```r
# Load the lars package and the diabetes dataset
library(reshape2)
library(lars)
```

```
## Loaded lars 1.3
```

```r
data(diabetes)
library(GGally)
```

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

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(ggplot2)
library(gridExtra)

library("rstan") # observe startup messages
```

```
## Loading required package: StanHeaders
```

```
##
## rstan version 2.32.3 (Stan version 2.26.1)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
## For within-chain threading using `reduce_sum()` or `map_rect()` Stan functions,
## change `threads_per_chain` option:
## rstan_options(threads_per_chain = 1)
```

```r
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
#data
X_matrix <- diabetes$x
class(X_matrix) <- "matrix"
y_vector <- diabetes$y

X_design <- cbind(1, X_matrix)

data.raw <- read.table(file='https://hastie.su.domains/Papers/LARS/diabetes.data', header=T)
```

x matrix has been standardized to have unit L2 norm in each column and zero mean

There are 10 explanatory variables, including age (age), sex (sex), body mass index (bmi) and mean arterial blood pressure (map) of 442 patients as well as six blood serum measurements (tc, ldl, hdl, tch, ltg and glu).

https://garthtarr.github.io/mplot/reference/diabetes.html

tc

Total cholesterol (mg/dL)? Desirable range: below 200 mg/dL

ldl

Low-density lipoprotein ("bad" cholesterol)? Desirable range: below 130 mg/dL

hdl

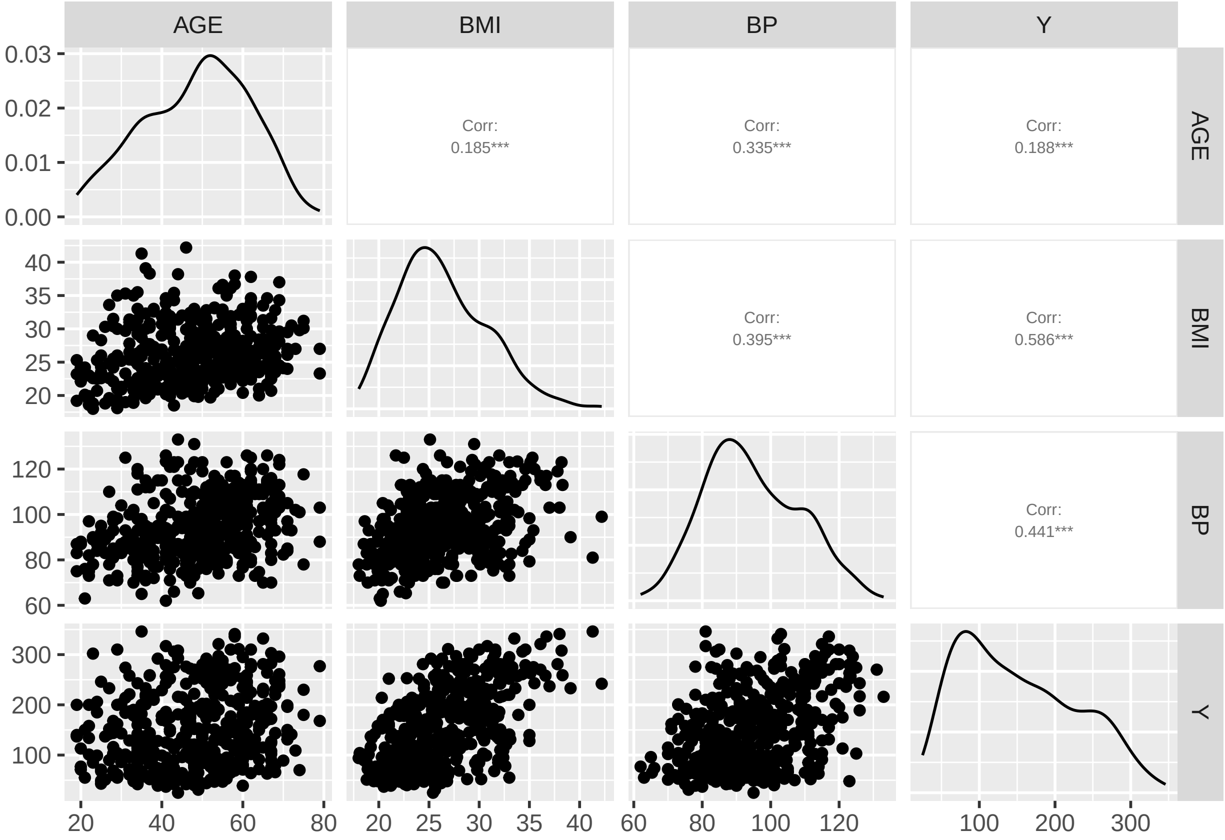
High-density lipoprotein ("good" cholesterol)? Desirable range: above 40 mg/dL

serum concentration of lamorigine (LTG), glucose (GLU)

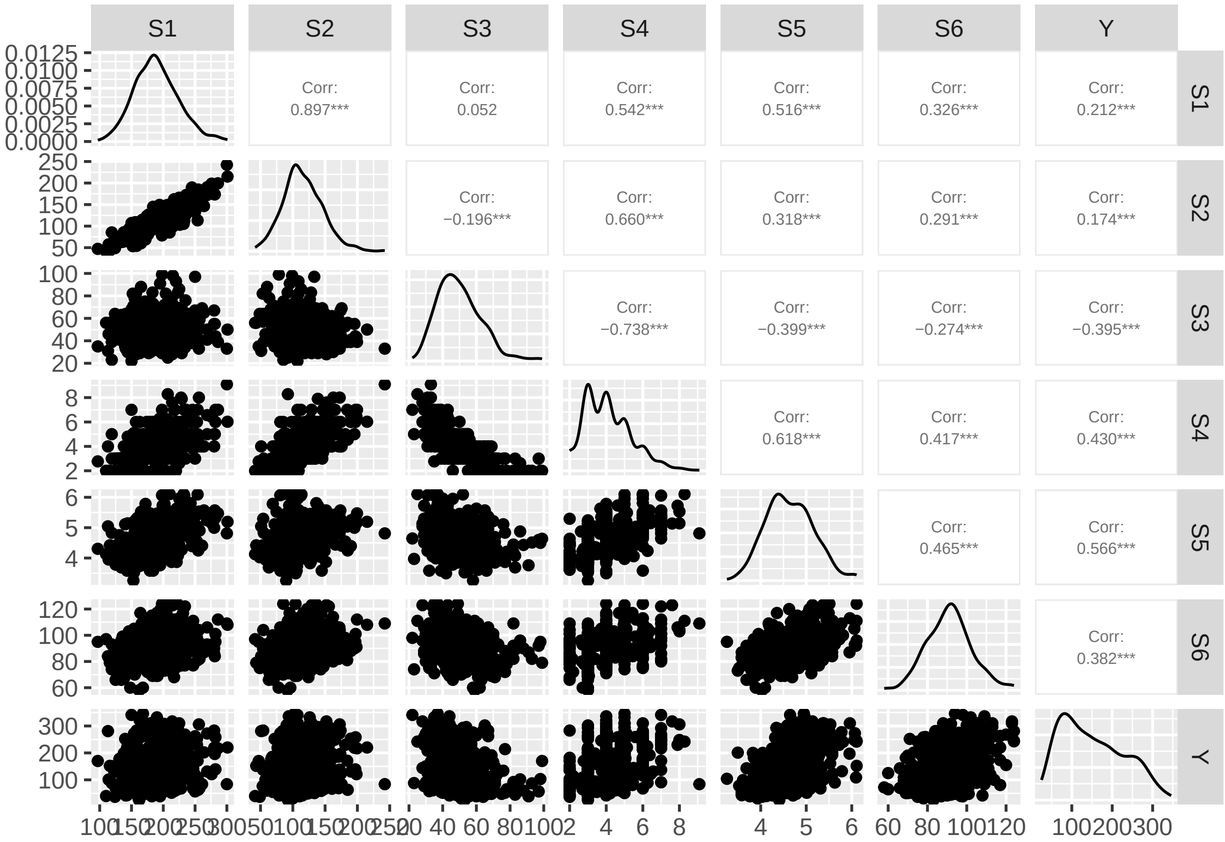
age age in years sex bmi body mass index bp average blood pressure s1 tc, total serum cholesterol s2 ldl, low-density lipoproteins s3 hdl, high-density lipoproteins s4 tch, total cholesterol / HDL s5 ltg, possibly log of serum triglycerides level s6 glu, blood sugar level

https://hastie.su.domains/Papers/LARS/

```
ggpairs(data.raw[c(1,3,4,11)],
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)
```



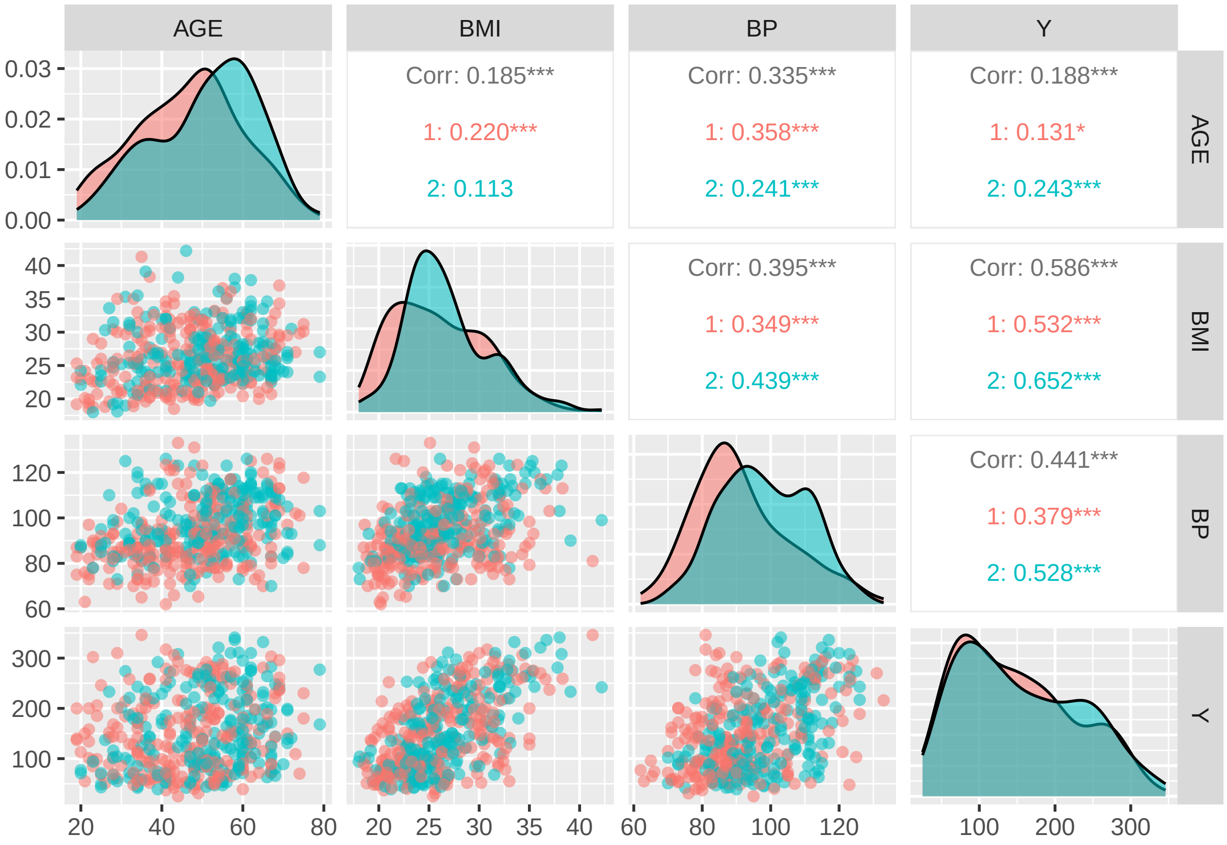
```
ggpairs(data.raw[c(5:10,11)],
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)
```



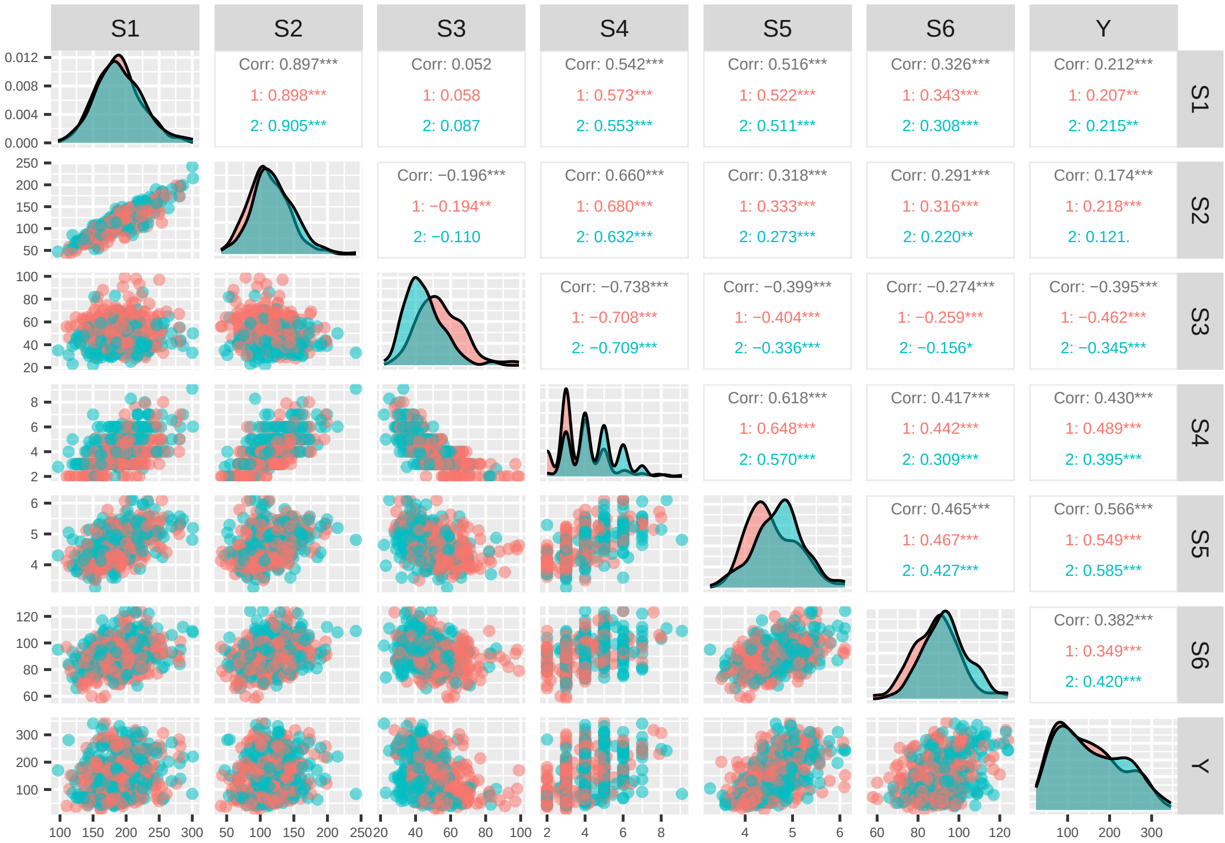
There is no obvious pattern in these variables. Except for S4 (tch value), seems multi mode,

Sex is a nature categorical variable. Examine whether difference between male and female.

```
ggpairs(data.raw, columns=c(1,3,4,11),
    ggplot2::aes(color=factor(SEX), alpha=0.7),
    upper=list(continuous=wrap("cor", size=3)),
    progress = FALSE)
```



```r
ggpairs(data.raw, columns=5:11,
    ggplot2::aes(color=factor(SEX), alpha=0.7),
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```



From S1 to S6, seems no obvious difference between sex.

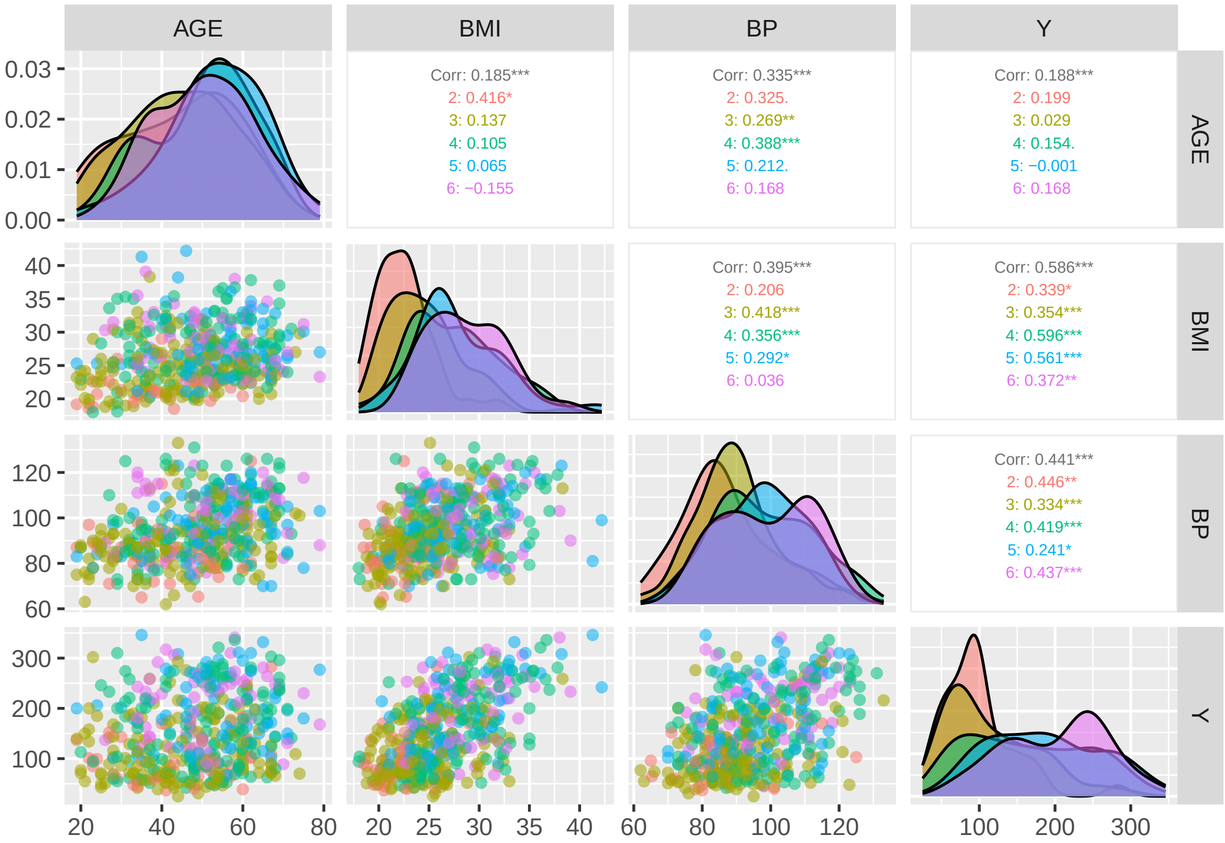
maybe some in BP,

Now we consider the multi mode if S4 (tch value), we re-encode it

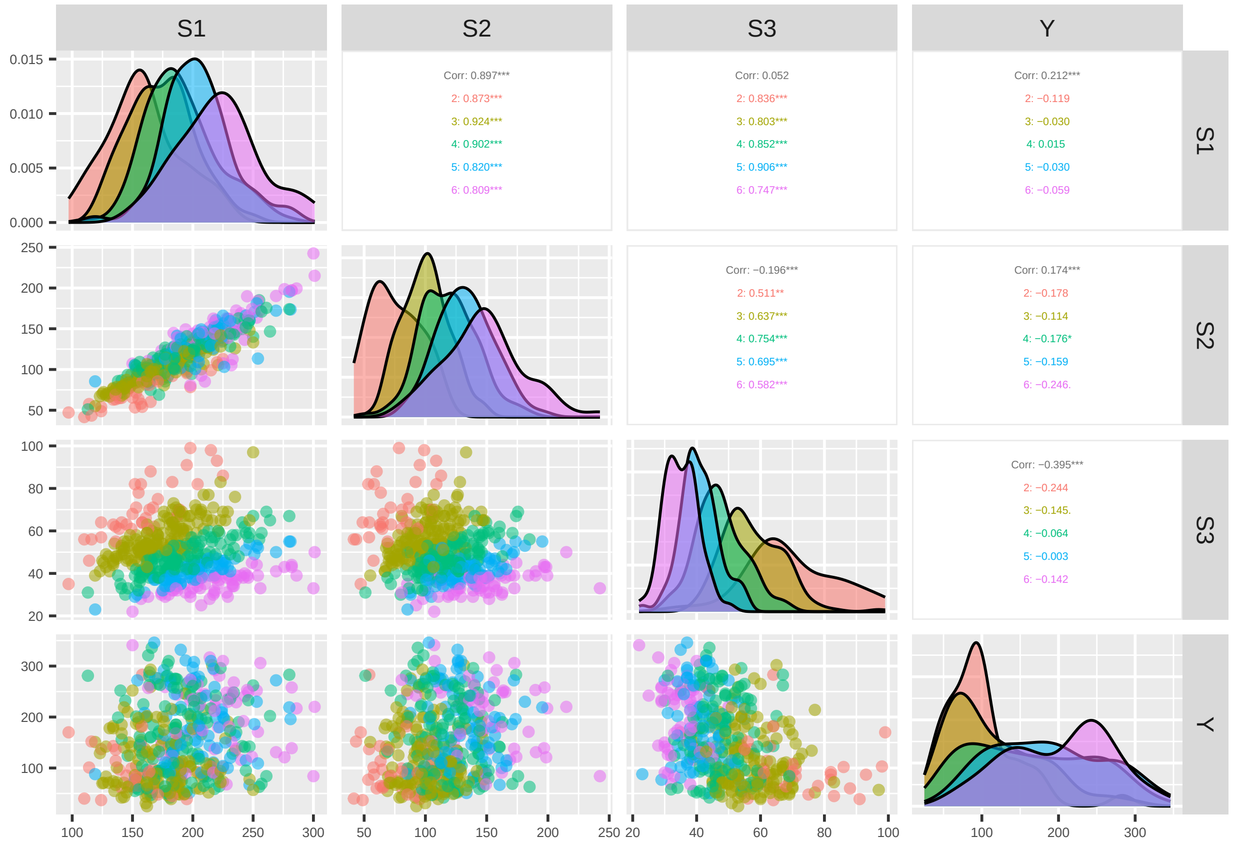
```
data.raw$S4.new <- 1
data.raw$S4.new[which(data.raw$S4<3)] <- 2
data.raw$S4.new[which((data.raw$S4>=3)&(data.raw$S4<4))] <- 3
data.raw$S4.new[which((data.raw$S4>=4)&(data.raw$S4<5))] <- 4
data.raw$S4.new[which((data.raw$S4>=5)&(data.raw$S4<6))] <- 5
data.raw$S4.new[which(data.raw$S4>=6)] <- 6
```

```
data.raw$S4.new.new <- data.raw$S1/data.raw$S3
#ggpairs(data.raw, columns=c(11,14))
```

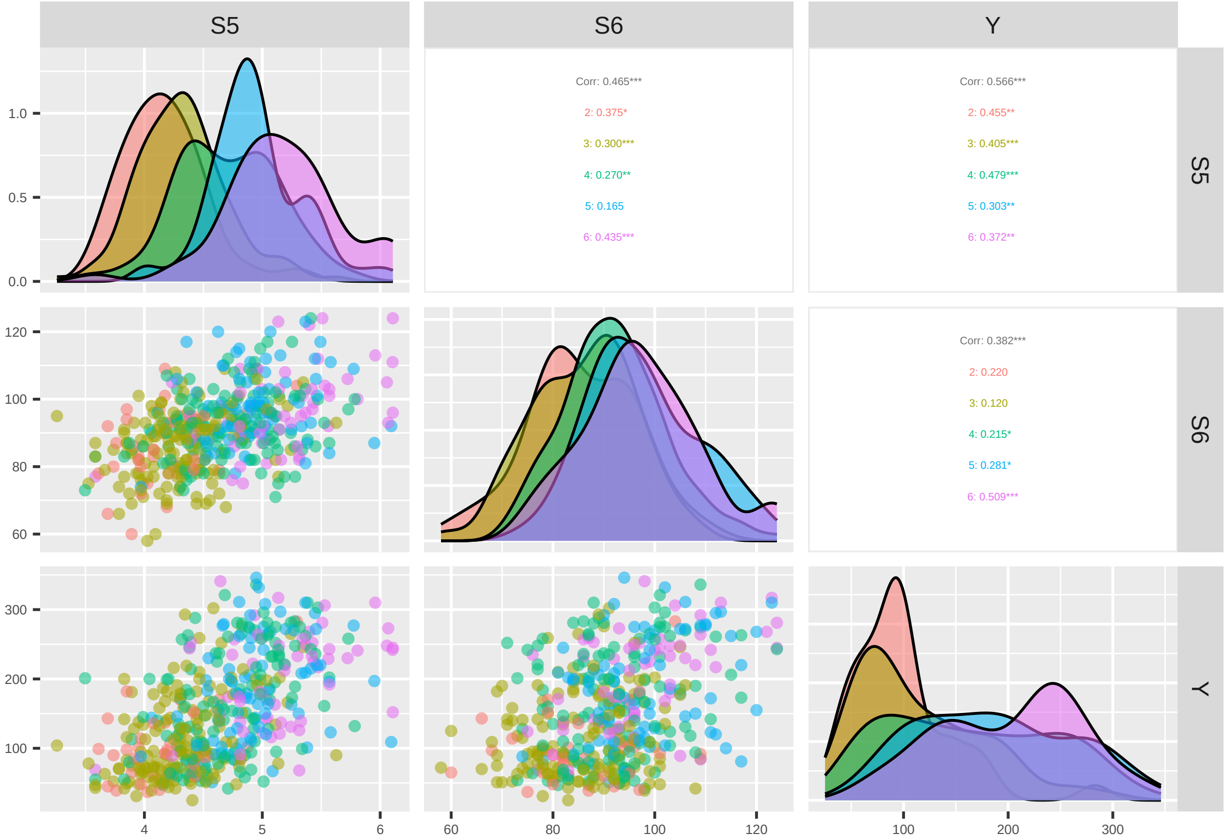
```
ggpairs(data.raw, columns=c(1,3,4,11),
    ggplot2::aes(color=factor(S4.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)
```



```
ggpairs(data.raw, columns=c(5,6,7,11),
    ggplot2::aes(color=factor(S4.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=1.5)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```



```r
ggpairs(data.raw, columns=c(9:11),
    ggplot2::aes(color=factor(S4.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=1.5)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```



different in BMI, BP, S1, S2, S3, S5, S6, even Y

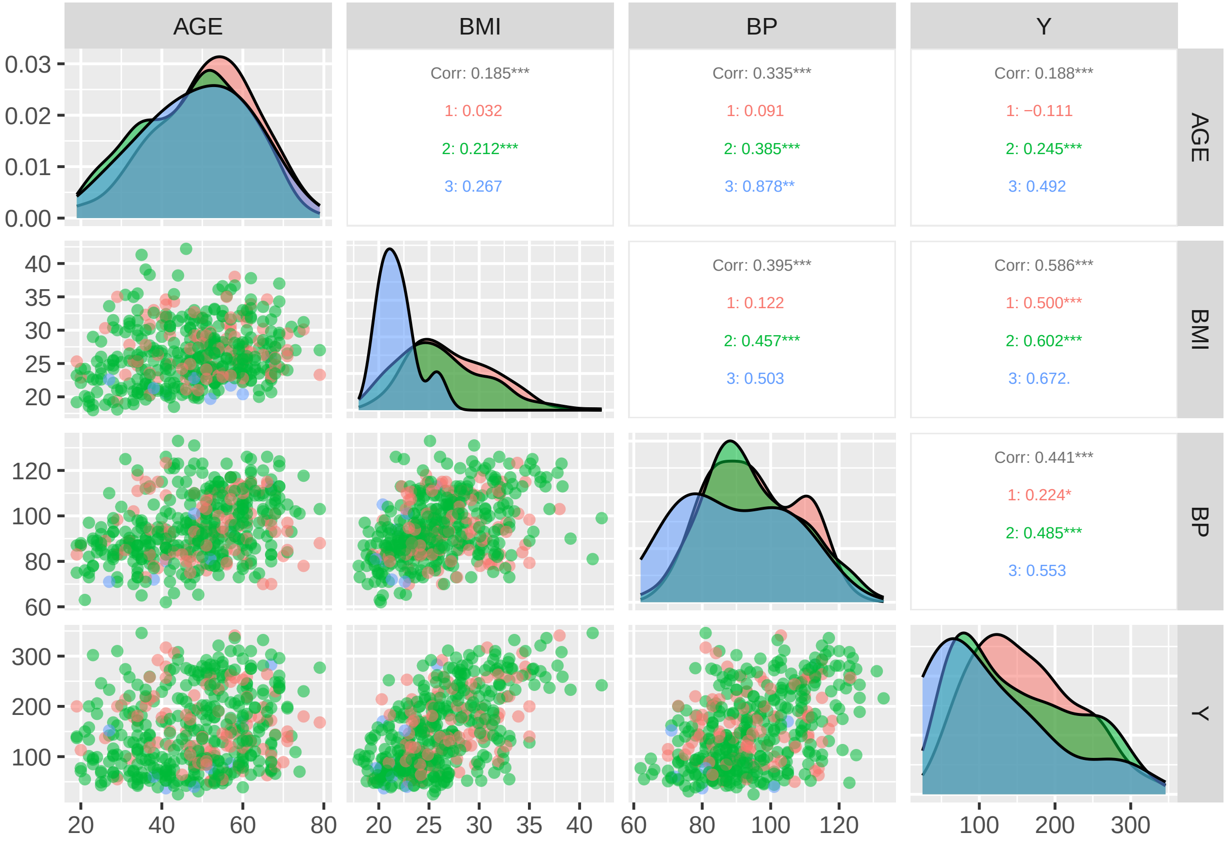
correspond to different value of tch, seems different mean, and variance of y in difference tch value is different.

S1-S3, S2-S3„ random intercept, random slope model

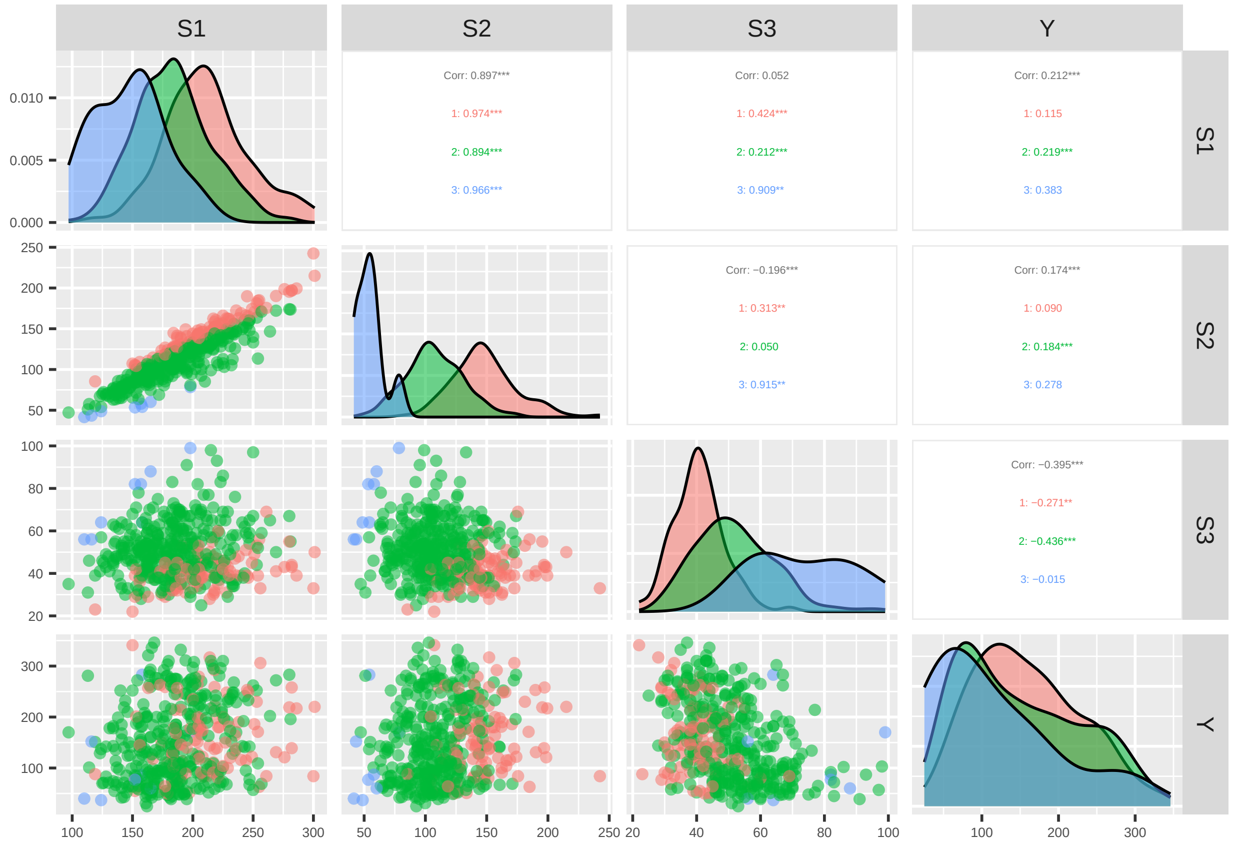
S2-S2 highly correlated. S1-Y, S2-Y, S3-Y weakly obvious random intercept, random slope

```r
ggpairs(data.raw, columns=c(1,3,4,11),
    ggplot2::aes(color=factor(S2.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)

ggpairs(data.raw, columns=c(5:11),
    ggplot2::aes(color=factor(S2.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=1.5)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```

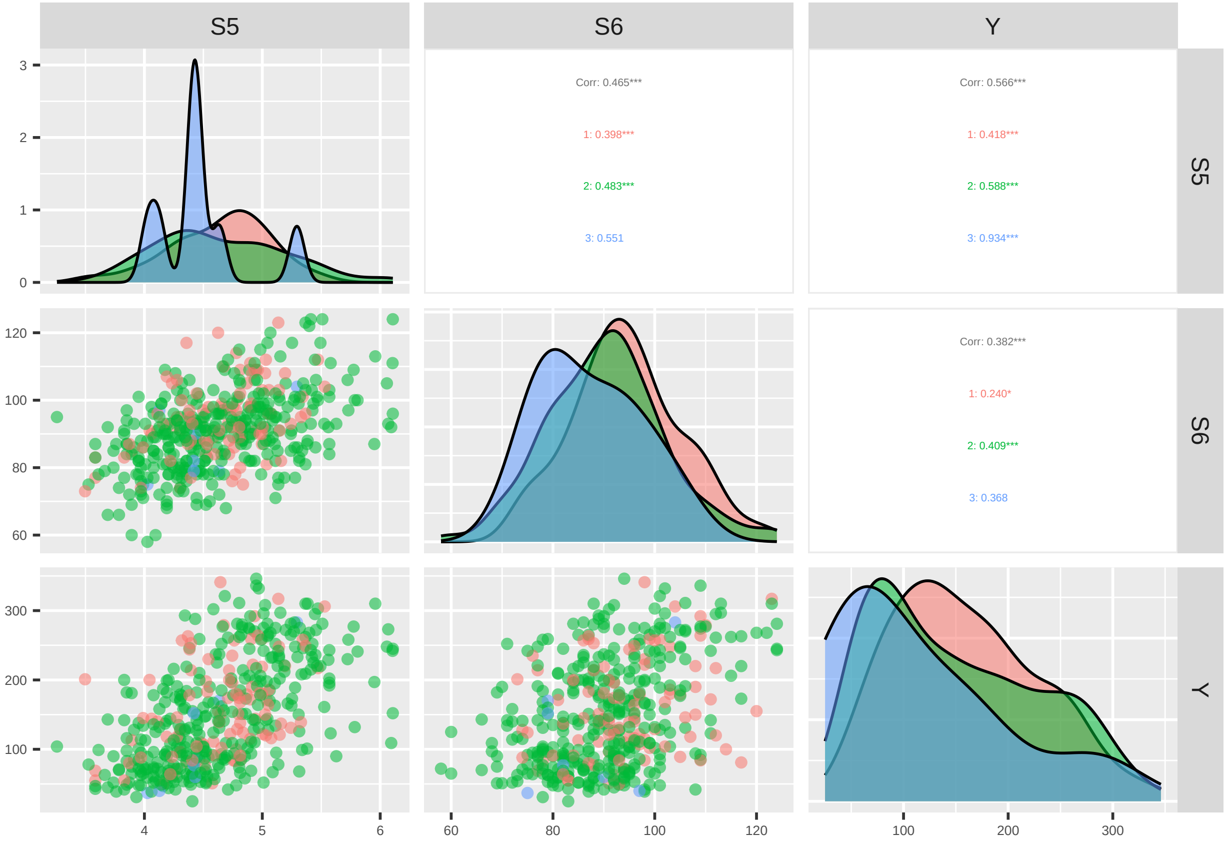
```r
data.raw$S2.new <- round(data.raw$S1/data.raw$S2)

ggpairs(data.raw, columns=c(1,3,4,11),
    ggplot2::aes(color=factor(S2.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=2)),
    progress = FALSE)
```



```
ggpairs(data.raw, columns=c(5,6,7,11),
    ggplot2::aes(color=factor(S2.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=1.5)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```



```
ggpairs(data.raw, columns=c(9:11),
    ggplot2::aes(color=factor(S2.new),
                 alpha = 0.7),
    upper=list(continuous=wrap("cor", size=1.5)),
    progress = FALSE)+
theme(axis.text.x = element_text(size=5),
      axis.text.y = element_text(size=5))
```



s1 tc, total serum cholesterol s2 ldl, low-density lipoproteins s3 hdl, high-density lipoproteins s4 tch, total cholesterol / HDL s5 ltg, possibly log of serum triglycerides level s6 glu, blood sugar level

use S4.new grouping, in each group, have different slope model should be like i: i-th group j: j-th obs in i-th group

fix effect: age sex(as factor) bmi bp(map) S6(glu) (same beta among all levels of S4.new)

random slope S1(tc), S2.new.new(tc/ldl–avoid high corr), S3(hdl), S5(ltg) (different beta among levels of S4.new)

grouping: S4.new

examine cor

model formula

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 bmi + \beta_4 bp + \beta_5 glu \\
&\quad + \gamma_{i1} tc + \gamma_{i2} ldl.new + \gamma_{i3} hdl + \gamma_{i4} ltg + e_i
\end{aligned}
$$

in matrix form ?? #TODO

e_i is error, the same inside group, different between groups model

$$y_{ij} \sim N(x_{1ij}\beta_{fix} + x_{2ij}\gamma_i, \sigma_i^2)$$

.

$$params: \beta_{fix},\ \gamma_i(\gamma_1 ... \gamma_5),\ \sigma_i^2(\sigma_1^2 ... \sigma_5^2)$$

.

$$params\,prior:$$

$$\beta_{fix}|(?) \sim\ N(\mu_{fix}, g(X_1'X_1)^{-1}(\sigma^2)).\text{(I tend to add } \sigma^2\text{)}$$

$$\gamma_i \sim iid\ N(\mu_\gamma, \Sigma_\gamma)\text{(drawing 5)}$$

$$\sigma_i \sim iid\ LN(\mu_\sigma, \tau_\sigma)\text{(drawing 5)}$$

.

$$hyper\,prior:$$

all indep

$$p(\mu_{fix}, \mu_\gamma, \Sigma_\gamma, \mu_\sigma, \tau_\sigma) = p(\mu_{fix})p(\mu_\gamma)p(\Sigma_\gamma)p(\mu_\sigma)p(\tau_\sigma)$$

$$p(\mu_{fix}) \propto 1,\ p(\mu_\sigma) \propto 1,\ p(\mu_\gamma) \propto 1$$

$$\tau_\sigma \sim \frac{1}{1+\sigma^2} I(\sigma > 0)$$

$$\Sigma_\gamma \sim \text{LKJ correlation matrix prior}$$

try sampling

//sample size 442 int<lower=1> n; //number of S4.new (group levels) int<lower=1> n_gp; //number of fix (contain intercept) int<lower=1> p_fix; //number of rand slope int<lower=1> p_rand;

//response real y[n]; //design matrix of fix age sex(as factor) bmi bp(map) S6(glu) matrix[n, p_fix] X1; //matrix of random slope S1(tc), S2.new.new(tc/ldl–avoid high corr), S3(hdl), S5(ltg) matrix[n, p_rand] X2; //which group int group[n];

//g-prior for fix int<lower=0> g;

```r
data.raw <- read.table(file='https://hastie.su.domains/Papers/LARS/diabetes.data', header=T)

data.raw$S4.new <- 1
data.raw$S4.new[which(data.raw$S4<3)] <- 1
data.raw$S4.new[which((data.raw$S4>=3)&(data.raw$S4<4))] <- 2
data.raw$S4.new[which((data.raw$S4>=4)&(data.raw$S4<5))] <- 3
data.raw$S4.new[which((data.raw$S4>=5)&(data.raw$S4<6))] <- 4
data.raw$S4.new[which(data.raw$S4>=6)] <- 5

data.raw$SEX <- data.raw$SEX-1
data.raw$S2.new <- data.raw$S1/data.raw$S2

data_list <- list(
  n = dim(data.raw)[1],
  n_gp = length(unique(data.raw$S4.new)),
  p_fix = 5,
  p_rand = 5,

  y = data.raw$Y,
  X1 = data.raw[c(1,2,3,4,10)],
  X2 = cbind(1, data.raw[c(5,9,12,13)]),

  group = data.raw$S4.new,
  g = 1
)
```

```
model_hier <- stan_model(file='./prior_hier.stan')


# Create a data list for Stan
set.seed(123)


# Fit the model to the data
stan_fit_hier <- sampling(model_hier,
                     data = data_list,
                     chains = 4,
                     iter = 5000)

# Print a summary of the results
print(stan_fit_hier)

# Plot the posterior distributions
plot(stan_fit_hier,
     pars=c("beta_fix", "gamma_rand", "sigma_gp"))
```