LDA final project

```r
library(foreign)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand
```

```r
library(nlme)
```

```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:lme4':
##
##     lmList

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```r
library(RLRsim)
library(CompRandFld)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(gee)
library(geepack)
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
#import the data
my.data <- read.csv("../final_data/frmgham2.csv")
head(my.data)
```

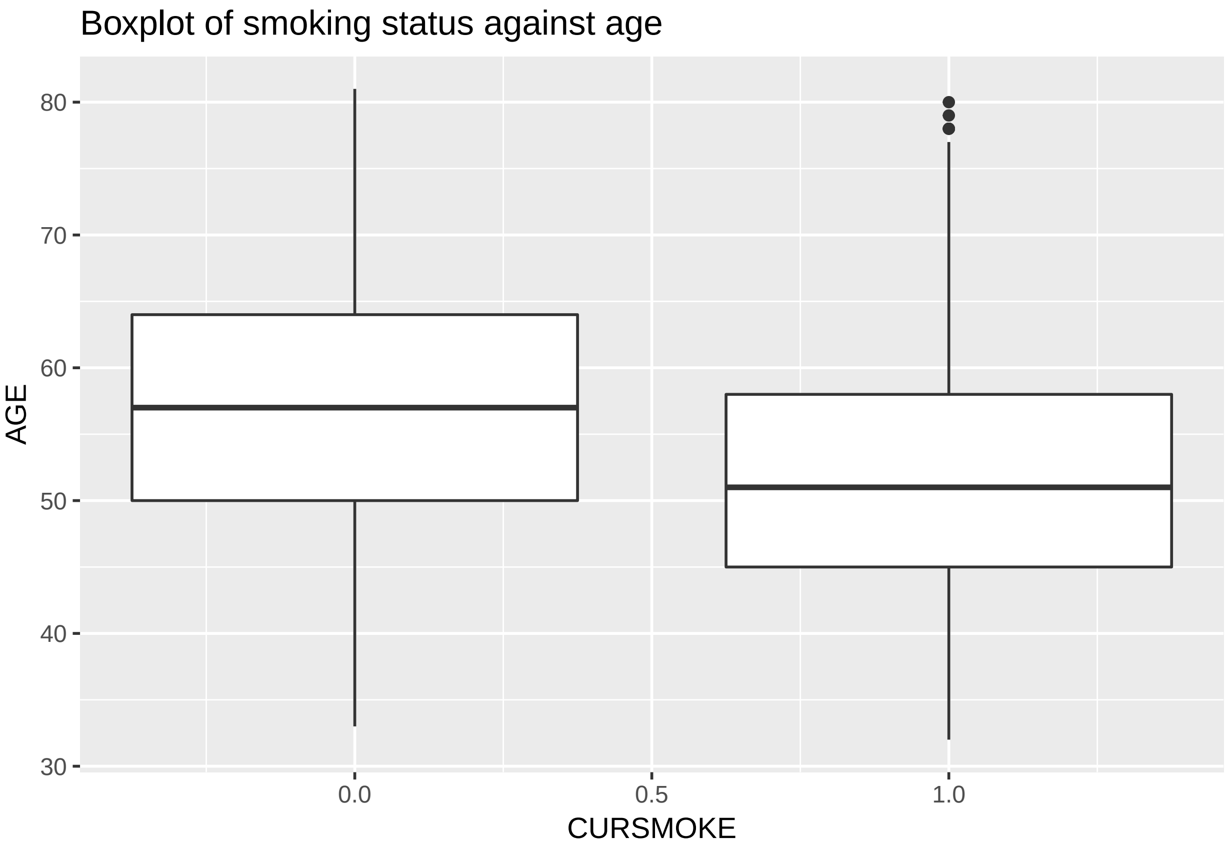
```
##   RANDID SEX TOTCHOL AGE SYSBP DIABP CURSMOKE CIGPDAY   BMI DIABETES
## 1   2448   1     195  39 106.0  70.0        0       0 26.97        0
## 2   2448   1     209  52 121.0  66.0        0       0    NA        0
## 3   6238   2     250  46 121.0  81.0        0       0 28.73        0
## 4   6238   2     260  52 105.0  69.5        0       0 29.43        0
## 5   6238   2     237  58 108.0  66.0        0       0 28.50        0
## 6   9428   1     245  48 127.5  80.0        1      20 25.34        0
##   BPMEDS HEARTRTE GLUCOSE educ PREVCHD PREVAP PREVMI PREVSTRK PREVHYP TIME
## 1      0       80      77    4       0      0      0        0       0    0
## 2      0       69      92    4       0      0      0        0       0 4628
## 3      0       95      76    2       0      0      0        0       0    0
## 4      0       80      86    2       0      0      0        0       0 2156
## 5      0       80      71    2       0      0      0        0       0 4344
## 6      0       75      70    1       0      0      0        0       0    0
##   PERIOD HDLC LDLC DEATH ANGINA HOSPMI MI_FCHD ANYCHD STROKE CVD HYPERTEN
## 1      1   NA   NA     0      0      1       1      1      0   1        0
## 2      3   31  178     0      0      1       1      1      0   1        0
## 3      1   NA   NA     0      0      0       0      0      0   0        0
## 4      2   NA   NA     0      0      0       0      0      0   0        0
## 5      3   54  141     0      0      0       0      0      0   0        0
## 6      1   NA   NA     0      0      0       0      0      0   0        0
##   TIMEAP TIMEMI TIMEMIFC TIMECHD TIMESTRK TIMECVD TIMEDTH TIMEHYP
## 1   8766   6438     6438    6438     8766    6438    8766    8766
## 2   8766   6438     6438    6438     8766    6438    8766    8766
## 3   8766   8766     8766    8766     8766    8766    8766    8766
## 4   8766   8766     8766    8766     8766    8766    8766    8766
## 5   8766   8766     8766    8766     8766    8766    8766    8766
## 6   8766   8766     8766    8766     8766    8766    8766    8766
```

```
summary(my.data$AGE)
```

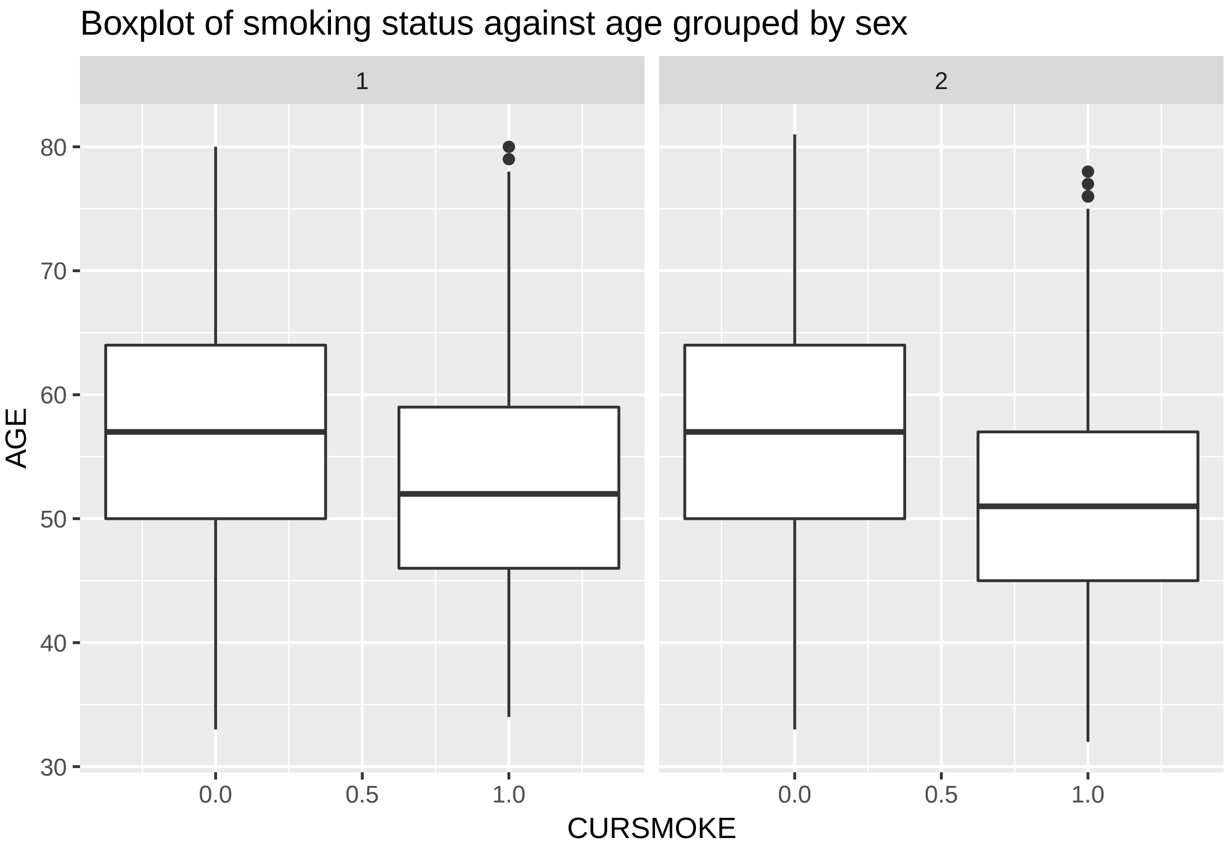
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   32.00   48.00   54.00   54.79   62.00   81.00
```

EDA of relationship between age and smoking status

```
boxplot.1 <- ggplot(data = my.data, aes(x = CURSMOKE, y = AGE, group=CURSMOKE))+
  geom_boxplot()+
  ggtitle('Boxplot of smoking status against age')
boxplot.1
```



```
boxplot.2 <- ggplot(data = my.data, aes(x = CURSMOKE, y = AGE, group=CURSMOKE))+
  geom_boxplot()+
  facet_wrap(~ SEX,ncol=2)+
  ggtitle('Boxplot of smoking status against age grouped by sex')
boxplot.2
```

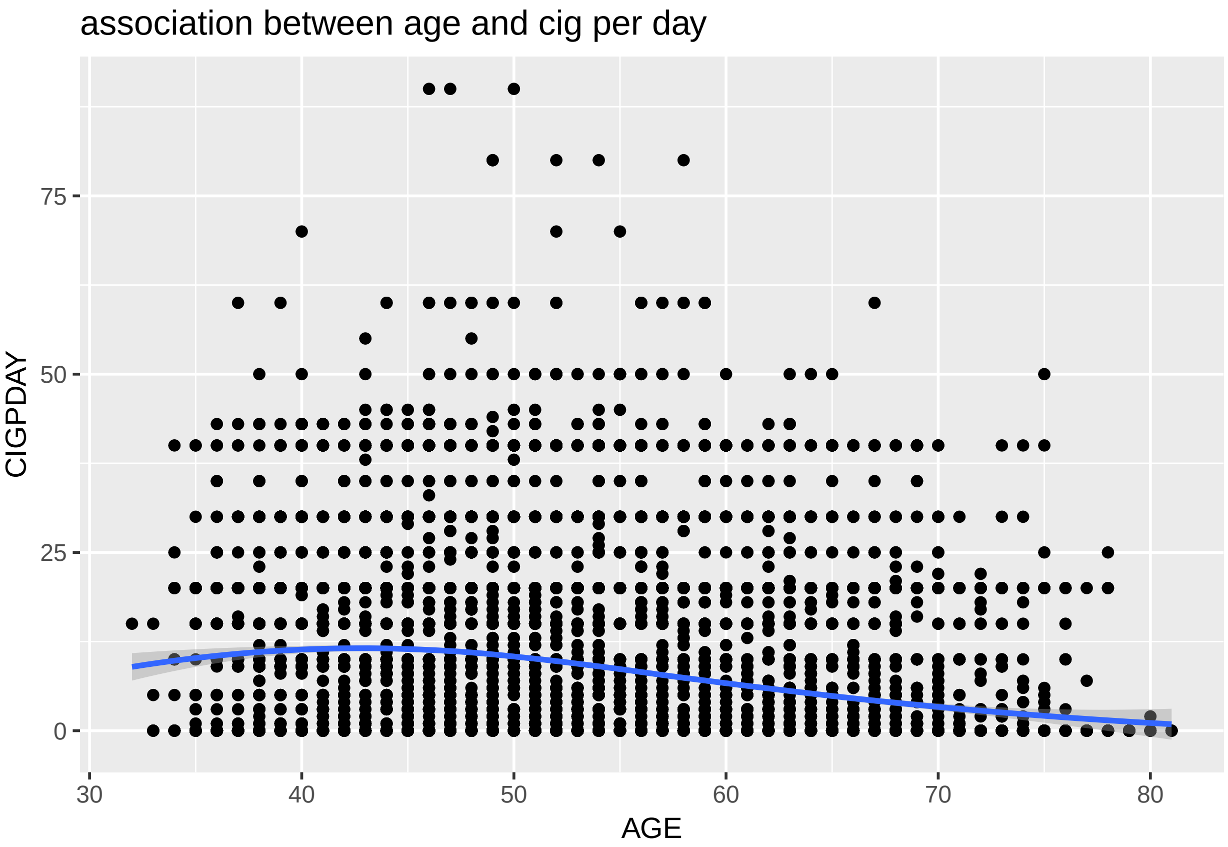


EDA of relationship between number of cigarettes smoked per day and age

```
spaghettiplot.1 <- ggplot(data=my.data, aes(x = AGE, y = CIGPDAY))+
  geom_point()+
  geom_smooth(method='loess')+
  ggtitle('association between age and cig per day')
spaghettiplot.1
```

```
## Warning: Removed 79 rows containing non-finite values (stat_smooth).
```

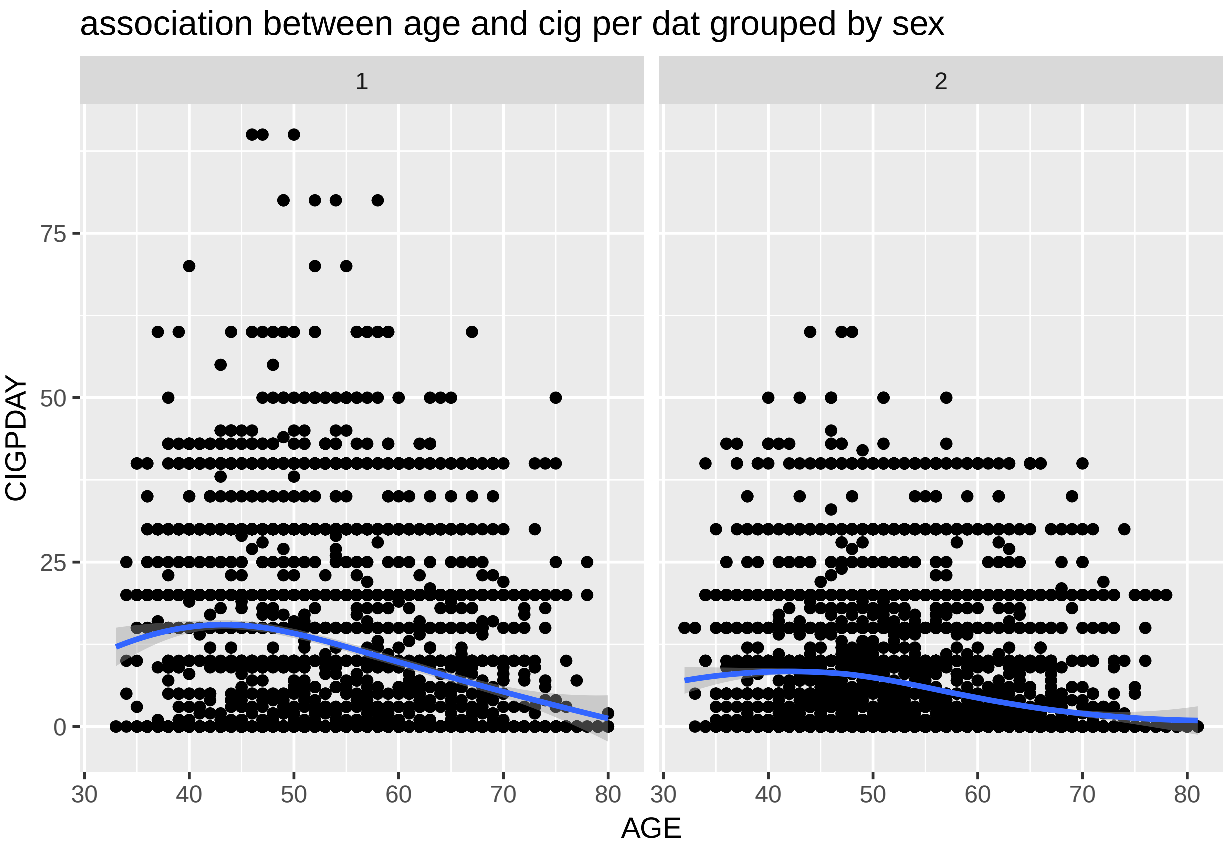
```
## Warning: Removed 79 rows containing missing values (geom_point).
```



```
spaghettiplot.2 <- ggplot(data=my.data, aes(x = AGE, y = CIGPDAY))+
  geom_point()+
  geom_smooth(method='loess')+
  facet_wrap(~SEX, ncol=2)+
  ggtitle('association between age and cig per dat grouped by sex')
spaghettiplot.2
```

```
## Warning: Removed 79 rows containing non-finite values (stat_smooth).
```

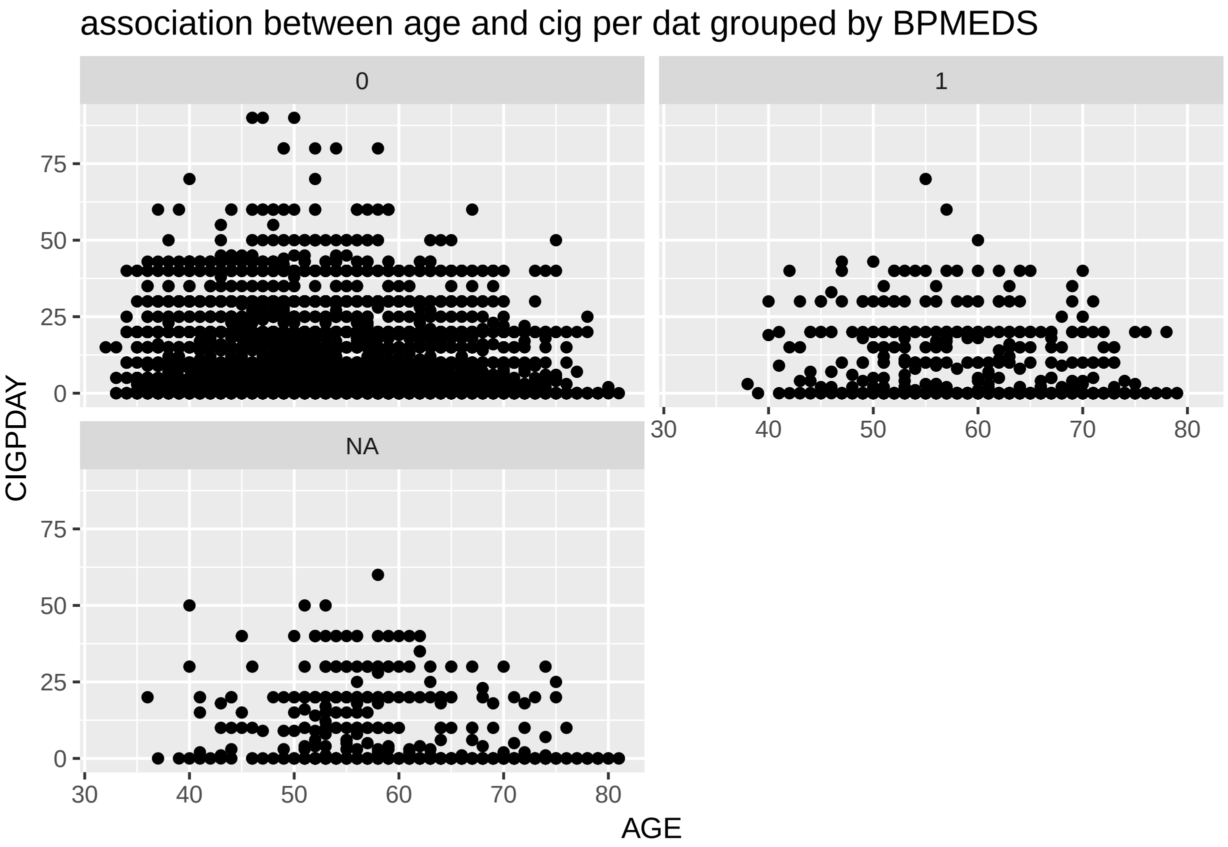
```
## Warning: Removed 79 rows containing missing values (geom_point).
```



BPMEDS adjusted

```
BPMEDS <- ggplot(data=my.data, aes(x = AGE, y = CIGPDAY, group = RANDID))+
  geom_point()+
  facet_wrap(~BPMEDS, ncol=2)+
  ggtitle('association between age and cig per dat grouped by BPMEDS')
BPMEDS
```

```
## Warning: Removed 79 rows containing missing values (geom_point).
```

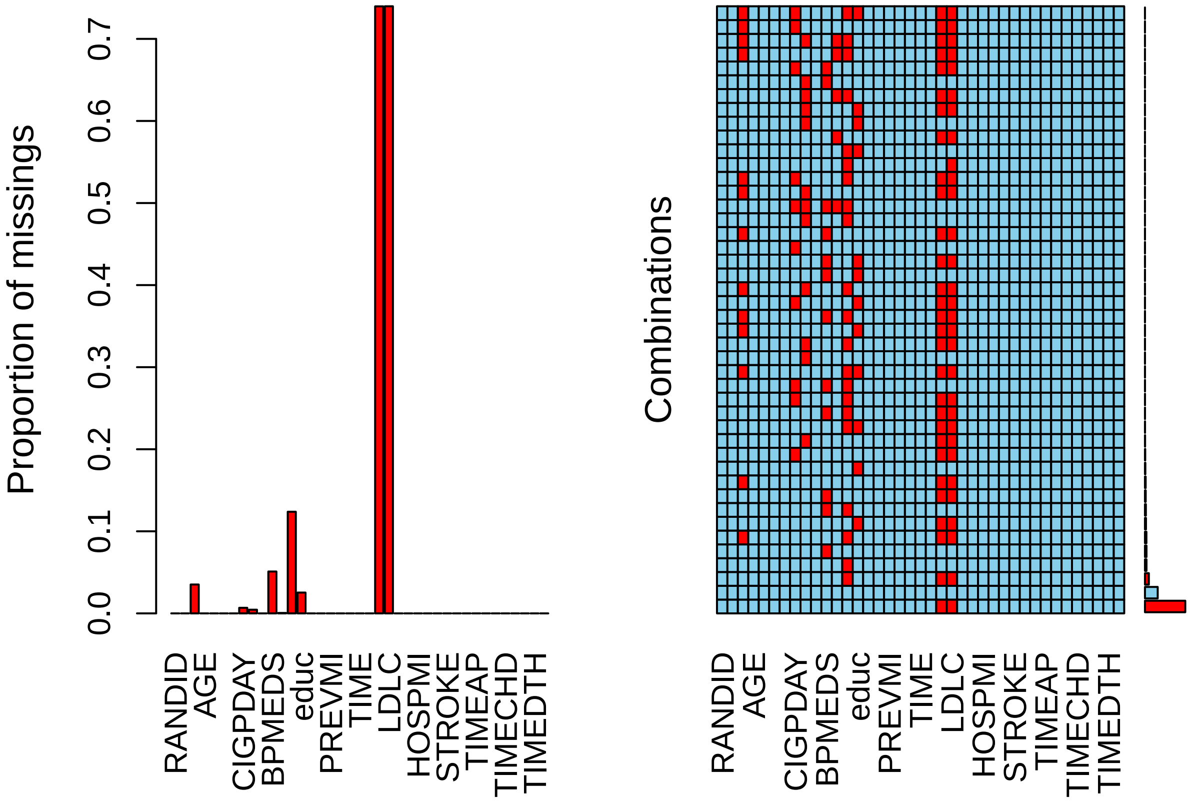


We find that there are some missing data in BPMEDS.

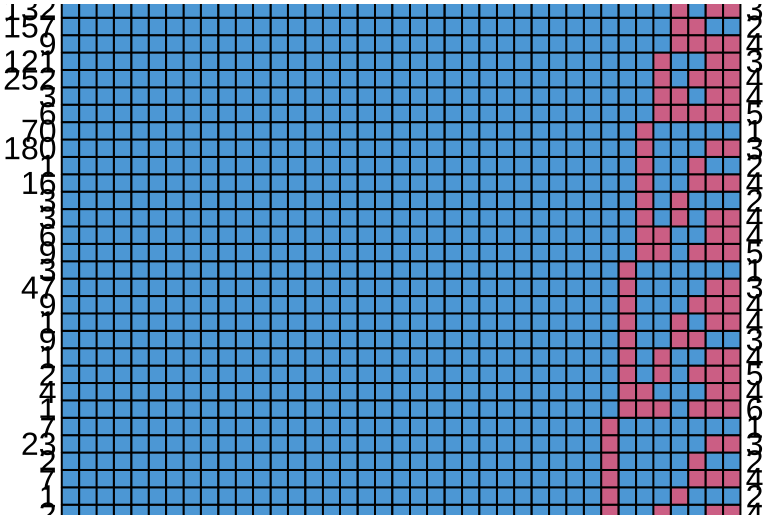
MISSING DATA ANALYSIS

```
VIM::aggr(my.data, prop=T, numbers=T)
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```



```
mice::md.pattern(my.data)
```



```
##      RANDID SEX AGE SYSBP DIABP CURSMOKE DIABETES PREVCHD PREVAP PREVMI
## 2236      1   1   1     1     1        1        1       1      1      1
## 7074      1   1   1     1     1        1        1       1      1      1
## 267       1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 683       1   1   1     1     1        1        1       1      1      1
## 267       1   1   1     1     1        1        1       1      1      1
## 132       1   1   1     1     1        1        1       1      1      1
## 157       1   1   1     1     1        1        1       1      1      1
## 9         1   1   1     1     1        1        1       1      1      1
## 121       1   1   1     1     1        1        1       1      1      1
## 252       1   1   1     1     1        1        1       1      1      1
## 3         1   1   1     1     1        1        1       1      1      1
## 6         1   1   1     1     1        1        1       1      1      1
## 70        1   1   1     1     1        1        1       1      1      1
## 180       1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 16        1   1   1     1     1        1        1       1      1      1
## 3         1   1   1     1     1        1        1       1      1      1
## 3         1   1   1     1     1        1        1       1      1      1
## 6         1   1   1     1     1        1        1       1      1      1
## 9         1   1   1     1     1        1        1       1      1      1
## 3         1   1   1     1     1        1        1       1      1      1
## 47        1   1   1     1     1        1        1       1      1      1
## 9         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 9         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 2         1   1   1     1     1        1        1       1      1      1
## 4         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 7         1   1   1     1     1        1        1       1      1      1
## 23        1   1   1     1     1        1        1       1      1      1
## 2         1   1   1     1     1        1        1       1      1      1
## 7         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 2         1   1   1     1     1        1        1       1      1      1
## 4         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 1         1   1   1     1     1        1        1       1      1      1
## 2         1   1   1     1     1        1        1       1      1      1
##           0   0   0     0     0        0        0       0      0      0
##       PREVSTRK PREVHYP TIME PERIOD DEATH ANGINA HOSPMI MI_FCHD ANYCHD
## 2236         1       1    1      1     1      1      1       1      1
## 7074         1       1    1      1     1      1      1       1      1
## 267          1       1    1      1     1      1      1       1      1
## 1            1       1    1      1     1      1      1       1      1
## 683          1       1    1      1     1      1      1       1      1
## 267          1       1    1      1     1      1      1       1      1
## 132          1       1    1      1     1      1      1       1      1
## 157          1       1    1      1     1      1      1       1      1
## 9            1       1    1      1     1      1      1       1      1
## 121          1       1    1      1     1      1      1       1      1
## 252          1       1    1      1     1      1      1       1      1
## 3            1       1    1      1     1      1      1       1      1
## 6            1       1    1      1     1      1      1       1      1
## 70           1       1    1      1     1      1      1       1      1
## 180          1       1    1      1     1      1      1       1      1
## 1            1       1    1      1     1      1      1       1      1
## 16           1       1    1      1     1      1      1       1      1
## 3          1       1    1      1     1      1      1       1      1
## 3          1       1    1      1     1      1      1       1      1
## 6          1       1    1      1     1      1      1       1      1
## 9          1       1    1      1     1      1      1       1      1
## 3          1       1    1      1     1      1      1       1      1
## 47         1       1    1      1     1      1      1       1      1
## 9          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 9          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 2          1       1    1      1     1      1      1       1      1
## 4          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 7          1       1    1      1     1      1      1       1      1
## 23         1       1    1      1     1      1      1       1      1
## 2          1       1    1      1     1      1      1       1      1
## 7          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 2          1       1    1      1     1      1      1       1      1
## 4          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 1          1       1    1      1     1      1      1       1      1
## 2          1       1    1      1     1      1      1       1      1
##            0       0    0      0     0      0      0       0      0
##      STROKE CVD HYPERTEN TIMEAP TIMEMI TIMEMIFC TIMECHD TIMESTRK TIMECVD
## 2236      1   1        1      1      1        1       1        1       1
## 7074      1   1        1      1      1        1       1        1       1
## 267       1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 683       1   1        1      1      1        1       1        1       1
## 267       1   1        1      1      1        1       1        1       1
## 132       1   1        1      1      1        1       1        1       1
## 157       1   1        1      1      1        1       1        1       1
## 9         1   1        1      1      1        1       1        1       1
## 121       1   1        1      1      1        1       1        1       1
## 252       1   1        1      1      1        1       1        1       1
## 3         1   1        1      1      1        1       1        1       1
## 6         1   1        1      1      1        1       1        1       1
## 70        1   1        1      1      1        1       1        1       1
## 180       1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 16        1   1        1      1      1        1       1        1       1
## 3         1   1        1      1      1        1       1        1       1
## 3         1   1        1      1      1        1       1        1       1
## 6         1   1        1      1      1        1       1        1       1
## 9         1   1        1      1      1        1       1        1       1
## 3         1   1        1      1      1        1       1        1       1
## 47        1   1        1      1      1        1       1        1       1
## 9         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 9         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 2         1   1        1      1      1        1       1        1       1
## 4         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 7         1   1        1      1      1        1       1        1       1
## 23        1   1        1      1      1        1       1        1       1
## 2         1   1        1      1      1        1       1        1       1
## 7         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 2         1   1        1      1      1        1       1        1       1
## 4         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 1         1   1        1      1      1        1       1        1       1
## 2         1   1        1      1      1        1       1        1       1
##           0   0        0      0      0        0       0        0       0
##      TIMEDTH TIMEHYP HEARTRTE BMI CIGPDAY educ TOTCHOL BPMEDS GLUCOSE HDLC
## 2236       1       1        1   1       1    1       1      1       1    1
## 7074       1       1        1   1       1    1       1      1       1    0
## 267        1       1        1   1       1    1       1      1       0    1
## 1          1       1        1   1       1    1       1      1       0    1
## 683        1       1        1   1       1    1       1      1       0    0
## 267        1       1        1   1       1    1       1      0       1    1
## 132        1       1        1   1       1    1       1      0       1    0
## 157        1       1        1   1       1    1       1      0       0    1
## 9          1       1        1   1       1    1       1      0       0    0
## 121        1       1        1   1       1    1       0      1       1    0
## 252        1       1        1   1       1    1       0      1       0    0
## 3          1       1        1   1       1    1       0      0       1    0
## 6          1       1        1   1       1    1       0      0       0    0
## 70         1       1        1   1       1    0       1      1       1    1
## 180        1       1        1   1       1    0       1      1       1    0
## 1          1       1        1   1       1    0       1      1       0    1
## 16         1       1        1   1       1    0       1      1       0    0
## 3          1       1        1   1       1    0       1      0       1    1
## 3          1       1        1   1       1    0       1      0       1    0
## 6          1       1        1   1       1    0       0      1       1    0
## 9          1       1        1   1       1    0       0      1       0    0
## 3          1       1        1   1       0    1       1      1       1    1
## 47         1       1        1   1       0    1       1      1       1    0
## 9          1       1        1   1       0    1       1      1       0    0
## 1          1       1        1   1       0    1       1      0       1    0
## 9          1       1        1   1       0    1       1      0       0    1
## 1          1       1        1   1       0    1       0      1       1    0
## 2          1       1        1   1       0    1       0      1       0    0
## 4          1       1        1   1       0    0       1      1       1    0
## 1          1       1        1   1       0    0       0      1       0    0
## 7          1       1        1   0       1    1       1      1       1    1
## 23         1       1        1   0       1    1       1      1       1    0
## 2          1       1        1   0       1    1       1      1       0    1
## 7          1       1        1   0       1    1       1      1       0   0
## 1          1       1        1   0       1    1       1      0       1   1
## 2          1       1        1   0       1    1       0      1       1   0
## 4          1       1        1   0       1    1       0      1       0   0
## 1          1       1        1   0       1    0       1      1       1   1
## 1          1       1        1   0       1    0       1      1       1   0
## 1          1       1        0   1       1    1       1      1       1   0
## 1          1       1        0   1       1    1       0      1       0   0
## 1          1       1        0   0       1    1       1      1       0   0
## 1          1       1        0   0       1    1       0      1       0   0
## 2          1       1        0   0       0    1       1      0       0   1
##            0       0        6  52      79  295     409    593    1440 8600
##       LDLC
## 2236    1     0
## 7074    0     2
## 267     1     1
## 1       0     2
## 683     0     3
## 267     1     1
## 132     0     3
## 157     1     2
## 9       0     4
## 121     0     3
## 252     0     4
## 3       0     4
## 6       0     5
## 70      1     1
## 180     0     3
## 1       1     2
## 16      0     4
## 3       1     2
## 3       0     4
## 6       0     4
## 9       0     5
## 3       1     1
## 47      0     3
## 9       0     4
## 1       0     4
## 9       1     3
## 1       0     4
## 2       0     5
## 4       0     4
## 1       0     6
## 7       1     1
## 23      0     3
## 2       1     2
## 7       0     4
## 1       1     2
## 2       0     4
## 4       0     5
## 1       1     2
## 1       0     4
## 1       0     3
## 1       0     5
## 1       0     5
## 1       0     6
## 2       1     5
##      8601 20075
```

model of EDA:

(1) Is there a relationship between age and smoking status? Does this relationship differ by sex?

```
model.q1 <- gee(CURSMOKE ~ AGE + as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                 + PREVHYP + TIMEDTH,
                 id = RANDID,
                 data = my.data,
                 family=binomial,
                 corstr = "unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##       (Intercept)               AGE   as.factor(SEX)2 as.factor(educ)2
##      5.2810549757     -0.0570318456     -0.7246241605     0.0503157656
## as.factor(educ)3 as.factor(educ)4               BMI         DIABETES
##     -0.2263406725     -0.2104931713     -0.0940121564    -0.1114780579
##          HEARTRTE           PREVCHD          PREVSTRK          PREVHYP
##      0.0173911758     -0.0569530253     -0.2633153119    -0.2252617826
##           TIMEDTH
##     -0.0001021991
```

(2) Is there a relationship between the number of cigarettes smoked per day and age? Does this relationship differ by sex?

```
model.q2 <- lmer(CIGPDAY ~ AGE + as.factor(SEX) + as.factor(educ) + (1|RANDID),
                  data = my.data)
summary(model.q2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: CIGPDAY ~ AGE + as.factor(SEX) + as.factor(educ) + (1 | RANDID)
##    Data: my.data
##
## REML criterion at convergence: 81293.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5718 -0.2999 -0.0874  0.1423  6.6072
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  RANDID   (Intercept) 97.85    9.892
##  Residual             36.59    6.049
## Number of obs: 11258, groups:  RANDID, 4320
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      22.40907    0.67179  33.357
## AGE              -0.19323    0.01054 -18.341
## as.factor(SEX)2  -6.23951    0.33015 -18.899
## as.factor(educ)2  1.13964    0.39174   2.909
## as.factor(educ)3 -0.52383    0.47294  -1.108
## as.factor(educ)4 -0.74875    0.53854  -1.390
##
## Correlation of Fixed Effects:
##              (Intr) AGE    a.(SEX as.()2 as.()3
## AGE          -0.890
## as.fc(SEX)2 -0.246 -0.021
## as.fctr(d)2 -0.336  0.127 -0.055
## as.fctr(d)3 -0.236  0.071 -0.092  0.351
## as.fctr(d)4 -0.251  0.065  0.076  0.300  0.243
```

If we think cig per day as count data, it follows poisson distribution. Then we can fit GEE model as well:

```
model.q2_1 <- gee(CIGPDAY ~ AGE + as.factor(SEX) + as.factor(educ),
                  data = my.data,
                  id = RANDID,
                  family=poisson,
                  corstr = "unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)              AGE  as.factor(SEX)2 as.factor(educ)2
##       4.31931466      -0.03504695      -0.70838799       0.07722876
## as.factor(educ)3 as.factor(educ)4
##      -0.09995154      -0.10845836
```

model including age and sex

(1) The relationship between current smoking status and systolic blood pressure.

```
model.p1 <- gee(SYSBP ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                id = RANDID,
                data = my.data,
                na.action = "na.omit",
                corstr = "unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##       89.0711437       -2.0414490        0.8846592        1.2687106
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##       -0.3063701       -2.9072926       -4.0416336
```

```
summary(model.p1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = SYSBP ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -56.456788 -14.552902  -2.892551  11.072266 147.850985
##
##
## Coefficients:
##                    Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)      90.3389980 1.51292615 59.711439  1.41867735 63.678326
## CURSMOKE         -1.5149317 0.46476110 -3.259592  0.45715635 -3.313815
## AGE               0.8653711 0.02377869 36.392712  0.02294929 37.707974
## as.factor(SEX)2   1.4262673 0.54709991  2.606959  0.57967657  2.460454
## as.factor(educ)2 -0.7481716 0.64772860 -1.155070  0.69947330 -1.069621
## as.factor(educ)3 -3.1296696 0.77855475 -4.019845  0.82616759 -3.788178
## as.factor(educ)4 -4.3249661 0.88481499 -4.887989  0.91529976 -4.725191
##
## Estimated Scale Parameter:  438.277
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5844383 0.3901881
## [2,] 0.5844383 1.0000000 0.4769521
## [3,] 0.3901881 0.4769521 1.0000000
```

```
#fit exchangeable model
model.p1.1 <- gee(SYSBP ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                id = RANDID,
                data = my.data,
                na.action = "na.omit",
                corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##       89.0711437       -2.0414490        0.8846592        1.2687106
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##       -0.3063701       -2.9072926       -4.0416336
```

```
summary(model.p1.1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
```

```
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = SYSBP ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "exchangeable")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -56.668486 -14.716778  -3.057859  10.920645 147.715717
##
##
## Coefficients:
##                    Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)      91.2341128 1.42488120 64.029277  1.38513683 65.866498
## CURSMOKE         -1.5600368 0.45243405 -3.448098  0.45462033 -3.431516
## AGE               0.8526327 0.02201558 38.728610  0.02223794 38.341350
## as.factor(SEX)2   1.4816755 0.57044893  2.597385  0.58269918  2.542779
## as.factor(educ)2 -0.8753434 0.67387802 -1.298964  0.70286876 -1.245387
## as.factor(educ)3 -3.2052279 0.81201172 -3.947268  0.83228281 -3.851128
## as.factor(educ)4 -4.3914410 0.92370116 -4.754179  0.91574686 -4.795475
##
## Estimated Scale Parameter:  438.4663
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5912929 0.5912929
## [2,] 0.5912929 1.0000000 0.5912929
## [3,] 0.5912929 0.5912929 1.0000000
```

Compare the naive SE and robust SE we can see that exchangeable model is reasonable.

(2) The relationship between current smoking status and diastolic blood pressure.

```r
model.p2 <- gee(DIABP ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                id = RANDID,
                data = my.data,
                na.action = "na.omit",
                corstr = "unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##      81.65916072      -1.76253559       0.06074128      -1.49015111
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##      -0.07960126      -0.96361582      -1.38809040
```

```r
summary(model.p2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = DIABP ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -53.815474  -7.934818  -1.165221   6.612862  66.053172
##
##
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept)      82.35453144 0.83838302 98.2301994  0.81533606 101.0068553
## CURSMOKE         -1.25577390 0.25866115 -4.8548996  0.26295278  -4.7756631
## AGE               0.04492311 0.01318527  3.4070687  0.01301711   3.4510820
## as.factor(SEX)2  -1.37262875 0.29612384 -4.6353199  0.31463807  -4.3625640
## as.factor(educ)2 -0.20708873 0.35071588 -0.5904743  0.38142348  -0.5429365
## as.factor(educ)3 -1.00717792 0.42133624 -2.3904374  0.43642586  -2.3077870
## as.factor(educ)4 -1.50398297 0.47863392 -3.1422407  0.49916228  -3.0130140
##
## Estimated Scale Parameter:  134.5107
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5561530 0.3490568
## [2,] 0.5561530 1.0000000 0.3838573
## [3,] 0.3490568 0.3838573 1.0000000
```

```
#fit exchangeable model
model.p2.2 <- gee(DIABP ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                id = RANDID,
                data = my.data,
                na.action = "na.omit",
                corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##      81.65916072      -1.76253559       0.06074128      -1.49015111
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##      -0.07960126      -0.96361582      -1.38809040
```

```
summary(model.p2.2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = DIABP ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "exchangeable")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -53.712361  -8.154391  -1.371213   6.476612  66.065595
##
##
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.    Robust z
## (Intercept)      83.83108830 0.80779562 103.777597  0.80310636 104.3835449
## CURSMOKE         -1.26041472 0.25595570  -4.924347  0.26302503  -4.7919954
## AGE               0.02165278 0.01258223   1.720901  0.01276702   1.6959931
## as.factor(SEX)2  -1.32576606 0.30843753  -4.298329  0.31690863  -4.1834331
## as.factor(educ)2 -0.36871011 0.36456603  -1.011367  0.38394652  -0.9603163
## as.factor(educ)3 -1.10545681 0.43896055  -2.518351  0.44024370  -2.5110111
## as.factor(educ)4 -1.54781045 0.49907624  -3.101351  0.50285473  -3.0780469
##
## Estimated Scale Parameter:  134.6896
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5293801 0.5293801
## [2,] 0.5293801 1.0000000 0.5293801
## [3,] 0.5293801 0.5293801 1.0000000
```

(2) The relationship between current smoking status and serum total cholesterol.

```
model.p3 <- gee(TOTCHOL ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                data = my.data,
                id = RANDID,
                na.action = "na.omit",
                corstr = "unstructured")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##      192.6834832        1.6590353        0.7348726       12.0862772
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##        0.9107182        2.0078913        0.6363909
```

```
summary(model.p3)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = TOTCHOL ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "unstructured")
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -148.090340  -29.650053   -2.060511   27.146185  399.388552
##
##
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)      192.2062030 3.16914487 60.6492322  3.14988352 61.0200986
## CURSMOKE           2.2459560 0.97321653  2.3077660  0.99901818  2.2481633
## AGE                0.7164736 0.04953822 14.4630466  0.04903112 14.6126292
## as.factor(SEX)2   12.7309870 1.18566784 10.7373976  1.22285054 10.4109101
## as.factor(educ)2   1.0663777 1.40446547  0.7592766  1.49560440  0.7130079
## as.factor(educ)3   2.1301734 1.68938577  1.2609159  1.78623276  1.1925509
## as.factor(educ)4   0.9692929 1.91708122  0.5056087  1.85486438  0.5225681
##
## Estimated Scale Parameter:  1956.654
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.6715907 0.4279521
## [2,] 0.6715907 1.0000000 0.4618901
## [3,] 0.4279521 0.4618901 1.0000000
```

```
#fit exchangeable model
model.p3.3 <- gee(TOTCHOL ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                data = my.data,
                id = RANDID,
                na.action = "na.omit",
                corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
##      (Intercept)         CURSMOKE              AGE  as.factor(SEX)2
##      192.6834832        1.6590353        0.7348726       12.0862772
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##        0.9107182        2.0078913        0.6363909
```

```
summary(model.p3.3)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
```

```
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = TOTCHOL ~ CURSMOKE + AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, na.action = "na.omit", corstr = "exchangeable")
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -145.536794  -30.920172   -3.262033   26.189308  396.017266
##
##
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)      207.8996413 2.98275961 69.7004346  3.06646447 67.7978314
## CURSMOKE           2.4215646 0.94657181  2.5582471  0.99702199  2.4287976
## AGE                0.4588722 0.04588231 10.0010701  0.04755714  9.6488594
## as.factor(SEX)2   12.5161015 1.23707672 10.1174820  1.22507684 10.2165849
## as.factor(educ)2  -0.2971410 1.46200503 -0.2032421  1.49968224 -0.1981360
## as.factor(educ)3   1.2072445 1.76174866  0.6852535  1.78958428  0.6745949
## as.factor(educ)4   0.3065030 2.00272497  0.1530430  1.86613064  0.1642452
##
## Estimated Scale Parameter:  1963.143
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.6430173 0.6430173
## [2,] 0.6430173 1.0000000 0.6430173
## [3,] 0.6430173 0.6430173 1.0000000
```

Using mixed effect model using cig per day instead of smoking status:

```
#saturated model
model.saturated <- lmer(CIGPDAY ~ as.factor(SEX) + AGE
                            + BPMEDS + as.factor(educ)
                            + TOTCHOL + BMI + GLUCOSE + DIABETES + HEARTRTE + PREVAP
                            + PREVCHD + PREVMI + PREVSTRK +STROKE+ PREVHYP + (1|RANDID),
                              na.action = 'na.omit',
                              data = my.data)
summary(model.saturated)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## CIGPDAY ~ as.factor(SEX) + AGE + BPMEDS + as.factor(educ) + TOTCHOL +
##     BMI + GLUCOSE + DIABETES + HEARTRTE + PREVAP + PREVCHD +
##     PREVMI + PREVSTRK + STROKE + PREVHYP + (1 | RANDID)
##    Data: my.data
##
## REML criterion at convergence: 67692.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4444 -0.3150 -0.1078  0.2067  6.2548
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  RANDID   (Intercept) 95.78    9.787
##  Residual             35.77    5.981
## Number of obs: 9310, groups:  RANDID, 4213
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)        23.237579   1.393589  16.675
## as.factor(SEX)2    -7.062484   0.339110 -20.827
## AGE                -0.178024   0.013735 -12.962
## BPMEDS              0.147068   0.340662   0.432
## as.factor(educ)2    0.602137   0.400578   1.503
## as.factor(educ)3   -0.984441   0.482907  -2.039
## as.factor(educ)4   -1.015103   0.548951  -1.849
## TOTCHOL             0.011407   0.002557   4.461
## BMI                -0.313257   0.035227  -8.893
## GLUCOSE            -0.010167   0.004203  -2.419
## DIABETES           -0.254321   0.570383  -0.446
## HEARTRTE            0.070622   0.008232   8.579
## PREVAP             -3.058426   0.985947  -3.102
## PREVCHD             0.948740   1.050262   0.903
## PREVMI             -2.594434   0.894896  -2.899
## PREVSTRK           -1.029740   0.917968  -1.122
## STROKE              0.943808   0.586941   1.608
## PREVHYP            -0.205972   0.240530  -0.856

##
## Correlation matrix not shown by default, as p = 18 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
#using variables that selected
model.mixed2 <- lmer(CIGPDAY~ AGE + as.factor(SEX) + SYSBP
                      + DIABP + TOTCHOL + as.factor(educ)
                      + (1|RANDID),
                      data = my.data,
                      na.action = "na.omit")
summary(model.mixed2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## CIGPDAY ~ AGE + as.factor(SEX) + SYSBP + DIABP + TOTCHOL + as.factor(educ) +
##     (1 | RANDID)
##    Data: my.data
##
## REML criterion at convergence: 78607.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6343 -0.3025 -0.1015  0.1642  6.6280
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev.
##  RANDID   (Intercept) 97.30    9.864
##  Residual             36.52    6.043
## Number of obs: 10868, groups:  RANDID, 4306
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)      21.750845   1.049612  20.723
## AGE              -0.199583   0.012148 -16.429
## as.factor(SEX)2  -6.448164   0.332812 -19.375
## SYSBP            -0.001364   0.006670  -0.205
## DIABP            -0.018833   0.011514  -1.636
## TOTCHOL           0.011959   0.002373   5.039
## as.factor(educ)2  1.055542   0.392720   2.688
## as.factor(educ)3 -0.625323   0.474072  -1.319
## as.factor(educ)4 -0.827192   0.539360  -1.534
##
## Correlation of Fixed Effects:
##             (Intr) AGE    a.(SEX SYSBP  DIABP  TOTCHO as.()2 as.()3
## AGE         -0.490
## as.fc(SEX)2 -0.144  0.019
## SYSBP        0.023 -0.460 -0.064
## DIABP       -0.452  0.314  0.076 -0.679
## TOTCHOL     -0.368 -0.095 -0.094  0.047 -0.154
## as.fctr(d)2 -0.225  0.110 -0.055  0.008  0.001  0.002
## as.fctr(d)3 -0.161  0.051 -0.092  0.028 -0.004 -0.007  0.351
## as.fctr(d)4 -0.175  0.044  0.074  0.032 -0.005 -0.002  0.301  0.244
```