Stat 141XP – Week Three – Session Two
Topic: Logistic Regression for Rare Events

The objective of today's lecture is to show what you should do when…

- Your outcome is binary
- You want to use logistic regression to make a predictive model
- But, you are facing a situation with rare events (too few observations in a category)

We will try to reach this objective by:

1. **Watching a video on logistic regression for rare events** https://www.youtube.com/watch?v=fVbrUz6V_uk

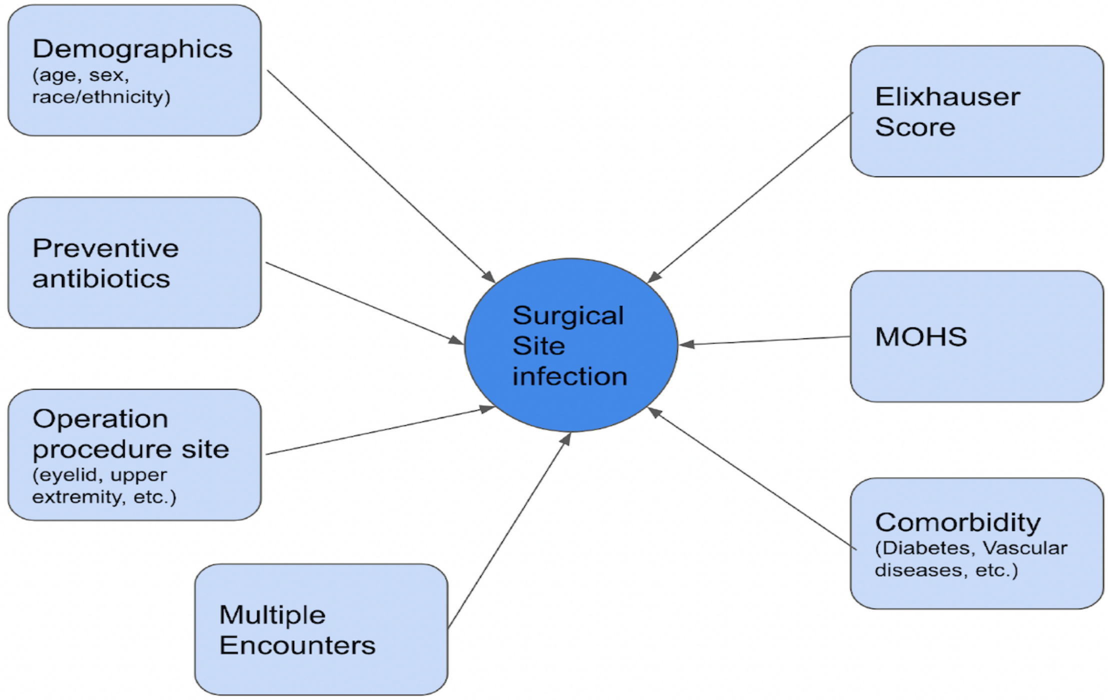
2. Going through a final presentation by a stat team that used "logistic regression with rare events to answer the client's question who specializes in plastic eye surgery". (question two)
3. Using ChatGPT to help with running logistic regression for rare events (question three)

Questions one to four will be "stat homework two". It will be due on Sunday April 23 eleven PM

Question one. Summarize what you learned from this YouTube that you did not know before**.**

Question two. Refer to the final presentation entitled “Surgical Site Infection” and answer the following questions: You are given the link to this document.

1. What is the use of the following schematic?


Demographics (age, sex, race/ethnicity)
Preventive antibiotics
Operation procedure site (eyelid, upper extremity, etc.)
Multiple Encounters
Surgical Site infection
Elixhauser Score
MOHS
Comorbidity (Diabetes, Vascular diseases, etc.)

a. What is the outcome? Is it numerical or categorical. If categorical identify the levels. What do the different levels show? **See page 3 in the report**

b. What are the predictors. Identify the numerical and categorical predictors and number of levels. **See page 3 in the report**

c. Google the definition of the terms Elix vw score, MOHS, and comorbidity in medicine

d. The schematic is supposed to help the client see the big picture and have an overall understanding of the problem to be solved. Refer to page three and read the problem statement. Do you think the team did a good job of aligning the schematic drawn with the statement of the problem?

5. Examine the EDA done on pages 4-12. What do you find in common about the given contingency tables.

6. Using the data called full_v5.csv, run logistic regression for rare events; using the "logistf library." Create the output as well as the table reported on page 21. You need to create a table with exponentiated coefficients and confidence interval. You can also find the R codes in the R-markdown files in the data folder given to you under Wednesday for week three.

7. Interpret the coefficient and confidence interval for immunocompromise within context.

8. Use the sjPlot library to draw the plot of odds for the output you created in part seven.

9. Create the contingency table for the outcome (antibiotics_after_procedure_less_thirty) and category2. Is there any relationship between these contingency tables and wide confidence intervals for the relevant odds ratios. Yes or No and explain why.

10. Identify the factors that contribute to surgical site infection.

11. Refer to the confusion matrix given on page 13 Use the relevant formulas to calculate accuracy, sensitivity, specificity, precision, recall, and F1-score. Define them all within context. (You can find these formula in the lecture for Monday – week three)

12. Comment on the ROC curve given on page 13 created for the model you created in. part 6.

11. Do you find this analysis effective? Yes or no and explain.

12. Complete the following table

accuracy	sensitivity	specificity	recall	Precision	F1-score	AUC
						0.542

The second part of this report or the random intercept model is the topic of next week.

Question three: Go to Chat GPT and aske the following question:

Ask Chatgpt the following question:

a) Once Chatgpt shows you how to create the data set, use the R codes given to create the relevant outputs and interpret what you find within context.

b) Then run logistic regression using regular glm and compare results

c) Draw the plot of the odds and summarize the findings within context for rare logistic method.