Framingham Heart Study

CHUHAN

12/3/2018

Contents

For this analysis, we are interested to describe the smoking habits of the participants in the Framingham Heart study as they age and the impact of smoking on certain health outcomes. The Framingham heart study asks participants about their smoking habits at each visit. In particular, participants are asked if they are currently smoking at this visit (0 = Not a current smoker, 1 = Current smoker), which we will refer to as current smoking status. In addition, participants also report the number of cigarettes they are smoking per day. A more complete description of each of variables in the Framingham Heart study can be found in the Framingham Heart Study Longitudinal Data Documentation.

Part1

we are interested to answer the following questions: (1) Is there a relationship between age and smoking status? Does this relationship differ by sex? (2) Is there a relationship between the number of cigarettes smoked per day and age? Does this relationship differ by sex?

Data Preparation

The dataset contains 11,627 observations on 4,434 participants - each participant could have up to three observations depending on the number of exams each subject attended.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.2.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nlme)
```

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

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

```
library(mice)
```

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

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

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

##
## Attaching package: 'lme4'

## The following object is masked from 'package:nlme':
##
##     lmList
```

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

```
##
## Attaching package: 'gridExtra'

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

```
smoking=read.csv("frmgham2.csv")%>%
  janitor::clean_names()%>%
  dplyr::select(randid, period, age, sex, cursmoke, cigpday,bpmeds, educ, bmi, diabetes,heartrte, prevs
  mutate(sex = as.factor(sex),
         sex = fct_recode(sex, man = "1", woman = "2"),
         educ = as.factor(educ),
         educ = recode(educ, '1'='0-11 years', '2'='High School Diploma, GED',
                      '3'='Some College, Vocational School',
                      '4'='College (BS, BA) degree or more'),
         bmi=ifelse(bmi<18.5,"underweight",ifelse(bmi<25,"normal","overweight")),
         bmi=as.factor(bmi),
         diabetes = as.factor(diabetes),
         prevstrk = as.factor(prevstrk),
         prevhyp = as.factor(prevhyp))
names(smoking)
```

```
##  [1] "randid"   "period"   "age"      "sex"      "cursmoke" "cigpday"
##  [7] "bpmeds"   "educ"     "bmi"      "diabetes" "heartrte" "prevstrk"
## [13] "prevhyp"  "sysbp"    "diabp"    "totchol"
```

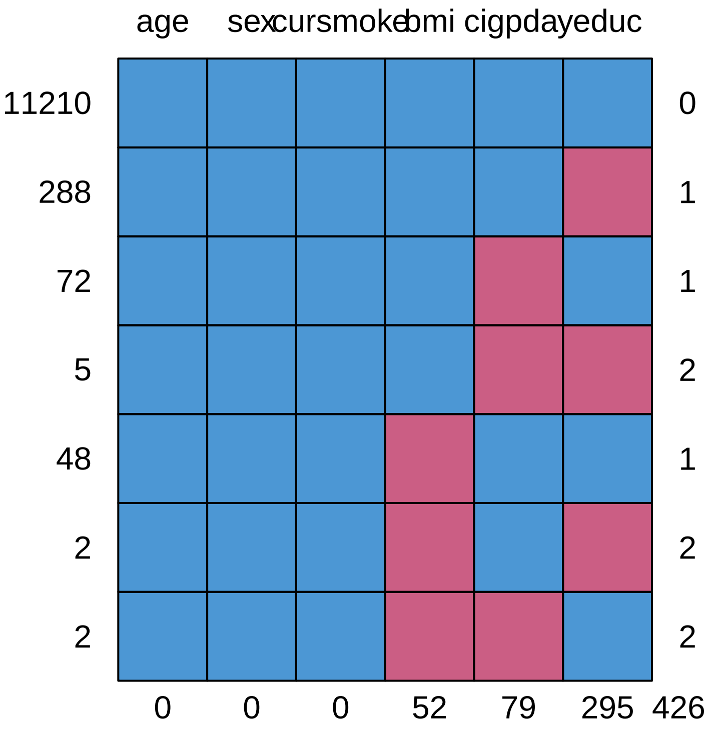
```
dim(smoking)
```

```
## [1] 11627    16
```

Missing machanism

We first check missing pattern of the variables we thought might be related to the smoking status. It shows that we have 11210 complete observations without any missing information on age, sex, somking status, number of cigarettes per day, bmi as well as educational level.

```
md.pattern(smoking[,c(3,4,6,8,9,5)], plot = TRUE)
```



```
##       age sex cursmoke bmi cigpday educ    
## 11210   1   1        1   1       1    1   0
## 288     1   1        1   1       1    0   1
## 72      1   1        1   1       0    1   1
## 5       1   1        1   1       0    0   2
## 48      1   1        1   0       1    1   1
## 2       1   1        1   0       1    0   2
## 2       1   1        1   0       0    1   2
##         0   0        0  52      79  295 426
```

Next, we checked the missing machanism for variable **cigpday**. The results gave us the support to our assumption that missing values in **cigpday** are missing at random(MAR). We could ues mice pakage in R to impute the missing values.

```
spre_cigpd <- smoking%>%
select(randid, cigpday, period)%>%
  spread(period, cigpday)%>%
  mutate(state2 = ifelse(is.na(`2`), 1, 0), state3 = ifelse(is.na(`3`), 1, 0))
logis <- glm(state2~`1`, data= spre_cigpd,family = binomial)
summary(logis)
```

```
##
## Call:
## glm(formula = state2 ~ `1`, family = binomial, data = spre_cigpd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7054  -0.5245  -0.4825  -0.4825   2.1015
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.091709   0.059444  -35.19  < 2e-16 ***
## `1`          0.011824   0.003639    3.25  0.00116 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3257.2  on 4401  degrees of freedom
## Residual deviance: 3247.1  on 4400  degrees of freedom
##   (32 observations deleted due to missingness)
## AIC: 3251.1
##
## Number of Fisher Scoring iterations: 4
```

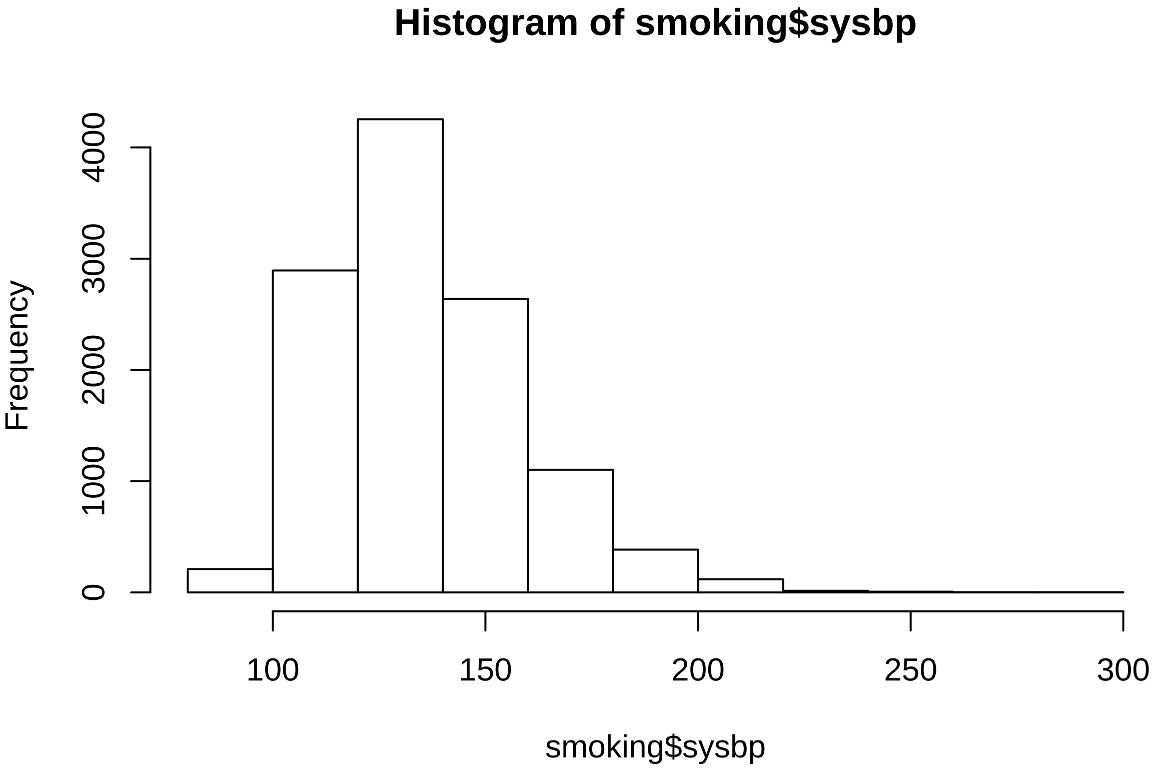
```
logis1 <- glm(state3~`2`+`1`, data= spre_cigpd,family = binomial)
summary(logis1)
```

```
##
## Call:
## glm(formula = state3 ~ `2` + `1`, family = binomial, data = spre_cigpd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8198  -0.6544  -0.6177  -0.6177   1.9002
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.559719   0.052744 -29.571   <2e-16 ***
## `2`         -0.001649   0.005197  -0.317   0.7510
## `1`          0.010698   0.005600   1.910   0.0561 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3714.4  on 3866  degrees of freedom
## Residual deviance: 3706.9  on 3864  degrees of freedom
##   (567 observations deleted due to missingness)
## AIC: 3712.9
##
## Number of Fisher Scoring iterations: 4
```

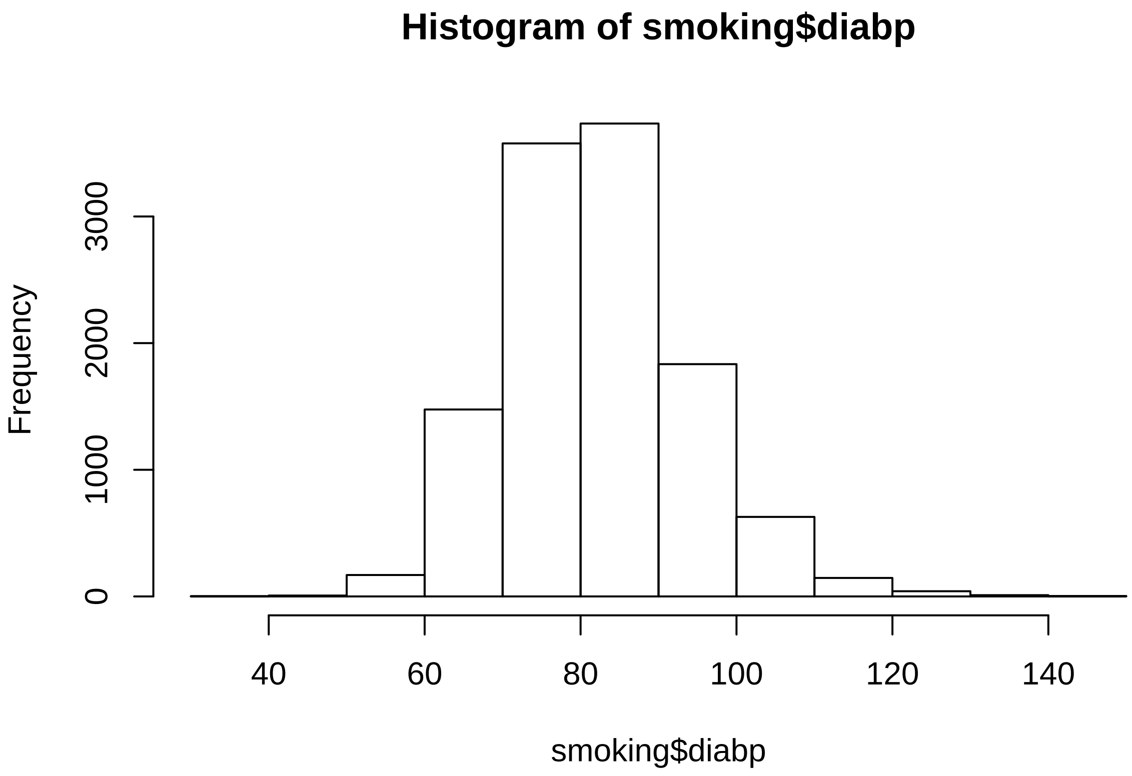
Removing outliers

Specifically, we filtered out observations with systolic blood pressure over 250, diastolic blood pressure over 150, and total cholesterol over 500 based on our findings from literature review[1]. We excluded a total of 27 observations in this step.

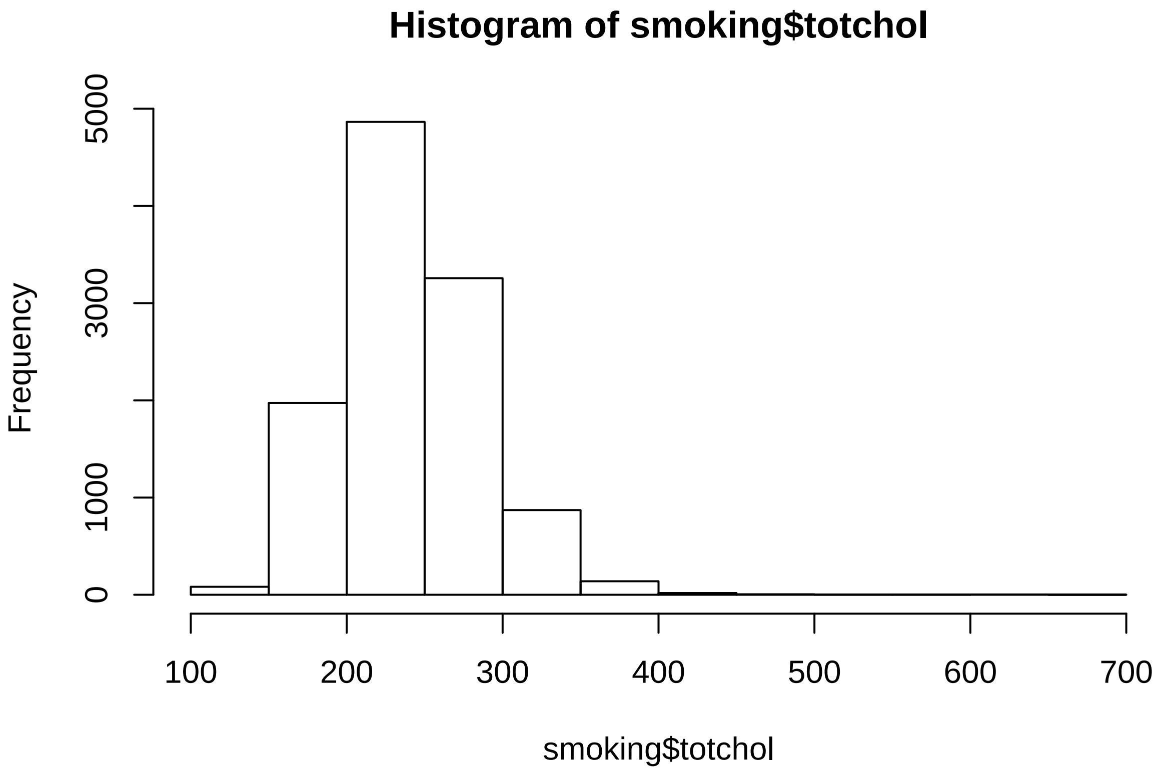
```
hist(smoking$sysbp)
```



```
hist(smoking$diabp)
```



```
hist(smoking$totchol)
```



```
sysbp.ol=smoking[which(smoking$sysbp>=250),]$randid
diabp.ol=smoking[which(smoking$diabp>=150),]$randid
totchol.ol=smoking[which(smoking$totchol>=500),]$randid
```

```
ol.id=unique(c(sysbp.ol,diabp.ol,totchol.ol));ol.id
```

```
##  [1]  610021 1080920 5080716 5807368 8303090  482553 1189726 2577634
##  [9] 5178346 6033947 7411567
```

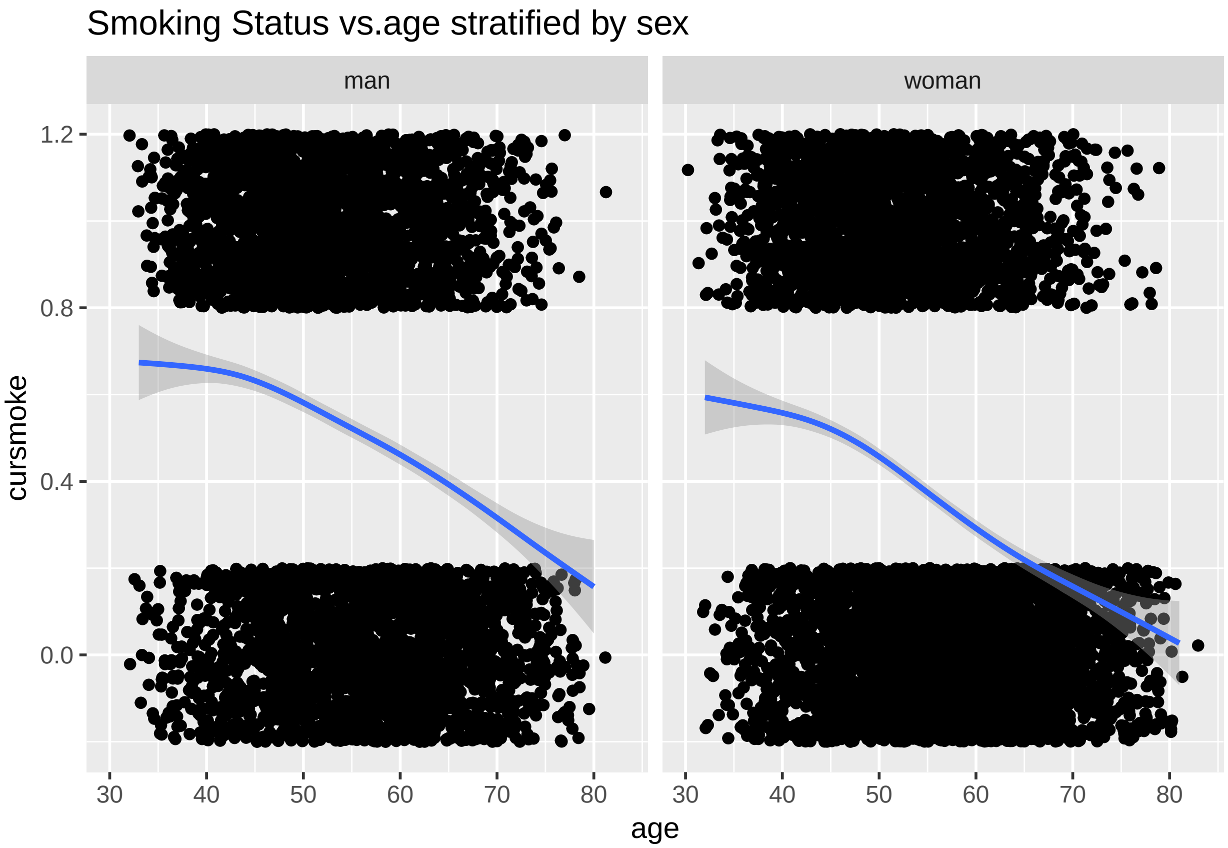
```
smoking2=smoking %>%
  filter(! randid  %in% ol.id)
```

Exploration of potential confounders

Potential confounders were selected through literature review at first. We found that hypertension, bmi and education level might associate with participants'age, sex as well as their smoking status. Thus we mainly focused on these variables here and visualized their impact on the relationship between age and sex with smoking.

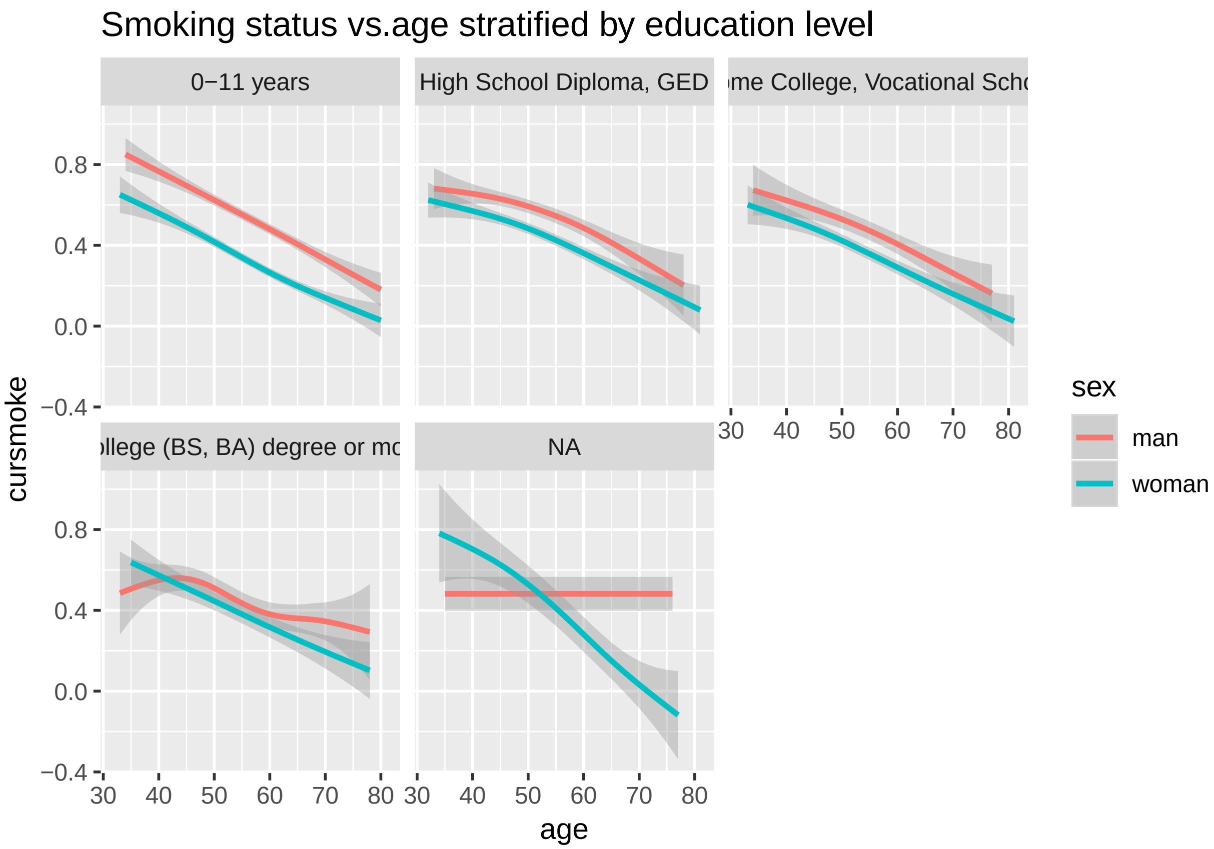
```
confounder <-smoking2 %>%
  select(randid,cursmoke,age,sex, educ, bmi,prevhyp,cigpday) %>%
  filter(!is.na(bmi))

#gender
ggplot(confounder,aes(x=age,y=cursmoke))+
  geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~sex) +labs(title="Smoking Status vs.age stratified by sex")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



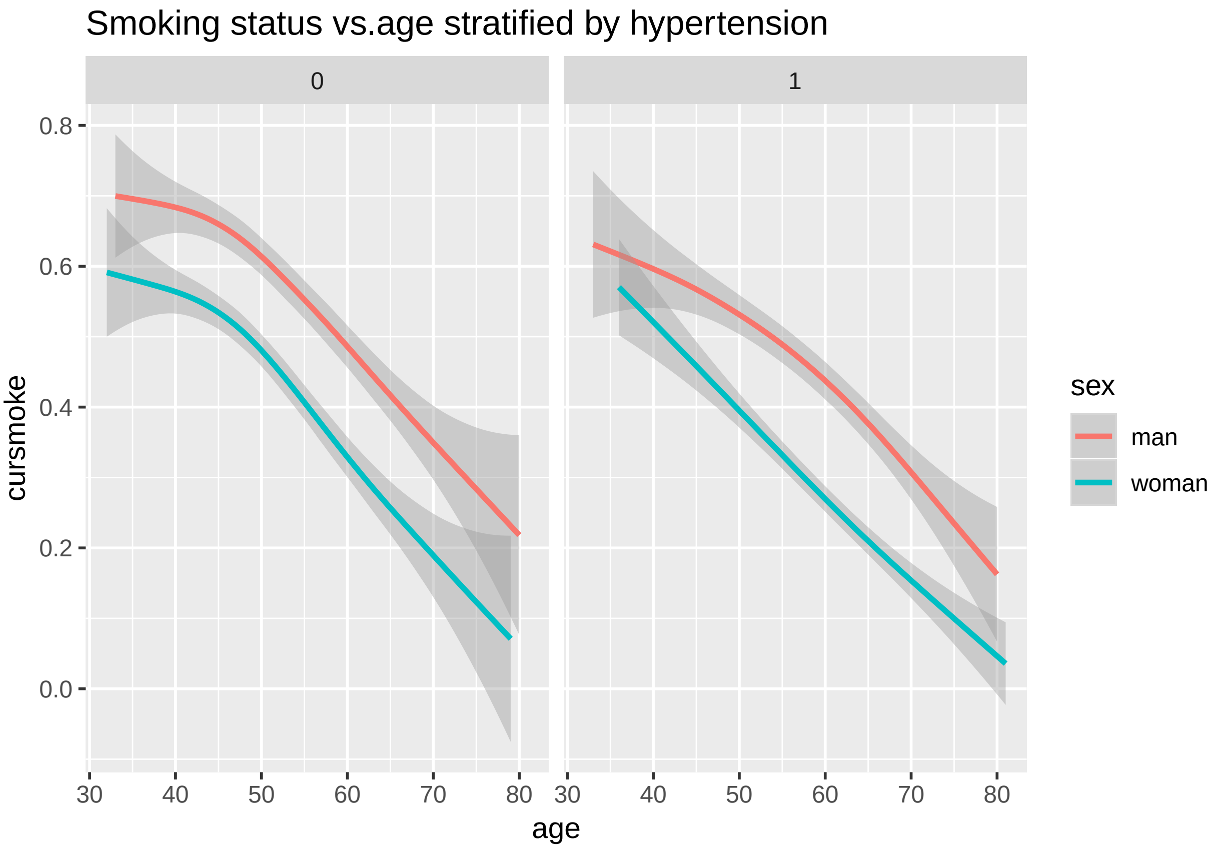
```r
#educ
ggplot(confounder,aes(x=age,y=cursmoke,color=sex))+
  #geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~educ)+labs(title="Smoking status vs.age stratified by education level")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



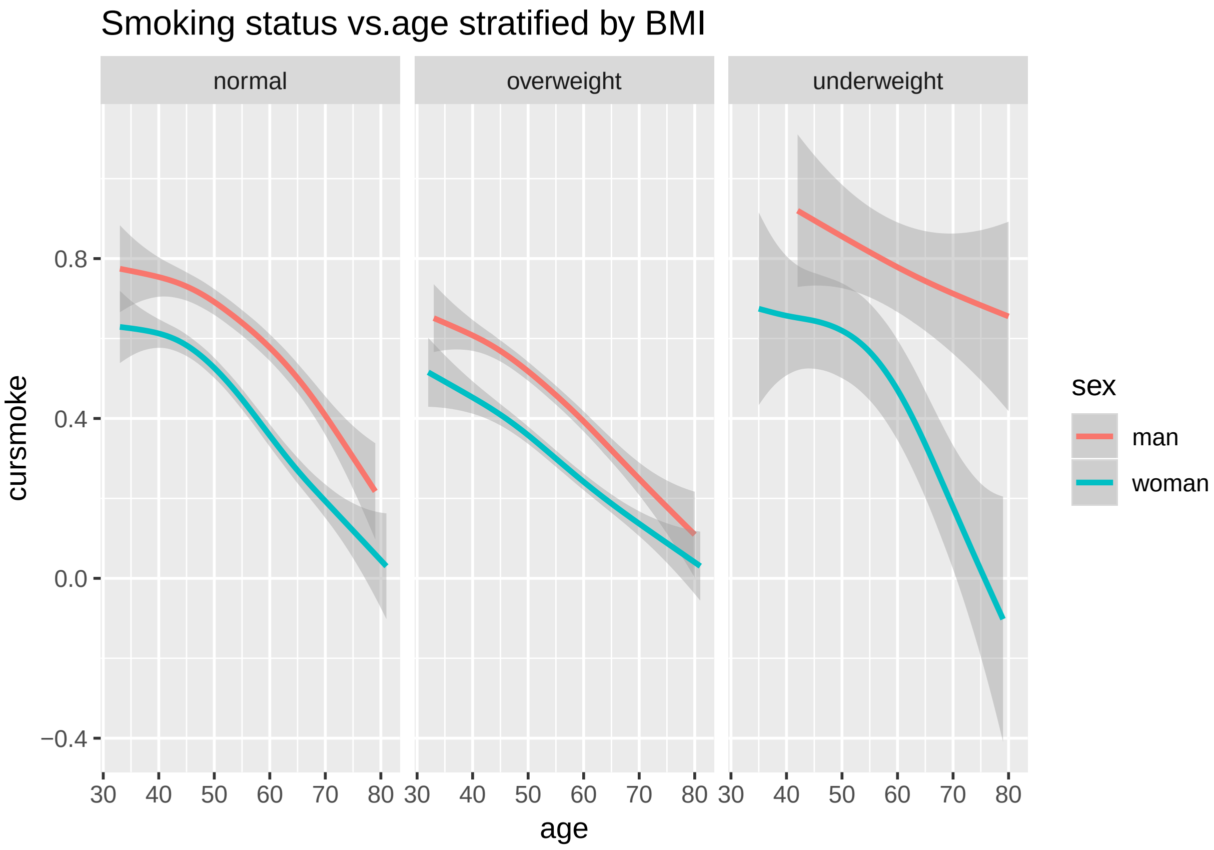
```r
#hypr
ggplot(confounder,aes(x=age,y=cursmoke,color=sex))+
  #geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~prevhyp)+labs(title="Smoking status vs.age stratified by hypertension")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
#bmi
ggplot(confounder,aes(x=age,y=cursmoke,color=sex))+
  #geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~bmi)+labs(title="Smoking status vs.age stratified by BMI")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Part2

Relationship between age and smoking status

Based on the above plots, the trends under different levels of **education**, **hypertension** and **bmi** seem to distinguish from each other. Thus we built a confounder select function based on the rule of thumb criteria to decide whether to include these potential confounders in the model. This function would fit a model with and without the potential confounder separately. Then the coefficient of the "confounded" variable will be checked to see whether it is still within the 95% confidence interval of the coefficient in model without this variable.

```
# confounder detect function
find_cnfd <- function(data, Y, method = method) {
  colnames(data)[Y] <- "Y"
  data_cf <- data %>%
    select(-age,-sex,-Y,-randid)
  cf_tbl <- tibble()
    glme1 <- glmer(Y ~ age+sex + (1|randid), data = data, family = method)
    est1 <- summary(glme1)$coefficients[2]
    ci_upper1 <- est1 + qnorm(0.975) * summary(glme1)$coefficients[5]
    ci_lower1 <- est1 - qnorm(0.975) * summary(glme1)$coefficients[5]
    est2 <- summary(glme1)$coefficients[3]
    ci_upper2 <- est2 + qnorm(0.975) * summary(glme1)$coefficients[6]
    ci_lower2 <- est2 - qnorm(0.975) * summary(glme1)$coefficients[6]

     for(i in 1:ncol(data_cf)){
      term_i <- names(data_cf)[i]
      glme2 <-  glmer(Y ~ age+sex + data_cf[,i] + (1|randid), data=data, family = method)
      cf1 <- summary(glme2)$coefficients[2]
      cf2 <- summary(glme2)$coefficients[3]
      cf_tbl_i <- tibble(potential_confounder = term_i, est_of_age = cf1, est_of_sex=cf2, ci_upper1 = c
      mutate(confounder = (cf1 > ci_upper1 | cf1 < ci_lower1) | (cf2 > ci_upper2 | cf1 < ci_lower1))
      cf_tbl <- rbind(cf_tbl, cf_tbl_i)
     }
      cf_tbl <- cf_tbl %>% select(potential_confounder, est_of_age,est_of_sex, confounder)
      return(list(table = cf_tbl, CI_age = c(ci_lower1, ci_upper1), CI_sex = c(ci_lower2, ci_upper2)))
}


#find_cnfd(data =confounder, Y=2, method = "binomial")
```

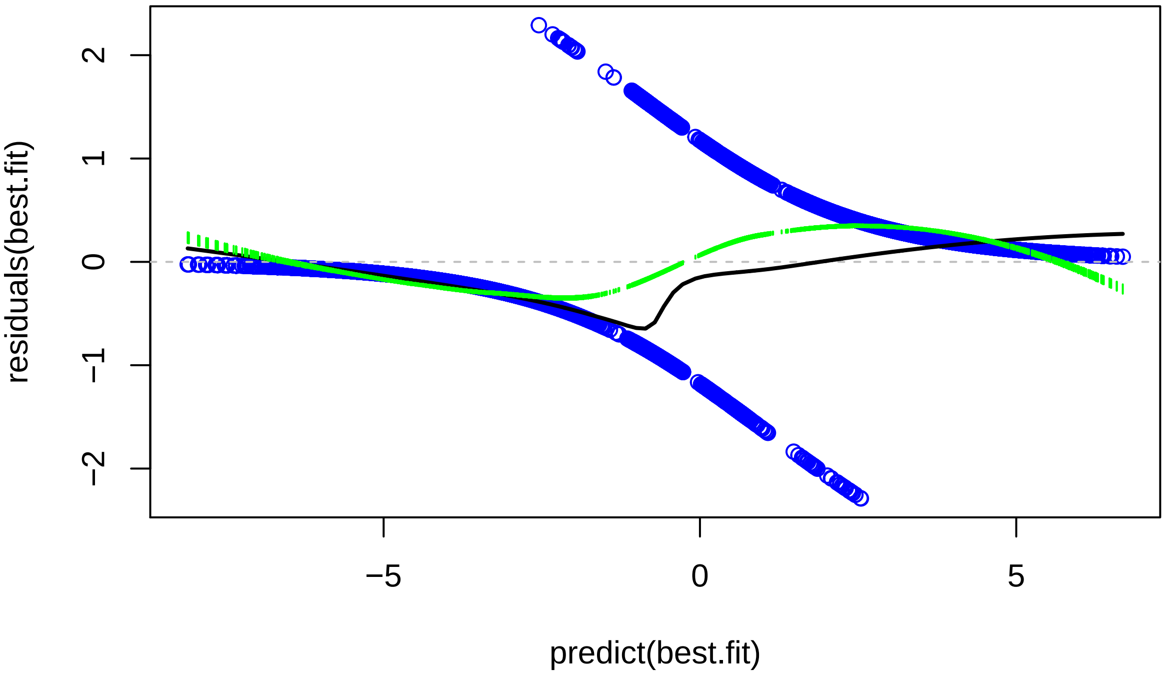
No other confounder was found to impact the relationship between age and sex with current smoking status. Thus we only included age, sex and their interaction as covariates. To study the association between age(or sex) and smoking status, we fitted several logistic regression models with mixed effect. Then, we seleced the final model based on AIC criteria.

```
#random intercept without interaction
fit1=glmer(cursmoke ~ age + sex + (1|randid), data =smoking2, family = binomial)
#random intercept with interaction
fit2=glmer(cursmoke ~ age*sex + (1|randid), data =smoking2, family = binomial)
#random intercept and slope without interaction
fit3=glmer(cursmoke ~ age + sex + (age|randid), data =smoking2, family = binomial)
#random intercept and slope without interaction
fit4=glmer(cursmoke ~ age*sex + (age|randid), data =smoking2, family = binomial)

# model selection
tibble(model = c("model1","model2","model3", "model4"), AIC = c(summary(fit1)$AIC[1],summary(fit2)$AIC[
summary(fit3)$AIC[1],summary(fit4)$AIC[1]))
```

```
## # A tibble: 4 x 2
##   model     AIC
##   <chr>   <dbl>
## 1 model1 10829.
## 2 model2 10819.
## 3 model3 10823.
## 4 model4 10827.
```

model2 with random intercept and interaction term ends with the lowest AIC value, so we chose model2 as our final model. Next, we made a residual plot to visualize the model fitting.

```
best.fit=fit2
plot(predict(best.fit),residuals(best.fit),col=c("blue","red")[smoking2$cursmoke])
abline(h=0,lty=2,col="grey")
lines(lowess(predict(best.fit),residuals(best.fit)),col="black",lwd=2)
rl=loess(residuals(best.fit)~predict(best.fit))
y=predict(rl,se=TRUE)
segments(predict(best.fit),y$fit+2*y$se.fit,predict(best.fit),y$fit-2*y$se.fit,col="green")
```



```
summary(best.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cursmoke ~ age * sex + (1 | randid)
##    Data: smoking2
##
##      AIC      BIC   logLik deviance df.resid
##  10819.4  10856.1  -5404.7  10809.4    11595
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5688 -0.1429 -0.0527  0.1970  3.5723
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  randid (Intercept) 33.98    5.829
## Number of obs: 11600, groups:  randid, 4423
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  13.62210    0.84266  16.166  < 2e-16 ***
## age          -0.23696    0.01448 -16.364  < 2e-16 ***
## sexwoman     -6.78258    1.08345  -6.260 3.85e-10 ***
## age:sexwoman  0.05427    0.01663   3.264   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    sexwmn
## age         -0.974
## sexwoman    -0.831  0.808
## age:sexwomn  0.758 -0.785 -0.950
```

We filled out the estimates of coefficients from summary and got:

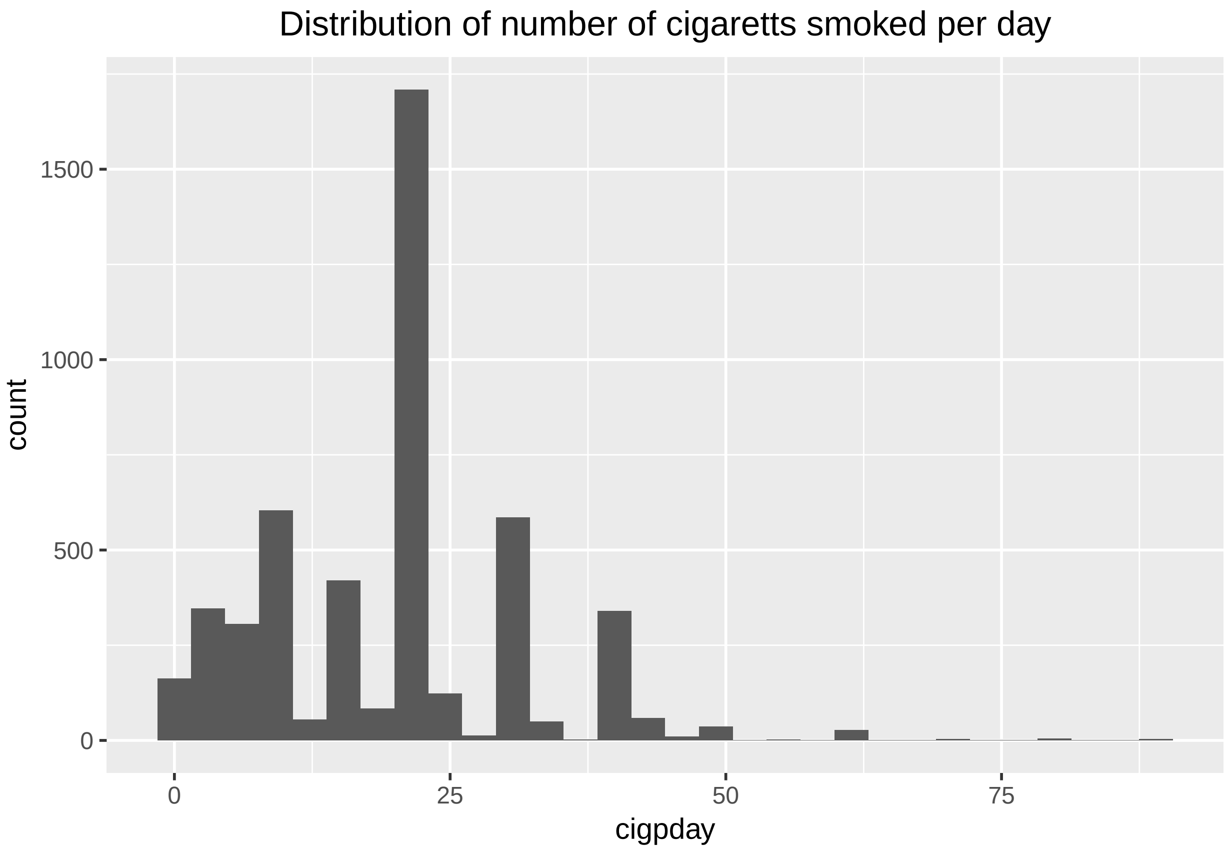
$$logit(P(cursmoker_{ij} = 1)) = 13.6221 - 0.2370 * age_{ij} - 6.7826 * sex_{ij} + 0.0543 * age_{ij} * sex_{ij} + 33.98$$

Interpretation: For every one year increase in age, the odds of smoking for men will decrease by 21.09% (p-value < 0.001), while the odds of smoking for women will decrease by 16.70% (p-value < 0.001).

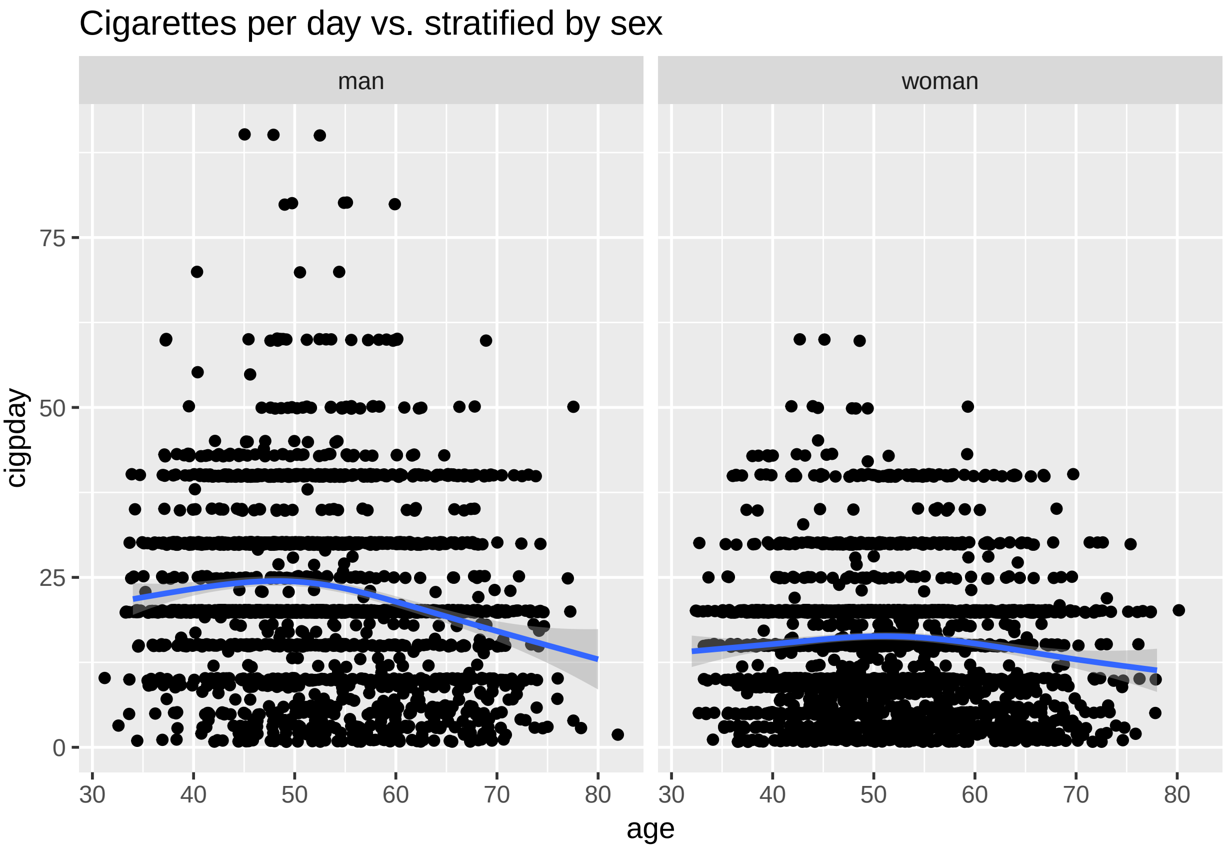
Relationship between the number of cigarettes smoked per day and age

For the association between age(or sex) and number of cigarettes per day, we only focused on the subjects who were potential smokers. The results from confounder select function are also in favor of excluding non-smokers in this part of analysis.

```
smoking.cigar = smoking2 %>%
  filter(cursmoke!=0)

smoking.cigar %>%
  ggplot(aes(x = cigpday)) + geom_histogram(bins = 30) + ggtitle("Distribution of number of cigaretts sm
```



```
ggplot(subset(confounder,cursmoke==1),aes(x=age,y=cigpday))+
  geom_jitter(height = 0.2, width = 3)+
  geom_smooth()+facet_wrap(~sex) +labs(title="Cigarettes per day vs. stratified by sex")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
confounder %>%
  find_cnfd(data =., Y = 8, method = "poisson")
```

```
## $table
## # A tibble: 4 x 4
##   potential_confounder est_of_age est_of_sex confounder
##   <chr>                     <dbl>      <dbl> <lgl>
## 1 cursmoke                0.00383     -0.470 TRUE
## 2 educ                   -0.0182      -2.02  FALSE
## 3 bmi                    -0.0180      -1.97  FALSE
## 4 prevhyp                -0.0178      -1.97  FALSE
##
## $CI_age
## [1] -0.01953388 -0.01672945
##
## $CI_sex
## [1] -2.194873 -1.736427
```

We fitted Poisson regression model with mixed effect to study the relationship between age and sex with the number of cigarettes smoked per day. Next, we seleced the final model based on AIC criteria.

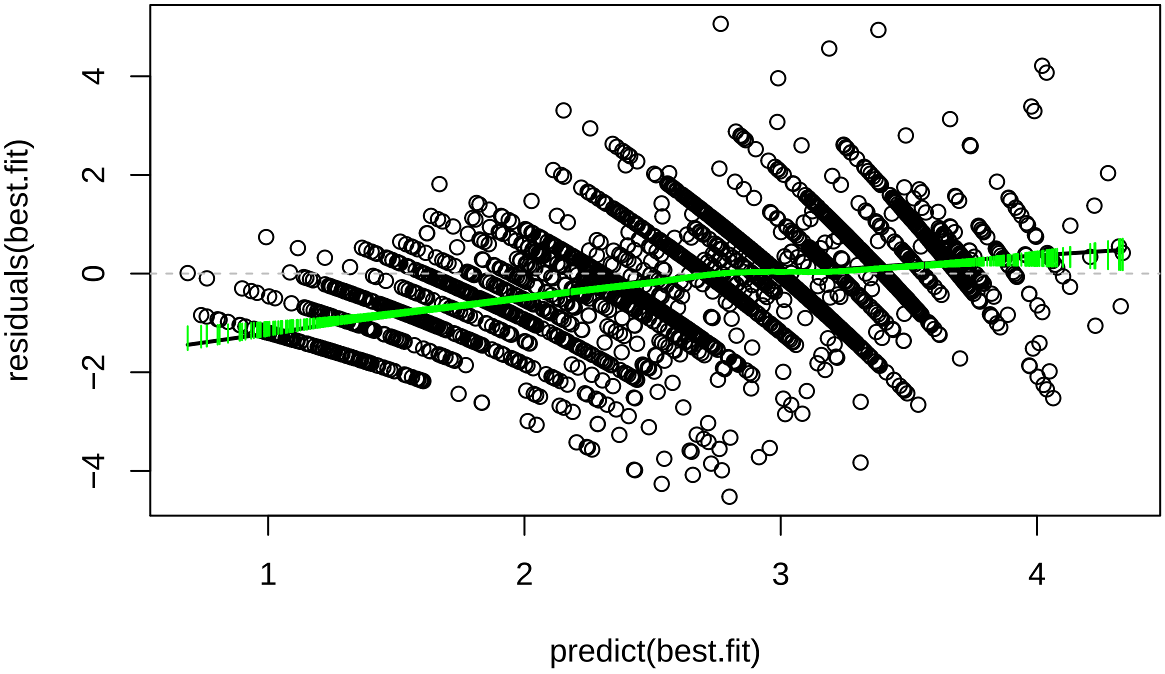
```
#random intercept without interaction
fit5=glmer(cigpday ~ age + sex + (1|randid), data =smoking.cigar, family = poisson)
#random intercept with interaction
fit6=glmer(cigpday ~ age*sex + (1|randid), data =smoking.cigar, family = poisson)
#random intercept and slope without interaction
fit7=glmer(cigpday ~ age + sex + (age|randid), data =smoking.cigar, family = poisson)
#random intercept and slope without interaction
```

```
fit8=glmer(cigpday ~ age*sex + (age|randid), data =smoking.cigar, family = poisson)

# model selection
tibble(model = c("model1","model2","model3", "model4"), AIC = c(summary(fit5)$AIC[1],summary(fit6)$AIC[
summary(fit7)$AIC[1],summary(fit8)$AIC[1]))
```

```
## # A tibble: 4 x 2
##   model      AIC
##   <chr>    <dbl>
## 1 model1 36427.
## 2 model2 36386.
## 3 model3 35906.
## 4 model4 35890.
```

model4 with random intercept and slope as well as interaction term ends with the lowest AIC value, so we chose model4 as our final model. The residual plot for model4 was shown below.

```
best.fit=fit8
plot(predict(best.fit),residuals(best.fit))
abline(h=0,lty=2,col="grey")
lines(lowess(predict(best.fit),residuals(best.fit)),col="black",lwd=2)
rl=loess(residuals(best.fit)~predict(best.fit))
y=predict(rl,se=TRUE)
segments(predict(best.fit),y$fit+2*y$se.fit,predict(best.fit),y$fit-2*y$se.fit,col="green")
```



```
summary(best.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: cigpday ~ age * sex + (age | randid)
##    Data: smoking.cigar
##
##      AIC      BIC   logLik deviance df.resid
##   35889.6  35935.2 -17937.8  35875.6     4935
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5622 -0.5434 -0.0024  0.3102  6.0475
##
## Random effects:
##  Groups Name        Variance  Std.Dev. Corr
##  randid (Intercept) 1.8604009 1.36397
##         age         0.0008349 0.02889  -0.90
## Number of obs: 4942, groups:  randid, 2290
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.200370   0.069983   45.73  < 2e-16 ***
## age          -0.004579   0.001391   -3.29    0.001 ***
## sexwoman     -0.925654   0.102932   -8.99  < 2e-16 ***
## age:sexwoman  0.008790   0.002058    4.27 1.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    sexwmn
## age         -0.960
## sexwoman    -0.634  0.610
## age:sexwomn  0.608 -0.637 -0.961
## convergence code: 0
## Model failed to converge with max|grad| = 0.225671 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

model4:

$$log(E(cigpday_{ij})) = \beta_0 + (\beta_1 + b_{1i}) * age_{ij} + \beta_2 * sex_{ij} + \beta_3 * age_{ij} * sex_{ij} + b_{0i}$$

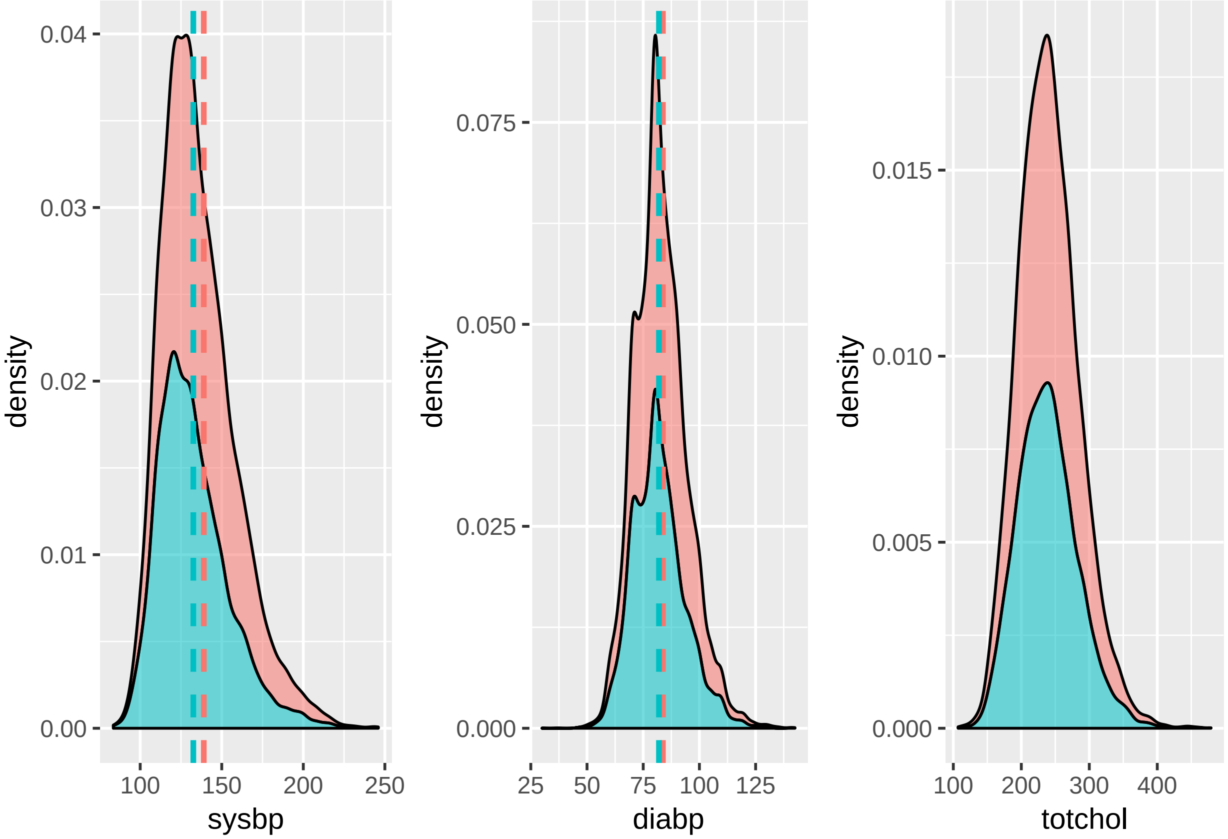
Interpretation: The expected number of cigarettes smoked per day will decrease by 0.9954 (p-value = 0.001) with one year increase in age for men. For women, the expected number of cigarettes smoked per day with one year increase in age will decrease by 1.0042 (p-value $<$ 0.001).

Exploration of Health Outcomes

```
mean.sysbp=smoking2 %>%
  group_by(cursmoke) %>%
  summarise(mean=mean(sysbp))
p1=ggplot(data=smoking2,aes(x=sysbp,fill=as.factor(cursmoke),alpha=1/10))+geom_density(position = "stac

mean.diabp=smoking2 %>%
  group_by(cursmoke) %>%
  summarise(mean=mean(diabp))

p2=ggplot(data=smoking2,aes(x=diabp,fill=as.factor(cursmoke),alpha=1/10))+geom_density(position = "stac

mean.totchol=smoking2 %>%
  group_by(cursmoke) %>%
  summarise(mean=mean(totchol))
p3=ggplot(data=smoking2,aes(x=totchol,fill=as.factor(cursmoke),alpha=1/10))+geom_density(position = "st
library(gridExtra)
plot2=grid.arrange(p1,p2,p3,nrow=1)
```

```
## Warning: Removed 408 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2 rows containing missing values (geom_vline).
```



For the relationship between smoking status and certain health outcomes(systolic blood pressure, diastolic blood pressure, serum total cholesterol), we could also fit linear mixed effect models and use AIC criteria to check if we need random slope for variables of interest. Additionally, in the mixed effect models, we could use bootstrapping for calculating p-values for fixed effect.