LDA final project

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

Contents

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. 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 also used QIC to compare models with unstructured and exchangeable correlation structures.

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: Model: cursmoke ~ age + as.factor(sex) + age:as.factor(sex) + as.factor(educ) + bmi + diabetes + heartrte+ prevchd + prevstrk + prevhyp + timedth, id = randid, corstr = 'exchangeable' (show beta's?)

From the significance of age and sex, we can find that both age and sex are significant under the significance level of 0.05, with both p-values less than 0.001. Also, for the interaction term (age:as.factor), the p-value is also significant. Thus, sex could differ this relationship, 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 –0.0594, for sex is –1.57, for interaction is 0.0156. The odds of currently smoking status is 5.8% lower than the odds of currently not smoking with the age increases 1 unit in men. Meanwhile for women, the log odds ratio of smoking against non-smoking is –1.6138, which means that the odds ratio of smoking against non-smoking is 19.9% with the age increases 1 unit.

(2) model for number of cigarettes per day

Model: cigpday ~ age + as.factor(sex) + age:as.factor(sex) + as.factor(educ) + diabetes + heartrte + prevchd + prevstrk + preyp + timedth, id = randid, family=poisson, corstr = 'exchangeable'

From the significance of age and sex, we can find that both age and sex are significant under the significance level of 0.05, with both p-values less than 0.001. Also, for the interaction term (age:as.factor), the p-value is 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 –0.0266, for sex is –1.21, for interaction is 0.0082. From the model we can see that with age increases 1 unit, the number of cigarettes per day will decrease 2.62% at a population level for men. For women, the log odds ratio of smoking against non-smoking is –1.228. Thus, cigarettes per day will decrease 70.7% with one unit of age increases at a population level for women.

(3) systolic blood pressure

Model: sysbp ~ cursmoke + age + factor(sex) + bmi + diabetes + heartrte + prevchd + prevstrk + death (show beta's?)

Our model using current smoking status after adjusting for age, sex, BMI, diabetes, heart rate, and prevalence of congenial heart disease, prevalence of stroke and death to predict systolic blood pressure had QIC of 68843.21. Current smoking status was highly significant in the model (p = 0.029). Those who did not smoke have an average systolic blood pressure over time of 40.12, and those who did smoke have an average systolic blood pressure of 3.

(4) diastolic blood pressure

Model: diabp ~ cursmoke + factor(sex) + factor(educ) + bmi + diabetes + heartrte + prevstrk +death

Our model using current smoking status and adjusting for education, BMI, diabetes, heart rate, and prevalence of stroke and death to predict diastolic blood pressure had QIC of 53509.72. Current smoking status was highly significant in the model (p = 0.000). Those who did not smoke have an average diastolic blood pressure over time of 47.35, and those who did smoke have an average diastolic blood pressure of 46.25.

(5) serum total cholesterol: Model: totchol ~ cursmoke + age + factor(sex) + bmi + diabetes + heartrte + prevhyp

Our model using current smoking status and adjusting for age, sex, BMI, diabetes , heart rate, and prevalence of hypertension to predict total cholesterol had QIC of 84596.03. Current smoking status was significant in the model (p = 0.04). Those who did not smoke have an average total cholesterol over time of 162.1, and those who did smoke have an average cholesterol of 164.47.

CONCLUSION: A conclusion specifically answering the objective of the analysis.

REFERENCES:

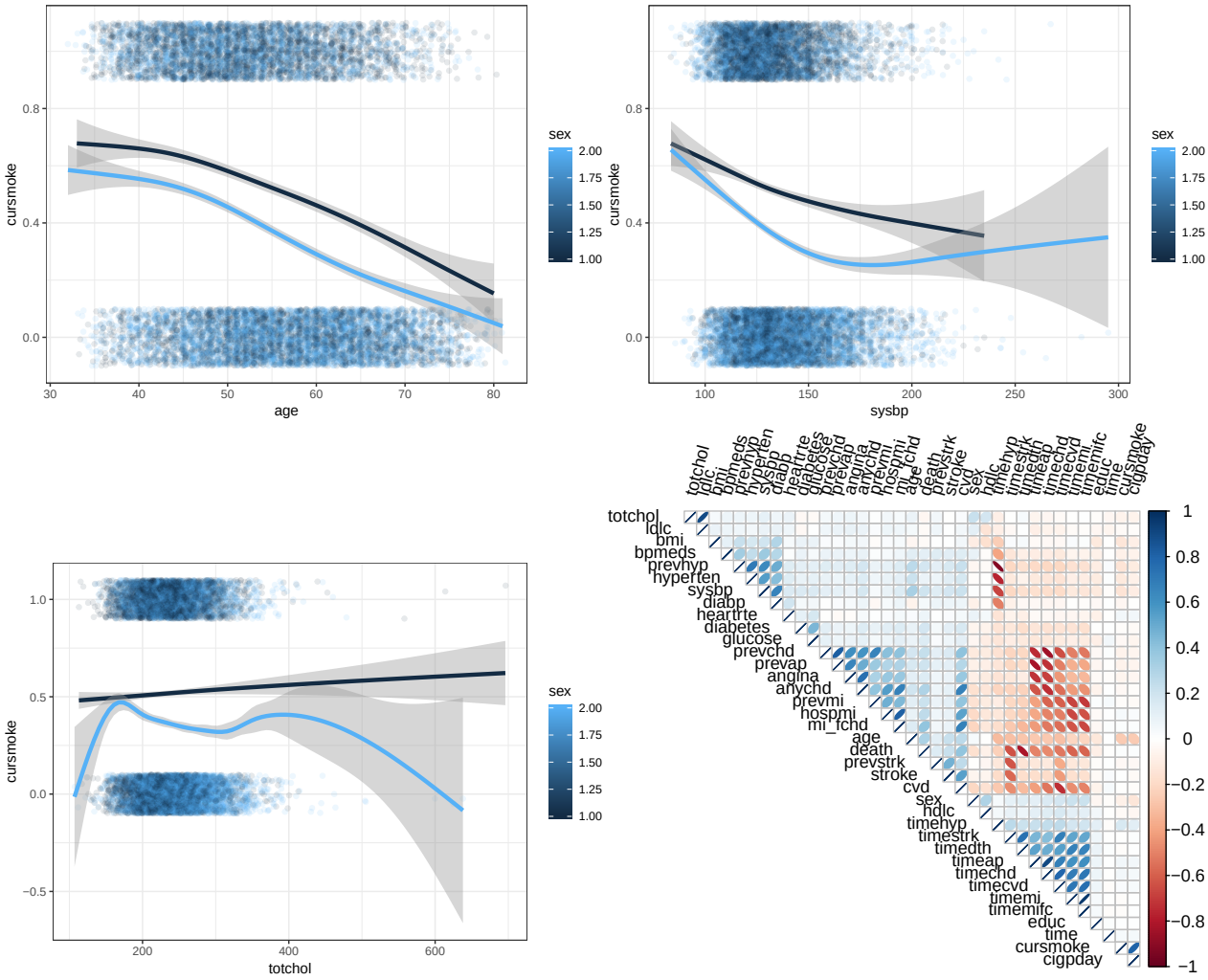
[1] diabetes (https://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/pdfs/fs_smoking_diabetes_508.pdf)

[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



Code

```
# Exploratory

## Smoking vs. Age, Sex
## (1): Smoking ~ age, sex
### Is there a relationship between age and smoking status?
#ANS: Yes, the proportion of smokers decreases with the age.

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
```

```r
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

#### 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 ther

#### 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
```

```r
## (1) The relationship between current smoking status and systolic 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 highe

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 prop
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) %>%
```

```r
  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 hig

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 hasnonlinera relationship with total cholesterol for women; proprtions inc

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()

smoke_age_sex
smoke_sysbp_sex
smoke_tc_sex

## 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

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

## (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)
```

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

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


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

model.q2_1 <- gee(CIGPDAY ~ AGE + as.factor(SEX) + as.factor(educ)
                + 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) + 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)

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)
```