EGR 328 Dr. Lowman
23 January 2024

Lab #1
Introduction to Making Inferences with Data

Learning Objectives/Activities:

1) Access and download data from a freely available online repository
2) Read and visualize table data using MATLAB
3) Build an inverse model to estimate the number of new infections per day using data and knowledge from news articles
4) Propose a forward model based on the available data
5) Compare model performance with existing data
6) Evaluate sources of error and uncurtaining in data and model assumptions

Part A. Introduction to working with data

The purpose of this initial section of the lab is for you to gain familiarity with a new dataset by inspecting the contents of the data table, reading the data into MATLAB, and visualizing the data through exploratory figures. The followings steps and comprehension questions will walk you through the initial steps that you should complete whenever you work with a new dataset.

Procedure:

1) Download a sample of daily COVID-19 tracking data for the entire US or for a state of your choice from the COVID tracking website data download page. (https://covidtracking.com/data/download). Save the Excel sheet as a .csv file to a folder on your laptop or Google Drive.
2) Open the data as an Excel sheet and look at the different column headings. Can you interpret what data represent? What are the units? How were the data collected? (Note: All datasets provide a descriptor of the data collected in a dataset. For the COVID-19 tracking data, the data definitions are provided here: https://covidtracking.com/about-data/data-definitions.)
3) Reading data requires an understanding of how your data is formatted. This is important because often we use programming languages or advanced software to analyze and make inferences about the data we are working with. In this class, we will primarily use MATLAB to work with our data. Review the documentation on the readmatrix and importdata built-in functions for reading in column-oriented data from a file (https://www.mathworks.com/help/matlab/ref/readmatrix.html and https://www.mathworks.com/help/matlab/ref/importdata.html). Use one of these functions to read in your COVID data as a matrix called ‘A’.
4) Store the data from the ‘hospitalizedCurrently’ column to a vector called ‘hospCurr’ in MATLAB. Check that you selected the correct column by displaying the first 10 entries and comparing with your spreadsheet.
    a. Enter: >> hospCurr(1:10)
5) Store the data from ‘date’ column to a vector called ‘dates’. Check that you selected the correct column by displaying the first 10 entries and comparing with your spreadsheet.
6) Use the MATLAB plot built-in function to plot current hospitalizations versus time (https://www.mathworks.com/help/matlab/ref/plot.html). What do you notice about this figure? Does it make sense?
7) Let’s make the x-axis easier to understand for humans. To do this, we are going to reformat the ‘dates’ vector that we created to correspond with MATLAB’s built-in feature to account for time. This requires us to transform the data two times: first into a character string, second into a datenumber (https://www.mathworks.com/help/matlab/ref/datenum.html). Follow these steps:

  a. >> date_str = string(dates); % convert the ‘dates’ vector from numeric data to a string (or text data)
  b. >> formatIn = ‘yyyymmdd’; % describe how the date is formatted
  c. >> dates_num = datenum(date_str,formatIn); % convert the date_str to MATLAB’s internal date format

8) Add a title and axes labels to your figure. Congratulations! You are now on your way to becoming an expert at working with data!

Comprehension Questions:

1) List five of the column headers from your dataset Excel file and provide a brief explanation (1-2 sentences each) explaining the data provided (e.g., What do the data represent? What are the units? How were the data collected?). (***Example****: The ‘date’ column provides the date of the recorded COVID-19 data in YYYYMMDD format.*)
2) What is the data format of your COVID-19 data? What does it mean?
3) Did you run into any errors while reading in your file to MATLAB? What were they and how did you address them?
4) How is your data matrix ‘A’ structured within MATLAB?
5) How can you check that you have correctly assigned a column of data to a vector within MATLAB?
6) How does MATLAB keep track of time internally?
7) Based on what you did today, write up step-by-step procedure for how to read in a new dataset and initially visualize it using MATLAB.
8) What do you notice about the plot of hospitalizations over time? Any interesting features?

Lab Credit:

- **Submit your final plot and your responses to the Comprehension Questions above on Canvas to receive credit for completing Part A of Lab #1.**

Part B. Building your first inverse model.

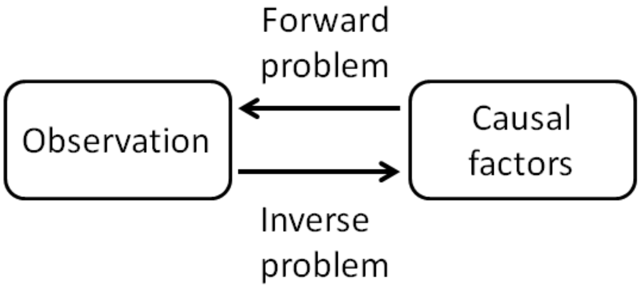
The purpose of this section of the lab is to walk you through building your first inverse model to estimate infection rates based on available COVID data. To do this we need a bit more information than what is provided in the dataset. Specifically, we need to figure out a way to relate the number of infections to other relevant data available in the table. Back in March 2020 when epidemiologists first noticed COVID cases on the rise, they used data collected from cruise ships in order to determine an infection rate. We will do the same thing here. The following steps and comprehension questions will walk you through the creation of your first inverse model.

Procedure:

1) Read the *Nature* article titled “What the cruise-ship outbreaks reveal about COVID-19” (https://www.nature.com/articles/d41586-020-00885-w#ref-CR2) and the Centers for Disease Control and Prevention (CDC) Report titled “Public Health Responses to COVID-19 Outbreaks on Cruise Ships — Worldwide, February–March 2020” (https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e3.htm)
2) Based on the articles, determine what data columns from the COVID-19 tracking dataset you can use to infer the number of persons infected.
3) The CDC article states that on the Diamond Princess cruise ship 712 passengers had positive test results and 9 of those individuals passed away. Use this information to infer a mortality rate. Would you expect this mortality rate to be the same in the general population?
4) Estimate the number of infections for your dataset using the mortality rate that you inferred from the cruise ship data.
5) Compare your inferred number of new infections to the number of new positive tests by plotting the data together on the same figure. Make sure to include a legend.

Comprehension Questions:

1) What additional information (outside of the dataset you downloaded) are you using to determine the number of people infected in COVID-19 in the general population?
2) Which data column will you use as a predictor for number of infections?
3) Explain why the mortality rate is or is not the same on the Diamond Princess cruise ship versus the general population. Does the current CDC estimate of a mortality rate align with what you found for the Diamond Princess cruise ship?
4) Compare your inferred number of new infections to the number of new positive tests. Should these two be related? If so, how? Do you see in the figure what you expect? What factors could impact the relationship between positive tests and the infection rate?
5) How do your estimates of number of infections compare with the epidemiologist’s models?
6) Consider the figure below that demonstrates in a simple way the difference between an inverse problem and a forward problem. You can also read the short blog post that it links to. Explain how the problem of inferring the number of infected fits within this simple description of an inverse problem.
7) Discuss a forward problem that you could tackle using the COVID-19 tracking data. How could you validate your results (i.e., check them against an independent dataset)?



Source: https://web.archive.org/web/20201130161929/https://miller-blog.com/inverse-problem-part-1/

Lab Credit:

- **Submit your final plot and your responses to the Comprehension Questions above on Canvas to receive credit for completing Part B of Lab #1.**