Smoking in the Framingham Heart Study

P8157 - Analysis of Longitudinal Data Final Project

Jingxuan He(jh3909), Chuhan Zhou(cz2493), Huijuan Zhang(hz2510), Yeyi Zhang(yz3300)

Objective

In this study, we are interested to describe the smoking patterns of adults. We would like to study the relationship between age and smoking and identify confounders that would impact this relationship. Also, we aim to quantify the effect of smoking on certain health outcomes and grasp a general idea about how smoking habits affect the physical health of the study population.

Study Design

We used data from Framingham Heart Study, a cohort study of the etiology of cardiovascular disease among a population of free-living subjects in the community of Framingham, Massachusetts. The dataset contains 11,627 observations on 4,434 participants - each participant could have up to three observations depending on the number of exams the subject attended. A more complete description of each of the variables in the Framingham Heart study can be found in the Framingham Heart Study Longitudinal Data Documentation.

Methods

To start with, we did the exploratory analysis to get a total sense of data. We firstly investigated the pattern of missing data and imputed data based on the assumption of missing at random. Imputed data was used to fit the marginal model. Then we checked the distribution and outliers for each health outcome, including systolic blood pressure, diastolic blood pressure, and total cholesterol level. Specifically, we grouped patients with systolic blood pressure over 250, diastolic blood pressure over 150, and total cholesterol over 500 as outliers through literature [1]. We excluded a total of 27 observations from the data in this step. Next, we centered the age by mean age(55) of the study population. Potential confounders were selected through literature review and we found that hypertension[2], serum total cholesterol level[3] and education level[4] might be well associated with both age and smoking status, thus we included these factors as confounders to determine the relationship between age and smoking status(part I) . In studying the relationship between health outcomes and smoking status(part II), we treated anti-hypertensive medication, prevalent coronary heart disease (CHD)[2] and diabetes status[5] as confounders in addition to age and sex. For serum total cholesterol, we only put prevalent CHD[2] as a confounder. In both parts of analysis, we included interaction between smoking status and centered age in the model.

In the first part of analysis, we regarded smoking status (0 = Not a current smoker, 1 = Current smoker) and smoking habit (the number of cigarettes smoked per day) as the outcomes, and restricted our study population to current smokers to account for the underlying smoking pattern. Scatterplots with locally weighted smoothers were used to explore their relationship with age respectively, with and without adjusting for confounders. Then we applied a random effects model to perform the longitudinal analysis.

In the second part of analysis, we firstly investigated the within-subject correlation over time and proposed potential correlation structures by plotting autocorrelation function (ACF) versus time lag between visits. We then fitted marginal models with the potential correlation structures to study the relationship by using complete cases in data. Furthermore, we used whole data to fit both random intercept model and random slope and intercept model to take variation between subjects into consideration. For random effect models, we used the normal approximation to calculate coefficients' p-values as well as using the likelihood ratio test to test the significance of random intercept. The final model was selected based on Akaike information criterion (AIC).

Results

For the first part of analysis, we explored the effect of sex on the relationship between smoking status and age(see Figure 1a). It shows female has an overall lower proportion of smokers than male after controlling for age. Then, we moved on to the effect of sex on the relationship between the number of cigarettes smoked per day and age (see Figure 1b). It demonstrated that female smokers usually smoke less cigarettes per day than male smokers after controlling for age. Additionally, male aged between 40-60 tend to have more extremely addicted smokers, with over 50 cigarettes per day, than female. For most female smokers, however, the number of cigarettes smoked per day is less than 20. The significance results from the random intercept models are basically consistent with our findings above. Sex is a strong modifier to affect both current smoking status and the number of cigarettes smoked. After modeling with generalized linear mixed effect model, we could find that the odds ratio for women versus men to be a current smoker is 0.28 with a 95% C.I. of (0.20, 0.39), controlling for age, education level, prevalent diseases. The number of cigarettes smoked per day for female smokers is 0.63 with a 95% C.I. of (0.60, 0.67) less than male smokers, controlling for the same confounders (see Table 1a).

For the second part of analysis, in the marginal model, we selected unstructured model and exchangeable model to account for within-subject correlation based on the results of ACF, with only a slight correlation difference between 6-year and 12-year time lag. AIC values for each model were calculated and we decided to stick with the exchangeable structure for its convenience. Overall, the marginal model yielded similar interpretation for the relationship between health outcomes and smoking status.

Furthermore, both random intercept model and random slope and intercept model were fitted for each of the outcomes. It is worth noting that all the variances for random slope were very small, just take less than 5% of the variances for random intercept, so that we ignore the effect of random slope and continue with the random intercept model. A further test for significance of random intercept again verified our decision of random intercept model. The interaction between age and smoking status had a significant impact on the relationship between total cholesterol and smoking status. In random effect model, systolic blood pressure for current smokers is -1.21 unit lower than non-smokers' with a 95% C.I. of (-2.08, -0.34), controlling for other factors. Current smokers have -0.94 unit lower diastolic blood pressure compared to non-smoker's at 6-year before with a 95% C.I. of (-1.44, -0.44), controlling for other factors. Current smokers have 2.55 unit higher serum total cholesterol compared to non-smoker's at 6-year before with a 95% C.I. of (0.75, 4.35), controlling for other factors (see Table 1b with p-values). For this part, we end up with random intercept models because marginal models only using complete cases will cause loss of information and using period is not as accurate as age.

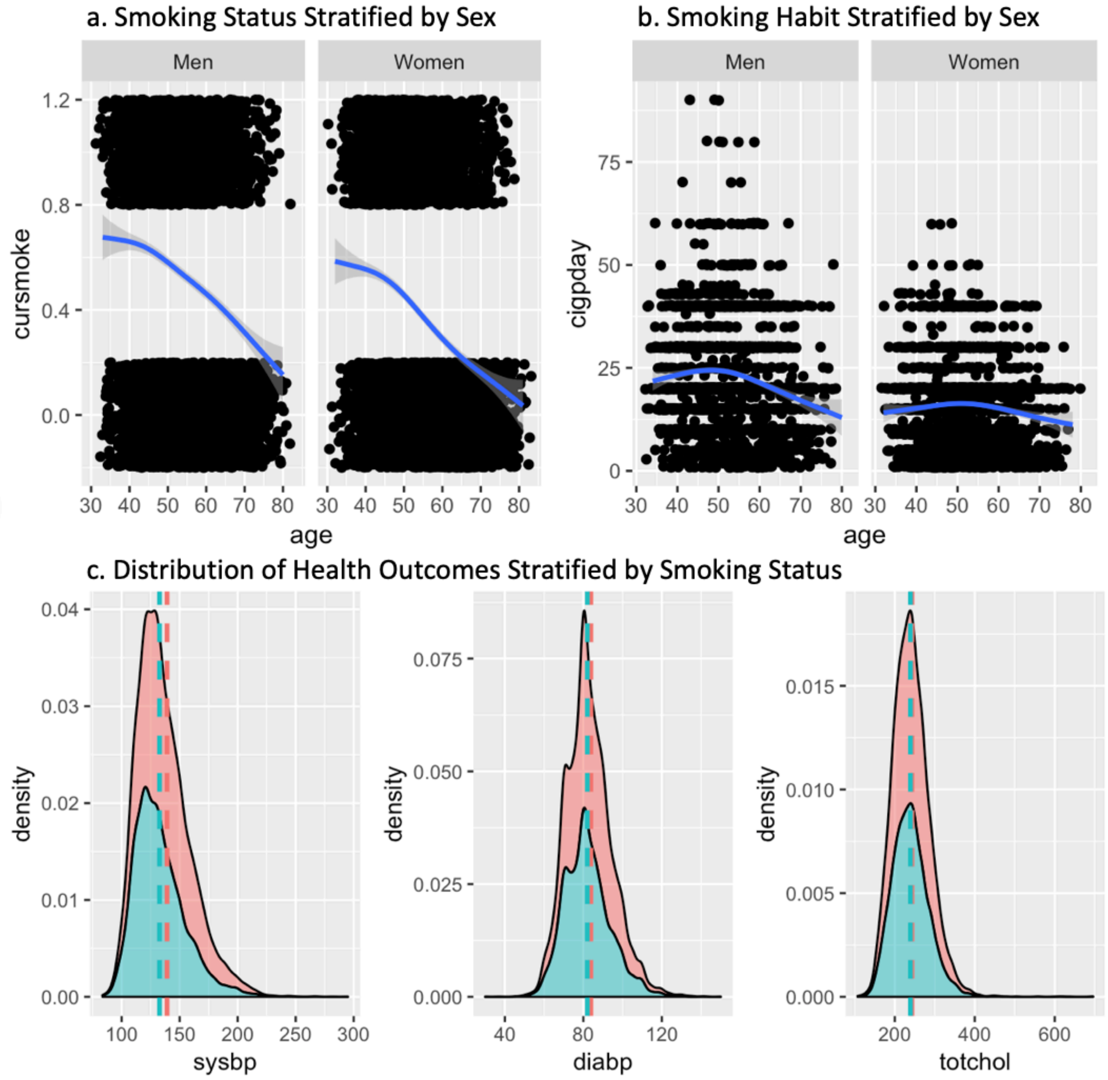
Conclusion

We found that age and gender have negative relationship with smoking habit after adjusted for confounders. Systolic and diastolic blood pressure are negatively associated with current smoking status, while Serum total cholesterol is positively associated with current smoking status after adjusting for confounders. In future analysis, we can consider the effect of the extent of smoking on subjects' physical health. Moreover, we are interested in including this as a part of results and also exploring the related smoking behaviors systematically, like quitting and relapse

Appendix

Figure:

Figure1. Descriptive Statistics for Variables of Interest



a. Smoking Status Stratified by Sex

b. Smoking Habit Stratified by Sex

c. Distribution of Health Outcomes Stratified by Smoking Status

Table:

Table 1a. Coefficient and Significance Level of Covariates in PART I Models

Model			intercept	c.age	totchol	sex	high school; some college; college or more	prevhyp	c.age*sex
Outcome	REM	Correlation Structure							
Cursmoke (binary)	MM	Unstr	-0.32	-0.14 ***	0.0031*	-1.26 ***	0.023;-0.36;-0.47.	-0.34**	0.026*
Cigpday (count)	MM	Unstr	2.73***	-0.00035	0.00077***	-0.46***	0.11**;0.056;0.00037	0.023	0.0088 ***

Significant codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Table 1b. Coefficients and AIC of PART II Models

Model			intercept	cursmoke	age_base	period	c.age	sex	bpmeds	diabetes	pervchd	cursmoke*c_age	Period*sex
Outcome	MM/REM	Correlation Structure											
SBP	MM	Unstr	84.43	-1.54	0.84	4.57	------	1.36	4.82	4.52	1.41	------	-0.17
		Exc	84.74	-1.67	0.84	4.53	------	1.35	4.89	4.17	1.30	------	-0.18
	REM	intercept	135.75 ***	-1.21**	------	------	0.78***	1.50**	5.06***	5.00***	1.68*	0.011	
DBP	MM	Unstr	79.97	-1.29	0.090	-0.17	------	-1.57	1.72	-0.70	-1.35	------	-0.087
		Exc	80.07	-1.36	0.094	-0.25	------	-1.60	1.60	-0.89	-1.44	------	-0.077
	REM	intercept	84.44***	-0.94***	------	------	-0.085 ***	-1.38***	2.02***	-1.11*	-0.78	0.18***	
STC	MM	Unstr	191.27	1.80	0.91	-1.38	------	-0.83	------	-------	-2.44	------	6.15
		Exc	191.77	1.73	0.95	-2.42	------	-0.82	------	------	-2.51	------	6.13
	REM	intercept	233.83 ***	2.55**	------	------	0.37***	12.20 ***	------	------	-4.14**	0.28***	

SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure,STC: Serum Total Cholesterol,MM: Marginal Model, REM: Random Effects Model,Unstr: Unstructured,Exc: Exchangeable, AR(1): Autoregressive. Significant codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Models:

$$\begin{aligned}&\text{logit}\Big(\mathrm{P}\big(Cursmoke_{ij} = 1\big|\mathrm{c.\,age}_{ij}, b_{0i}\big)\Big)\\ &\qquad = \beta_0 + b_{0i} + \beta_1 * c.\,age + \beta_2 I\{woman\} + \beta_3 * c.\,age * I\{woman\} + \beta_4 * totchol\\ &\qquad + \beta_5 I\{high\,school\} + \beta_6 I\{vocational\ school\} + \beta_7 I\{college\ or\ more\}\\ &\qquad + \beta_8 I\{prevhyp\} + \varepsilon_{ij}\end{aligned}$$

$where\ b_{0i} \sim N(0, \sigma_0^2),\ \varepsilon_{ij}{\sim}N(0, \sigma^2),\ \ b_{0i}\ and\ \varepsilon_{ij}\ are\ independent$

$$\begin{aligned}&\log\big(Cursmoke_{ij}\big|\mathrm{c.\,age}_{ij}, b_{0i}\big)\\ &\qquad = \beta_0 + b_{0i} + \beta_1 X_{c.age} + \beta_2 I\{woman\} + \beta_3 X_{c.age} * I\{woman\} + \beta_4 X_{totchol}\\ &\qquad + \beta_5 I\{high\,school\} + \beta_6 I\{vocational\ school\} + \beta_7 I\{college\ or\ more\}\\ &\qquad + \beta_8 I\{prevhyp\}\ + \varepsilon_{ij}\end{aligned}$$

$where\ b_{0i} \sim N(0, \sigma_0^2),\ \varepsilon_{ij}{\sim}N(0, \sigma^2),\ \ b_{0i}\ and\ \varepsilon_{ij}\ are\ independent$

$$\begin{aligned}E(Sysbp_{ij}|b_{0i}) &= \beta_0 + b_{0i} + \beta_1 I\{cursmoke\} + \beta_2 I\{woman\} + \beta_3 X_{c.age} * I\{cursmoke\}\\ &+ \beta_4 I\{bpmeds\} + \beta_5 I\{diabetes\} + \beta_6 X_{c.age} + \beta_7 I\{prevchd\} + \varepsilon_{ij}\end{aligned}$$

$where\ b_{0i} \sim N(0, \sigma_0^2),\ \varepsilon_{ij}{\sim}N(0, \sigma^2),\ \ b_{0i}\ and\ \varepsilon_{ij}\ are\ independent$

$$\begin{aligned}E(Diabp_{ij}|b_{0i}) &= \beta_0 + b_{0i} + \beta_1 I\{cursmoke\} + \beta_2 I\{woman\} + \beta_3 X_{c.age} * I\{cursmoke\}\\ &+ \beta_4 I\{bpmeds\} + \beta_5 I\{diabetes\} + \beta_6 X_{c.age} + \beta_7 I\{prevchd\}\ + \varepsilon_{ij}\end{aligned}$$

$where\ b_{0i} \sim N(0, \sigma_0^2),\ \varepsilon_{ij}{\sim}N(0, \sigma^2),\ \ b_{0i}\ and\ \varepsilon_{ij}\ are\ independent$

$$\begin{aligned}&E(Totchol_{ij}|b_{0i})\\ &\qquad = \beta_0 + b_{0i} + \beta_1 I\{cursmoke\} + \beta_2 I\{woman\} + \beta_3 X_{c.age} * I\{cursmoke\} + \beta_4 X_{c.age}\\ &\qquad + \beta_5 I\{prevhyp\} + \varepsilon_{ij}\end{aligned}$$

$where\ b_{0i} \sim N(0, \sigma_0^2),\ \varepsilon_{ij}{\sim}N(0, \sigma^2),\ \ b_{0i}\ and\ \varepsilon_{ij}\ are\ independent$

Reference:

[1] http://www.bloodpressureuk.org/microsites/kyn/Home/AboutKYN/BPbasics/Thenumbers

[2] A. Virdis, C. Giannarelli, M. Fritsch Neves, S. Taddei, L. Ghiadoni. Cigarette Smoking and Hypertension. *Current Pharmaceutical Design*, Volume 16, Issue 23, **2010**

[3] Irvine H. Page, M.D.; Lena A. Lewis, Ph.D.; Mohammed Moinuddin, Ph.D. Effect of Cigarette Smoking on Serum Cholesterol and Lipoprotein Concentrations. *JAMA.* **1959**;171(11):1500-1502.

[4] Vida Maralani, Understanding the links between education and smoking. *Social Science Research*, Volume 48, November **2014**, Pages 20-34.

[5] Ian H. de Boer, Sripal Bangalore, etc. Diabetes and Hypertension: A Position Statement by the American Diabetes Association. *Diabetes Care* **2017** Sep; 40(9): 1273-1284.

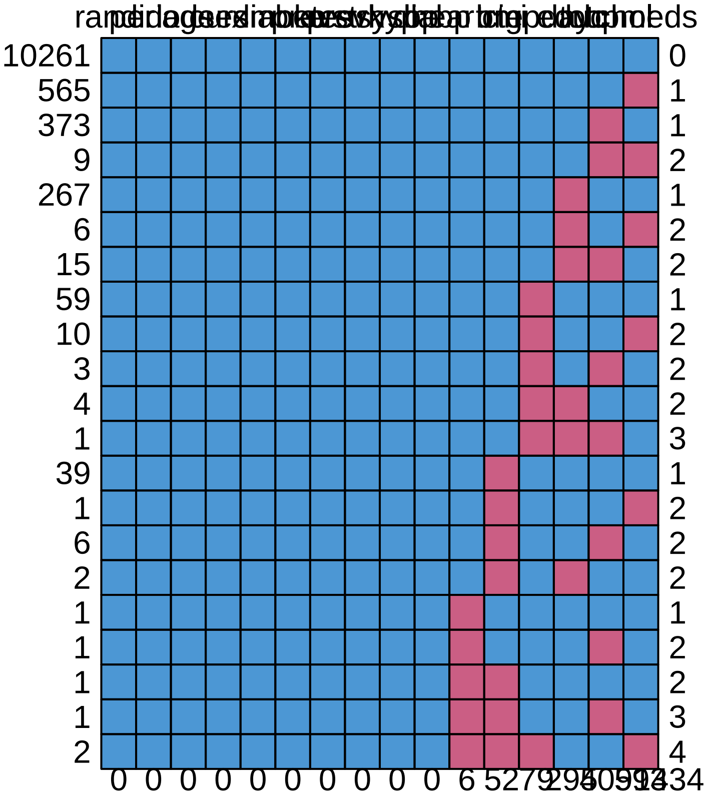
Framingham Heart study - EDA Part

CHUHAN

12/3/2018

Missing Data Iumputation

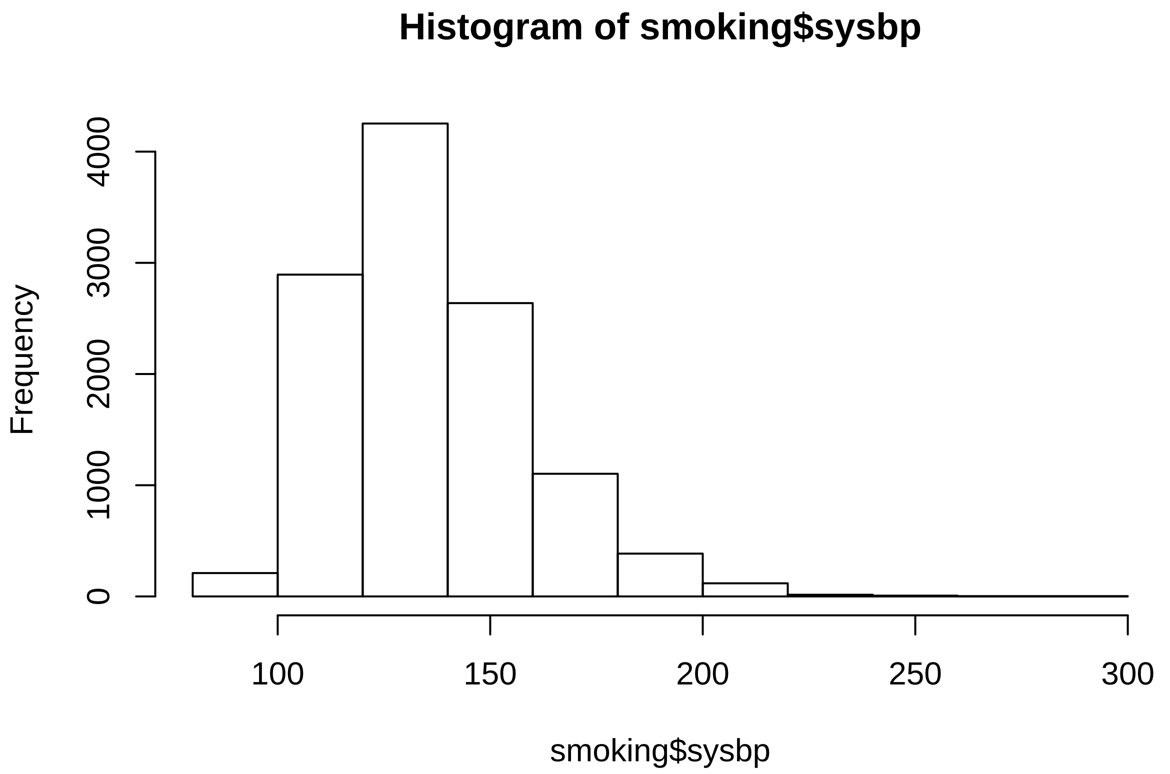
```
md.pattern(smoking, plot = TRUE)
```



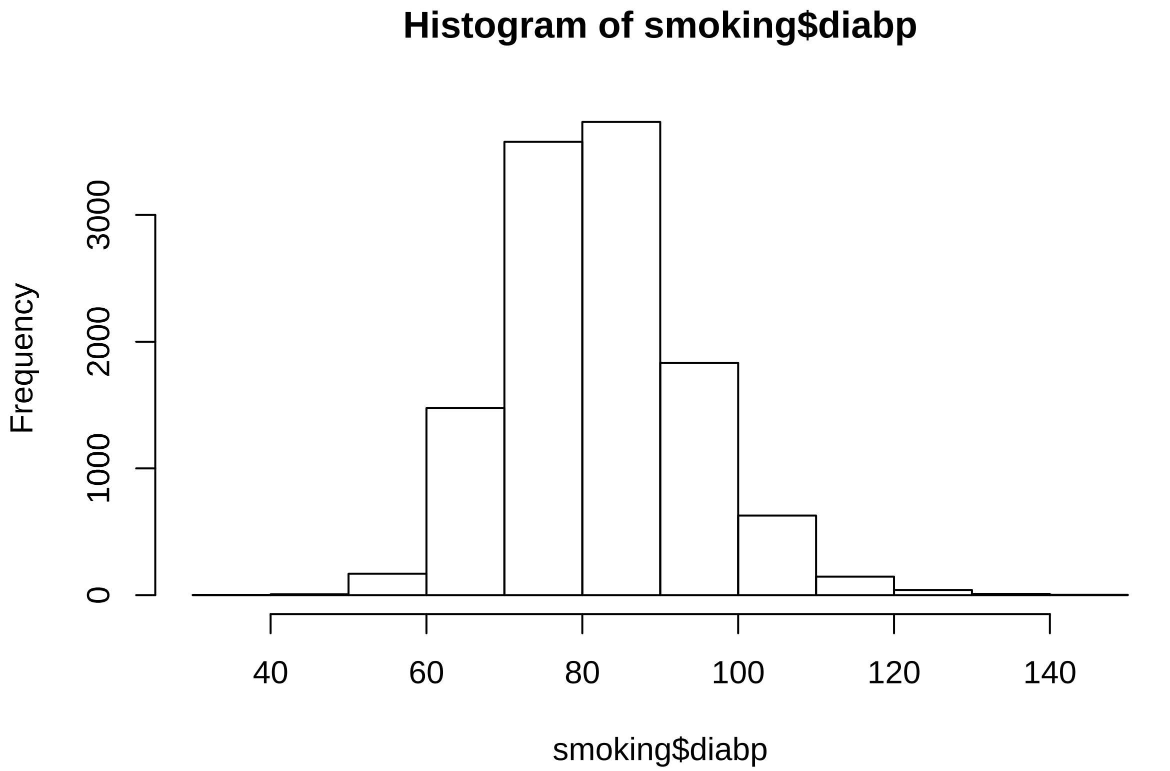
```
##        randid period age sex cursmoke diabetes prevstrk prevhyp sysbp diabp
## 10261       1      1   1   1        1        1        1       1     1     1
## 565         1      1   1   1        1        1        1       1     1     1
## 373         1      1   1   1        1        1        1       1     1     1
## 9           1      1   1   1        1        1        1       1     1     1
## 267         1      1   1   1        1        1        1       1     1     1
## 6           1      1   1   1        1        1        1       1     1     1
## 15          1      1   1   1        1        1        1       1     1     1
## 59          1      1   1   1        1        1        1       1     1     1
## 10          1      1   1   1        1        1        1       1     1     1
## 3           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
## 39          1      1   1   1        1        1        1       1     1     1
## 1           1      1   1   1        1        1        1       1     1     1
## 6           1      1   1   1        1        1        1       1     1     1
## 2           1      1   1   1        1        1        1       1     1     1
## 1          1      1   1   1        1        1        1       1     1     1
## 1          1      1   1   1        1        1        1       1     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
##       heartrte bmi cigpday educ totchol bpmeds
## 10261        1   1       1    1       1      1    0
## 565          1   1       1    1       1      0    1
## 373          1   1       1    1       0      1    1
## 9            1   1       1    1       0      0    2
## 267          1   1       1    0       1      1    1
## 6            1   1       1    0       1      0    2
## 15           1   1       1    0       0      1    2
## 59           1   1       0    1       1      1    1
## 10           1   1       0    1       1      0    2
## 3            1   1       0    1       0      1    2
## 4            1   1       0    0       1      1    2
## 1            1   1       0    0       0      1    3
## 39           1   0       1    1       1      1    1
## 1            1   0       1    1       1      0    2
## 6            1   0       1    1       0      1    2
## 2            1   0       1    0       1      1    2
## 1            0   1       1    1       1      1    1
## 1            0   1       1    1       0      1    2
## 1            0   0       1    1       1      1    2
## 1            0   0       1    1       0      1    3
## 2            0   0       0    1       1      0    4
##              6  52      79  295     409    593 1434
```

Removing outliers

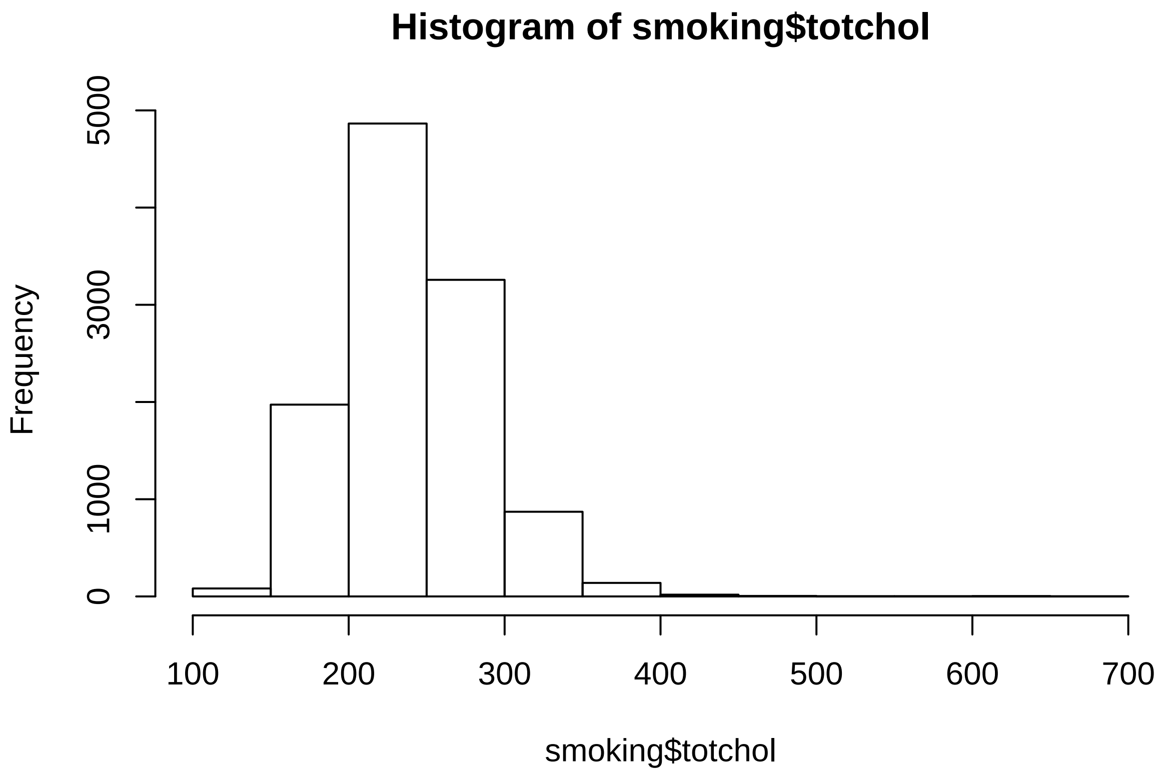
```
hist(smoking$sysbp)
```



```
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
```

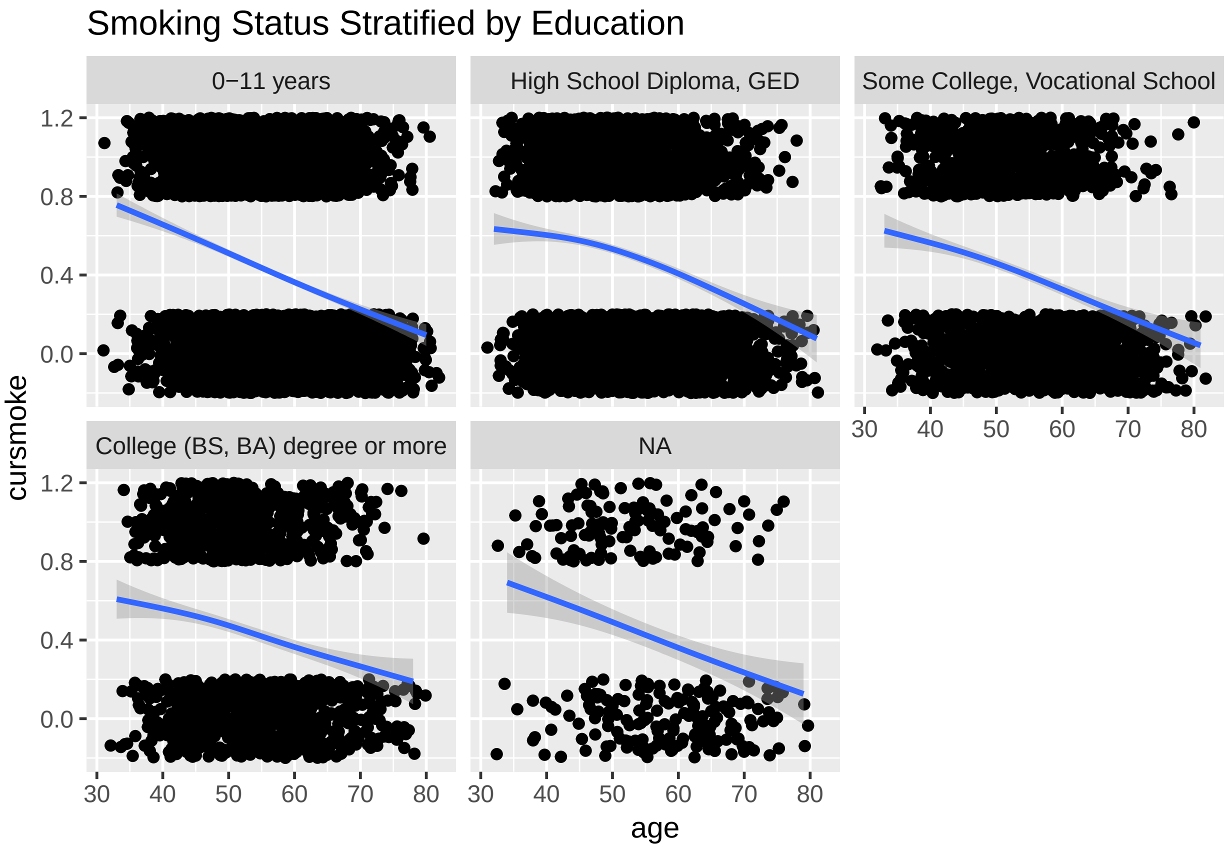
```
##  [1]  610021 1080920 5080716 5807368 8303090  482553 1189726 2577634
##  [9] 5178346 6033947 7411567
```

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

Exploration of Smoking Status

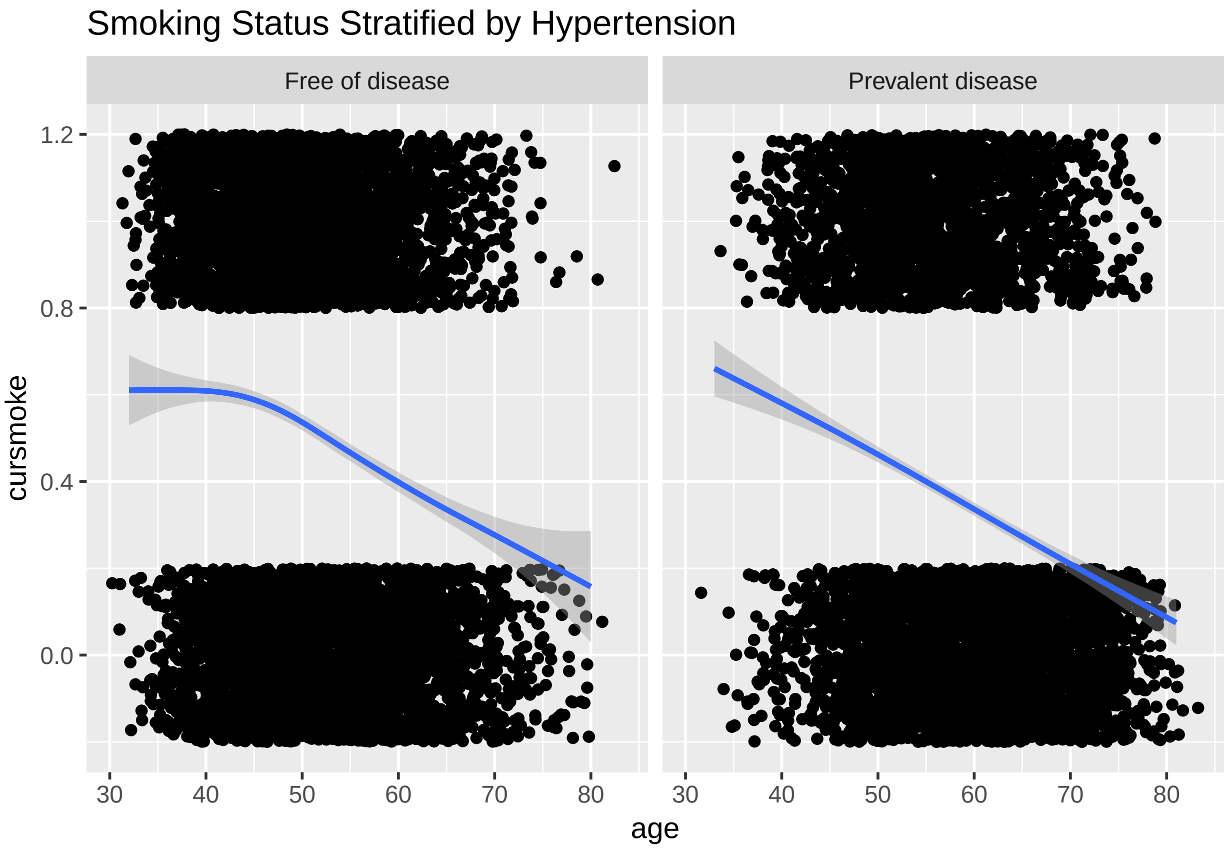
```
#gender
p1p1=ggplot(smoking2,aes(x=age,y=cursmoke))+
  geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~sex) +labs(title="Smoking Status Stratified by Sex")

p1p2=ggplot(subset(smoking2,cursmoke==1),aes(x=age,y=cigpday))+
  geom_jitter(height = 0.2, width = 3)+
  geom_smooth()+facet_wrap(~sex) +labs(title="Smoking Habit Stratified by Sex")

#educ
ggplot(smoking2,aes(x=age,y=cursmoke))+
  geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~educ)+labs(title="Smoking Status Stratified by Education")
```

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



```r
#hypr
ggplot(smoking2,aes(x=age,y=cursmoke))+
  geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+facet_wrap(~prevhyp)+labs(title="Smoking Status Stratified by Hypertension")
```

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

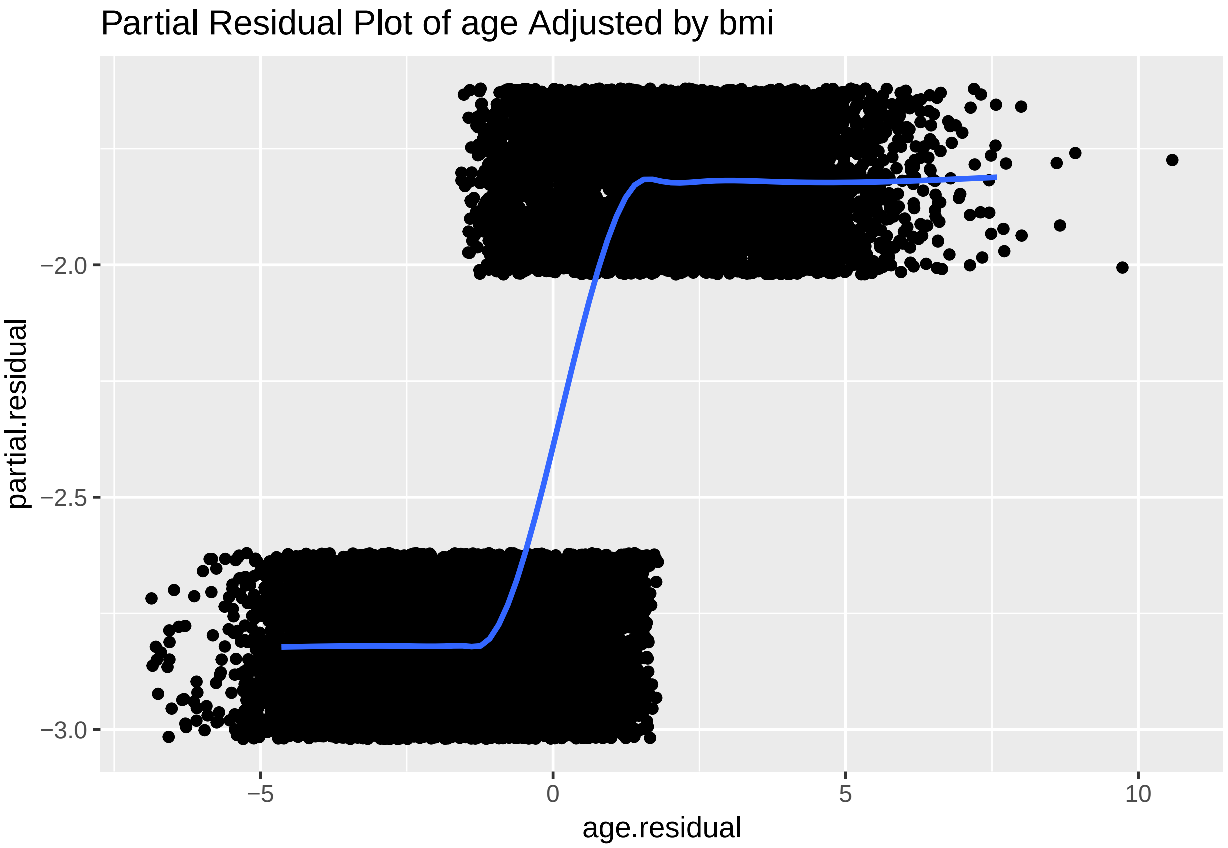


```
#bmi
fit.0=glm(cursmoke~age+sex,data=smoking2,family = binomial)
fit.1=glm(cursmoke~educ+bmi+prevhyp+age+sex,data=smoking2,family = binomial)
summary(fit.1)
```

```
##
## Call:
## glm(formula = cursmoke ~ educ + bmi + prevhyp + age + sex, family = binomial,
##     data = smoking2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9702  -1.0096  -0.6528   1.1158   2.4073
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            5.628887   0.211647  26.596  < 2e-16
## educHigh School Diploma, GED           0.037883   0.049595   0.764   0.4450
## educSome College, Vocational School -0.256815   0.059879  -4.289 1.80e-05
## educCollege (BS, BA) degree or more -0.283114   0.066839  -4.236 2.28e-05
## bmi                                  -0.093633   0.005673 -16.506  < 2e-16
## prevhypPrevalent disease             -0.118991   0.044505  -2.674   0.0075
## age                                  -0.054691   0.002411 -22.683  < 2e-16
## sexWomen                             -0.709911   0.041574 -17.076  < 2e-16
##
## (Intercept)                          ***
## educHigh School Diploma, GED
## educSome College, Vocational School ***
## educCollege (BS, BA) degree or more ***
## bmi                                 ***
## prevhypPrevalent disease            **
## age                                 ***
## sexWomen                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15414  on 11259  degrees of freedom
## Residual deviance: 14061  on 11252  degrees of freedom
##   (340 observations deleted due to missingness)
## AIC: 14077
##
## Number of Fisher Scoring iterations: 4
```

```
beta=matrix(fit.1$coefficients[1:6])
x <- model.matrix(fit.1)[,c(1:6)]
y <- x %*% beta
par.res <- smoking2$cursmoke-y[1]


residual.plot=data.frame(cbind(fit.0$residuals,par.res))
colnames(residual.plot)=c("age.residual","partial.residual")

ggplot(residual.plot,aes(x=age.residual,y=partial.residual)) +
  geom_jitter(height = 0.2, width = 3) +
  geom_smooth()+labs(title="Partial Residual Plot of age Adjusted by bmi")
```

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



```
#plot1=grid.arrange(p1p1,p1p2,widths=c(1,1))
```

Exploration of Health Outcomes

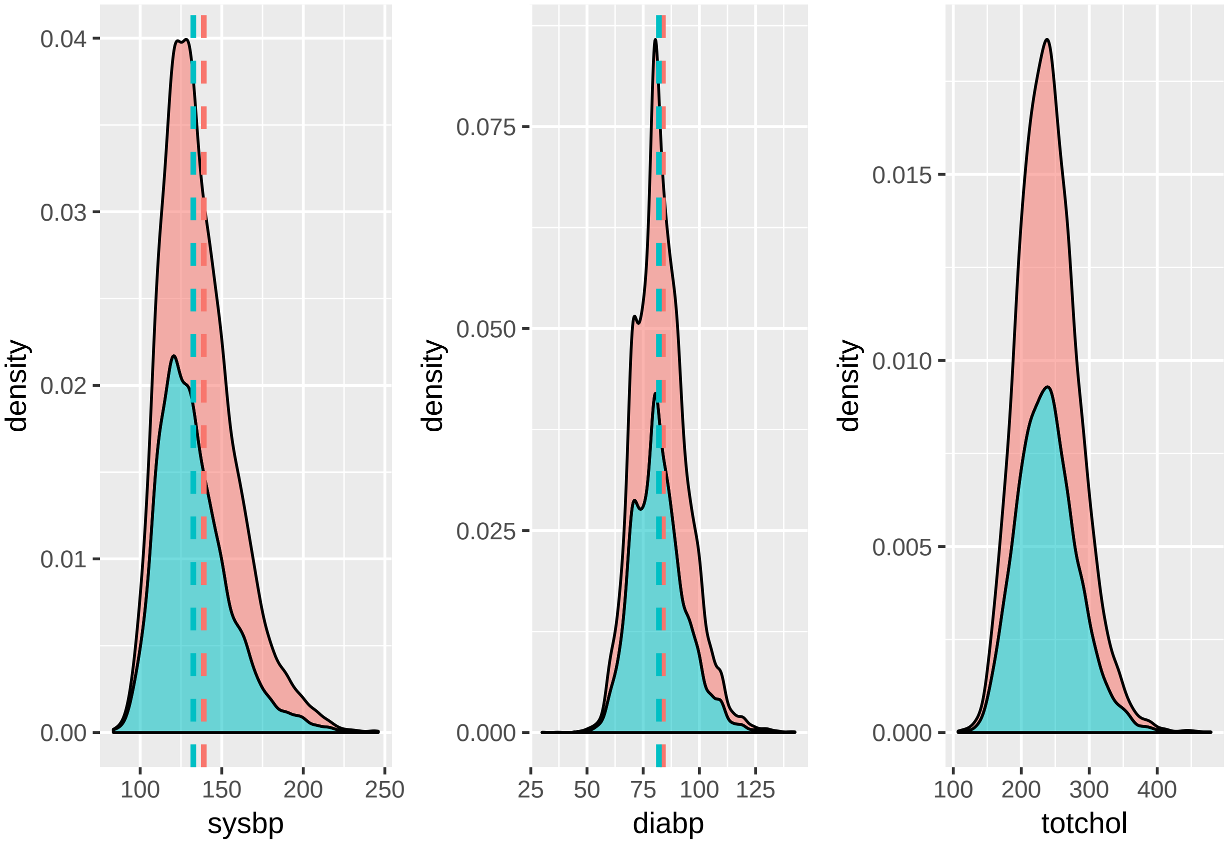
```
mean.sysbp=smoking2 %>%
  group_by(cursmoke) %>%
  summarise(mean=mean(sysbp))
p1=ggplot(data=smoking2,aes(x=sysbp,fill=as.factor(cursmoke),alpha=1/10))+geom_density(position = "stac

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 = "sta
library(gridExtra)
```

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

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

```
##     combine
```

```
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).
```



Framingham Heart Study - Modeling Part

12/17/2018

Contents:

Read data.

```
smoking = read.csv('../frmgham2.csv') %>%
  janitor::clean_names() %>%
  # only keep variavles we're interested in
  dplyr::select(randid, period, age, sex, cursmoke, cigpday, bpmeds, educ, bmi, diabetes,
                diabetes, heartrte, glucose,
                prevap, prevmi, prevstrk, prevchd, prevhyp, sysbp, diabp, totchol) %>%
  mutate(sex = as.factor(sex),
         sex = recode(sex, '1'='Men', '2'='Women'),
         bpmeds = as.factor(bpmeds),
         bpmeds = recode(bpmeds, '0'='Not currently used', '1'='Current Use'),
         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'),
         diabetes = as.factor(diabetes),
         diabetes = recode(diabetes, '0'='Not a diabetic', '1'='Diabetic'),
         prevchd = as.factor(prevchd),
         prevchd = recode(prevchd, '0'='Free of disease', '1'='Prevalent disease'),
         prevhyp = as.factor(prevhyp),
         prevhyp = recode(prevhyp, '0'='Free of disease', '1'='Prevalent disease'))

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))

smoking.ol = smoking %>%
  filter(! randid  %in% ol.id) %>%
  mutate(c.age = age - 55)
```

Descriptive statistics.

```
smoking %>%
  group_by(sex) %>%
  summarise(mean(age))
```

```
## # A tibble: 2 x 2
##   sex   `mean(age)`
##   <fct>       <dbl>
## 1 Men          54.5
## 2 Women        55.0
```

```
smoking %>%
  group_by(sex) %>%
  filter(!is.na(cigpday)) %>%
  summarise(mean(cigpday))
```

```
## # A tibble: 2 x 2
##   sex   `mean(cigpday)`
##   <fct>           <dbl>
## 1 Men             11.6
## 2 Women            5.67
```

Part I

Search for potential confounders by rule of thumb

```
#names(smoking.ol)

conf = data.frame(matrix(nrow = 0, ncol = 2))
for (i in c(8:ncol(smoking.ol)-1)){
  fit.org = glmer(cursmoke ~ age + sex + (1|randid),
                  data = smoking.ol, family = binomial, nAGQ=0)
  a = as.vector(summary(fit.org)$coefficients[2:3,1])
  fit.conf = glmer(cursmoke ~ age + sex + smoking.ol[,i] + (1|randid),
                   data = smoking.ol, family = binomial, nAGQ=0)
  b = as.vector(summary(fit.conf)$coefficients[2:3,1])
  conf = rbind(conf, ((b-a)/a)*100)
}

colnames(conf) <- c('age', 'sex')
row.names(conf) <- names(smoking.ol)[8:ncol(smoking.ol)-1]
conf
```

```
##                  age          sex
## bpmeds   -1.69706110  1.242898223
## educ      1.32761956  2.282422125
## bmi      -2.10613892  8.816749587
## diabetes -0.61020441  0.108575654
## heartrte  2.33501383  4.492940943
## glucose  -6.86679044  0.018007096
## prevap   -1.49655545  1.406969393
## prevmi   -1.06359899  2.136138270
## prevstrk -0.05728646 -0.004290296
## prevchd  -2.05779748  2.202559114
## prevhyp  -5.23141658  0.361767658
## sysbp    -3.46811734 -0.625489604
## diabp    -0.57167858  0.969523886
## totchol  -0.06924931  3.866558528
```

Mixed Effects Model

```
fit.bin = glmer(cursmoke ~ c.age*sex+totchol+educ+prevhyp+(1|randid),
                data = smoking.ol, family = binomial, nAGQ=0)
summary(fit.bin)$coefficients
```

```
##                                        Estimate  Std. Error     z value
## (Intercept)                        -0.320093321 0.356223421  -0.8985746
## c.age                              -0.140749680 0.009603514 -14.6560599
## sexWomen                           -1.259724522 0.166361810  -7.5721977
## totchol                             0.003128688 0.001367425   2.2880143
## educHigh School Diploma, GED        0.022768287 0.196843702   0.1156668
## educSome College, Vocational School -0.360260452 0.237859303  -1.5145947
## educCollege (BS, BA) degree or more -0.470526471 0.267690712  -1.7577243
## prevhypPrevalent disease           -0.343851337 0.126110030  -2.7265978
## c.age:sexWomen                      0.025612617 0.012941877   1.9790497
##                                        Pr(>|z|)
## (Intercept)                        3.688793e-01
## c.age                              1.232187e-48
## sexWomen                           3.669624e-14
## totchol                            2.213669e-02
## educHigh School Diploma, GED       9.079166e-01
## educSome College, Vocational School 1.298751e-01
## educCollege (BS, BA) degree or more 7.879443e-02
## prevhypPrevalent disease           6.399098e-03
## c.age:sexWomen                     4.781041e-02
```

```
smoking.cigar = smoking.ol %>%
  filter(cursmoke == !0)
fit.poi = glmer(cigpday ~ c.age*sex+totchol+educ+prevhyp
                +(1|randid), data = smoking.cigar, family = poisson,nAGQ = 0)
summary(fit.poi)$coefficients
```

```
##                                         Estimate   Std. Error
## (Intercept)                         2.7277706787 0.0467676695
## c.age                              -0.0003505435 0.0010382579
## sexWomen                           -0.458
## totchol                             0.000
## educHigh School Diploma, GED        0.102
## educSome College, Vocational School  0.055
## educCollege (BS, BA) degree or more  0.0003721177 0.0487926040
## prevhypPrevalent disease            0.0233427561 0.0146009019
## c.age:sexWomen                      0.0088016603 0.0015254369
##
## (Intercept)                        58.32
## c.age                              -0.33
## sexWomen                          -14.894798141 3.562590e-50
## totchol                             4.743490903 2.100664e-06
```

One of the main scientific questions of interest for part I was the relationship between AGE and smoking status and number of cigarettes. The results from your model are missing for age in the write-up (you only have results for sex) (-5 points)

We were also interested in if sex modifies the relationship between age and smoking. This interaction term is not included in your model (-3 points)

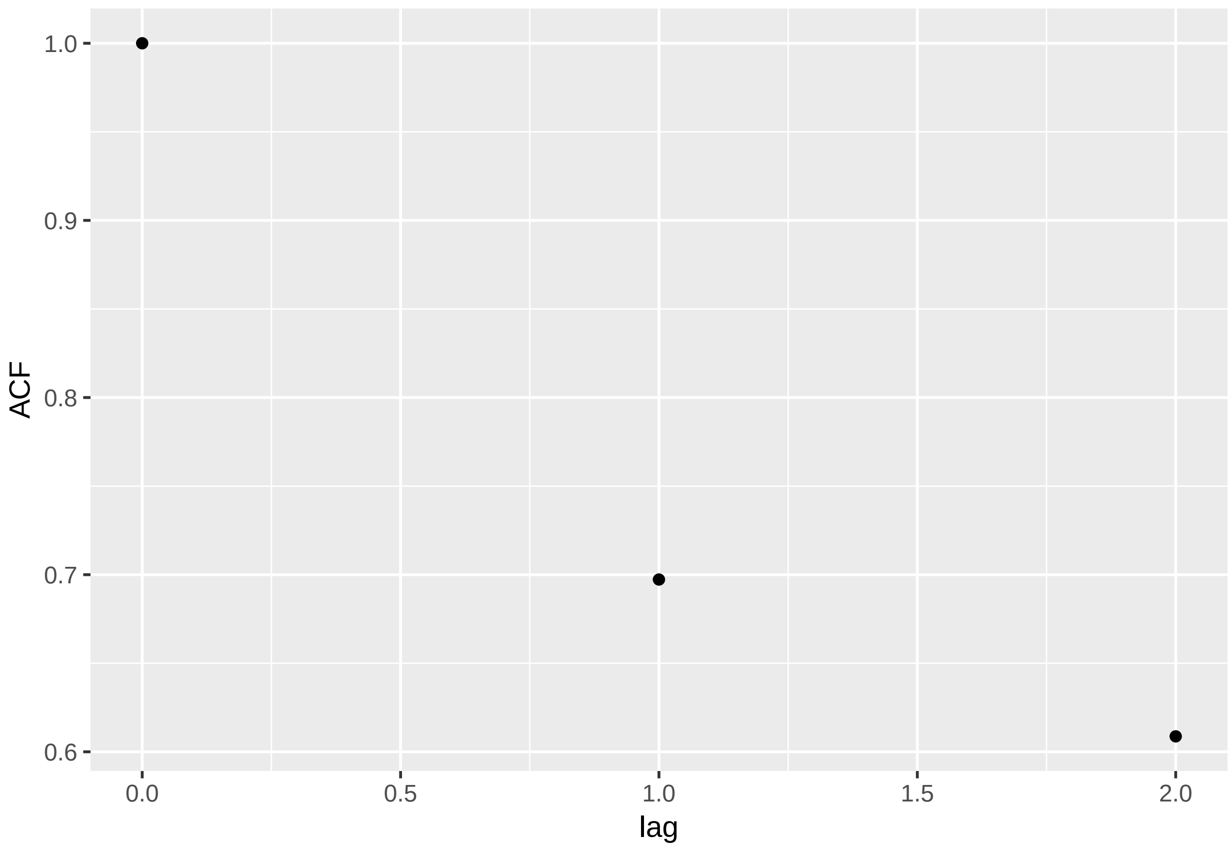
```
## educHigh School Diploma, GED          2.901297251 3.716212e-03
## educSome College, Vocational School   1.239826108 2.150397e-01
## educCollege (BS, BA) degree or more   0.007626519 9.939150e-01
## prevhypPrevalent disease              1.598720153 1.098828e-01
## c.age:sexWomen                        5.769927566 7.930561e-09
```

Part II

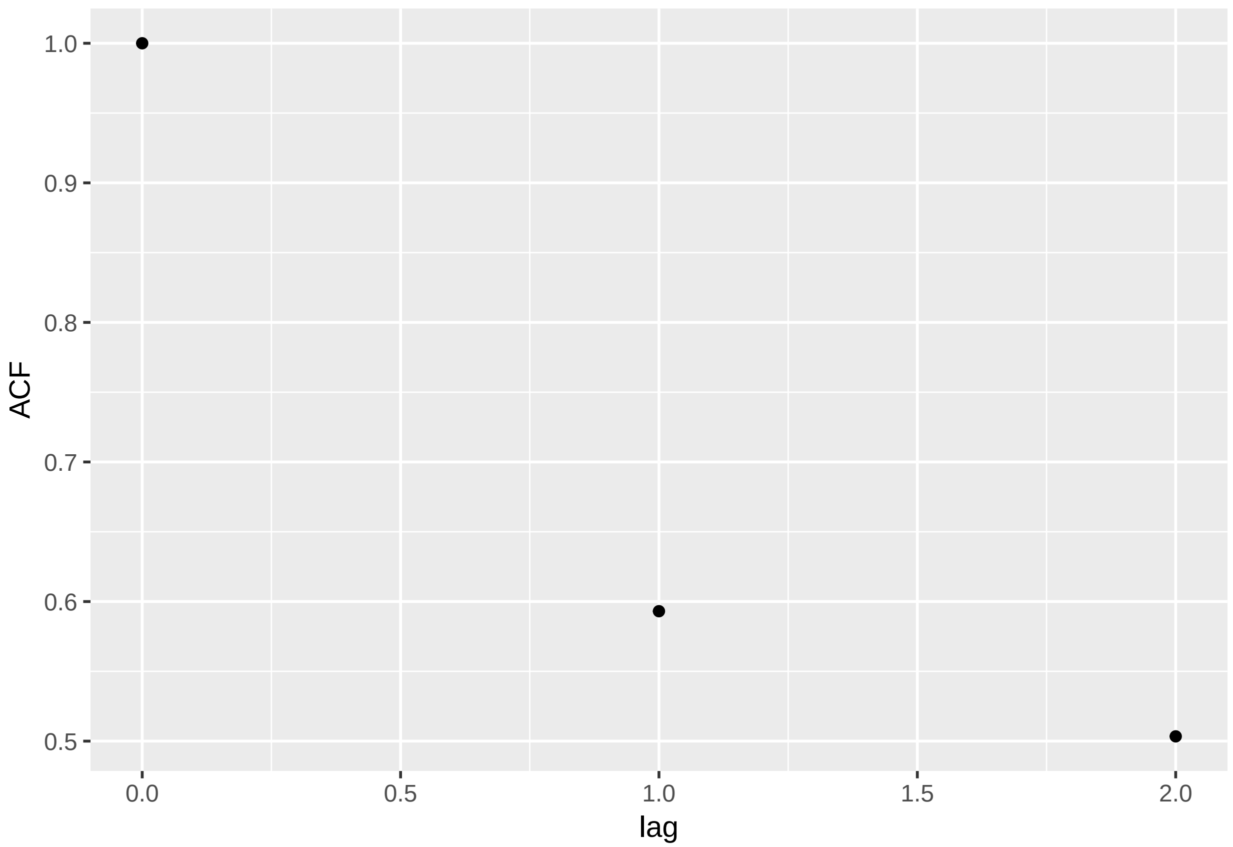
Marginal Model

```
smoking.complete0 = read.csv('../complete.v1.csv') %>%
  dplyr::select(-X) %>%
  group_by(randid) %>%
  mutate(n = sum(period)) %>%
  filter(n == 6)

smoking.complete = smoking.complete0 %>%
  dplyr::select(randid, cursmoke, age, period, sex, bpmeds, diabetes, prevchd,
                sysbp, diabp, totchol) %>%
  mutate(cursmoke = recode(cursmoke, '0'='Not smoker', '1'='Smoker'),
         sex = recode(sex, '1'='Men', '2'='Women'),
         bpmeds = recode(bpmeds, '0'='Not used', '1'='Used'),
         diabetes = recode(diabetes, '0'='are not diabetic', '1'='Diabetic'),
         prevchd = recode(prevchd, '0'='Free of disease', '1'='Prevalent disease')) %>%
  filter(! randid  %in% ol.id)

smoking.margin = smoking.complete %>%
  filter(period == 1) %>%
  dplyr::select(randid, age_base = age) %>%
  right_join(smoking.complete, by = "randid")
```

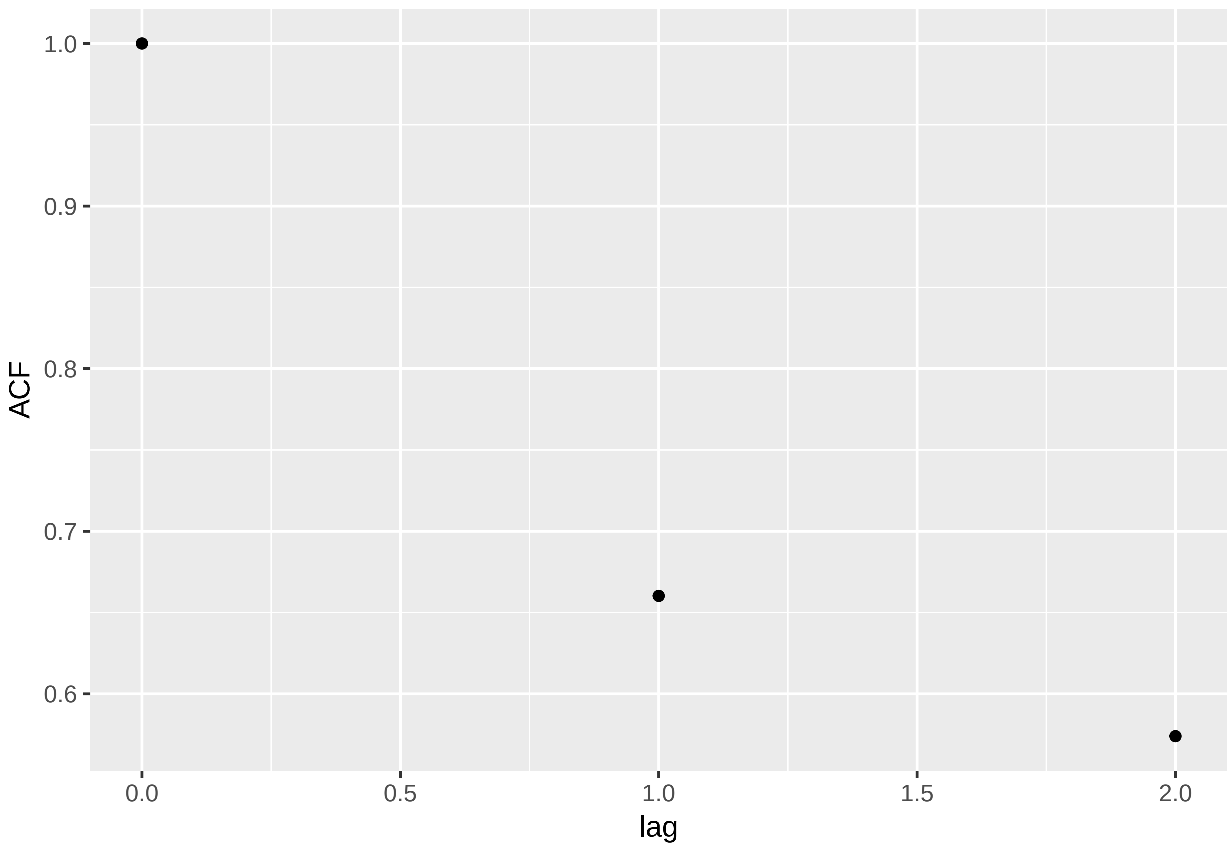
```
############# (1) systolic blood pressure #############
##fit the saturated model
saturated.fit = gls(sysbp ~ as.factor(period), data = smoking.margin)

##autocorrelation function
acf.out = ACF(saturated.fit, form = ~ 1| randid)

ggplot(acf.out, aes(x = lag, y = ACF)) +
  geom_point()
```



```
############# (2) diastolic blood pressure #############
##fit the saturated model
saturated.fit = gls(diabp ~ as.factor(period), data = smoking.margin)

##autocorrelation function
acf.out = ACF(saturated.fit, form = ~ 1| randid)

ggplot(acf.out, aes(x = lag, y = ACF)) +
  geom_point()
```



```
############# (3) serum total cholesterol #############
##fit the saturated model
saturated.fit = gls(totchol ~ as.factor(period), data = smoking.margin)

##autocorrelation function
acf.out = ACF(saturated.fit, form = ~ 1| randid)

ggplot(acf.out, aes(x = lag, y = ACF)) +
  geom_point()
```



Thus, we choose unstructured and exchangeable correlation structure.

```
############# (1) systolic blood pressure #############
# correlation structure: unstructured
fit1.un = gls(sysbp ~ cursmoke*period + age_base + sex + bpmeds + diabetes + prevchd,
              data = smoking.margin,
              correlation = corSymm(form = ~1|randid))
coef(summary(fit1.un))
```

```
##                               Value  Std.Error    t-value       p-value
## (Intercept)              84.8978565 1.96072958 43.2991156  0.000000e+00
## cursmokeSmoker           -2.1314295 0.83193838 -2.5620041  1.042215e-02
## period                    4.3486540 0.23181443 18.7592032  3.873065e-77
## age_base                  0.8423805 0.03674582 22.9245271 2.824594e-113
## sexWomen                  0.9930834 0.60725095  1.6353757  1.020031e-01
## bpmedsUsed                4.8249918 0.62883529  7.6729024  1.845383e-14
## diabetesDiabetic          4.5302290 0.98710839  4.5893937  4.501831e-06
## prevchdPrevalent disease  1.4430727 0.83062854  1.7373262  8.236171e-02
## cursmokeSmoker:period     0.2930963 0.35346019  0.8292202  4.070005e-01
```

```
# correlation structure: exchangeable
fit1.cs = gls(sysbp ~ cursmoke*period + age_base + sex + bpmeds + diabetes + prevchd,
               data = smoking.margin,
               correlation = corCompSymm(form = ~1|randid))
coef(summary(fit1.cs))
```

```
##                               Value  Std.Error    t-value       p-value
## (Intercept)              85.2498068 1.95230559 43.6662208  0.000000e+00
## cursmokeSmoker           -2.3339404 0.80183039 -2.9107657  3.613741e-03
## period                    4.2900770 0.21763262 19.7124725  7.939635e-85
## age_base                  0.8400527 0.03668416 22.8996012 4.859118e-113
## sexWomen                  0.9782209 0.60670264  1.6123564  1.069173e-01
## bpmedsUsed                4.9038797 0.62544239  7.8406577  4.959560e-15
## diabetesDiabetic          4.1835050 0.96160595  4.3505398  1.372078e-05
## prevchdPrevalent disease  1.3330960 0.81517181  1.6353559  1.020073e-01
## cursmokeSmoker:period     0.3303391 0.33521002  0.9854689  3.244186e-01
```

```
############# (2) diastolic blood pressure #############
# correlation structure: unstructured
fit2.un = gls(diabp ~ cursmoke*period + age_base + sex + bpmeds + diabetes + prevchd,
              data = smoking.margin,
              correlation = corSymm(form = ~1|randid))
coef(summary(fit2.un))
```

```
##                                Value  Std.Error   t-value      p-value
## (Intercept)              81.20135094 1.08247991 75.014187 0.000000e+00
## cursmokeSmoker           -3.96259460 0.48926443 -8.099086 6.212427e-16
## period                   -0.78240502 0.13667140 -5.724716 1.067211e-08
## age_base                  0.09093227 0.02012327  4.518762 6.294679e-06
## sexWomen                 -1.78416407 0.33197716 -5.374358 7.865886e-08
## bpmedsUsed                1.78020098 0.37089810  4.799704 1.613007e-06
## diabetesDiabetic         -0.70044871 0.57117378 -1.226332 2.201038e-01
## prevchdPrevalent disease -1.31718014 0.48242142 -2.730352 6.338233e-03
## cursmokeSmoker:period     1.36214393 0.20917297  6.512046 7.785108e-11
```

```
# correlation structure: exchangeable
fit2.cs = gls(diabp ~ cursmoke*period + age_base + sex + bpmeds + diabetes + prevchd,
               data = smoking.margin,
               correlation = corCompSymm(form = ~1|randid))
coef(summary(fit2.cs))
```

```
##                                Value  Std.Error   t-value      p-value
## (Intercept)              81.29530835 1.07971341 75.293414 0.000000e+00
## cursmokeSmoker           -4.03192096 0.47340276 -8.516894 1.885574e-17
## period                   -0.85551737 0.12991724 -6.585095 4.784286e-11
## age_base                  0.09493386 0.02013374  4.715164 2.449220e-06
## sexWomen                 -1.80087852 0.33239605 -5.417870 6.177346e-08
## bpmedsUsed                1.67044397 0.36847583  4.533388 5.874841e-06
## diabetesDiabetic         -0.88723745 0.56049061 -1.582966 1.134622e-01
## prevchdPrevalent disease -1.40010914 0.47536216 -2.945353 3.233619e-03
## cursmokeSmoker:period     1.36779982 0.20032792  6.827804 9.145254e-12
```

```
############# (3) serum total cholesterol #############
# correlation structure: unstructured
fit3.un = gls(totchol ~ cursmoke*period + age_base + sex + prevchd,
              data = smoking.margin,
              correlation = corSymm(form = ~1|randid))
coef(summary(fit3.un))
```

```
##                               Value  Std.Error   t-value      p-value
## (Intercept)              184.947542 4.25292383 43.487151 0.000000e+00
## cursmokeSmoker            -2.850882 1.78533828 -1.596830 1.103365e-01
## period                     1.114636 0.47999857  2.322166 2.024485e-02
## age_base                   0.933565 0.07992506 11.680505 2.610948e-31
## sexWomen                   11.844684 1.32327677  8.951026 4.176673e-19
## prevchdPrevalent disease  -3.281589 1.79537171 -1.827805 6.760990e-02
## cursmokeSmoker:period      2.711610 0.75055503  3.612807 3.044488e-04
```

```
# correlation structure: exchangeable
fit3.cs = gls(totchol ~ cursmoke*period + age_base + sex + prevchd,
              data = smoking.margin,
              correlation = corCompSymm(form = ~1|randid))
coef(summary(fit3.cs))
```

```
##                                Value  Std.Error   t-value      p-value
## (Intercept)              185.83502652 4.24503436 43.777037 0.000000e+00
## cursmokeSmoker            -2.81787885 1.76309001 -1.598261 1.100177e-01
## period                     0.09617635 0.46793169  0.205535 8.371585e-01
## age_base                   0.96952784 0.07980784 12.148279 1.035398e-33
## sexWomen                  11.41409199 1.32232782  8.631817 7.000129e-18
## prevchdPrevalent disease  -3.39664669 1.78292061 -1.905103 5.679661e-02
## cursmokeSmoker:period      2.67213311 0.73772284  3.622137 2.936878e-04
```

Compare marginal models

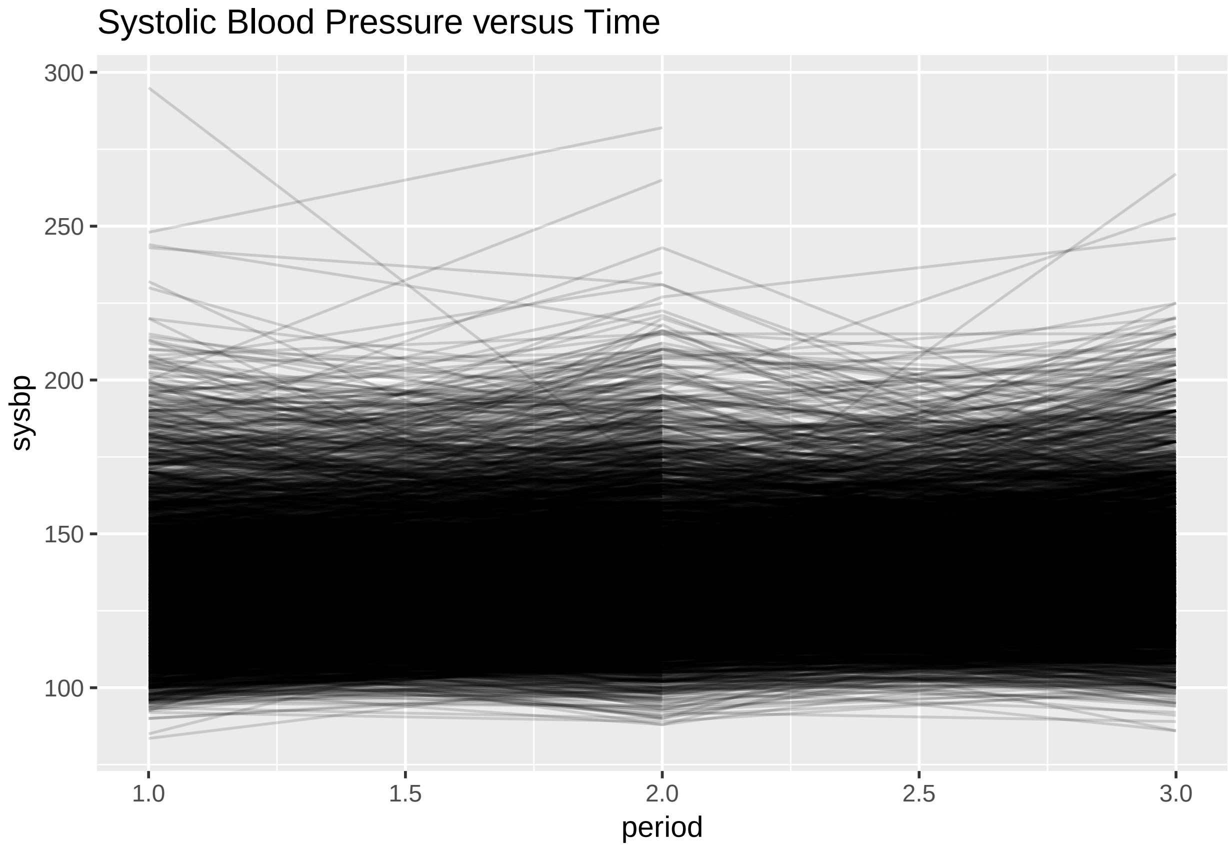
```
aics = data.frame(matrix(nrow = 0, ncol = 6))
aics = rbind(aics, as.vector(c(AIC(fit1.un),AIC(fit1.cs),AIC(fit2.un),
                               AIC(fit2.cs),AIC(fit3.un),AIC(fit3.cs))))
colnames(aics) <- c('fit1.un', 'fit1.cs', 'fit2.un', 'fit2.cs', 'fit3.un', 'fit3.cs')
aics
```

```
##    fit1.un  fit1.cs fit2.un  fit2.cs  fit3.un  fit3.cs
## 1 81049.12 81176.13 70697.5 70767.51 96212.01 96270.03
```

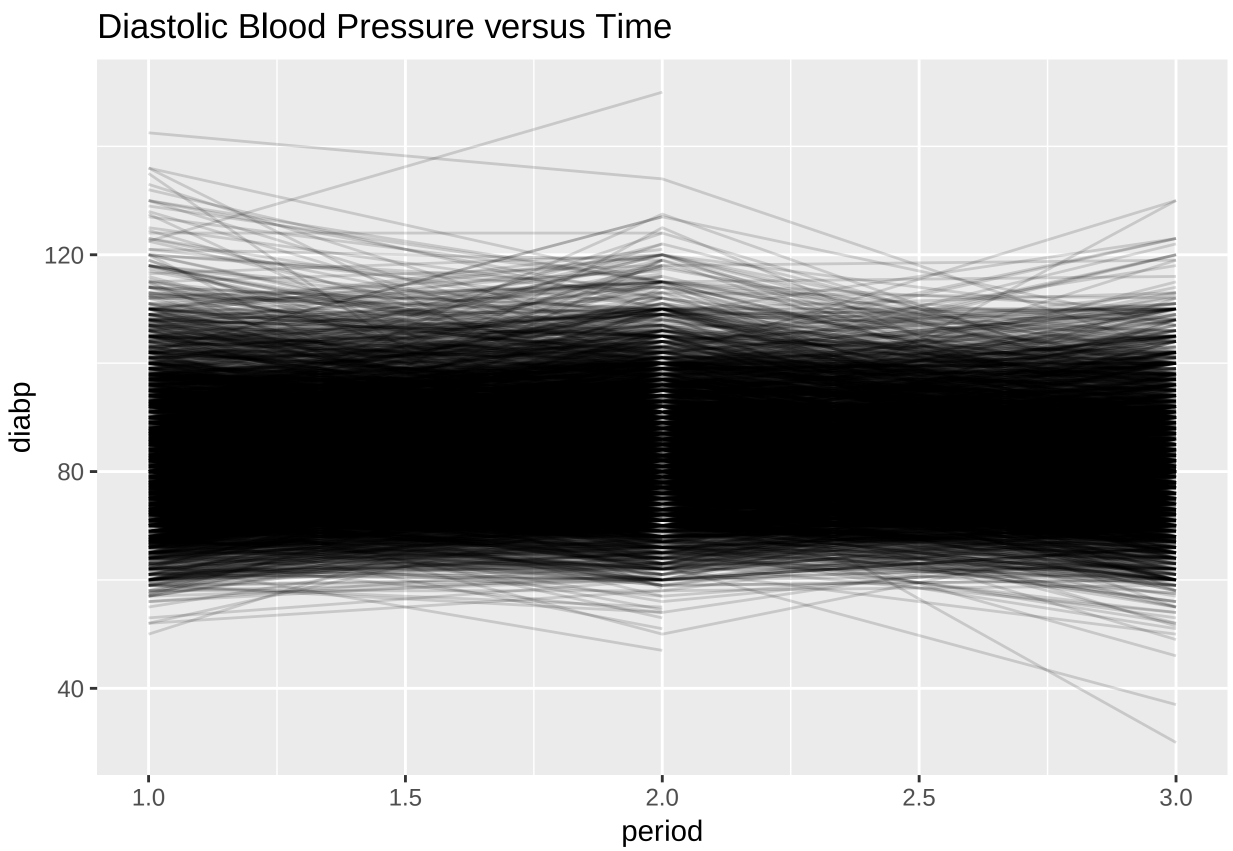
Mixed Effects Model

```
smoking1.ol = smoking %>%
  mutate(cursmoke = as.factor(cursmoke),
         cursmoke = recode(cursmoke, '0'='Not current smoker', '1'='Current smoker'))
```

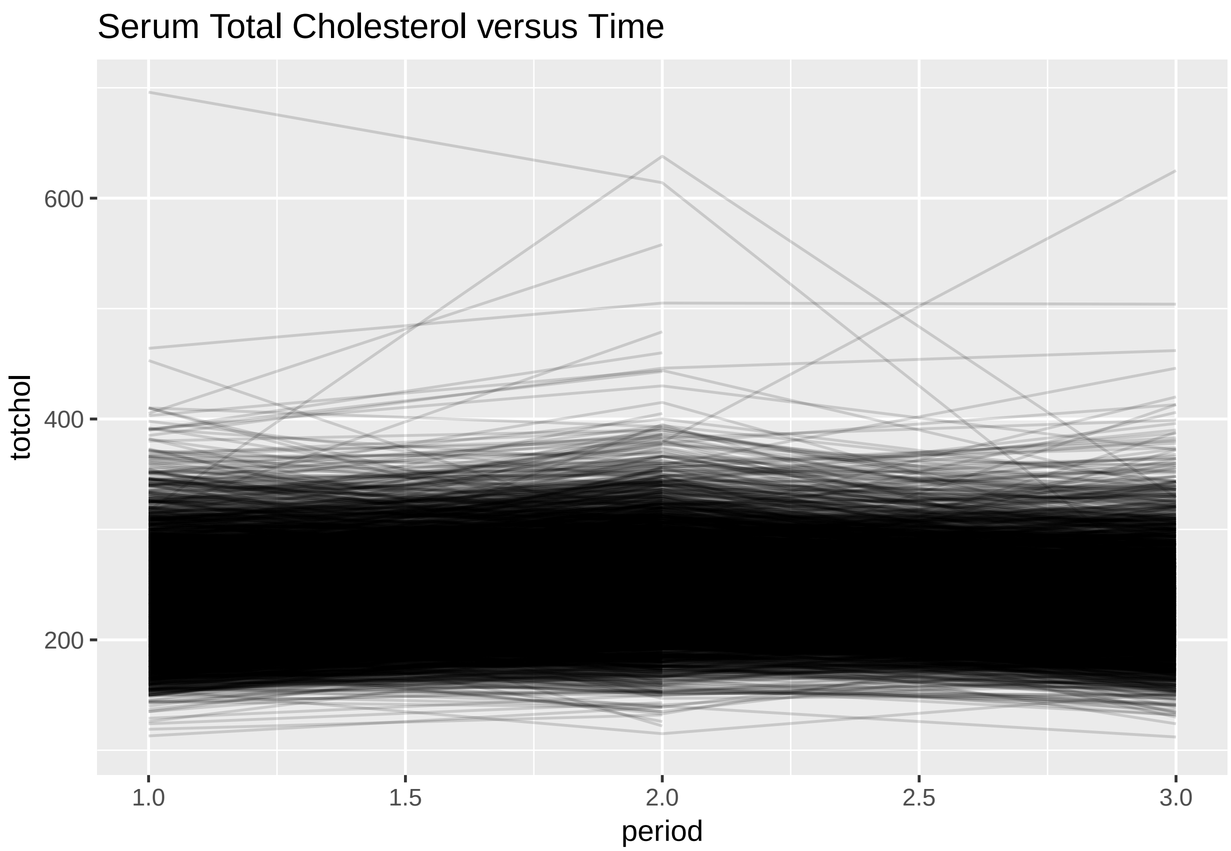
```
ggplot(data = smoking1.ol, aes(x = period, y = sysbp, group = randid)) +
  geom_line(alpha=0.15) +
  ggtitle('Systolic Blood Pressure versus Time')
```



```
ggplot(data = smoking1.ol, aes(x = period, y = diabp, group = randid)) +
  geom_line(alpha=0.15) +
  ggtitle('Diastolic Blood Pressure versus Time')
```



```
ggplot(data = smoking1.ol, aes(x = period, y = totchol, group = randid)) +
  geom_line(alpha=0.15) +
  ggtitle('Serum Total Cholesterol versus Time')
```



Thus, we propose a random intercept and slope model. Also include a smoking status by age interaction.

```
smoking2 = smoking1.ol %>%
  filter(! randid  %in% ol.id) %>%
  mutate(c.age = age - 55)
#mean(smoking2$age) # 54.78302
```

```
############## (1) sysbp ##############
fit.slope.sysbp = lmer(sysbp ~ cursmoke*c.age + sex + bpmeds + diabetes + prevchd
                       + (c.age|randid), data = smoking2)
summary(fit.slope.sysbp)$varcor
```

```
##  Groups   Name        Std.Dev. Corr
##  randid   (Intercept) 16.33282
##           c.age        0.29977 1.000
##  Residual             12.55219
```

```
fit.int.sysbp = lmer(sysbp ~ cursmoke*c.age + sex + bpmeds + diabetes + prevchd
                     + (1|randid), data = smoking2)
summary(fit.int.sysbp)$coefficients
```

```
##                                  Estimate Std. Error     t value
## (Intercept)                   135.7467627 0.48809084 278.1178242
## cursmokeCurrent smoker         -1.2075890 0.44504016  -2.7134382
## c.age                           0.7791444 0.02790010  27.9262259
## sexWomen                        1.5007207 0.56133201   2.6734992
## bpmedsCurrent Use               5.0554481 0.61292433   8.2480786
## diabetesDiabetic                4.9958428 0.92689725   5.3898561
## prevchdPrevalent disease       1.6816289 0.75946667   2.2142235
## cursmokeCurrent smoker:c.age   0.0106271 0.03902215   0.2723352
```

```
# random intercept only
coefs.sysbp = data.frame(coef(summary(fit.int.sysbp)))
# p.values
2 * (1 - pnorm(abs(coefs.sysbp$t.value)))
```

```
## [1] 0.000000e+00 6.658898e-03 0.000000e+00 7.506445e-03 2.220446e-16
## [6] 7.051411e-08 2.681341e-02 7.853643e-01
```

```
############## (2) diabp ##############
fit.slope.diabp = lmer(diabp ~ cursmoke*c.age + sex + bpmeds + diabetes + prevchd
                       + (c.age|randid), data = smoking2)
summary(fit.slope.diabp)$varcor
```

```
##  Groups   Name        Std.Dev. Corr
##  randid   (Intercept) 8.74510
##           c.age       0.16547  0.184
##  Residual             7.50911
```

```
fit.int.diabp = lmer(diabp ~ cursmoke*c.age + sex + bpmeds + diabetes + prevchd
                     + (1|randid), data = smoking2)
summary(fit.int.diabp)$coefficients
```

```
##                                   Estimate Std. Error    t value
## (Intercept)                    84.44198853 0.27321889 309.063509
## cursmokeCurrent smoker         -0.94022634 0.25593621  -3.673675
## c.age                          -0.08473978 0.01616762  -5.241328
## sexWomen                       -1.38495010 0.31120587  -4.450270
## bpmedsCurrent Use               2.01535787 0.35977937   5.601649
## diabetesDiabetic               -1.10733641 0.53851209  -2.056289
## prevchdPrevalent disease       -0.78180941 0.44102236  -1.772721
## cursmokeCurrent smoker:c.age    0.18167727 0.02281520   7.962991
```

```
# random intercept only
coefs.diabp = data.frame(coef(summary(fit.int.diabp)))
#p.values
2 * (1 - pnorm(abs(coefs.diabp$t.value)))
```

```
## [1] 0.000000e+00 2.390872e-04 1.594253e-07 8.576249e-06 2.123217e-08
## [6] 3.975468e-02 7.627502e-02 1.776357e-15
```

```
############## (3) totchol ##############
fit.slope.totchol = lmer(totchol ~ cursmoke*c.age + sex + prevchd + (c.age|randid),
                         data = smoking2)
summary(fit.slope.totchol)$varcor
```

```
##  Groups   Name        Std.Dev. Corr
##  randid   (Intercept) 34.3701
##           c.age        1.0195  0.014
##  Residual             25.0122
```

```
fit.int.totchol = lmer(totchol ~ cursmoke*c.age + sex + prevchd + (1|randid),
                       data = smoking2)
summary(fit.int.totchol)$coefficients
```

```
##                                   Estimate Std. Error    t value
## (Intercept)                  233.8303284 1.02373060 228.410021
## cursmokeCurrent smoker         2.5494978 0.91666199   2.781285
## c.age                          0.3679352 0.05513289   6.673606
## sexWomen                      12.1959286 1.19577122  10.199216
## prevchdPrevalent disease      -4.1422804 1.54541330  -2.680371
## cursmokeCurrent smoker:c.age   0.2752498 0.07912999   3.478451
```

```
# random intercept only
coefs.totchol = data.frame(coef(summary(fit.int.totchol)))
#p.values
2 * (1 - pnorm(abs(coefs.totchol$t.value)))
```

```
## [1] 0.000000e+00 5.414426e-03 2.495915e-11 0.000000e+00 7.354069e-03
## [6] 5.043201e-04
```

Thus, use a random intercept model.

Test for significance of random intercept

```
# model without the random effect
fit.lm = lm(sysbp ~ cursmoke*c.age + sex + bpmeds + diabetes + prevchd, data = smoking2)

# test for the significance of the random intercept
exactLRT(fit.int.sysbp, fit.lm)
# data:
# LRT = 2864.2, p-value < 2.2e-16

exactLRT(fit.int.diabp, fit.lm)
# data:
# LRT = 14863, p-value < 2.2e-16

# model without the random effect
fit.lm.totchol = lm(totchol ~ cursmoke*c.age + sex + prevchd, data = smoking2)
exactLRT(fit.int.totchol, fit.lm.totchol)
# data:
# LRT = 4106.5, p-value < 2.2e-16
```