```
bystats(Prostate$hg, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(22, 29, 22, 110, 29, 110, 22, 22) of Prostate$hg by 0
##
##     N     Mean               Median        Mode      SD
## 0   "502" "13.4464544167082" "13.69921875" "numeric" "1.95110289213966"
## ALL "502" "13.4464544167082" "13.69921875" "numeric" "1.95110289213966"
##     0%            25%            50%           75%           100%
## 0   "5.8994140625" "12.298828125" "13.69921875" "14.69921875"
"21.19921875"
## ALL "5.8994140625" "12.298828125" "13.69921875" "14.69921875"
"21.19921875"
```

```
bystats(Prostate$sz, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(23, 29, 23, 110, 29, 110, 23, 23) of Prostate$sz by 0
##
##     N     Missing Mean               Median Mode      SD                0%
25%
## 0   "497" "5"     "14.6297786720322" "11"   "numeric" "12.324729212138"
"0" "5"
## ALL "497" "5"     "14.6297786720322" "11"   "numeric" "12.324729212138"
"0" "5"
##     50%  75%  100%
## 0   "11" "21" "69"
## ALL "11" "21" "69"
```

```
bystats(Prostate$sg, 0, fun=function(x) c(Mean=mean(x), Median=median(x),
Mode=mode(x), SD=sd(x), quantile(x)))
```

```
##
##  c(24, 29, 24, 110, 29, 110, 24, 24) of Prostate$sg by 0
##
##     N     Missing Mean               Median Mode      SD
0%
## 0   "491" "11"    "10.3095723014257" "10"   "numeric" "2.01876149655325"
"5"
## ALL "491" "11"    "10.3095723014257" "10"   "numeric" "2.01876149655325"
"5"
##     25% 50%  75%  100%
## 0   "9" "10" "11" "15"
## ALL "9" "10" "11" "15"
```

Examining correlataions

Using the subset function to create a dataframe with only the continuous variables of interest from the Prostrate dataset

```
Prostate_ContinuousVars <-  subset(Prostate, select = c(age, wt, sbp, dbp,
hg, sz, sg))
View(Prostate_ContinuousVars)
```

Obtaining the Pearson and Spearman correlation matrices using the cor() function with complete.obs option to remove rows with missing data for any of the continuous variables selected

```
cor_pearson <- cor(Prostate_ContinuousVars, method = c("pearson"),
use="complete.obs")
cor_spearman <- cor(Prostate_ContinuousVars, method = c("spearman"),
use="complete.obs")
```

Using the round() function on the output of the cor() function to round the results to 2 decimal places.

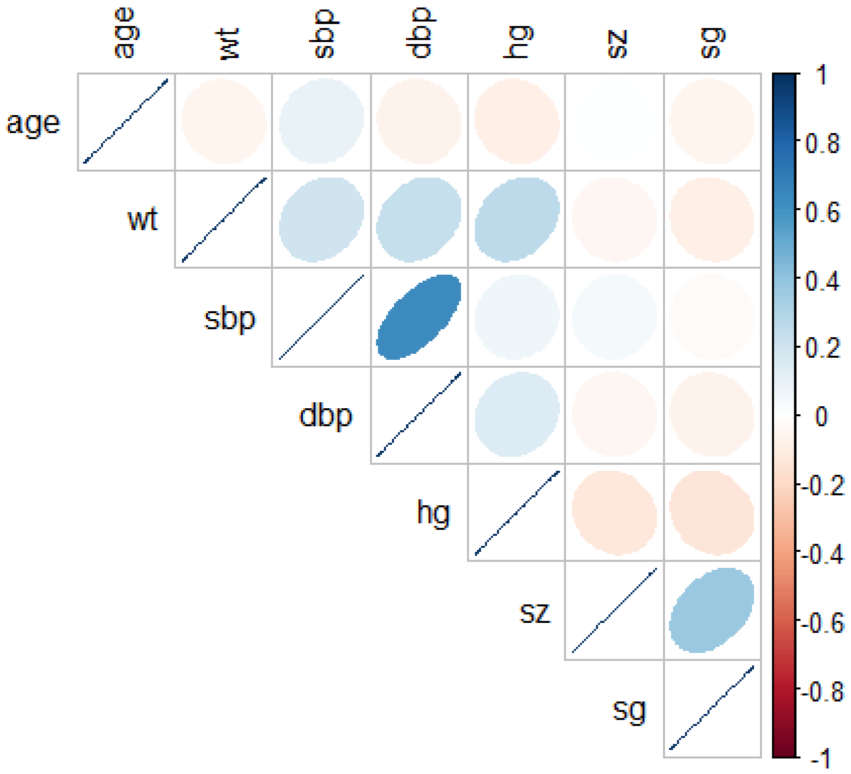
```
round(cor_pearson, 2)

##       age    wt   sbp   dbp    hg    sz    sg
## age  1.00 -0.06  0.10 -0.07 -0.09  0.01 -0.06
## wt  -0.06  1.00  0.21  0.23  0.26 -0.05 -0.09
## sbp  0.10  0.21  1.00  0.63  0.06  0.05 -0.03
## dbp -0.07  0.23  0.63  1.00  0.15 -0.04 -0.07
## hg  -0.09  0.26  0.06  0.15  1.00 -0.13 -0.14
## sz   0.01 -0.05  0.05 -0.04 -0.13  1.00  0.38
## sg  -0.06 -0.09 -0.03 -0.07 -0.14  0.38  1.00

round(cor_spearman, 2)

##       age    wt   sbp   dbp    hg    sz    sg
## age  1.00 -0.03  0.07 -0.10 -0.13 -0.03 -0.03
## wt  -0.03  1.00  0.19  0.21  0.26 -0.01 -0.08
## sbp  0.07  0.19  1.00  0.57  0.07  0.07 -0.03
## dbp -0.10  0.21  0.57  1.00  0.16 -0.01 -0.05
## hg  -0.13  0.26  0.07  0.16  1.00 -0.14 -0.12
## sz  -0.03 -0.01  0.07 -0.01 -0.14  1.00  0.36
## sg  -0.03 -0.08 -0.03 -0.05 -0.12  0.36  1.00
```

Plotting the correlation outputs (correlograms)

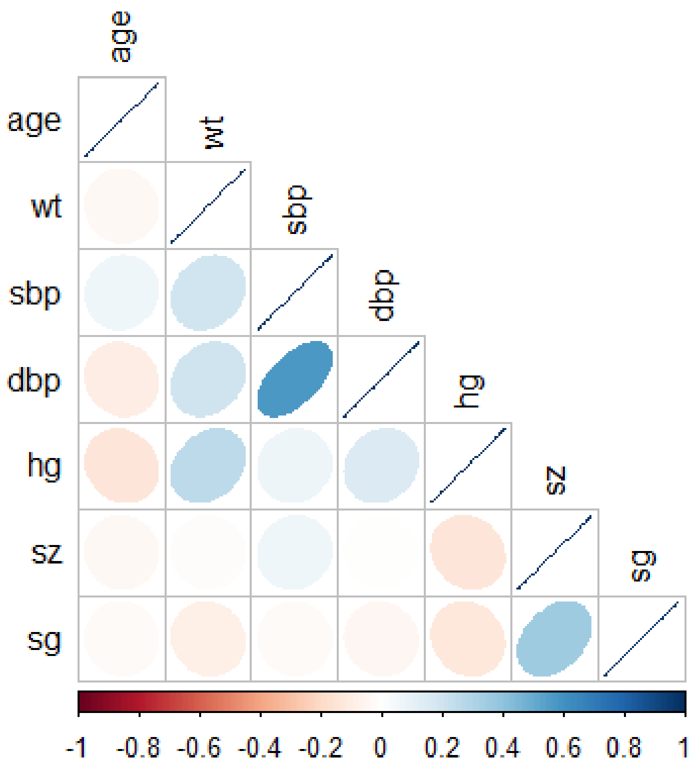
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(corr=cor_pearson, type="upper", method="ellipse", tl.col = "black")
```



```
corrplot(corr=cor_spearman, type="lower", method="ellipse", tl.col = "black")
```



Round the correlation matrices to 1 decimal place for easy interpretation

```
round(cor_pearson, 1)

##      age   wt sbp  dbp   hg   sz   sg
## age  1.0 -0.1 0.1 -0.1 -0.1  0.0 -0.1
## wt  -0.1  1.0 0.2  0.2  0.3  0.0 -0.1
## sbp  0.1  0.2 1.0  0.6  0.1  0.0  0.0
## dbp -0.1  0.2 0.6  1.0  0.1  0.0 -0.1
## hg  -0.1  0.3 0.1  0.1  1.0 -0.1 -0.1
## sz   0.0  0.0 0.0  0.0 -0.1  1.0  0.4
## sg  -0.1 -0.1 0.0 -0.1 -0.1  0.4  1.0

round(cor_spearman, 1)

##      age   wt sbp  dbp   hg   sz   sg
## age  1.0  0.0 0.1 -0.1 -0.1  0.0  0.0
## wt   0.0  1.0 0.2  0.2  0.3  0.0 -0.1
## sbp  0.1  0.2 1.0  0.6  0.1  0.1  0.0
## dbp -0.1  0.2 0.6  1.0  0.2  0.0  0.0
## hg  -0.1  0.3 0.1  0.2  1.0 -0.1 -0.1
## sz   0.0  0.0 0.1  0.0 -0.1  1.0  0.4
## sg   0.0 -0.1 0.0  0.0 -0.1  0.4  1.0
```