LDA final project

Kaitlin Maciejewski, Morgan de Ferrante, Bingnan Li, Volha Tryputsen

Contents

Exploratory

Smoking vs. Age, Sex

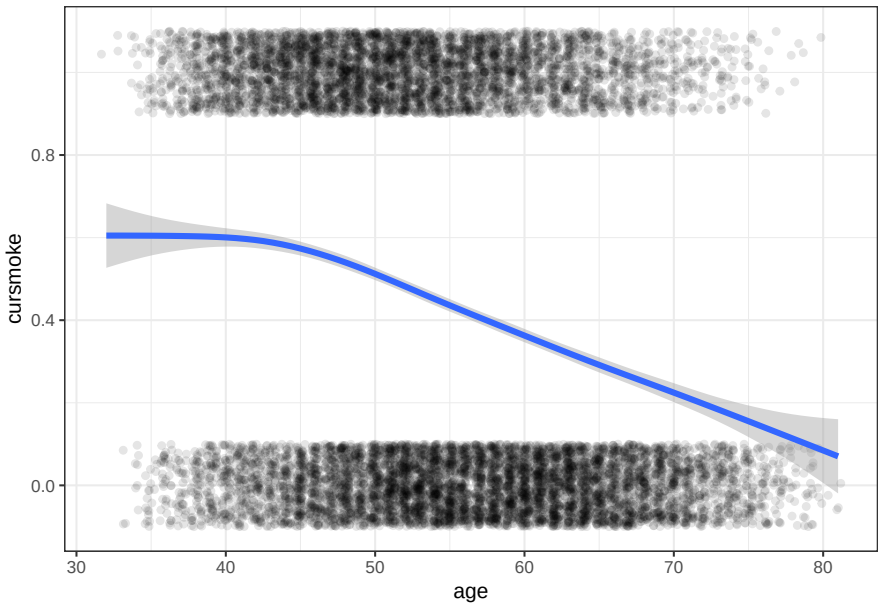
(1): Smoking ~ age, sex

Is there a relationship between age and smoking status?

ANS: Yes, the proportion of smokers decreases with the age.

```
data %>%
  select(cursmoke, age) %>%
```

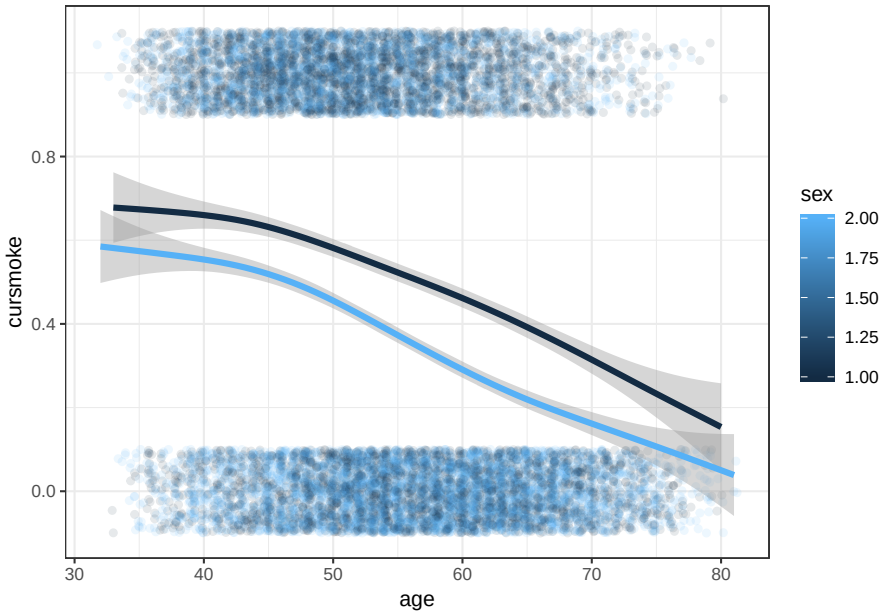
```
  ggplot(aes(x = age, y = cursmoke)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



Does this relationship differ by sex?

ANS: There is a higher proportion of smoker among men compared to women as both age ,but there is no interaction between age and sex.

```
data %>%
  select(cursmoke, age, sex) %>%
  ggplot(aes(x = age, y = cursmoke, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



(2) number of cigarettes ~ age, sex

Is there a relationship between the number of cigarettes smoked per day and age?

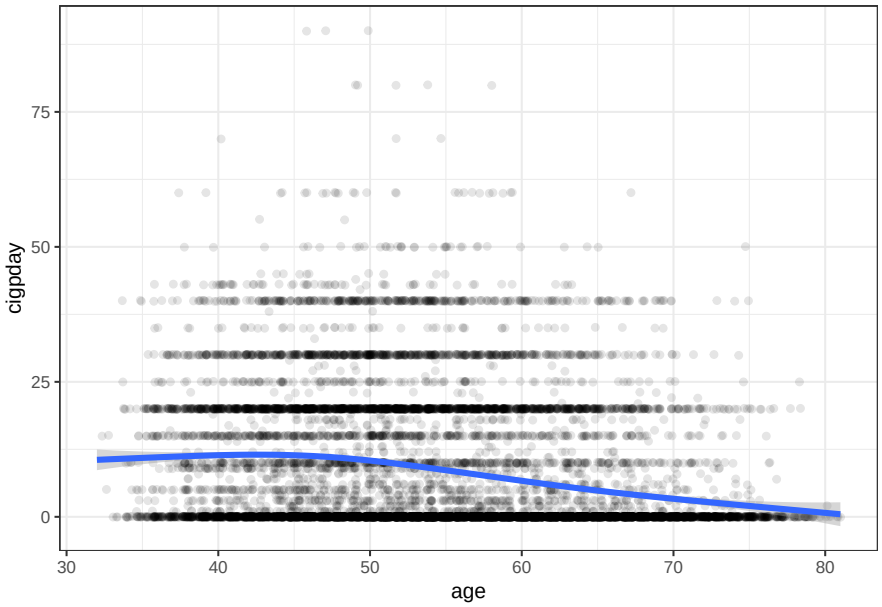
ANS: Yes, number of sigarets smoked per day stays constant for 30-50 years old and decreases with age after 50 years old.

All

```
data %>%
  select(cigpday, age) %>%
  ggplot(aes(x = age, y = cigpday)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```

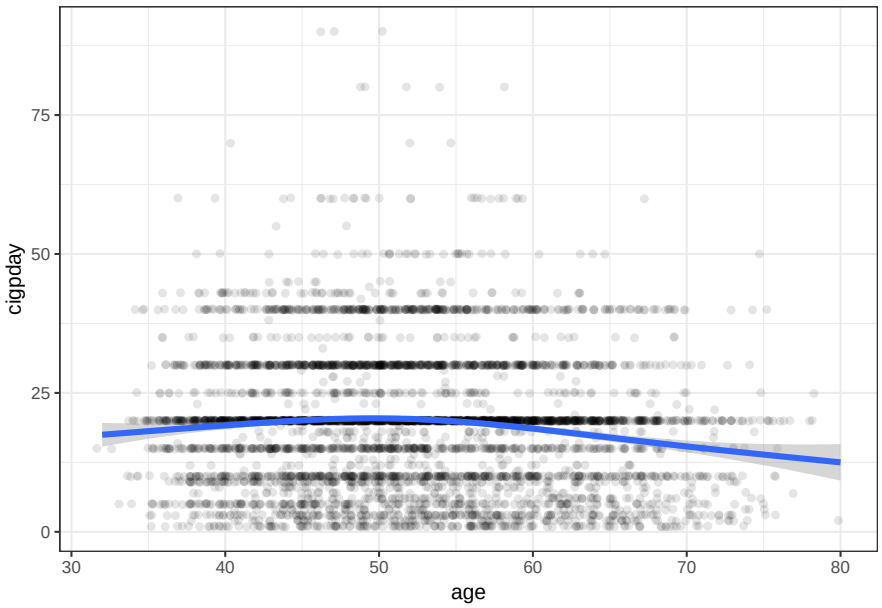
```
## Warning: Removed 79 rows containing non-finite values (stat_smooth).
```

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



Smokers only

```
data %>%
  select(cigpday, age) %>%
  filter(cigpday > 0) %>%
  ggplot(aes(x = age, y = cigpday)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



Does this relationship differ by sex?

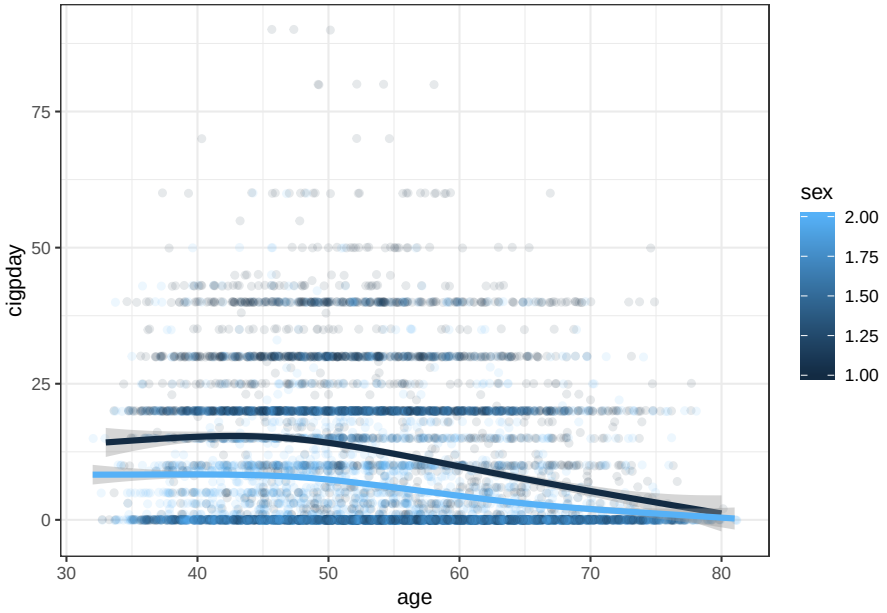
ANS: There is sex effect (men smoke higer number of sigarets per day than women across age), but there is no sex and age interaction.

All

```
data %>%
  select(cigpday, age, sex) %>%
  ggplot(aes(x = age, y = cigpday, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```

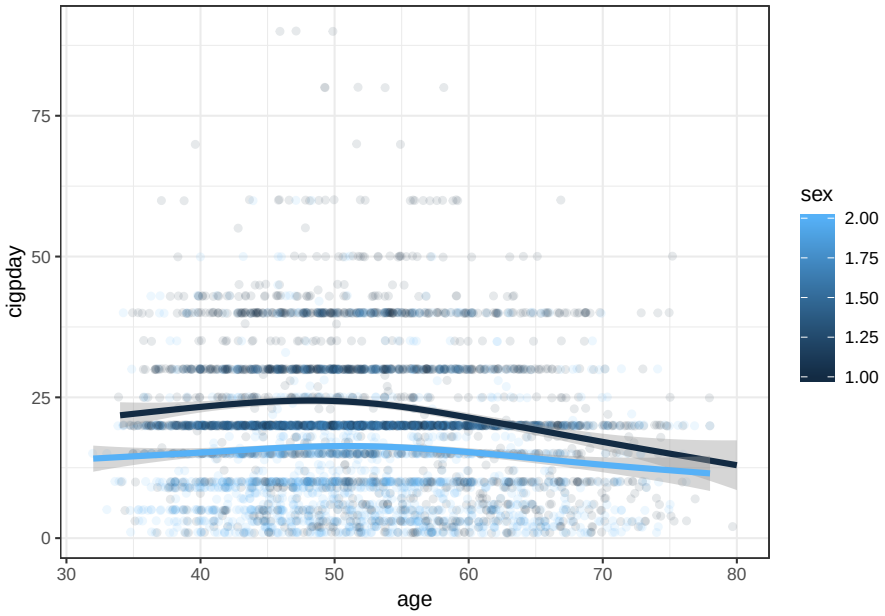
```
## Warning: Removed 79 rows containing non-finite values (stat_smooth).
```

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



Smokers only

```
data %>%
  select(cigpday, age, sex) %>%
  filter(cigpday > 0) %>%
  ggplot(aes(x = age, y = cigpday, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



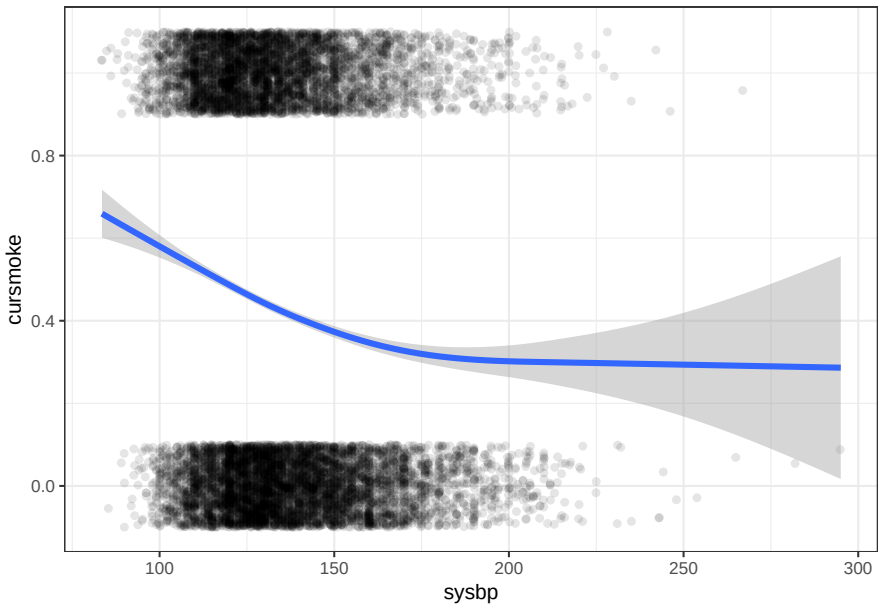
Smoking vs. health outcomes

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

smoking ~ sysbp

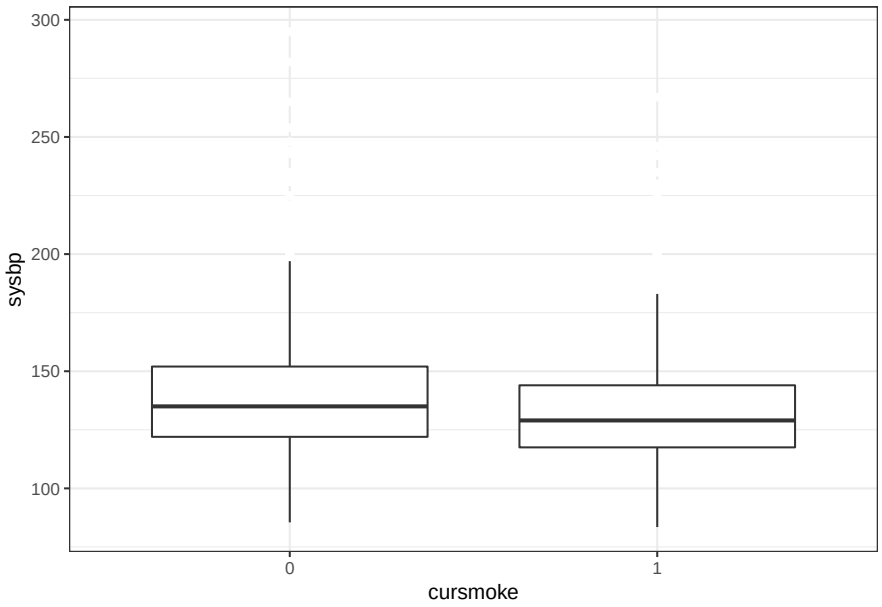
ANS: Proportion of smokers decreases with increase of systolic blood presure

```
data %>%
  select(cursmoke, sysbp) %>%
  ggplot(aes(x = sysbp, y = cursmoke)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



ANS: slightly higher sysbp for non-smokers

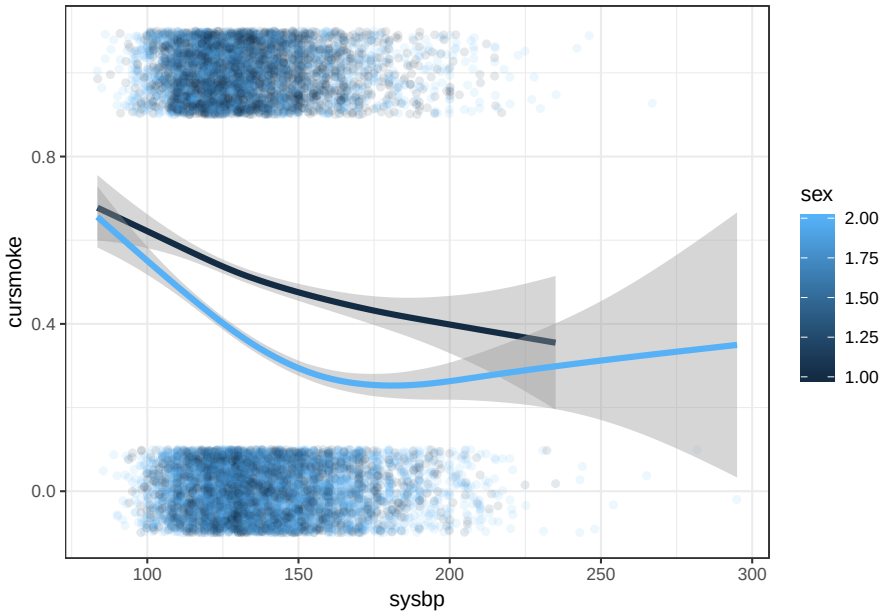
```
data %>%
  select(cursmoke, sysbp) %>%
  mutate(cursmoke = factor(cursmoke)) %>%
  ggplot(aes(y = sysbp, x = cursmoke)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```



smoking ~ sysbp, sex

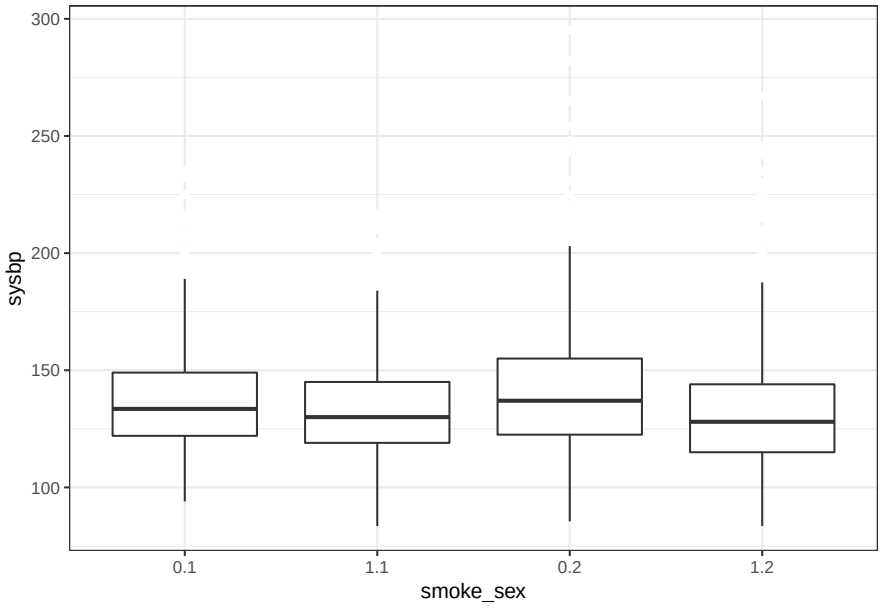
ANS: Proportion of smokers decreases with increase of systolic blood presure; the proportion is higher for men (sex effect).

```
data %>%
  select(cursmoke, sysbp, sex) %>%
  ggplot(aes(x = sysbp, y = cursmoke, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



ANS: no differences in sysbp between male and female smokers and non-smokers

```
data %>%
  select(cursmoke, sex, sysbp) %>%
  mutate(cursmoke = factor(cursmoke),
           smoke_sex = interaction(cursmoke, sex)) %>%
  ggplot(aes(y = sysbp, x = smoke_sex)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```

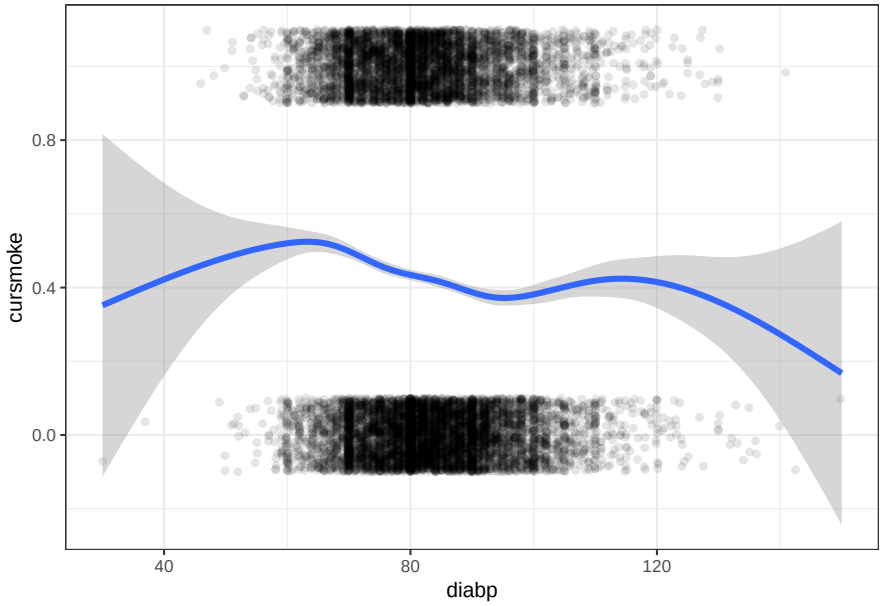


(2) The relationship between current smoking status and diastolic blood pressure.

smoking ~ diabp

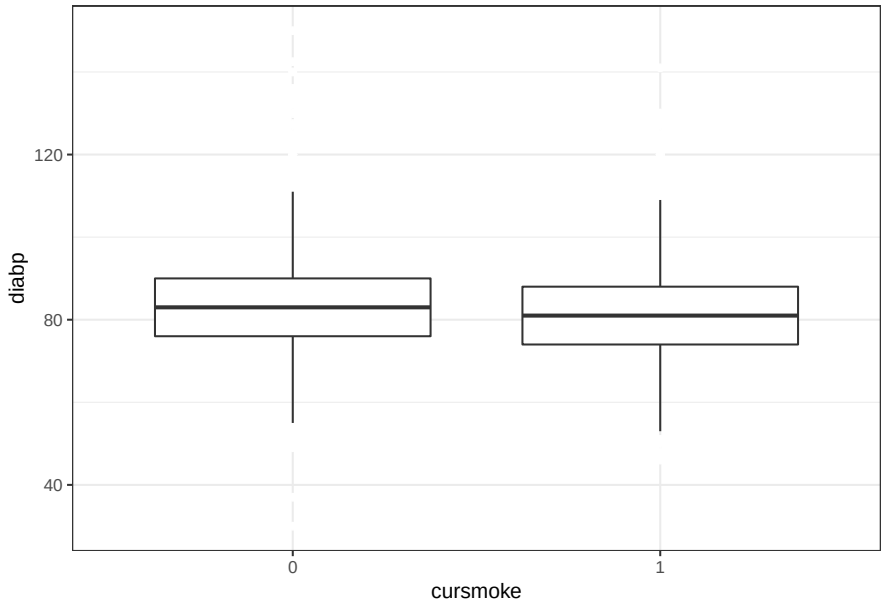
ANS: Proportion of smokers decreases with increase of diastolic blood presure for BP=100 ad then proportion increases again (latter could be due to not enough data)

```
data %>%
  select(cursmoke, diabp) %>%
  ggplot(aes(x = diabp, y = cursmoke)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



ANS: no difference

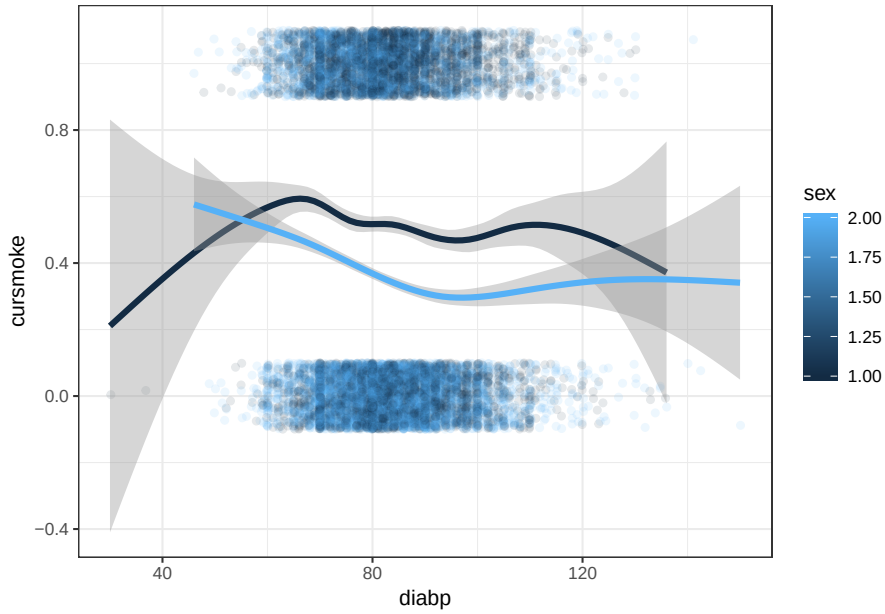
```
data %>%
  select(cursmoke, diabp) %>%
  mutate(cursmoke = factor(cursmoke)) %>%
  ggplot(aes(y = diabp, x = cursmoke)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```



smoking ~ diabp, sex

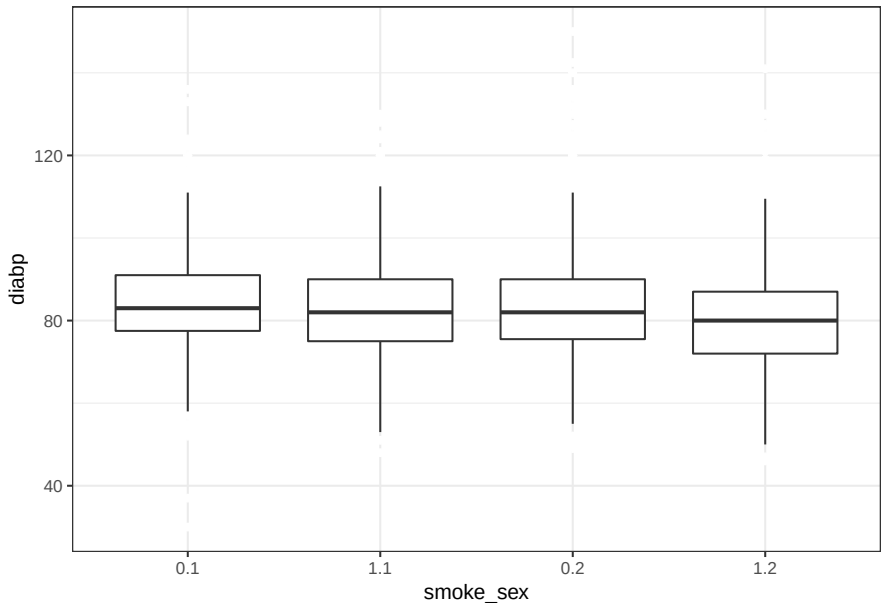
ANS: Proportion of smokers decreases with increase of diastolic blood presure; the proprtions are higher for men (sex effect).

```
data %>%
  select(cursmoke, diabp, sex) %>%
  ggplot(aes(x = diabp, y = cursmoke, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```



ANS: no difference

```r
data %>%
  select(cursmoke, sex, diabp) %>%
  mutate(cursmoke = factor(cursmoke),
           smoke_sex = interaction(cursmoke, sex)) %>%
  ggplot(aes(y = diabp, x = smoke_sex)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```



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

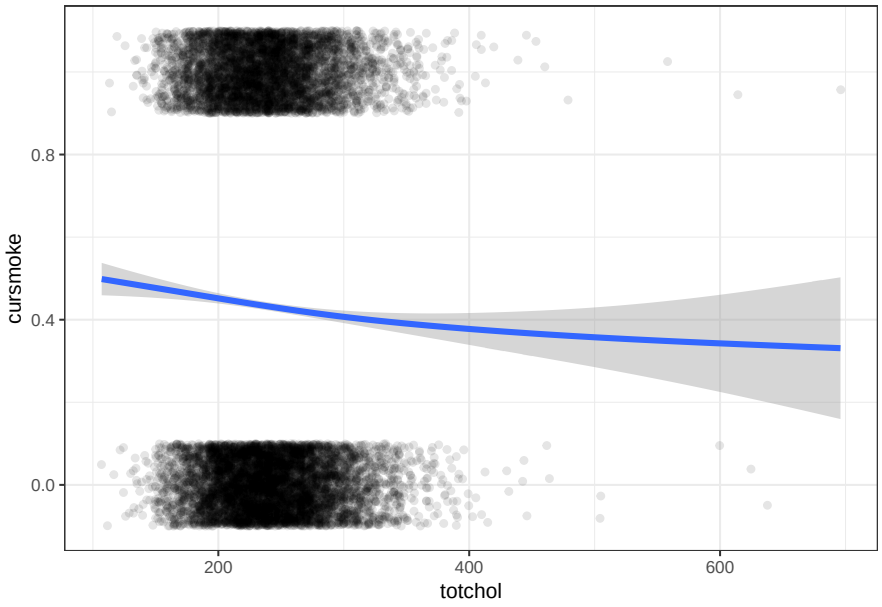
smoking ~ totchol

ANS: Proportion of smokers slightly decreases with increase of total cholesterol values

```r
data %>%
  select(cursmoke, totchol) %>%
  ggplot(aes(x = totchol, y = cursmoke)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```

```
## Warning: Removed 409 rows containing non-finite values (stat_smooth).
```

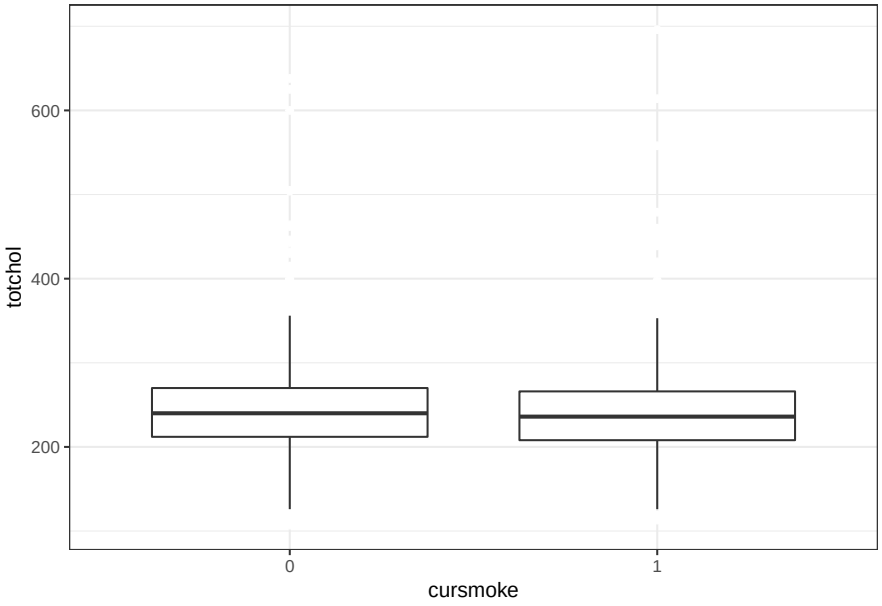
```
## Warning: Removed 409 rows containing missing values (geom_point).
```



ANS: no difference

```
data %>%
  select(cursmoke, totchol) %>%
  mutate(cursmoke = factor(cursmoke)) %>%
  ggplot(aes(y = totchol, x = cursmoke)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```

```
## Warning: Removed 409 rows containing non-finite values (stat_boxplot).
```



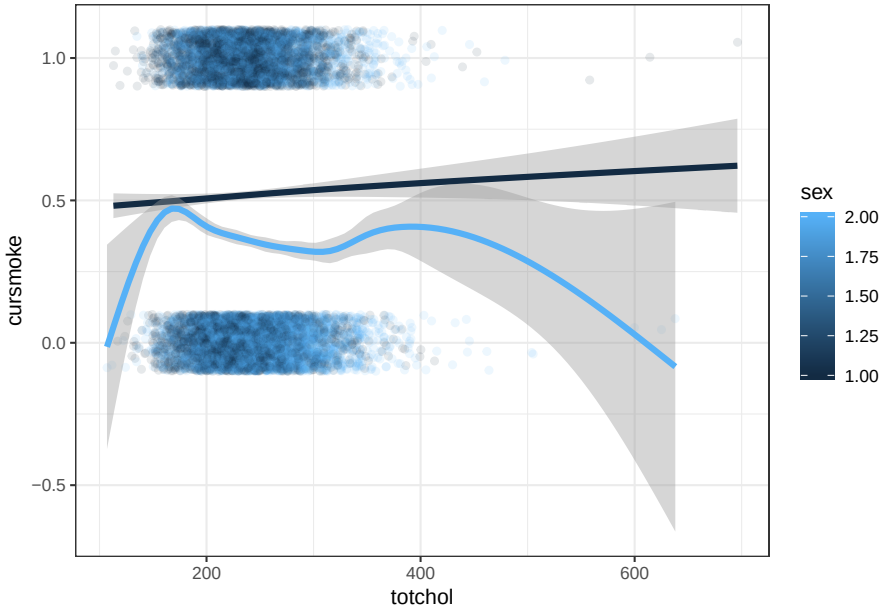
smoking ~ totchol, sex [!!!]

ANS: Proportion of smokers hasnonlinera relationship with total cholesterol for women; proprtions increases with increase in total cholesterol for men (sex by totchol interaction effect).

```r
data %>%
  select(cursmoke, totchol, sex) %>%
  ggplot(aes(x = totchol, y = cursmoke, group = sex, color = sex)) +
  geom_jitter(height = 0.1, alpha = 0.1) +
  geom_smooth(lwd = 1.5) +
  theme_bw()
```

```
## Warning: Removed 409 rows containing non-finite values (stat_smooth).
```

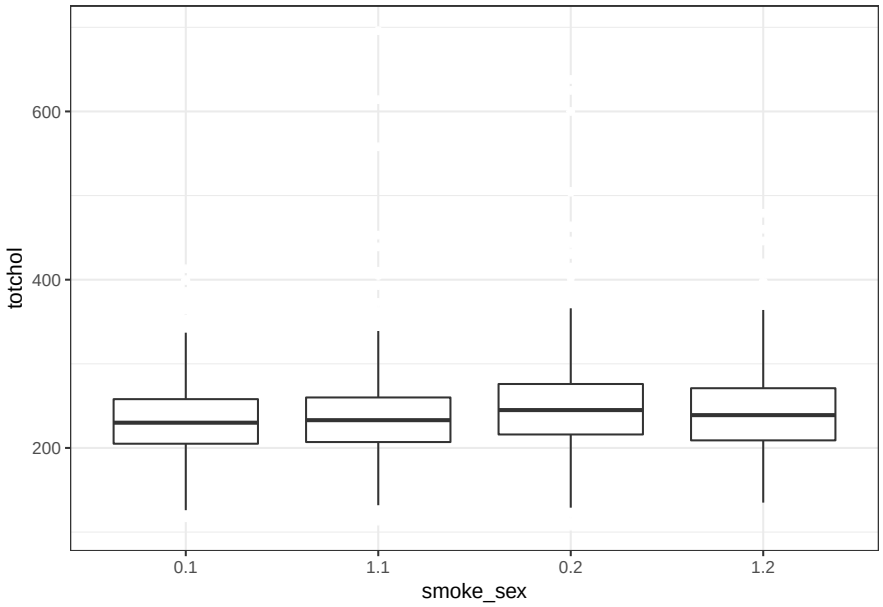
```
## Warning: Removed 409 rows containing missing values (geom_point).
```



ANS: no difference

```r
data %>%
  select(cursmoke, sex, totchol) %>%
  mutate(cursmoke = factor(cursmoke),
           smoke_sex = interaction(cursmoke, sex)) %>%
  ggplot(aes(y = totchol, x = smoke_sex)) +
  geom_boxplot(outlier.colour = "white") +
  theme_bw()
```

```
## Warning: Removed 409 rows containing non-finite values (stat_boxplot).
```



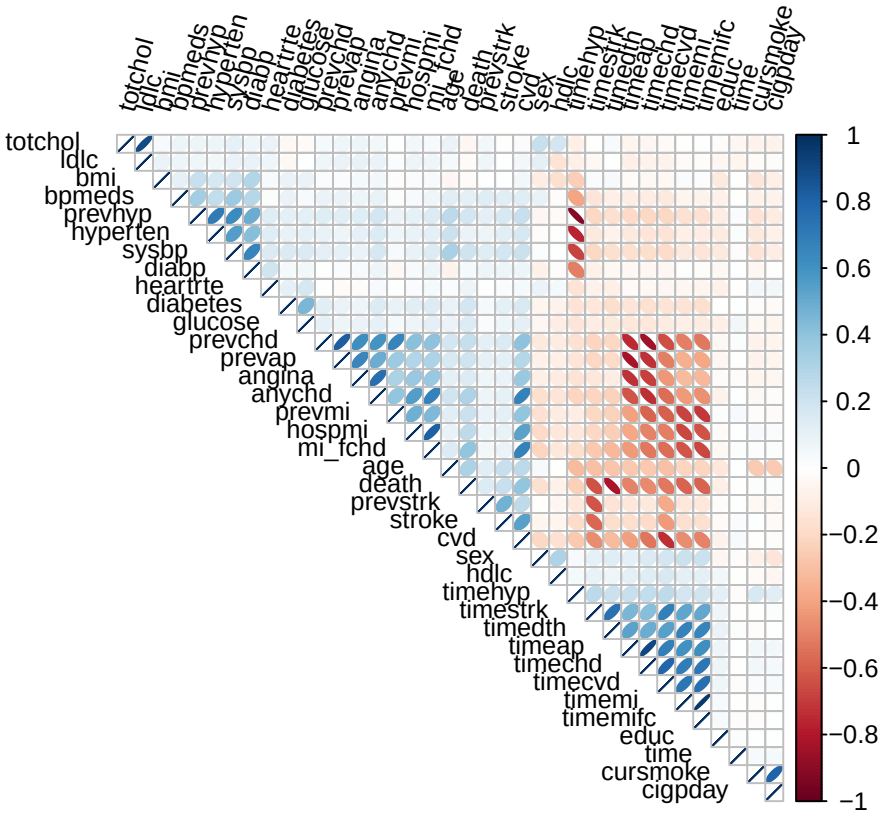
Cor plot

```
library(dplyr)

corr <- data[,c(-1,-21)] %>% cor(., use = "complete.obs")

library(corrplot)

corrplot(corr, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 75,
         tl.offset = 1, tl.cex= .8, method = "ellipse")
```



Missingness

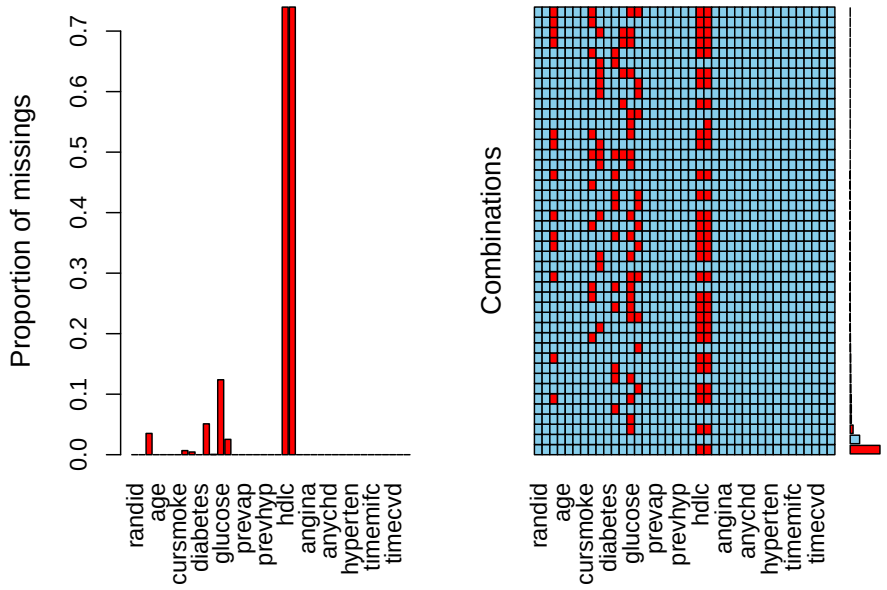
```
prop <- round(colSums(is.na(data))/dim(data)[1], 3)

knitr::kable(sort(prop, decreasing = TRUE)[1:9], col.names = "Proportion of NAs")
```

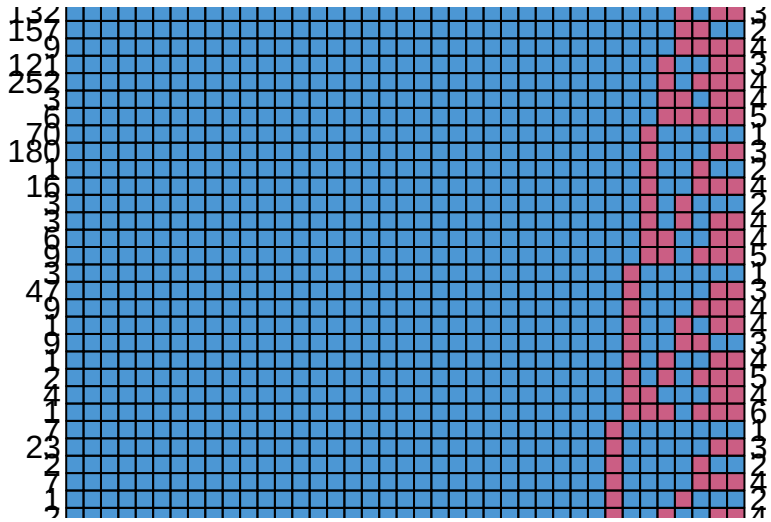
	Proportion of NAs
hdlc	0.740
ldlc	0.740
glucose	0.124
bpmeds	0.051
totchol	0.035
educ	0.025
cigpday	0.007
bmi	0.004
heartrte	0.001

```
#MISSING DATA ANALYSIS
VIM::aggr(data, prop=T, numbers=T)
```

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




```
mice::md.pattern(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
```

```
# HDLC and LDLC have a lot of missing data, and these variables should be eliminated
```

```r
prob.data <- data %>%
  group_by(period) %>%
  summarise(sysbp_prob = sum(sysbp, na.rm = TRUE)/n())
prob.data
```

```
## # A tibble: 3 x 2
##   period sysbp_prob
##    <int>      <dbl>
## 1      1       133.
## 2      2       137.
## 3      3       140.
```

```r
table(data$period)
```

```
##
##    1    2    3
## 4434 3930 3263
```

Models

```r
library(lme4)
library(dplyr)
```

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

```r
summary(gee::gee(cursmoke ~ age + sex + as.factor(educ) + bmi + diabetes+ heartrte + prevchd + prevstrk
```

```
##      (Intercept)              age              sex as.factor(educ)2
##     6.0056791362    -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
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ age + sex + as.factor(educ) + bmi +
##     diabetes + heartrte + prevchd + prevstrk + prevhyp + timedth,
##     id = randid, family = binomial, corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.8962482 -0.3999547 -0.2019900  0.4654084  0.9569903
##
```

```
##
## Coefficients:
##                       Estimate   Naive S.E.     Naive z  Robust S.E.
## (Intercept)       5.865249e+00 2.876839e-01  20.3878246 2.918503e-01
## age              -5.085698e-02 2.523157e-03 -20.1560927 2.409602e-03
## sex              -7.161378e-01 5.853646e-02 -12.2340470 6.021884e-02
## as.factor(educ)2  8.329407e-02 6.878131e-02   1.2109985 7.132044e-02
## as.factor(educ)3 -1.824354e-01 8.341765e-02  -2.1870119 8.622144e-02
## as.factor(educ)4 -2.120259e-01 9.400901e-02  -2.2553781 9.703148e-02
## bmi              -8.381699e-02 6.455452e-03 -12.9839063 6.897218e-03
## diabetes         -6.038094e-02 9.758132e-02  -0.6187756 1.015290e-01
## heartrte          8.932497e-03 1.404441e-03   6.3601797 1.481180e-03
## prevchd          -2.590671e-01 7.830011e-02  -3.3086429 9.019862e-02
## prevstrk         -1.886466e-01 1.712697e-01  -1.1014592 1.758071e-01
## prevhyp          -4.865095e-02 4.122989e-02  -1.1799922 4.099808e-02
## timedth          -8.971112e-05 1.418953e-05  -6.3223476 1.414685e-05
##                      Robust z
## (Intercept)       20.0967721
## age              -21.1059707
## sex              -11.8922537
## as.factor(educ)2   1.1678849
## as.factor(educ)3  -2.1158935
## as.factor(educ)4  -2.1851244
## bmi              -12.1522886
## diabetes          -0.5947164
## heartrte           6.0306610
## prevchd           -2.8721845
## prevstrk          -1.0730321
## prevhyp           -1.1866640
## timedth           -6.3414186
##
## Estimated Scale Parameter:  0.9748127
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7563191 0.5136670
## [2,] 0.7563191 1.0000000 0.5713635
## [3,] 0.5136670 0.5713635 1.0000000
```

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

```
summary(gee::gee(cigpday ~ age + sex + as.factor(educ),  id = randid, corstr = "unstructured"))
```

```
##      (Intercept)              age              sex as.factor(educ)2
##       32.8474189       -0.2787825       -5.9230065        0.7894811
## as.factor(educ)3 as.factor(educ)4
##       -0.7873846       -0.9535842
```

```
##
##  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::gee(formula = cigpday ~ age + sex + as.factor(educ), id = randid,
##     corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -16.938455  -7.957787  -4.031599   7.120862  77.643497
##
##
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)      28.3465052 0.87303381  32.468966  0.88378195  32.074094
## age              -0.1969624 0.01219203 -16.155019  0.01052473 -18.714256
## sex              -6.0131712 0.31430105 -19.131884  0.33919691 -17.727671
## as.factor(educ)2  1.1048806 0.37423991   2.952332  0.39926096   2.767314
## as.factor(educ)3 -0.4829763 0.45051511  -1.072054  0.43622459  -1.107174
## as.factor(educ)4 -0.7195971 0.51224189  -1.404799  0.55221861  -1.303102
##
## Estimated Scale Parameter:  133.1198
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7707103 0.4613932
## [2,] 0.7707103 1.0000000 0.5843157
## [3,] 0.4613932 0.5843157 1.0000000
```

While answering these questions, please account for any confounders that you have evidence may impact the relationship between age and sex with smoking.

Next you are interested in the relationship between certain health outcomes and smoking status. In particular you are interested in:

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

```
summary(gee::gee(sysbp ~ cursmoke + age + sex + as.factor(educ), id = randid, na.action = "na.omit", co
```

```
##      (Intercept)         cursmoke              age              sex
##       87.8024330       -2.0414490        0.8846592        1.2687106
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##       -0.3063701       -2.9072926       -4.0416336
```

```
##
##  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::gee(formula = sysbp ~ cursmoke + age + sex + as.factor(educ),
##     id = randid, 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)      88.9127308 1.71558828 51.826380  1.65360529 53.769017
## cursmoke         -1.5149317 0.46476110 -3.259592  0.45715635 -3.313815
## age               0.8653711 0.02377869 36.392712  0.02294929 37.707974
## sex               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
```

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

```
summary(gee::gee(diabp ~ cursmoke + age + sex + as.factor(educ), id = randid, na.action = "na.omit", co
```

```
##      (Intercept)         cursmoke              age              sex
##      83.14931183      -1.76253559       0.06074128      -1.49015111
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##      -0.07960126      -0.96361582      -1.38809040
```

```
##
##  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::gee(formula = diabp ~ cursmoke + age + sex + as.factor(educ),
##     id = randid, 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)      83.72716018 0.94598765 88.5076670  0.92664562 90.3551025
## cursmoke         -1.25577390 0.25866115 -4.8548996  0.26295278 -4.7756631
```

```
## age               0.04492311 0.01318527  3.4070687  0.01301711  3.4510820
## sex              -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
```

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

```
summary(gee::gee(totchol ~ cursmoke + age + sex + as.factor(educ), id = randid, na.action = "na.omit", c
```

```
##      (Intercept)         cursmoke              age              sex
##      180.5972060        1.6590353        0.7348726       12.0862772
## as.factor(educ)2 as.factor(educ)3 as.factor(educ)4
##        0.9107182        2.0078913        0.6363909
```

```
##
##  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::gee(formula = totchol ~ cursmoke + age + sex + as.factor(educ),
##     id = randid, 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)      179.4752160 3.62072716 49.5688319  3.65765697 49.0683565
## cursmoke           2.2459560 0.97321653  2.3077660  0.99901818  2.2481633
## age                0.7164736 0.04953822 14.4630466  0.04903112 14.6126292
## sex               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
```

```
model_1<-glmer(cursmoke~ age + sex + sysbp + diabp + totchol + as.factor(educ) + (1|randid), family=bin
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 1.19128 (tol
## = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
knitr::kable(summary(model_1)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	16.730	0.965	17.332	0.000
age	-0.206	0.010	-20.134	0.000
sex	-3.760	0.324	-11.597	0.000
sysbp	-0.001	0.005	-0.193	0.847
diabp	-0.027	0.008	-3.341	0.001
totchol	0.005	0.002	3.029	0.002
as.factor(educ)2	0.922	0.284	3.245	0.001
as.factor(educ)3	-0.330	0.326	-1.014	0.311
as.factor(educ)4	-0.481	0.372	-1.293	0.196

```
glmer(cursmoke~ age + sex + totchol + as.factor(educ) + bmi + diabetes+ heartrte + prevchd + prevstrk +
```

models from bingnan . . .

model of EDA:

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

```
my.data <- read.csv("../final_data/frmgham2.csv")
library(gee)

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")
```

```
##      (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
```

```
summary(model.q1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = CURSMOKE ~ AGE + as.factor(SEX) + as.factor(educ) +
##     BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK + PREVHYP +
##     TIMEDTH, id = RANDID, data = my.data, family = binomial,
##     corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.8962482 -0.3999547 -0.2019900  0.4654084  0.9569903
##
##
## Coefficients:
##                      Estimate   Naive S.E.     Naive z  Robust S.E.
## (Intercept)       5.149112e+00 2.742443e-01  18.7756353 2.776882e-01
## AGE              -5.085698e-02 2.523157e-03 -20.1560927 2.409602e-03
## as.factor(SEX)2  -7.161378e-01 5.853646e-02 -12.2340470 6.021884e-02
## as.factor(educ)2  8.329407e-02 6.878131e-02   1.2109985 7.132044e-02
## as.factor(educ)3 -1.824354e-01 8.341765e-02  -2.1870119 8.622144e-02
## as.factor(educ)4 -2.120259e-01 9.400901e-02  -2.2553781 9.703148e-02
## BMI              -8.381699e-02 6.455452e-03 -12.9839063 6.897218e-03
## DIABETES         -6.038094e-02 9.758132e-02  -0.6187756 1.015290e-01
## HEARTRTE          8.932497e-03 1.404441e-03   6.3601797 1.481180e-03
## PREVCHD          -2.590671e-01 7.830011e-02  -3.3086429 9.019862e-02
## PREVSTRK         -1.886466e-01 1.712697e-01  -1.1014592 1.758071e-01
## PREVHYP          -4.865095e-02 4.122989e-02  -1.1799922 4.099808e-02
## TIMEDTH          -8.971112e-05 1.418953e-05  -6.3223476 1.414685e-05
##                     Robust z
## (Intercept)       18.5427819
## AGE              -21.1059707
## as.factor(SEX)2  -11.8922537
## as.factor(educ)2   1.1678849
## as.factor(educ)3  -2.1158935
## as.factor(educ)4  -2.1851244
## BMI              -12.1522886
## DIABETES          -0.5947164
## HEARTRTE           6.0306610
## PREVCHD           -2.8721845
## PREVSTRK          -1.0730321
## PREVHYP           -1.1866640
## TIMEDTH           -6.3414186
##
```

```
## Estimated Scale Parameter:  0.9748127
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7563191 0.5136670
## [2,] 0.7563191 1.0000000 0.5713635
## [3,] 0.5136670 0.5713635 1.0000000
```

(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")
```

```
##      (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
```

```
summary(model.q2_1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = CIGPDAY ~ AGE + as.factor(SEX) + as.factor(educ),
##     id = RANDID, data = my.data, family = poisson, corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -21.229171  -7.140456  -4.339729   7.097624  77.590901
##
##
## Coefficients:
##                     Estimate  Naive S.E.    Naive z Robust S.E.
## (Intercept)       3.75330634 0.087902347  42.698591 0.076431587
## AGE              -0.02440602 0.001514284 -16.117201 0.001307541
## as.factor(SEX)2  -0.74791541 0.038827423 -19.262556 0.039577460
## as.factor(educ)2  0.10746849 0.044349757   2.423204 0.045214123
## as.factor(educ)3 -0.05830472 0.057813176  -1.008502 0.058516443
## as.factor(educ)4 -0.08779348 0.061727955  -1.422265 0.064819322
##                     Robust z
## (Intercept)       49.1067432
## AGE              -18.6655853
## as.factor(SEX)2  -18.8975090
## as.factor(educ)2   2.3768787
## as.factor(educ)3  -0.9963818
## as.factor(educ)4  -1.3544338
##
## Estimated Scale Parameter:  16.02383
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7413885 0.4848711
## [2,] 0.7413885 1.0000000 0.6452621
## [3,] 0.4848711 0.6452621 1.0000000
```

model including age and sex

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

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

```
##      (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
```

```r
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
```

```r
#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")
```

```
##      (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.

```
model.p2 <- gee(DIABP ~ CURSMOKE + AGE + as.factor(SEX)
                + as.factor(educ),
                id = RANDID,
                data = my.data,
```

```
                na.action = "na.omit",
                corstr = "unstructured")
```

```
##      (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)
```

```
##
##  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")
```

```
##      (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")
```

```
##      (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")
```

```
##      (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
```

```
#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
```

models from morgan

```
library(dplyr)

# data %>%
#   group_by(randid) %>%
#   summarise(sum_death = sum(death)) %>%
#   filter(sum_death > 0 )

# library(lme4)
#
# final_data <- data %>%
#   select(cursmoke, age, sex, educ,totchol,bmi,diabetes,heartrte,prevchd, prevstrk, prevhyp, timedth ,
#   mutate(sex = factor(sex),
#          educ = factor(educ))
#
# fit <- glmer(cursmoke ~ age + factor(sex) + factor(educ)  + bmi +
#              diabetes + heartrte + prevchd + prevstrk + prevhyp  + (1 | randid),
#              family = binomial,
#              na.action = "na.omit")
#
#
# fit <- gee(cursmoke ~ age + factor(sex) + factor(educ)  + bmi +
#                diabetes + heartrte + prevchd + prevstrk + prevhyp + timedth,
#            id = randid,
```

```
#             family = "binomial",
#             na.action = "na.omit")

##################
# Initial Models #
##################


# totchol_fit <- gee(totchol ~ cursmoke + age + factor(sex) + factor(educ)  + bmi +
#               diabetes + heartrte + prevchd + prevhyp +  prevstrk + death,
#             id = randid,
#             family = "gaussian",
#             na.action = "na.omit")
#
# round(2 * pnorm(abs(coef(summary(totchol_fit))[,5]), lower.tail = FALSE), 3)
#
#
# sysbp_fit <- gee(sysbp ~ cursmoke + age + factor(sex) + factor(educ)  + bmi +
#                       diabetes + heartrte + prevchd + prevstrk + death,
#                     id = randid,
#                     family = "gaussian",
#                     na.action = "na.omit")
# round(2 * pnorm(abs(coef(summary(sysbp_fit))[,5]), lower.tail = FALSE), 3)
#
# diabp_fit <- gee(diabp ~ cursmoke + age + factor(sex) + factor(educ)  + bmi +
#                       diabetes + heartrte + prevchd + prevstrk +death,
#                     id = randid,
#                     family = "gaussian",
#                     na.action = "na.omit")
#
# round(2 * pnorm(abs(coef(summary(diabp_fit))[,5]), lower.tail = FALSE), 3)


################################################
# Models After Removing Non Significant Terms #
################################################

totchol_fit <- gee(totchol ~ cursmoke + age + factor(sex) + bmi +
                     diabetes + heartrte  + prevhyp ,
                   id = randid,
                   family = "gaussian",
                   corstr = "unstructured",
                   na.action = "na.omit")
```

```
##  (Intercept)     cursmoke          age factor(sex)2          bmi
##  162.1032803    2.3672102    0.6268036   12.3710689    0.7814908
##     diabetes     heartrte      prevhyp
##   -5.8194286    0.1891382    5.2800067
```

```
knitr::kable(summary(totchol_fit)$coefficients[,c(1,4,5)], digits = 3)
```

	Estimate	Robust S.E.	Robust z
(Intercept)	149.257	5.279	28.274
cursmoke	3.430	1.009	3.399

	Estimate	Robust S.E.	Robust z
age	0.648	0.053	12.309
factor(sex)2	13.474	1.219	11.050
bmi	1.404	0.144	9.748
diabetes	-6.628	2.638	-2.513
heartrte	0.117	0.032	3.626
prevhyp	3.451	0.903	3.820

```
knitr::kable(round(2 * pnorm(abs(coef(summary(totchol_fit))[,5]), lower.tail = FALSE), 3))
```

	x
(Intercept)	0.000
cursmoke	0.001
age	0.000
factor(sex)2	0.000
bmi	0.000
diabetes	0.012
heartrte	0.000
prevhyp	0.000

```
QIC(totchol_fit)
```

```
##      QIC
## 84588.89
```

```
sysbp_fit <- gee(sysbp ~ cursmoke + age + factor(sex)  + bmi +
                   diabetes + heartrte + prevchd + prevstrk + death,
                 id = randid,
                 family = "gaussian",
                 corstr = "unstructured",
                 na.action = "na.omit")
```

```
##  (Intercept)     cursmoke          age factor(sex)2          bmi
##   40.1191453   -1.1246140    0.7236371    2.9078810    1.3070103
##     diabetes     heartrte      prevchd     prevstrk        death
##    4.8530466    0.2464253    2.4405511    8.6116353    7.0492511
```

```
knitr::kable(summary(sysbp_fit)$coefficients[,c(1,4,5)], digits = 3)
```

	Estimate	Robust S.E.	Robust z
(Intercept)	42.702	2.163	19.740
cursmoke	-0.702	0.433	-1.619
age	0.741	0.022	33.618
factor(sex)2	3.272	0.542	6.042
bmi	1.373	0.062	22.301
diabetes	3.664	1.127	3.250
heartrte	0.175	0.016	11.231
prevchd	1.670	0.876	1.906
prevstrk	4.894	2.004	2.443
death	7.650	0.639	11.967

```
knitr::kable(round(2 * pnorm(abs(coef(summary(sysbp_fit))[,5]), lower.tail = FALSE), 3))
```

	x
(Intercept)	0.000
cursmoke	0.105
age	0.000
factor(sex)2	0.000
bmi	0.000
diabetes	0.001
heartrte	0.000
prevchd	0.057
prevstrk	0.015
death	0.000

```
QIC(sysbp_fit)
```

```
##      QIC
## 68837.26
```

```
diabp_fit <- gee(diabp ~ cursmoke + factor(sex) + factor(educ) + bmi +
                   diabetes + heartrte  + prevstrk +death,
                 id = randid,
                 family = "gaussian",
                 corstr = "unstructured",
                 na.action = "na.omit")
```

```
##   (Intercept)      cursmoke  factor(sex)2 factor(educ)2 factor(educ)3
##    47.3480563    -1.0968017    -0.8594048     0.7880155     0.6921165
## factor(educ)4           bmi      diabetes      heartrte      prevstrk
##     0.5395935     0.9078065    -1.6162433     0.1550842     4.3067909
##         death
##     2.8662538
```

```
knitr::kable(summary(diabp_fit)$coefficients[,c(1,4,5)], digits = 3)
```

	Estimate	Robust S.E.	Robust z
(Intercept)	48.560	1.142	42.511
cursmoke	-0.607	0.243	-2.501
factor(sex)2	-0.591	0.295	-1.999
factor(educ)2	0.864	0.344	2.513
factor(educ)3	0.769	0.400	1.924
factor(educ)4	0.470	0.450	1.043
bmi	0.971	0.035	28.015
diabetes	-1.731	0.598	-2.895
heartrte	0.112	0.009	12.846
prevstrk	1.842	1.207	1.526
death	3.219	0.336	9.582

```
knitr::kable(round(2 * pnorm(abs(coef(summary(diabp_fit))[,5]), lower.tail = FALSE), 3))
```

	x
(Intercept)	0.000

	x
cursmoke	0.012
factor(sex)2	0.046
factor(educ)2	0.012
factor(educ)3	0.054
factor(educ)4	0.297
bmi	0.000
diabetes	0.004
heartrte	0.000
prevstrk	0.127
death	0.000

```
QIC(diabp_fit)
```

```
##      QIC
## 53507.67
```