Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES

Lecture 28

Linear Regression: Understanding Research Question and Data Source

Indian Institute of Technology Guwahati

Jan-May 2023

How to attack the data analysis/model fitting:

- Understand the **research question(s)**. Understand the **data** you have.
  1. How the **data was collected**?
  2. What type of the **study design used**: *Randomized or Observational; Prospective or Retrospective etc.*
  3. Can you make **connection with the primary research question and the data**? Is the **research question feasible** based on the data you have?
  4. Are there **secondary research questions**?
  5. What are the **potential source of bias**? Sample (data) **may not be** a representative of the target (source) population.
  6. Are there any **confounders**? **Confounders**: A confounder (also known as

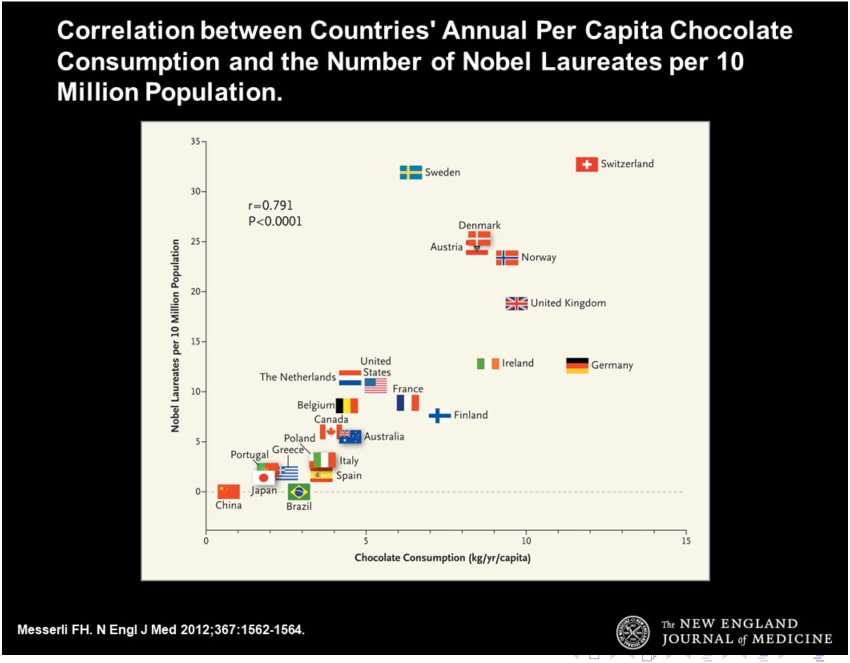
     confounding variable, confounding factor) is **a variable that influences both the dependent variable and independent variable**, causing a spurious association. **Confounding is a causal concept**, and as such, cannot be described in terms of correlations or associations.
  7. Are the **number of observations/individuals** in the data **sufficient**? Is it a **small data**?

How to attack the data analysis/model fitting:

- Do the **scatter plot**(s): response vs. input variable(s).
- **Fit the regression** model(s) (or other type of model(s)).
- **Interpret** the output from the fitted models:
  1. Are all the results expected? Whether the **results go well with existing domain** (basic science) knowledge?
  2. If not, what are the **reasons behind the aberration** from the expected results.
- **Check the diagnostics** for model assumptions. If you find problem, go back and correct (if you can) the chosen model; or, take decision about the outliers/influential points.
- Always remember: *“Essentially, all models are wrong, but some are useful” – Box*

Examples:

- A researcher wants to see **whether the occupation ‘bartender’ is a source of lung cancer.** Is the research question correct to you?
- Does **Caesarean section (C-section) increase** the chance of **childhood asthma** in a prospective study in Delhi/Beijing?



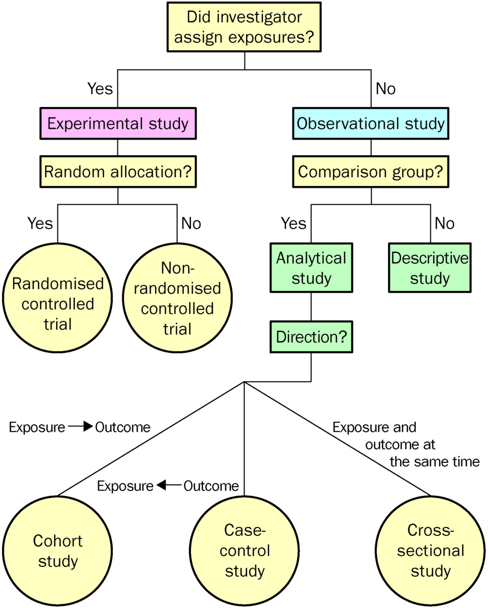
Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Messerli FH. N Engl J Med 2012;367:1562-1564.

The NEW ENGLAND JOURNAL of MEDICINE

- The **New England Journal of Medicine (NEJM)** is the most **prestigious journal of medical science**. The impact factor of NEJM is 176.1 (2021)!
- **Example: Bad science or fraud science??**
- The above article[1] is an example of **spurious correlation**.
- My intuition: if you **replace the x-axis with per capita cows** you may see the **same pattern!** Since most of the chocolate producing Europian countries have many per capita cows.

[1] Franz, H. Messerli (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. The New England Journal of Medicine, 367, 16.



- **Different types of Studies**

- The approach, interpretation, issue of bias and role of confounding factor will be different depending upon the study-design that generates the data.

- Reference: Grimes, D. A., & Schulz, K. F. (2002). An overview of clinical research: the lay of the land. The Lancet, 359(9300), 57-61.