LDA final project

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

OBJECTIVE: The goal of the analysis is to describe the smoking habits of the participants in the Framingham Heart study as they age. In particular, we are interested in describing the relationship between smoking status and related covariates.

STUDY DESIGN: The Framingham Heart Study is a long-term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. It was conducted to identify risk factors and their joint effect on smoking. Data collected includes laboratory, clinic, questionnaire, and adjudicated event data. Each participant has 1 to 3 observation periods, approximately 6 years apart. There are 11,627 observations on the 4,434 participants.

METHODS: We are interested in the relationship between age and smoking status, age and number of cigarettes smoked per day, and the relationship between smoking status and health outcomes including systolic blood pressure, diastolic blood pressure, and serum total cholesterol. We are also interested to see if the relationships differ by sex, and wish to account for confounders.
Exploratory analysis included using graphs to investigate the relationship of covariates of interest over time, looking at correlation between variables in the data set, and conducting literature search to identify potential confounders[1,2,3]. In addition, some variables in the dataset measured similar outcomes – for example prevalent coronary heart disease included angina, myocardial infarction, coronary insufficiency, so there was no need to include individual measures of angina, or myocardial infarction - or were highly correlated, suggesting possible multicollinearity.
The variables selected for consideration in model building were outcomes of interest: smoking status, cigarettes per day, systolic blood pressure, diastolic blood pressure, and serum total cholesterol; age and sex, due to interactions seen in exploratory analysis; and possible confounders BMI, diabetes, heart rate, prevalent coronary heart disease, prevalent stroke, death.
In formal analysis we used longitudinal generalized estimating equations (GEEs) models. Although we dealt with incomplete and unequally spaced data, we decided in favor of marginal modeling approach after experiencing multiple convergence issues during model building exercise under MCAR assumption. However, we realize and discussed within the group that missing data could be a result of MAR or NMAR mechanisms and then conditional likelihood-based model or marginal model after imputation could be more appropriate. To determine which covariates we should adjust for in our full models, we took into account overall significance of confounders for the effect of smoking status and our three variables of interest. We used a cut-off of p = 0.10 for inclusion into final model. We also used QIC to compare models with unstructured and exchangeable correlation structures. The QIC values were very similar for both structures so we use exchangeable for interpretability and parsimony.

RESULTS: Among subjects at first observation, age ranged from 32 to 70 years, with a mean of 49.93 years. Of the 4,434 participants, 1,944 were male and 2,490 were female. 2,181 were current smokers and 2,253 were not. Total cholesterol for all subjects ranged from 107 to 696, with a mean of 237. Diastolic blood pressure ranged from 48 to 142.5, with a mean of 83.08. Systolic blood pressure ranged from 83.5 to 295, with a mean of 132.9. Cigarettes smoked per day ranged from 0 to 70, with a mean of 8.97 for all subjects. For smokers, cigarettes smoked per day ranged from 1 to 70, with a mean of 18.37.

(1) model for current smoking status:

cursmoke = 5.947 - 0.064 age -1.827 I(sex = female) + 0.020 age x I(sex = female) + 0.063 I(educ = 2) -0.183 I(educ = 3) - 0.229 I(educ = 4) - 0.087 bmi + 0.009 heartrte - 0.265 prevchd

(Note: education 1 = 0-11 years; 2 = High School Diploma; GED 3=Some College, Vocational School; 4=College (BS, BA) degree or more)

We find that both age and sex are significant under the significance level of 0.05, with both p-values less than 0.001. The p value for interaction term (age:as.factor) is also significant. Thus, sex could effect this relationship, after adjusting for BMI, heart rate, and prevalence of congenial heart disease. The coefficient for age is –0.064, for sex is –1.827, for interaction is 0.020. The odds of currently smoking is 6.2% lower than the odds of currently not smoking with 1 year increase in age for men. Meanwhile for women, the odds ratio of smoking against non-smoking is 18.4% with 1 year increase in age, adjusted for other covariates.

(2) model for number of cigarettes per day

cigpday = 4.751 - 0.027 age - 1.520 I(sex = female) + 0.014 age x I(sex = female) + 0.087 I(educ = 2) - 0.074 I(educ = 3) - 0.069 I(educ = 4) - 0.041 bmi - 0.125 diabetes + 0.007 heartrte - 0.255 prevchd

We find that age and sex, and the interaction are significant under the significance level of 0.05, with both p-values less than 0.001. Thus, sex could differ cigarettes per day with age, after adjusting for BMI, diabetes, heart rate, and prevalence of congenial heart disease, prevalence of stroke and time to death. The coefficient for age is –0.027, for sex is –1.520, for interaction is 0.014. From the model we can see that with age increases 1 year, the number of cigarettes per day will decrease 2.67% at a population level for men, adjusted for other covariates. For women, the log odds ratio of smoking against non-smoking is -1.533. Thus, the relative mean number of cigarettes per day is 21.6% with 1 year increase in age at a population level for women, adjusted for other covariates.

(3) systolic blood pressure

sysbp = 43.000 - 0.772 I(cursmoke = yes) + 0.741 age + 3.321 I(sex = female) + 1.370 bmi + 3.173 diabetes + 0.172 heartrte + 1.553 prevchd + 4.201 prevstrk + 7.794 death

The above model using exchangeable correlation structure had QIC of 68837.55. Current smoking status was not significant in the model at the 0.05 significance level (p = 0.074), indicating current smoking status is not significantly associated with systolic blood pressure after adjusting for covariates. Those who did not smoke have an average systolic blood pressure over time of 43, and current smokers have an average systolic blood pressure of 42.228.

(4) diastolic blood pressure

diabp = 48.526 - 0.525 cursmoke - 0.518 I(sex = female) + 0.857 I(educ = 2) + 0.776 I(educ = 3) + 0.531 I(educ = 4) + 0.991 bmi - 2.072 diabetes + 0.106 heartrte + 1.049 prevstrk + 3.317 death

Our model with exchangeable correlation structure using current smoking status and adjusting for covariates to predict diastolic blood pressure had QIC of 53508. Current smoking status was significant in the model (p = 0.03). Those who did not smoke have an average diastolic blood pressure over time of 48.526, and those who did smoke have an average diastolic blood pressure of 48.001.

(5) serum total cholesterol:

totchol = 157.548 + 3.824 I(cursmoke = yes) + 0.402 age + 13.291 I(sex = female) + 1.656 bmi - 7.648 + 0.118 heartrte + 2.423 prevhyp

Our model using current smoking status and adjusting covariates with exchangeable correlation structure to predict total cholesterol had QIC of 84588.27. Current smoking status was highly significant in the model (p = 0). Those who did not smoke have an average total cholesterol over time of 157.548, and current smokers have an average cholesterol of 161.372.

CONCLUSION: Age and sex are both significantly associated with number of cigarettes per day and current smoking status. Overall, it appears that number of cigarettes per day decreases with age, and the odds of being a current smoker decreases with age. Sex does differ the relationship between age and cigarettes per day as well as age and current smoking status. Women had lower odds of being a current smoker than men, and women had on average fewer cigarettes per day. In using smoking status to predict systolic blood pressure after accounting for confounders, we found that smoking status was significant, and current smokers had on average lower systolic blood pressure. After building a model to predict diastolic blood pressure from smoking status, we found that smoking status was not significantly associated with diastolic blood pressure. In our model to predict serum total cholesterol, we found that smoking status was significant and current smokers had higher cholesterol on average.

REFERENCES:

[1] Smoking and diabetes, CDC, (https://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/pdfs/fs_smoking_diabetes_508.pdf)

[2] Education: The Effect of Education on Smoking Behavior: New Evidence from Smoking Durations of a. Sample of Twins. IZA DP No. 4796. March 2010. Pierre Koning (http://ftp.iza.org/dp4796.pdf)

[3] Smoking and cigarettes per day with BMI: Sneve M, Jorde R. Cross-sectional study on the relationship between body mass index and smoking, and longitudinal changes in body mass index in relation to change in smoking status: the Tromso Study. Scand J Public Health. 2008 Jun;36(4):397-407. doi: 10.1177/1403494807088453.

APPENDIX:

Figures

Figure 1: Smoking status vs covariates

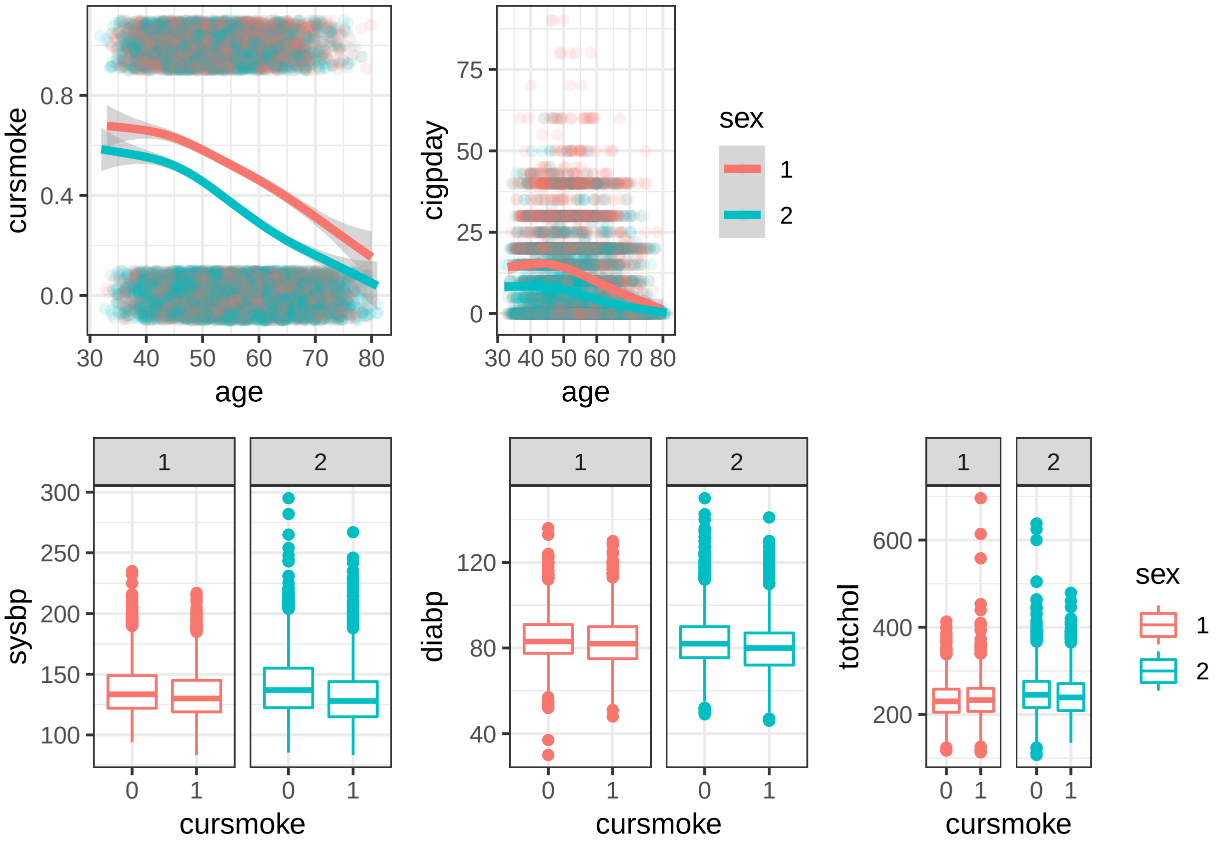
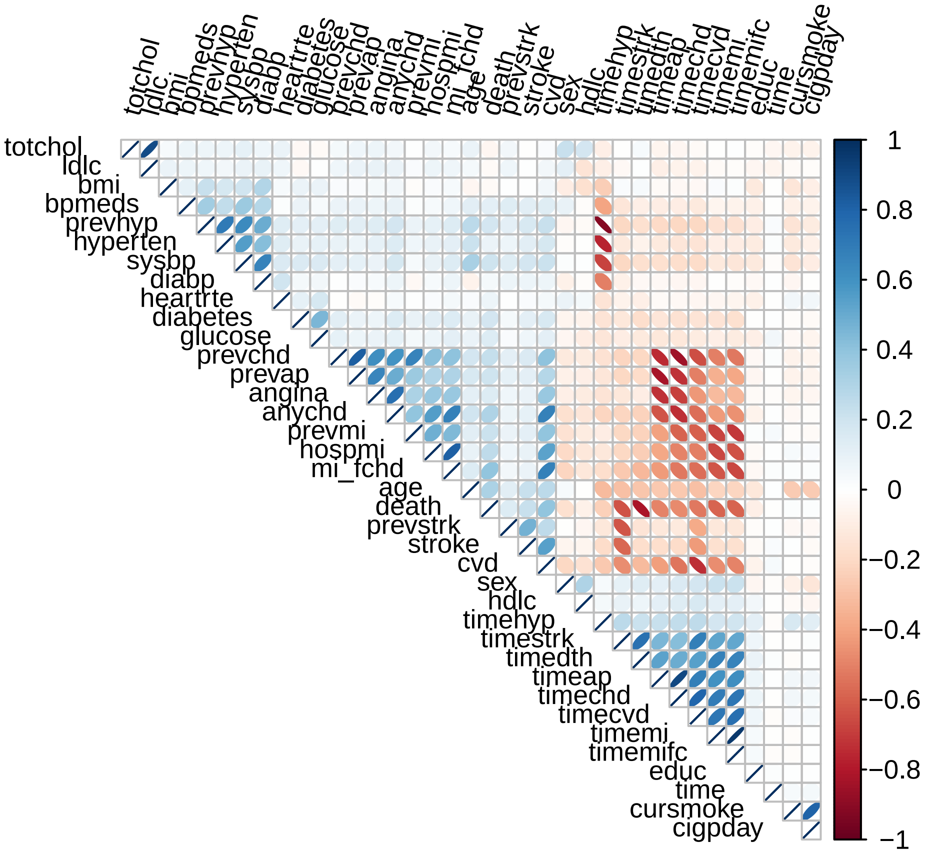


Figure 2: Correlation plot



Code

```
# Exploratory

library(psych)
knitr::kable(describe(data)[,c(2,3,4,5,8,9,10,13)], digits = 3)

## 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.

smoke_v_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.

smoke_age_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

n_c_age_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()

#### Smokers only
n_c_age_smoke = 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
n_c_age_s_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()


#### Smokers only

n_c_age_s_smoke = 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
```

```r
#blood pressure.
### smoking ~ sysbp
#ANS: Proportion of smokers decreases with increase of systolic blood
#presure

smoke_sbp = 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

smoke_sysbp_status = 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).

smoke_sysbp_sex = 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

smoke_sysbp_sex_status = 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)
smoke_dbp = 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()

smoke_dbp_box = 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).

smoke_dbp_s = 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()

smoke_dbp_s_bp = 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

smoke_tc = 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()

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

### smoking ~ totchol, sex [!!!]
# ANS: Proportion of smokers has non linear relationship with total
#cholesterol for women; proprtions increases with increase in total
```

```r
#cholesterol for men (sex by totchol interaction effect).

smoke_tc_sex = 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()

smoke_tc_sex_bp = 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()

library(ggplot2)

ggplot(data= data, aes(age, cigpday)) +
  geom_point(alpha = .6) +
  geom_smooth(method = 'lm', col = 'red') +
  geom_smooth(method = 'loess') +
  ggtitle("cigpday vs Age Scatterplot")

ggplot(data=data, aes(age, cigpday, group = randid)) +
  geom_path(alpha = .6) +
  geom_smooth(aes(group = NULL), method = 'lm', col = 'red') +
  geom_smooth(aes(group = NULL), method = 'loess') +
  ggtitle("cigpday vs Age Spaghetti Plot")

###

# data.residual <- data %>% na.omit()  %>%
# group_by(age) %>%
# mutate(mean.cigpday = mean(cigpday)) %>%
# ungroup() %>%
# mutate(residuals = cigpday - mean.cigpday) %>%
# group_by(randid) %>%
# mutate(median.residual = median(residuals)) %>%
# ungroup()
#
# data.stats <- c(min(data.residual$median.residual),
# quantile(data.residual$median.residual,
# c(.25, .5, .75)), max(data.residual$median.residual))
#
# data.id.select <- data.residual %>%
# filter(median.residual %in% data.stats)
#
# data.residual.plot <- ggplot() +
# geom_line(data = data.id.select,
# aes(x = age, y = cigpday, group = randid)) +
# ggtitle("cigpday by time, Selected from Residuals") +
# geom_smooth(data = data,
# aes(x = age, y = cigpday))

###

ggplot(data, aes(y = cursmoke, x = age)) + geom_jitter(height = 0.1) +
stat_summary(fun.y = 'mean', geom="line", col = 'red')

select(data, -c(randid, time,timeap:timehyp)) %>% GGally::ggcorr(.)


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")

corr2 <- data[,c(2,3,4,5,6,7,8,10,11,12,13,14)] %>% cor(., use = "complete.obs")
corrplot(corr2, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)


ggplot(data = data,
aes(x = as.factor(cursmoke), y = age)) +
geom_violin() +
  facet_wrap(~sex)

ggplot(data = data,
aes(x = cigpday, y = age)) +
geom_point() +
  facet_wrap(~sex) +
  geom_smooth()

## Cor plot

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")

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

```r
table(data$period)

# Summary

table(data$cursmoke, data$period)
data1 <- filter(data,period == "1")
summary(data1$age)
table(data1$sex)
data2 <- filter(data1,cursmoke == "yes")
summary(data1$cigpday)
summary(data2$cigpday)
summary(data1$totchol)
summary(data1$diabp)
summary(data1$sysbp)


# Models

### glmer ###

library(lme4)

smoke_stat_age <- glmer(cursmoke~age+educ + (1|randid), family=binomial, na.action = "na.omit")

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

smoke_stat_sex <- glmer(cursmoke~age+educ + as.factor(sex) + (1|randid), family=binomial, na.action = "

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



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

cigpday_age <- glmer(cigpday~age+educ +(1|randid), family=poisson, na.action = "na.omit")

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

ncig_gee <- glmer(cigpday~age+as.factor(sex) + educ+ (1|randid), family=poisson, na.action = "na.omit")

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


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

smoke_sys<-glmer(cursmoke~sysbp + (1|randid), family=binomial, na.action = "na.omit")

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

smoke_sys<-glmer(cursmoke~sysbp + as.factor(sex) + (1|randid), family=binomial, na.action = "na.omit")

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

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

smoke_dias<-glmer(cursmoke~diabp + (1|randid), family=binomial,  na.action = "na.omit")

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

smoke_dias<-glmer(cursmoke~diabp + as.factor(sex) + (1|randid), family=binomial,  na.action = "na.omit"

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

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

smoke_chol<-glmer(cursmoke~totchol + (1|randid), family=binomial, na.action = "na.omit")

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

smoke_chol<-glmer(cursmoke~totchol + as.factor(sex) + (1|randid), family=binomial,  na.action = "na.omi

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

###

my.data.complete <- data %>%
  dplyr::select(-c(hdlc,ldlc)) %>%
  na.omit()

model.saturated <- geepack::geeglm(formula = cursmoke ~ as.factor(sex) +
    age + age * as.factor(sex) + sysbp + diabp + sysbp * diabp +
    bpmeds + as.factor(educ) + totchol + bmi + glucose + diabetes +
    heartrte + prevap + prevchd + prevmi + prevstrk + prevhyp,
    family = binomial, data = my.data.complete, id = randid,
    corstr = ("unstructured"))

model <- aov(model.saturated)
car::vif(model)

### NEW MODELS WITH GLMER / LMER ###


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

model.q1 <- glmer(CURSMOKE ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                  + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                  + PREVHYP + TIMEDTH + (1|RANDID),
                  data = my.data,
                  family=binomial,
                  corstr = "exchangeable", nAGQ = 0)
summary(model.q1)
exp(coef(summary(model.q1)))
```

```
# diabetes, prevstrk, prevhyp not significant at a =.1

model.q1 <- glmer(CURSMOKE ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + HEARTRTE + PREVCHD + TIMEDTH + (1|RANDID),
                 data = my.data,
                 family="binomial",
                 corstr = "exchangeable", nAGQ = 0)
summary(model.q1)
exp(coef(summary(model.q1)))

# We can see from the model that with age increases 1, the log odds ratio of
#smoking against non-smoking is -0.051.
#
# Which means that with age incraeses 1, the odds ratio of smoking against
#non-smoking is 1.0523. The odds of currently smoking is 5% lower than the odds
#of currently not smoking with the age incraeses 1 unit.
#
# We can see that both age and sex have a p-value less than 0.05. Thus, sex can
# differ the relationship between age and smoking status. For women, the log
#odds # ratio of smoking against non-smoking is -0.76 as the age increases 1.
#Thus, for # women, the odds of smoking against non-smoking is 53.2% lower with the age
# increases 1 unit.

# (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)

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

model.q2_1 <- glmer(CIGPDAY ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ) + BMI + DIABET
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  family="poisson",
                  corstr = "exchangeable", nAGQ = 0)
summary(model.q2_1)
exp(coef(summary(model.q2_1)))

# prevhyp not sig

model.q2_1 <- glmer(CIGPDAY ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) +
                      as.factor(educ) + BMI + DIABETES + HEARTRTE + PREVCHD +
                      PREVSTRK + (1|RANDID)
                  + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  family="poisson",
                  corstr = "exchangeable", nAGQ = 0)
```

```
summary(model.q2_1)
exp(coef(summary(model.q2_1)))

# From the result we can see that the p-value of age and sex are both show
# significance. Thus, sex can differ the relationship between cigraettes per day.
#
# From the model that with age increases 1, the number of cigarettes per day
# will decrease 2.32% at a population level for men.
#
# For women, cigaretees per day will decrease 55.9% with one unit of age
# increases at a population level.

## (3) Totchol and cursmoke

totchol_fit <- glmer(TOTCHOL ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) +
                       as.factor(educ) + BMI + DIABETES + HEARTRTE + PREVCHD +
                       PREVSTRK + (1|RANDID)
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  family="poisson",
                  corstr = "exchangeable", nAGQ = 0)
summary(totchol_fit)

#prevstrk not sig

totchol_fit <- glmer(TOTCHOL ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) +
                       as.factor(educ) + BMI + DIABETES + HEARTRTE + PREVCHD +
                       (1|RANDID) + PREVHYP + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  family="poisson",
                  corstr = "exchangeable", nAGQ = 0)
summary(totchol_fit)


## (4) Sysbp and cursmoke

# treat as gaussian??

sysbp_fit <- lmer(SYSBP ~ AGE + as.factor(SEX) + as.factor(educ) + BMI +
                    DIABETES + HEARTRTE + PREVCHD + PREVSTRK + (1|RANDID)
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  corstr = "exchangeable", nAGQ = 0)

summary(sysbp_fit)


## (5) Diabp and cursmoke

# doesnt converge as GLMER
```

```
diabp_fit <- glmer(DIABP ~ AGE + as.factor(SEX) + as.factor(educ) + BMI +
                     DIABETES + HEARTRTE + PREVCHD + PREVSTRK + (1|RANDID)
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  na.action = "na.omit",
                  family=poisson,
                  corstr = "exchangeable", nAGQ = 0)

summary(diabp_fit)

#### Gee ####
## FINAL MODELS ##

## (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) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                 + PREVHYP + TIMEDTH,
                 id = RANDID,
                 data = my.data,
                 family=binomial,
                 corstr = "unstructured")
knitr::kable(summary(model.q1)$coefficients[,c(1,4,5)], digits = 3)
model.q1[["working.correlation"]]
QIC(model.q1)

model.q1 <- gee(CURSMOKE ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                 + PREVHYP + TIMEDTH,
                 id = RANDID,
                 data = my.data,
                 family=binomial,
                 corstr = "exchangeable")
knitr::kable(summary(model.q1)$coefficients[,c(1,4,5)], digits = 3)
round(2 * pnorm(abs(coef(summary(model.q1))[,5]), lower.tail = FALSE), 3)
model.q1[["working.correlation"]]
QIC(model.q1)

# remove nonsignificant
model.q1 <- gee(CURSMOKE ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + HEARTRTE + PREVCHD
                 + TIMEDTH,
                 id = RANDID,
                 data = my.data,
                 family=binomial,
                 corstr = "exchangeable")
knitr::kable(summary(model.q1)$coefficients[,c(1,4,5)], digits = 3)
round(2 * pnorm(abs(coef(summary(model.q1))[,5]), lower.tail = FALSE), 3)
model.q1[["working.correlation"]]
```

```r
QIC(model.q1)

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


# 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) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  id = RANDID,
                  family=poisson,
                  corstr = "unstructured")
knitr::kable(summary(model.q2_1)$coefficients[,c(1,4,5)], digits = 3)

model.q2_1[["working.correlation"]]
QIC(model.q2_1)


model.q2_1 <- gee(CIGPDAY ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + PREVSTRK
                 + PREVHYP + TIMEDTH,
                  data = my.data,
                  id = RANDID,
                  family=poisson,
                  corstr = "exchangeable")
knitr::kable(summary(model.q2_1)$coefficients[,c(1,4,5)], digits = 3)
round(2 * pnorm(abs(coef(summary(model.q2_1))[,5]), lower.tail = FALSE), 3)
model.q2_1[["working.correlation"]]
QIC(model.q2_1)

#remove nonsignificant at 0.1
model.q2_1 <- gee(CIGPDAY ~ AGE + as.factor(SEX) + AGE:as.factor(SEX) + as.factor(educ)
                 + BMI + DIABETES + HEARTRTE + PREVCHD + TIMEDTH,
                  data = my.data,
                  id = RANDID,
                  family=poisson,
                  corstr = "exchangeable")
knitr::kable(summary(model.q2_1)$coefficients[,c(1,4,5)], digits = 3)round(2 * pnorm(abs(coef(summary(mo
model.q2_1[["working.correlation"]]
QIC(model.q2_1)

##################
# 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 #
###############################################

## (3) Totchol and cursmoke

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

knitr::kable(summary(totchol_fit)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(totchol_fit))[,5]), lower.tail = FALSE), 3))
QIC(totchol_fit)

totchol_fit2 <- gee(totchol ~ cursmoke + age + factor(sex) + bmi +
                     diabetes + heartrte  + prevhyp ,
                   id = randid,
                   family = "gaussian",
                   corstr = "exchangeable",
                   na.action = "na.omit")

knitr::kable(summary(totchol_fit2)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(totchol_fit2))[,5]), lower.tail = FALSE), 3))
QIC(totchol_fit2)

## (4) Sysbp and cursmoke

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

knitr::kable(summary(sysbp_fit)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(sysbp_fit))[,5]), lower.tail = FALSE), 3))
QIC(sysbp_fit)

sysbp_fit2 <- gee(sysbp ~ cursmoke + age + factor(sex)  + bmi +
                   diabetes + heartrte + prevchd + prevstrk + death,
                 id = randid,
                 family = "gaussian",
                 corstr = "exchangeable",
                 na.action = "na.omit")

knitr::kable(summary(sysbp_fit2)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(sysbp_fit2))[,5]), lower.tail = FALSE), 3))
QIC(sysbp_fit2)

## (5) Diabp and cursmoke

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

knitr::kable(summary(diabp_fit)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(diabp_fit))[,5]), lower.tail = FALSE), 3))
QIC(diabp_fit)

diabp_fit2 <- gee(diabp ~ cursmoke + factor(sex) + factor(educ) + bmi +
                   diabetes + heartrte  + prevstrk +death,
                 id = randid,
                 family = "gaussian",
                 corstr = "exchangeable",
                 na.action = "na.omit")

knitr::kable(summary(diabp_fit2)$coefficients[,c(1,4,5)], digits = 3)
knitr::kable(round(2 * pnorm(abs(coef(summary(diabp_fit2))[,5]), lower.tail = FALSE), 3))
QIC(diabp_fit2)

###
```