Stat 403/650/890 Analysis Assignment 2:

Given the dataset derived from the 2019 SSC case study here: <https://ssc.ca/en/case-study/case-study-2-risk-cardiovascular-disease-among-osteoarthritis-patients>

... predict cardiovascular disease (CCCA\_121).

The data set you have is a random selection of 10,000 patients from the approximately 130,000 patients that were followed for 5 years. We are only looking at the first of three cycles.

**1) Describe the response variable.**

**Is it continuous / categorical / binary?**

**What does its data type imply about the modeling method that is likely to work best?**

2**) Describe the other variables.**

**Which are continuous? Fixed effects? Random effects?**

**Are there any variables that are NOT explanatory or a response? (e.g. IDs, weights)**

**3) Describe the missingness pattern.**

**Are there any missing values? Which variables?**

**Are the missing values randomly assigned, or are is there some observable pattern?**

**What are you going to do about the missing data?**

**4) How are you going to decide which variables to use. Note that only the prediction accuracy matters (lowest deviance/AIC/BIC)?**

**5) Describe your final model.**

**How well does it fit the response?**

**Are there any possible colliinearity / VIF issues?**

**How much better/worse is the model when the arthritis variable CCCA\_051 is in it?**

Not every one of these questions needs to be answered distinctly. If you do something that effectively answers more than one question, explain that.

Target word range 800-1100

Date Source:

Canadian Community Health Survey (CCHS) cycles

Background:

The Canadian Community Health Survey (CCHS) is a nationwide cross-sectional survey. This survey gathers health-related data for the Canadian population 12 years of age and over living in the 10 provinces and 3 territories, covering about 97% of the target population.

Variables:

It is often a good idea to cross-tabulate variables with the ‘Age’ variable (from the same cycle) to double check if the question was restricted to particular age groups. Similarly, cross-tabulating with the ‘Province’ variable often helps identify variables that were created from an ‘optional CCHS component.’

Main question:

1. Within Canadian adults (20-64 years of age), is having osteoarthritis associated with the developing heart disease?

For the purpose of this case study, assume that, from the literature, we know that the following variables are risk factors for the outcome and confounders in the above relationship: age, sex, ethnicity, education, household income, body mass index (BMI), access to a regular medical doctor, smoking habit, alcohol drinking habit, high-blood pressure, and diabetes. Also, assume that physical activity is suspected to be an intermediate factor between osteoarthritis and heart disease.

Variables:

Has heart disease CCCA\_121

Has arthritis or rheumatism CCCA\_051

Age DHHAGAGE

Sex DHHA\_SEX

Marital Status DHHAGMS

Cultural / racial origin SDCAGRAC

Immigrant status SDCAFIMM

Highest level of education EDUADR04

Total household income INCAGHH

Body mass index HWTAGBMI

BMI Note: underweight (<18.5), healthy weight (between 18.5 and 25), overweight (>25). (1)

Physical activity index PACADPAI

Has a regular medical doctor TWDA\_5

Type of smoker SMKADSTY

Type of drinker ALCADTYP

Has high blood pressure CCCA\_071

Has diabetes CCCA\_101

Has emphysema or chronic obstructive pulmonary disease (COPD) CCCA\_91B

Daily consumption - total fruits and vegetables FVCADTOT

Self-perceived stress GENA\_07

Province GEOAGPRV

Sampling weight - master weight WTSAM

Useful sample code:

## Load the .csv file into R. Store it as 'dat'

data = read.csv("cchs.csv")

### See if any given case has a missing value

### Keep only the complete cases

hasNAs = rep(NA,nrow(data))

for(k in 1:nrow(data))

{

hasNAs[k] = any(is.na(data[k,]))

}

data2 = data[,!hasNAs]

#### Binary Model with random effects

model = glmer(CCCA\_121 ~ INCAGHH + GEOAGPRV + GEOARPRV, data = data, family = binomial, weights = WTSAM)

### Binary model without

model = glm(CCCA\_121 ~ INCAGHH + GEOAGPRV + GEOARPRV, data = data, family = binomial, weights = WTSAM)

### Stepwise model selection

full\_model = glm(CCCA\_121 ~ "CCCA\_121 + CCCA\_051 + DHHAGAGE + DHHA\_SEX + DHHAGMS + SDCAGRAC + SDCAFIMM + EDUADR04 + INCAGHH + HWTAGBMI + PACADPAI + TWDA\_5 + SMKADSTY + ALCADTYP + CCCA\_071 + CCCA\_101 + CCCA\_91B + FVCADTOT + GENA\_07 + GEOAGPRV, weights=WTSAM, data=data)

### AIC

step\_model = stepAIC(full\_model, trace = FALSE, k=2)

### BIC

step\_model = stepAIC(full\_model, trace = FALSE, k=log(nrow(data)))

### Get the variance inflation factor of a model

vif(model)