MEDSCI 9506

Lecture 1

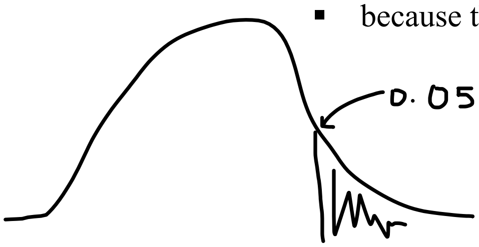
- Everything in R is either a vector or a function
    - analyze data with functions
    - 1- makes lists
    - most important thing is comments
        - # (space) within code makes a comment
    - () means its a function
    - ? function in the console
    - str = structure
    - tells you what kind of data you have
    - cant start a line with a number
    - all the vectors in a data frame must have the same number of elements (columns)
    - data frame is a special kind of list
        - list is a generic way of storing data; where you can put anything in it
    - as.matrix turns data frame into a matrix
        - $ subset out = only show me a portion of list/dataframe
    - data frame is a special list
    - every element is the same length
    - true = 1
    - false = 0
    - matrix must say what [row, column]
    - "12" is a character
    - 12 is an integer
    - 12.0 is a double

- dataframe(a,b.c) puts variables a,b,c into a datafram object
- apply(X, margin, function)
- X is the dataframe/matrix (must be all numbers)
- margin is row, column
- function: what you want to apply to those columns
- QC - quality control plots
    - residuals vs fitted: how far each point is from the line of best fit
    - Q-Q residuals: tells us if the fit from the line is normally distributed
    - scale-location: how far we are off; should be a straight line if it's a good dataset
    - residuals vs leverage: how far the furthest point is

MEDSCI 9506

Lecture 2

- Stack exchange → Resource
- R bloggers → Resource
- Vector [n]
  -
- Mat [r, c]
  - Ros, columns
  - [, c] → get all the rows in the column
- Can not subset data frame row only by column
- Ecologically fallacy → similar to Simpsons paradox
  - Group data can behave like ungrouped data
- What is a p-value
  - This is how we use p value
    - how confident are you in saying that two sets of data are different
    - likely hood that the correlation of the result is based on chance or coincidence
    - probability of rejecting the NULL
    - are the results statistically significant (0.05)
      - because its arbitrary
  - what it actually means
    - probability of rejecting the NULL (usually assuming no difference between groups) if the NULL is a reasonable model of the generating process for the data
  - NULL hypothesis (no difference) is the best?
    - No, it's not
    - Effectively there are an infinite number of reasonable NULL hypothesis
  - probability of the data being in the tail of the NULL distribution beyond some arbitrary value (0.05)
    - how far you are in the tail, doesn't really matter
      - it depends on the distribution
  - shouldn't say that x is more significant than y
    - because the model could be different or could be wrong



- false discovery rate or multiple test correlation
- # Benjani- Hochberg → multiple a p- value with a number

MEDSCI 9506

Lecture 3:

November 18, 2024

Pre session tasks

- Background
- Mood disorders
    - Bipolar disorder and major depressive disorder
    - Diagnosing bipolar disorder is complex
    - Overlapping symptoms with MDD
    - can lead to decade long delay in acute diagnosis
- RNA editing
    - A to I RNA editing
    - post transcriptional modification
    - potential as a biomarker
- objectives
    - validate an RNA editing based blood biomarkers panel combined with an AI algorithm to distinguish bw BD from MDD
- Methods
    - cohorts
    - study 1: internal development and validation cohort
    - study 2: external validations with 143 participatns
        - 100 MDD, 43 BD
- RNA editing biomarkers
    - eight genes
    - RNA extracted, sequences with NGS
    - analyzed for A to I editing patterns
- AI algorithm
    - developed using extra trees (ET) method
    - combined RNA editing biomarkers with covariates
        - sex
        - psychiatric treatments
    - evaluated using performance metrics
        - accuracy
        - sensitivity
        - specificity
        - AUC-ROC

- Key findings
    - diagnostic performance
    - internal validations
        - 0.901 AUC ROC
        - 80.6% sensitivity
        - 85.1% specificity
        - 83.7% accuracy
    - biomarker relevance
        - significant discrimination bw MDD and BD using RNA editing patterns
        - biomarkers linked to immune and neurological functions
- Implications
    - AI driven approach, earlier and more accurate BD diagnosis
    - improved diagnositc accuracy
    - tailored treatments
- limitations
    - small sample size; 388 participants
    - lack of differentiation among BD subtypes
    - potential confounding effects of medications
    - ongoing studies aim to further validate the test in a larger and drug naive population
- AUC ROC: area under the receiver operating characteristics curve
    - evaluating performance of binary classification models
    - y: true positive rate
    - x: false positive rate
    - shows trade off bw detect actual positives and minimizing false positives
    - 1 = perfect model, 0.5 random guessing, less than 0.5 worse than random guessing

Lecture notes

Oral exam

- What is the syntax
- how to make vector
- how to make a matric
- Subset a matrix and a vector
- How to make dataframe
- Given a simple for loop what will the output

  - How would you do it from 10 to 20
- Give simple function what till be output

Code Explanation

Cars Dataset

- summary(cars)
    - # This function provides a summary of the `cars` dataset, which includes key descriptive statistics like mean, median, and range for numerical variables. The `cars` dataset is preloaded in R and contains speed and stopping distances for cars.

Data Types and Structures

- v.x <- vector()
    - # Creates a generic empty vector `v.x` without specifying its type.
- v.y <- c(1:10)
    - # Creates a numerical vector `v.y` containing integers from 1 to 10.
- str(v.x)
    - # Displays the structure of `v.x`, showing its type and length.
- str(v.y)
    - # Displays the structure of `v.y`.
- m.x <- matrix(data = c(1:12), nrow = 3, ncol = 4, byrow = TRUE)
    - # Creates a 3x4 matrix `m.x` filled row-wise with integers 1 to 12.
- str(m.x)
    - # Displays the structure of `m.x`, including its dimensions and data type.

Data Frame Creation and Operations

- v.a <- c(14:17)
    - # A numerical vector containing integers from 14 to 17.
- v.b <- c("A", "b", "C", "D")
    - # A character vector with four elements.
- v.c <- c(2.5)
    - # A numerical vector initialized with a single value (should likely contain more values for matching length).
- df.abc <- data.frame(v.a, v.b, v.c)
    - # Combines the vectors into a data frame `df.abc`.
- str(df.abc)
    - # Displays the structure of `df.abc`, showing variable types and dimensions.
- sum(df.abc$v.a)
    - # Calculates the sum of the values in the `v.a` column of `df.abc`.

- List.ab5 <- list(v.a, v.b, v.c, v.5)
  - # Creates a list `List.ab5` containing multiple objects.
- as.matrix(df.abc)
  - # Converts the data frame `df.abc` to a matrix, coercing elements to a common data type.
- str(as.matrix(df.abc))
  - # Displays the structure of the matrix derived from `df.abc`.

Anscombe's Dataset Analysis

- data(anscombe)
  - # Loads the `anscombe` dataset, which includes preloaded data demonstrating relationships between variables.
- str(anscombe)
  - # Shows the structure of `anscombe`.
- apply(anscombe, 2, mean)
  - # Computes the mean for each column (2 indicates column-wise operation).
- cor(anscombe$x1, anscombe$y1)
  - # Computes the correlation between `x1` and `y1`.
- lm(anscombe$y1 ~ anscombe$x1)
  - # Fits a linear regression model predicting `y1` using `x1`.
- plot(anscombe$x1, anscombe$y1)
  - # Creates a scatterplot of `x1` and `y1`.
- abline(lm(anscombe$y1 ~ anscombe$x1))
  - # Adds a best-fit line to the scatterplot.

Matrix Operations

- ans.mx <- as.matrix(anscombe)
  - # Converts the `anscombe` dataset to a matrix format.
- class(ans.mx)
  - # Checks the class of `ans.mx`.
- ans.df <- as.data.frame(ans.mx)
  - # Converts the matrix back to a data frame.
- class(ans.df)
  - # Checks the class of the resulting data frame.

- summary(ans.df)
  - # Provides a summary of the data frame.
- ans.ls <- as.list(anscombe)
  - # Converts the `anscombe` dataset to a list format.
- class(ans.ls)
  - # Checks the class of the list.

Custom Matrix and Histogram

- new.mx <- matrix(runif(1000), nrow = 10, ncol = 100)
  - # Generates a 10x100 matrix filled with 1000 random numbers from a uniform distribution.
- hist(new.mx[1,], breaks = 9, main = "First row of matrix")
  - # Plots a histogram of the first row of the matrix.
- cs <- colSums(new.mx)
  - # Computes column sums of the matrix.
- hist(cs, breaks = 9, main = "Histogram of Column Sums", xlab = "Column sums", ylab = "Frequency")
  - # Plots a histogram of column sums.

Simpson's Paradox

- Friend1 <- read.csv("mydata.csv")
- Friend2 <- read.csv("mydata-2.csv")
  - # Reads two datasets from CSV files.
- cg <- rbind(Friend1, Friend2)
  - # Combines the datasets by stacking rows.
- plot(Friend1$X, Friend1$Y, main = "Friend1 plot")
  - # Plots `X` vs `Y` for `Friend1` with a title.
- abline(lm(Y ~ X, data = Friend1))
  - # Adds a trend line to the plot for `Friend1`.
- plot(cgX, cgY, main = "Combined plot")
  - # Plots the combined dataset with a trend line to demonstrate Simpson's Paradox.

CODE

- personality <- rnorm(100, 10, 4)
  - # Generates a random sample of 100 values from a normal distribution with a mean of 10 and standard deviation of 4, representing "personality."
- looks <- rnorm(100, 10, 3)
  - # Generates a random sample of 100 values from a normal distribution with a mean of 10 and standard deviation of 3, representing "looks."
- df.p1 <- data.frame(looks, personality)
  - # Creates a data frame `df.p1` with "looks" as the first column and "personality" as the second column.
- plot(df.p1$looks, df.p1$personality)
  - # Creates a scatterplot of "looks" (x-axis) against "personality" (y-axis).
- abline(15, -0.9)
  - # Adds a line with intercept 15 and slope -0.9 to the scatterplot. This line could represent a "threshold" or "benchmark."
- cor(df.p1)
  - # Computes the correlation matrix between "looks" and "personality," quantifying the linear relationship between these variables.

Why It Works:

- rnorm() draws values from a normal distribution.
- data.frame() organizes related variables into a tabular structure.
- plot() visually explores relationships between variables.
- cor() evaluates how strongly two variables are linearly related.

What is a p-value

This section explains the concept of p-values, which measure the probability of observing results as extreme as the data under the null hypothesis. It also explores how p-values behave under different conditions through simulation.

Code:

- nrow = 1000
- ncol = 6
- p.mat <- matrix(data = runif(nrow * ncol), nrow = nrow, ncol = ncol)

    - # Creates a 1000x6 matrix filled with random numbers uniformly distributed between 0 and 1.
- p.mat[, 4:6] <- runif(n = 3000, min = 0.5, max = 1.5)
    - # Replaces columns 4 to 6 with random numbers between 0.5 and 1.5.
- p.out <- apply(p.mat, 1, function(x) as.numeric(t.test(x[1:3], x[4:6])$p.value))
    - # For each row, performs a t-test comparing the first three columns with the last three columns and extracts the p-value.
- hist(p.out, breaks = 99, main = "Histogram of P value", xlab = "p-value", col = 'gray')
    - # Plots a histogram of the p-values obtained from the t-tests.
- abline(v = 0.05, col = 'red', lty = 2, lwd = 3)
    - # Adds a vertical red dashed line at 0.05, marking the threshold for significance.
- sum(p.out <= 0.05) / nrow
    - # Calculates the proportion of p-values below the 0.05 significance threshold.
- min(p.out)
    - # Finds the minimum p-value from the simulated tests.
- p.adjust(p.out1, method = "BH")
    - This adjusts the p-values using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR).

Comparing Sample Sizes

By increasing sample size or the number of columns being compared, the code demonstrates how statistical power increases, making it easier to detect small differences.

Key Observations:

- Larger sample sizes lead to p-value distributions concentrating closer to significant thresholds.
- This is a demonstration of how increasing statistical power reduces false negatives.

Benjamini-Hochberg Correction

Code:

- min(p.adjust(p.out1))

    - Adjusts p-values using the Benjamini-Hochberg (BH) method, which controls the FDR.
    - Returns the minimum adjusted p-value, reflecting the most significant finding after correction.

Why It Works:

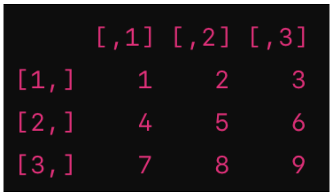
- The BH method ranks raw p-values, adjusting them to account for the multiple comparisons problem.

- Helps mitigate the risk of false discoveries in multiple testing scenarios.

How is a vector made

- Create a vector using c()
    - numeric_vector <- c(1, 2, 3, 4, 5)
    - character_vector <- c("apple", "banana", "cherry")

How is a matrix made

- matrix(data, nrow, ncol, byrow = TRUE)
    - byrow= False if filing column wise
    - mat <- matrix(1:9, nrow = 3, ncol = 3)
    - 

      ```
           [,1] [,2] [,3]
      [1,]    1    2    3
      [2,]    4    5    6
      [3,]    7    8    9
      ```
- Can use cbind() to combine column or rbind to combine rows
    - vec1 <- c(1, 2, 3)
    - vec2 <- c(4, 5, 6)
    - mat <- rbind(vec1, vec2)

How to subset a vector, matrix, or dataframe

-

Logic for a for loop

Logic of simple function

MEDSCI 9506

Dec 2, 2024

Lecture 7

Importance of a Research Question

- First interaction with your research
- Develop foundation for research activities
- Clearly defined research question needed to understand how to go about research

What is an Effective Research Question?

- A well-defined research question…
    - Describes the things that you are performing your research about (the population)
    - Clearly defines the purpose of your research (to compare or to describe)
    - Outlines the endpoints you are analyzing (the outcomes)
    - Sets up the setting in which you are performing the research (study design and/or time frame)

Defining the Purpose of Research

- Often, defining a clear research question is difficult because the exact purpose of the research is unclear or uncertain
- Some questions to ask are:
    - Am I trying to describe phenomena? Or am I trying to develop a deeper understanding of something that is well-defined? (i.e. Is this a causal question?)
    - What is the ultimate goal of this research?
    - Am I trying to understand something very specific or trying to understand a general phenomenon?

PICO[1]

- **P** – Population
- **I** – Intervention

- **C** – Comparison
- **O** – Outcome
- (**T** – Time or Type of question)
- (**S** – Study Design)

Examples

- "Among adults over the age of 18 living in Canada, what is the difference in the rate of coronary artery disease between those who smoke and those who do not?"
- "What is the difference in economic well-being between adults aged 18-25 living in Sweden vs Canada?"
- "What is the difference in electrical resistance in wiring made of copper vs tungsten?"

Can we make these better?

- "What is the impact of seat belts?"
- "What is the effect of binge drinking on academic performance?"
- "Among individuals, what are the effects of perceived cultural norms on mental wellness between migrants and non-migrants?"

Key Takeaways

- The research question is a vital part of the research process that set-ups the foundation for the research itself
- The PICOS format helps to clearly describe the What? Why? How? of your research
- Developing a clear research question helps clarify all other aspects of your research not only for others but for yourself as well!

What is a cause?

- Key considerations:
    - Complex
    - Complicated (sometimes)
    - Multifactorial
    - Unidirectional

Assessing Causality: An Overview

- The study of causal inference has a storied history

- Lots of debate, but mathematical formula a recent advent
- Many ways folks have developed to conceptualize causal mechanisms
    - Hill’s ‘criteria’
    - Sufficient-Component Cause
    - Potential Outcomes

Hill’s “Criteria”

- Temporality
- Strength of association
- Dose-response relationship (gradient)
- Consistency
- Specificity
- Plausibility
- Coherence
- Analogy
- Experimental evidence

Strength of Association

- Hill’s Criteria states that causal relationships have stronger associations
- Example: Smoking increases risk of dying from lung cancer by 10 times in non-smokers and 30 times for those who smoke heavily

Sufficient-Component Cause

- Sufficient: Each set of factors that may cause disease
- Necessary: Cause must be present for disease to occur
- Sufficient causes of Lung Cancer?
- Necessary causes of Lung Cancer?

Potential Outcomes and the Counterfactual

- THE analytical framework for quantifying causal effects
- Comparison of counterfactual in order to obtain some measure of causal association

Potential Outcomes and the Counterfactual

- Ask the question, if exposure ‘a’ had not occurred for the individual (or population) would the outcome have occurred?
- A set of scenarios for binary outcome:
    - 1) Doomed – outcome regardless of exposure
    - 2) Exposure causative – exposure causes outcome
    - 3) Exposure protective – exposure protects against outcome
    - 4) Immune – outcome never happens

Correlation vs Causation

- Difficulty with causal inference is identifying causal effect while separating out spurious correlation
- Causes are MULTIFACTORIAL!
- Question: Does marriage cause divorce?

Directed Acyclic Graphs: Intro

- Directed Acyclic Graphs useful for developing causal models
- Arrows indicate causal relationship; Cannot make cycle (hence, acyclic)

Confounding and RCT’s

- DAG’s particularly useful to identify potential confounders
- Confounder: A factor that CAUSES both exposure and outcome
- Why are RCT’s so valued?
- “Breaks” associations between confounder through randomization

Observational Studies

- Sometimes RCT’s not possible
- Problem: How do we know that we’ve accounted for all confounders?
- Follow the theory!

Importance of Theory Driven Research

- Strive to simulate RCT conditions as closely as possible
- Guidance of underlying theory important

1. Understand where potential sources of bias are (confounding)
2. Understand limitations in data
3. Understand whether causal statements may be made under current knowledge

The role of DAG's in Research

- DAG's provide visual representation of theory
- Helps to understand what factors are most important in answering a given research question
- Helps understand how to build statistical models in order to answer research question

What is Data/Big Data?

- Many sources, Much data (good and bad)
- As much data as there are, there are equally as many analyses (good and bad) that can be done

Statistics vs Machine Learning

- Statistics:
    - **Confirmatory** → provide evidence (or lack thereof) of some hypothesis or theory
    - Build models using simple functions
    - Focus on **inference** (hypothesis testing/confidence intervals)
    - Machine Learning:
    - **Exploratory** → reveals POSSIBLE structure in data
    - Build models using more complex functions
    - Focus on **prediction**

"Learning"

- Can be thought of as mathematical pattern recognition
- Can be supervised, reinforcement, or unsupervised
- Supervised → Given a set of predictors, what is the association with the outcome
    - If outcome is continuous → Regression
    - If outcome is categorical → Classification
- Performance measured by how good predictions are on new data (out-of-sample)

- Unsupervised → Let model identify which predictors are most important
- General process:
    - "train" model on a subset of data
    - "test" or "validate" model on the rest
- Called "training set" and "testing set" respectively
- Nothing stops you from fitting increasingly complex models
- HOWEVER there are a couple things to consider:
    - Bias-Variance Tradeoff
    - Overfitting
- These are related to generalizability

Bias-Variance Trade-off

- Bias = error in prediction
- Variance = variability in prediction
- Ideal case: low bias, low variance (i.e. high accuracy, high precision)
- DIFFICULT TO ACHIEVE!
- Usually a large tradeoff between bias and variance
    - As you build out the model to be less biased, it usually comes at the cost of variance (and vice versa)
- Principle can be seen in some machine learning methods:
    - Regularization: introduces bias into model to reduce variance (Lasso, Ridge regression)

Overfitting

- As you fit more and more complex models, it fits better and better to training data
- Is this always a good thing?
- Internal validity vs External validity
- Related to bias-variance tradeoff → overfitting leads to less generalizability leads to poor predictions

Commonly used Models

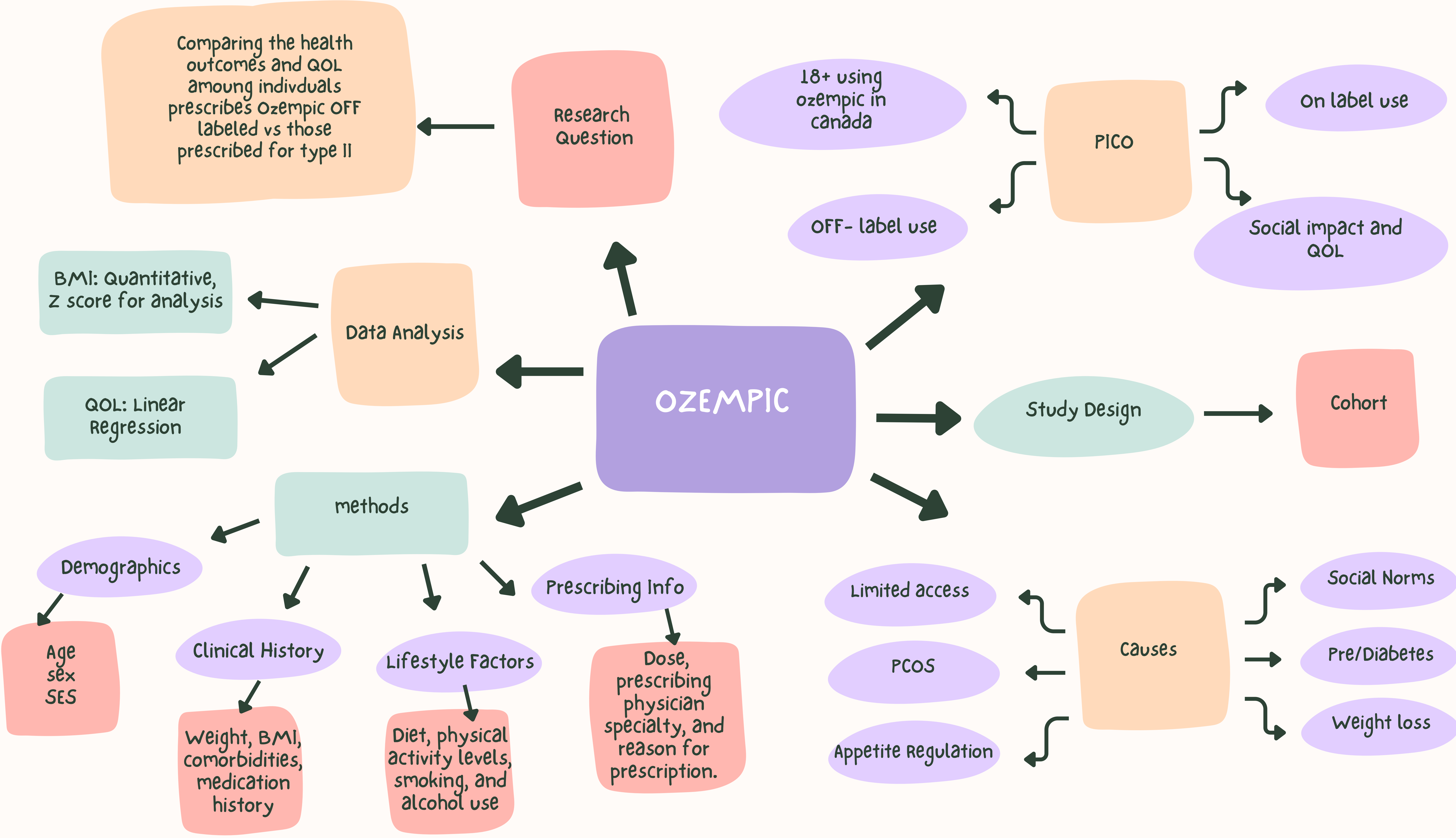
- Support Vector Machine
    - Primarily for classification, can be applied to large or small datasets
- Neural Network
    - For classification or regression, state-of-the-art for many tasks with abundant data where relationship between inputs and outputs must be "discovered." Not so good for data-poor tasks, difficult to interpret.
- Decision Trees
    - Intuitive models for regression or classification, naturally handle categorical variables, can work well when inputs have good theory behind them but relationships are nonlinear
- All of these models can capture nonlinear relationships.
- Some applications: image labeling, speech recognition, machine translation, spam filtering, ...

Thoughts and Considerations

- Machine Learning focuses on building flexible predictive models.
- When might machine learning models be useful over traditional statistical models?
- What kinds of predictions might be useful for you?
- Do you think your data could support the predictions?

Machine Learning vs AI

- Sometimes people may use these terms interchangeably
- THIS IS NOT EXACTLY THE CASE
- Artificial Intelligence: Theory and development around automation of computer systems to simulate human action
- Machine Learning can be considered a subset of AI → ML is an application of AI to recognize patterns in data
- ML requires historical data, whereas general AI is able to use historical data, present data, and predict data that is to come (like a human can)



OZEMPIC

Research Question

Comparing the health outcomes and QOL amoung indivduals prescribes Ozempic OFF labeled vs those prescribed for type II

PICO

18+ using ozempic in canada

OFF- label use

On label use

Social impact and QOL

Data Analysis

BMI: Quantitative, Z score for analysis

QOL: Linear Regression

Study Design

Cohort

methods

Demographics

Age
sex
SES

Clinical History

Weight, BMI, comorbidities, medication history

Lifestyle Factors

Diet, physical activity levels, smoking, and alcohol use

Prescribing Info

Dose, prescribing physician specialty, and reason for prescription.

Causes

Limited access

PCOS

Appetite Regulation

Social Norms

Pre/Diabetes

Weight loss