STA-230 Final Problem

A researcher (you) is interested in investigating the effect of mental health diagnosis during pregnancy on the prescription of opioids post-partum. It is expected that opioid prescription post-partum will vary substantially by delivery type – vaginal birth versus a Cesarean section.

In order to investigate the issue, we need to analyze insurance claims data. To this end, we have been given access to two files: a membership file (`member.csv`) and a claims file (`claims.csv`). Each file contains information on a set of insured women across calendar years 2017-2018.

The membership file is in long format with five variables: member id, year (2017, 2018), month (1-12), date of birth, and insurance plan type (HMO vs PPO). Each woman will have one line for each month/year during which she was insured.

The claims file is also in long format with four variables: member id, date of claim service, a CPT service code and an ICD diagnosis code.

Instructions to prepare our dataset are as follows:

Case inclusion for analysis

Our study is restricted to only those women who were at least 18 years old and at most 40 at the time of delivery. Additionally, to be sure that we have adequate information regarding prior mental health and subsequent opioid prescription, we require that a woman must be insured in the month of her delivery, for 3 months following her delivery, and for 9 calendar months prior to her delivery.

Primary outcome: opioid prescription

The outcome of interest is the prescription of an opioid within 90 days post-partum. All entries in the claims data have been adjusted to only include post-partum prescriptions so we do not need to worry about dates. opioid prescriptions are identified with the ICD code `J0745`

Delivery Date and Type

Delivery dates are identified with the ICD codes `O80` and `O82` (letter "O") and are stored in the claims file. These entries serve two purposes: using the delivery date for determining the mother's age at the time of birth and for determining if the delivery was a vaginal birth or a Cesarean.

Mental Health Disorder

The definition used to determine a mental health disorder is given by the ICD codes `F41.8` and `F32.3`. The first of these is related to anxiety disorders, while the second is associated with depression.

Migraine headache

Patients with a history of migraines are identified with ICD code of `G43.4`. Creating an indicator for this variable (i.e., marking which patients had previously filed a claim related to migraines) will be used to control for opioids that were potentially dispensed for migraines rather than related to delivery

Exploratory Plots

Using the data that you have modified, you should create 2 or 3 plots with ggplot that investigate the relationships between our outcome and our newly constructed covariates or between the covariates themselves.

Statistical Model

We must develop a statistical model where the outcome of interest is an opioid prescription, as defined above. Possible covariates to consider including are:

1. The age of the mother at time of delivery
2. Delivery type
3. Insurance plan type
4. Mental health diagnosis, either as a single indicator for any mental health diagnosis, one for anxiety and depression separately, or any combination therein.

Report

Once we have selected a final model, write a short summary of your findings. What variables were included in your model and how do you interpret the covariates? Does your model agree with the exploratory plots that you created? Does mental health diagnosis appear to have an effect on the prescriptions of periods post-postpartum?

Data

Here are the two data files provided

```
member <- read.csv("https://collinn.github.io/data/member.csv")
claims <- read.csv("https://collinn.github.io/data/claims.csv")
```

Along with a list of ICD codes for reference

ICD Codes	
F32.3	Depression
F41.8	Anxiety
G43.4	Migraine
J0745	opioid Prescription
Z34	Pregnancy Supervision (unused)
O80	Vaginal Birth
O82	Cesarean section

Rules

0. You may have groups of 1 to 3 individuals. You can speak with individuals within your group about the endeavor but nobody else in the course.

1. Course notes and internet are fair game

2. You *may not* use DASIL

3. You may ask the mentors questions. They cannot give you assistance, but they can answer if you are doing something correctly or not

4. You may ask me clarifying questions, which I will do my best to answer. You may also talk to me about what you are doing (sometimes explaining it out loud helps), but I will not be able to give any more direction than this

5. Final submission will be an R Markdown HTML file, just as we did in the labs. To this end, please consider the following:

    a. Do not print extraneous output. The report should be concise
    b. Leave comments in your code to describe what you are doing
    c. Include names of everybody in your group at the top

6. Due Thursday 12/14 at 5pm on pweb

Assessment

The grading scheme for this will be broken down as 60/40/10 between data processing, reporting, and code submission. A maximum of 100 points are available

Data Processing – 60 pts

- Inclusion Criteria (20 pts)

  – Age criteria (10 pts)
  – Coverage criteria (10 pts)

- Variable creation (40 pts)

  – Age at delivery
  – delivery type
  – opioid prescription status
  – Migraine status
  – Insurance type
  – Mental Health diagnosis

Report – 40 pts

- 2-3 Exploratory graphs with brief explanations as to what the plot is showing (10 pts)
- Create statistical model (10 pts)
- Justify model, interpret coefficients, give conclusion as to association between mental health diagnosis and opioid prescription (10 pts)
- Your document should be clearly organized with headers and labels

Code submission – 10pts

- Code should run without errors
- Comments describing what your code is doing
- No extra output included in final document

Hints

In no particular order:

1. This entire endeavor can be done using only packages used in class

2. Recall the difference between `inner_join()` and `left_join()`. What does each keep? You may need both

3. In nearly all data processing steps, you will want your data to be grouped by `id`

3.5. Several of the tasks you are asked to complete are specific iterations of more general problems. Think carefully about what you are trying to do when using a search engine for help

4. Use `%in%` instead of `==` for testing if ICD codes are contained within some vector

5. Many of these covariates can be made in any order – if you are getting stuck preparing one, try moving on to a different one

6. If you are running into problems, try working with a smaller subset of your data to verify that you are creating variables correctly. Here is a good candidate for doing so:

```
member_practice <- member[member$id == "PID240296", ]
claims_practice <- claims[claims$id == "PID240296", ]
```

7. Once you have processed all the variables you need, remove the columns in the dataset that you don't and use `unique()` to be sure that you only have one observation per row. Your final dataset should have 4240 total subjects

8. Here is an example of a function that might be useful. If you understand how it works, you could modify it to do something different

```r
library(lubridate)
happen_after_n <- function(y, m, d, n) {
  m <- m + 12*(y == 2018)
  event <- month(d) + 12*(year(d) == 2018)
  (max(m) - max(event)) >= n
}

# How might you use that function with this?
df <- data.frame(month = rep(1:12, times = 2),
                 year = rep(2017:2018, each = 12),
                 date1 = "2017-10-10",
                 date2 = "2018-10-10")
```