Short summaries of what I read so far.

Jaemin Oh

November 23, 2022

Contents

1 Causal Inference

1.1 Holland, 1986 [2]

The very basic of Rubin's model was explained, and the distinction between associational inference and the causal inference was provided.

Let U be the population and $u \in U$ be a unit. Three variables Y, W, and C are given, where Y is the variable to be analyzed (response variable), W is a (potential) cause of Y, and C is an attribute. They are functions from U to $\mathbb{R}$, and their distributions are given by the relative frequency on U. As a function, W and C are the same. However, they are different in the sense that we cannot do a randomized experiment with C and can do with W. For example, a clinical surgery can be W and the gender can be C. This property distinguishes the cause from the attribute.

In this setting, the associational inference is focused on $E(Y|W)$ or $E(Y|C)$. In other words, the discovery of the way that Y is related to W or C will be satisfactory. On the other hand, in causal inference, a direct comparison between treatment and control for each unit is required. This cannot be done in practice, since any unit cannot receive both treatments simultaneously. But we can do counterfactual imaginations that lead additional functions $\{Y_w\}_{w \in I}$ which is called potential outcomes. To overcome the practical issue and estimate the causal effect, the researcher should design the study to approximate randomized experiment, which is the simplest setting. Note that, in a randomized experiment, $Y_w =_d Y|W = w$ by consistency and ignorability (missing at random).

2 Spatial Data Analysis

2.1 Hierarchichal Modeling and Analysis for Spatial Data [1]

1. Overview of spatial data problems

   Spatial data has three possible different forms: point referenced data, areal data, and point pattern data. Let $D \subset \mathbb{R}^d$ be a set of locations. If the data can be described as $Y(s_i)$ where $s_i \in D$ and s_i is deterministic, then it belongs to the class of point referenced data. Instead of the exact location, imagine that the information in $B_i \in 2^D$ is given. We call this case as an areal data. When D is a random set, then it is a point pattern data.

   For point referenced data, it is natural to think that $Cov\left(Y(s_i), Y(s_j)\right)$ is a function of a distance between s_i and s_j. The most convenient approach is assuming

$$(Y(s_i), \ldots, Y(s_m)) \sim N_m\left(\mu, \Sigma\right)$$
$$(\Sigma)_{ij} = \sigma^2 e^{-\phi d_{ij}^{\kappa}} + \tau^2 I(i = j)$$

   where τ^2 is called a *nugget effect*.

References

[1] Sudipto Banerjee, Bradley P. Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2ed. edition, 2015.

[2] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.