A logistic regression model predicting mortality during ICU stay

Aim

To build a logistic regression model that predicts mortality (outcome) based on 4 predictors:

Gender

Age at admission

A diagnosis of ‘Coronary atherosclerosis of native coronary artery’ indicated by ICD9 code 41401

Maximum value of Troponin I during the first 24 hours in the ICU

Defining the population

In the MIMIC III dataset, there are patients who have multiple hospitalizations and multiple ICU stays per hospitalization. Since I would like to represent each patient with only one set of data points, I will define my population as the first ICU stay for each patient during their first hospitalization.

Another reason why I am picking this subset of data is, I want to build a model that predicts mortality in patients who are atherosclerotic but haven't been diagnosed as such when entering the hospital. Since an exploration of the 'DIAGNOSIS' column in 'ADMISSIONS' table did not yield any patients with atherosclerosis and since ICD9 codes are assigned at the time of discharge, it is reasonable to assume that these patients had not been diagnosed with atherosclerosis at the time of admission.

Within this subset of patients, I would like to determine the contribution of age, gender and a spike in Troponin I levels (which happens most commonly after a myocardial infarction), during their first ICU stay to mortality (the diagnosis of atherosclerosis might happen before subsequent ICU stays).

Validity of the model

This model is valid for patients who are atherosclerotic but haven’t been diagnosed as such at admission and have a heart attack during/ prior to their first ICU stay during their first hospitalization at this hospital.

The model is not valid for patients who

- have a prior history of atherosclerosis
- are diagnosed as atherosclerotic before subsequent ICU stays or hospitalizations
- are never diagnosed as atherosclerotic
- do not have a heart attack during or prior to their first ICU stay of their first hospitalization
- have a heart attack while known to be atherosclerotic/ during subsequent ICU stays or hospitalizations
- never have a heart attack
- have a history of one or more ICU stays or hospitalizations
- have two or more ICU stays within the same hospitalization

Defining the population (code)

```
#setting up the environment

library(tidyverse)

library(bigrquery)

library(rsample)

library(plotROC)

con <- DBI::dbConnect(drv = bigquery(), project = "learnclinicaldatascience")

admissions <- tbl(con, 'mimic3_demo.ADMISSIONS') %>% collect()

icustays <- tbl(con, "mimic3_demo.ICUSTAYS") %>% collect()

patients <- tbl(con, "mimic3_demo.PATIENTS") %>% collect()

diagnoses_icd <- tbl(con, "mimic3_demo.DIAGNOSES_ICD") %>% collect()

d_labitems <- tbl(con, "mimic3_demo.D_LABITEMS") %>% collect()

labevents <- tbl(con, "mimic3_demo.LABEVENTS") %>% collect() #defining the population

first_hospitalization <- admissions %>% group_by(SUBJECT_ID) %>% filter(ADMITTIME == min(ADMITTIME))
%>% ungroup() %>% select(SUBJECT_ID, HADM_ID, DIAGNOSIS, ADMITTIME, HOSPITAL_EXPIRE_FLAG)

analytic_dataset <- icustays %>% inner_join(first_hospitalization, by = c("SUBJECT_ID" = "SUBJECT_ID",
"HADM_ID" = "HADM_ID")) %>% group_by(SUBJECT_ID, HADM_ID) %>% filter(INTIME == min(INTIME)) %>%
ungroup()
```

Defining the outcome

The outcome is death during the first ICU stay of the first hospitalization for each patient. The ‘HOSPITAL_EXPIRE_FLAG’ in the ‘ADMISSIONS’ table will be used to define the outcome.

Defining the outcome (code)

```
analytic_dataset %<>% select(SUBJECT_ID, HADM_ID, ICUSTAY_ID,
death_outcome= HOSPITAL_EXPIRE_FLAG)
```

Defining the predictors

Gender (code)

```
gender <- patients %>% select(SUBJECT_ID, GENDER) %>%
mutate(male = case_when(GENDER == "M" ~ 1, TRUE ~ 0)) %>%
select(SUBJECT_ID, male)
```

Age at admission (code)

```
date_of_birth <- patients %>% select(SUBJECT_ID, DOB)
```

```
age_at_admission <- first_hospitalization %>% left_join(date_of_birth)
%>% mutate(age_at_admission = round(as.numeric((ADMITTIME -
DOB)/365.25))) %>% select(SUBJECT_ID, HADM_ID, age_at_admission)
```

Defining the predictors

One of the clinical predictors I will be using is a discharge diagnosis of ‘Coronary atherosclerosis of native coronary artery’ indicated by ICD9 code 41401. I would like to determine the role having atherosclerosis (a disease of the arteries characterized by the deposition of plaques of fatty material on their inner walls) plays in mortality during the first ICU stay of a first hospitalization.

ICD9 code 41401 (code)

```
atherosclerosis <- diagnoses_icd %>% filter(ICD9_CODE == "41401") %>%
distinct(SUBJECT_ID) %>% mutate(atherosclerosis = 1)
```

Missingness

Patients with no discharge diagnosis of atherosclerosis will be assigned a value of zero (0) for the ‘atherosclerosis’ column after joining the ‘atherosclerosis’ dataset with the analytic dataset.

Defining the predictors

The second clinical predictor I will be using is the maximum level of Troponin I during the first 24 hours in the ICU. Troponin I spikes during the first 12 hours following a myocardial infarction and is thus, a biomarker. I would like to determine the role having a heart attack plays in the mortality of a patient with undiagnosed atherosclerosis during the first ICU stay of a first hospitalization

Troponin I (code)

```
d_labitems %>% filter(str_detect(LABEL, pattern = regex("Troponin I", ignore_case =
TRUE)))
```

```
icu_admission_time <- icustays %>% select(SUBJECT_ID, HADM_ID, ICUSTAY_ID, INTIME)
```

```
tropI <- labevents %>% filter(ITEMID == 51002) %>% select(SUBJECT_ID, HADM_ID,
CHARTTIME, VALUENUM) %>% inner_join(icu_admission_time) %>% mutate(end_time =
INTIME + 24*3600) %>% filter(CHARTTIME >= INTIME & CHARTTIME <= end_time) %>%
group_by(SUBJECT_ID, HADM_ID, ICUSTAY_ID) %>% summarise(max_first24hr_tropI =
max(VALUENUM))
```

Analytic methodology

- Split the analytic dataset into training (70% of analytic dataset) and testing (30% of analytic dataset) datasets to avoid overfitting.

- Build a logistic regression model using the glm() function and the training dataset with death as the outcome and gender, age at admission, a discharge diagnosis of atherosclerosis and the maximum level of troponin I during the first 24 hours of first ICU stay of first hospitalization as predictors.

- Evaluate the model using the testing dataset.

```
#building the analytic dataset

analytic_dataset <- analytic_dataset %>% left_join(gender) %>% left_join(age_at_admission) %>% left_join(tropI) %>%
left_join(atherosclerosis) %>% mutate(atherosclerosis = case_when(is.na(atherosclerosis) ~ 0, TRUE ~ atherosclerosis))

#building training and testing datasets

set.seed(2020)

data_split <- initial_split(analytic_dataset, prop = 7/10)

training_data <- training(data_split)

testing_data <- testing(data_split)

#building the logistic regression model

model <- training_data %>% glm(formula = death_outcome ~ male + age_at_admission + max_first24hr_tropI + atherosclerosis,
family = "binomial")

summary(model)

#evaluating the model with training data

training_data$predicted_outcome <- predict(model, training_data, type = "response")

training_roc <- training_data %>% ggplot(aes(m = predicted_outcome, d = death_outcome)) + geom_roc(n.cuts = 10, labels=F,
labelround = 4) + style_roc(theme = theme_grey)

training_roc

calc_auc(training_roc)$AUC*100

#evaluating the model with testing data

testing_data$predicted_outcome <- predict(model, testing_data, type = "response")

testing_roc <- testing_data %>% ggplot(aes(m = predicted_outcome, d = death_outcome)) + geom_roc(n.cuts = 10, labels=F,
labelround = 4) + style_roc(theme = theme_grey)

testing_roc

calc_auc(testing_roc)$AUC*100
```

Type of prediction model

This prediction model is-

- Operational: Model can predict if an ICU bed will become available again (due to death) after a patient in the target population is admitted
- Research: Model can help study mortality in the target population and a better understanding of the causes of mortality might help prevent it
- Clinical: If there is a spike in troponin, I levels in 24 hours following ICU admission, the model can help determine the log odds of mortality when the patient is atherosclerotic versus when they are not.

This could lead to further investigation of the cause of troponin I spike (is it undiagnosed atherosclerosis or one of the other several causes of troponin I spike, like sepsis, kidney failure or chronic kidney disease, chemotherapy-related damage to the heart, pulmonary embolism, heart infection, myocarditis, heart damage from using recreational drugs or a traumatic injury to the heart?)

Implementation of the model

Users:

- Hospital administrators, to determine the availability of beds in the ICU.
- Research scientists, to study mortality
- Clinicians, to determine the role atherosclerosis may play in mortality and thus improve intervention strategies

The model is simple enough that it can be deployed using the Arden syntax natively in the EHR. Since the model uses a lab test as one of the predictors, data brittleness however, might be a concern for implementation.