**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, and looking at correlation between variables in the data set. We found that the proportion of smokers decreases with age, and there is a higher proportion of smoker among men compared to women as both age, but there is no interaction between age and sex. The number of cigarettes smoked per day stays constant for subjects 30-50 years old and decreases with age after 50 years old. There is sex effect (men smoke higher number of cigarettes per day than women across age), but there is no sex and age interaction. The Proportion of smokers decreases with increase of systolic blood pressure, and there is a slightly higher systolic bp for non-smokers. Proportion of smokers decreases with increase of systolic blood pressure; the proportion is higher for men (sex effect). Proportion of smokers decreases with increase of diastolic blood pressure; the proportions are higher for men (sex effect). Proportion of smokers slightly decreases with increase of total cholesterol values. Proportion of smokers has linear relationship with total cholesterol for women; proportions increases with increase in total cholesterol for men (sex by totchol interaction effect).

We found potential confounders through a literature search [references needed], and we used both common sense and a correlation plot to reduce the covariates used in the models. Some measures such as [list] measured similar outcomes, or were highly correlated, so this was taken into consideration to avoid multicollinearity in our models.

 We generated models using longitudinal generalized estimating equations (GEEs). To determine which covariates we should adjust for in our full models, we took into account overall significance of each related possibly confounder for the effect of smoking status our three variables of interest. We also used QIC to compare models with different correlation structure.

totchol ~ cursmoke + age + factor(sex) + bmi + diabetes + heartrte  + prevhyp

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

sysbp ~ cursmoke + age + factor(sex)  + bmi +  diabetes + heartrte + prevchd + prevstrk + death

 (maybe move models to results??)

**#RESULTS:**

**a) descriptive statistics for the data**

**b) a summary of your key findings including supporting numerical summaries (i.e. confidence intervals, pvalues, etc.)**

From EDA was concluded that there was association [graphs, p-values - appendix] between response and:

- demographic variables (age, sex, education, BMI),

- heart-related variables (heart rate, prevstrk, total cholesterol [with possible interaction with age -> test in the model])

- blood pressure variables (systolic and diastolic blood pressure, blood pressure medications, hyperten, prevhyp),

- diabetes-related variables (diabetes, glucose),

- prevalence variables (prevchd, prevap, death, angina, hospmi, mi\_fchd, cvd),

- time to event variables (timeap, timecvd, timehyp).

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. 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. 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.

**c) interpretations of your key findings (i.e. interpretations of coefficients).**

**#CONCLUSION:**

A conclusion specifically answering the objective of the analysis.

**#APPENDIX:**

(will add R code and knit later)

**#References:**

diabetes (https://www.cdc.gov/tobacco/data\_statistics/sgr/50th-anniversary/pdfs/fs\_smoking\_diabetes\_508.pdf)

VARIABLES/Confounders TO INCLUDE:

age - we want to see how smoking status changes as we get older

sex - exploratory data analysis showed there were differences

education - found in the literature

total cholesterol

BMI - confounder based on literature

diabetes (https://www.cdc.gov/tobacco/data\_statistics/sgr/50th-anniversary/pdfs/fs\_smoking\_diabetes\_508.pdf)

heart rate -  smoking increases heart rate (in literature)

prevchd since prevmi and prevap are correlaed

prevstroke

prevhyp

timedth

DONT INCLUDE:

systolic, diastolic bc we have hypertension

BPmeds  - not significant

LDL and HDL have too many NAs

glucose - correlated with diabetes