Background

Project 3 will be the analysis of a case-control study of Multiple Sclerosis (MS) patients and prior usage of Ipratropium and Salmeterol from the Saint Francis Medical Center in Illinois from 2014 to 2017. The publication citation and link are:

1. Ren, J., Ascencio, M., Raimondi, T., Rainville, E. C., Valenzuela, R. M., & Asche, C. V. (2019). Association Between Exposure of Ipratropium and Salmeterol and Diagnosis of Multiple Sclerosis: A Matched Case–control Study. *Clinical Therapeutics*, *41*(8), 1477-1485.

https://www-sciencedirect-com.libproxy.lib.unc.edu/science/article/pii/S0149291819302425?via%3Dihub .

Each case was matched to 10 controls with a common service year/quarter, age within 5 years, sex, race, and insurance payer type.

This project is designed to be completed in either R or SAS. There will be 5 parts, each weighted equally, for a total of 100 points on this project.

The important variables to use for this project include:

- `ms_cn`, which is whether a patient is a case (`ms_cn=1`) or control (`ms_cn=0`).
- `Match_pairs`, which is an ID of which cluster of case and controls an individual belongs to.
- `ipratro_yes`, which is 1 if the individual has taken at least 1 Ipratropium prescription in the past and 0 otherwise.
- `ipratro_cnt`, which is the total number of Ipratropium prescrptions an individual has filled.
- `SALME_yes`, which is 1 if the individual has taken at least 1 Salmeterol prescription in the past and 0 otherwise.
- `SALME_cnt`, which is the total number of Salmeterol prescrptions an individual has filled.
- `FH`, which is a family history of MS.
- `ANY_AL`, which is 1 if an individual has any alcohol use beyond social drinking.
- `AGE`, the age of an individual at assessment.
- `SEX`, which is 1 if male and 0 if female.
- `smoking`, which is 1 if the individual is an active smoker, 2 if the individual is a former smoker, and 3 if the individual never smoked.
- `payer_3cat`, a categorization of the individual's insurance payer status.
- `myeline_med`, an indicator that is 1 if an individual has taken other myeline-related medications.
- `Year`, the calendar year of assessment.

Section I: Data Cleaning

For the data cleaning process, you will create 3 datasets: `ms_cc`, `ms_cases`, and `ms_controls`. If you are using SAS, output your results to a sas7bdat file in the output folder. If you are using R, output your results to a CSV file in the output folder.

`ms_cc`

1. Keep only the important variables listed above for this project.
2. Add a new variable called `ipratro_salme`, which is 1 if the individual ever filled an Ipratropium or Salmeterol prescription, and 0 otherwise.
3. Add two new variables called `ipratro_many` and `salme_many` which is 1 if the individual filled 2 or more prescriptions of Iptratropium or Salmeterol, respectively.

`ms_cases`

1. This dataset should subset `ms_cc` such that only the cases are included.

`ms_controls`

1. This dataset should subset `ms_cc` such that only the controls are included.

Hints

For R, the `tidyverse` package can assist with much of the data cleaning process. Then, results can be written to a CSV file using the `write_csv()` function within `tidyverse`.

Section II: Case-Control Matching Assessment

The publication noted that patients were matched based on common service year/quarter, age within 5 years, sex, race, and insurance payer type. The original publication stated that this "controls confounding", which is incorrect (matching removes confounding bias for certain target populations, but not all; more details to come in EPID715).

Age is the only variable that is not perfectly matched on. Find the difference between each control and corresponding case age, and summarize this information with the average difference (`avg_diff`), the average absolute difference (`abs_diff`), and the mean squared difference (`msq_diff`).

Save the 3 variables using the provided names above into a dataset called `cc_match`. This dataset should have only 1 row and 3 columns.

Hints

For SAS, I was able to solve this in 2 lines of a DATA step using `MERGE` and `BY`. Remember that if two datasets have the same variable name, `MERGE` will keep only the left-most dataset's information, so you may need an inline `RENAME` statement.

For R, you may want to do what is called a left-join. A left join of dataset X and dataset Y on some ID variable would keep all observations in X, append the columns of matches of the ID variables from the Y

dataset. So, not all observations of dataset X will have a match (but will be present), and not all observations of dataset Y will be matched (so will not be present). The corresponding `tidyverse` function is `left_join`.

For both, once you have the control age and the case age, you can create a new variable with the difference between age of control and age of case (call this `diff_age`), and then add another variable which takes the absolute value of that difference (call this `abs_diff_age`).

- The average difference is simply the mean of `diff_age`.
- The average absolute difference is the mean of `abs_diff_age`.
- The mean squared difference is the mean of `diff_age**2` (square `diff_age`, and then take the mean).

Section III: Unadjusted Odds Ratio

Let's calculate the Unadjusted Odds Ratio by hand. Recall that the odds ratio will be the odds of exposure among cases divided by the odds of exposure among controls. For each of the following exposure definitions, find the unadjusted odds ratio and 95% (Wald-type) confidence interval. Save the exposure variable name (`variable_name`), ORs (`or_unadj`), and 95% CIs (`or_lcl`, `or_ucl`) to a dataset called `ms_or`. If the odds ratio is 0, set the 95% CI to be missing.

- `ipratro_selme`
- `ipratro_yes`
- `ipratro_many`
- `salme_yes`
- `salme_many`

Hints

For SAS, you have to do repeated work, just with a different variable each time; I suggest using a MACRO. You can find the unadjusted odds ratio by hand (i.e. using PROC IML) or by using PROC FREQ and the appropriate options. In addition, you can concatenate datasets together and perform a merge at the same time.

For R, it may be easiest to create a function which takes in a dataframe of counts and outputs the odds ratio and confidence interval by hand. Recall that the variance for a log odds ratio is

$$Var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

For concatenating, the `bind_rows()` function from `tidyverse` may be the easiest to use.

The output for this section will be a dataset with 5 rows (one row per exposure variable) and 4 columns (variable name, OR, and confidence interval).

Section IV: Adjusted Odds Ratio using GLMs

Using a Generalized Linear Model with binomial distribution and logit link, model the proportion of being a case or control as a function of `ipratro_salme` while controlling for family history, age, sex, any alcohol use, smoking status (set 1 as reference), and insurance payer status (set 1 as reference).

Output 2 datasets:

- `ms_params` which contains the model parameter names (`param_name`), model parameter estimates (`param_est`), and the model parameter standard errors (`param_se`). Remove any rows which are all 0's.
- `ms_glm` which contains the odds ratio for `ipratro_salme` (`oddsratio`), and its lower and upper 95% confidence intervals (`oddsratio_lcl`, `oddsratio_ucl`).

Hints

For SAS, use PROC GENMOD. The ESTIMATE statement I used is:

```
ESTIMATE "OR"INT 0 ipratro_salme 1 / E EXP;
```

but may vary depending on your implementation.

For R, use the `glm()` function. The `tidy()` function in the package `broom` takes a model summary object and transforms it into a dataframe, which may be helpful.

Section V: How Many Controls? A Bootstrap Simulation

In case-control studies, additional controls do not (asymptotically/over the long run) affect the point estimate, but they do help regarding precision. To prove this, we will perform a bootstrap-like simulation comparing when fewer controls are selected versus when all 10 controls are selected.

This simulation will be repeated for 1, 2, 3, 5, and 10 controls. For each simulation, create 5000 replicate datasets. Each replicate dataset should contain the given number of controls per `match_pairs`, selected with replacement. Then, concatenate the cases to each of the replicate datasets. Then, run PROC GENMOD or `glm` using the exact same model structure as Section IV. Lastly, summarize the information across all replicates by finding the exponentiated average log odds (take average of log odds, then exponentiate), standard deviation of the log odds, minimum odds ratio, and maximum odds ratio across all simulation runs.

You will submit 1 dataset called `sim_controls` containing 5 rows with the following information:

- `num_controls`, the number of controls in the current simulation. Note that you do need to repeat this analysis for 1, 2, 3, 5, and 10 controls.
- `avg_or`, the exponentiated average log odds across the current simulation runs
- `se_or`, the standard deviation of the log odds across the current simulation runs
- `min_or`, the minimum odds ratio across all simulation runs
- `max_or`, the maximum odds ratio across all simulation runs.

For grading, accuracy will be checked to 2 decimal places. I will use seed 700 for grading.

Hints

For SAS, PROC SURVEYSELECT allows you to sample within a strata of another variable using the STRATA statement. So, if you went within STRATA of `match_pairs`, then you would randomly select N controls per case, where N is what you specify in SAMPSIZE.

For R, to get a random sample of controls within strata of `match_pairs`, you can `group_by(match_pairs)` before using `slice_sample(n=...)`. Also, note that randomly selecting 5 observations 5000 times is equivalent to randomly selecting 25000 observations and then assigning which replicate it belongs to afterwards using `rep` or another function. For `glm`, the code is not totally intuitive for how to extract just the logodds estimate, so use the code below:

```
# Data has at least these columns: replicate, ms_cn, ipratro_salme, FH, age, sex, ANY_AL, smoking, payer_3cat
# Requires tidyverse and broom packages to be installed and connected
data |>
    group_by(replicate) |>
        mutate(
          logodds = glm(ms_cn ~ ipratro_salme + FH + age + sex + ANY_AL + factor(smoking) + factor(payer_3cat),
                        family=binomial()) |>
            tidy() |>
            filter(term == "ipratro_salme") |>
            pull(estimate)
        ) |>
        ungroup() |>
        summarise(
            # Fill me out with the 5 variables you need to calculate!
            ...
        )
```

For both, it will potentially be easier to create the replicates for controls first, and then worry about duplicating cases for each replicate afterwards. Larger simulations like this can take a little bit - anything longer than 5 minutes probably means something is wrong. 5000 replicates were chosen because that guarantees 2 correct decimal places around 90% of the time; feel free to use more or less replicates if you are off of the answer key. These are the numbers I got for a reduced number of iterations on the answer key:

Controls	Number Iterations	Mean	SE	LCL	UCL
1	500	0.43	0.29	0.20	1.09
2	500	0.46	0.21	0.28	0.87
3	500	0.47	0.16	0.31	0.75
5	500	0.48	0.13	0.35	0.72
10	500	0.49	0.10	0.37	0.67

Bonus Section

For extra points, write up to 5 criticisms you have of the case-control study data and publication in the README of Project 3. Each valid criticism will be worth 3 points, and if you supply more than 5 criticisms, only the first 5 will be graded.