Classical papers in statistics

Mentor: Ethan Ancell Mentee: Janice Kim

Our DRP focused on reading classical papers about statistical topics and discussing them. The topic selection was based on fundamental topics in statistics and the influence of the paper based on the number of references. Rather than analyzing or summarizing the paper, we focused on parts or concepts in the paper for the discussion. We had a total of seven topics: Lasso Regression, Support Vector Machine, Misspecification of the Model, Akaike information criterion, Longitudinal Data Analysis and Generalized Linear Model, Kaplan-Meier Estimator for incomplete observations, and False Discovery Rate Controlling.

The main topic of the lasso regression paper was the complementary role of lasso regression with better interpretability compared to the subset, ridge regression, and Garotte function. During the meeting, we also primarily discussed two different forms of the Lasso regression, adding penalty as an absolute value and imposing a tuning parameter as a construction, with the Lagrange multiplier as a constructor.

The Support Vector Machine paper talked about the construction of a hyperplane or soft margin for the separable and inseparable training data using an alpha weight factor that critically affects the margin decision. The authors also compared with other classification methods including linear classifier, KNN, and LeNet to SVM, using data from the US Postal database. Our meeting topic was a multivariate normal distribution and its property that implies independence among variables when covariance is 0, which isn't always true for other distributions.

For the misspecification paper, we talked about the consistency of the Maximum Likelihood Estimator (MLE). Even though there is a misspecification between the unknown true distribution and assumed family of distribution, MLE converges the theta* that has the smallest Kullback Leibler information criterion (KLIC) that measures the divergence. The distribution of MLE also asymptotically converges to the normal distribution regardless of the unknown true distribution, providing valuable information in estimation.

In the fourth and fifth meetings, we talked about the usefulness of the Akaike Information Criterion (AIC) regarding model selection among models with different numbers of parameters with parameter k. Other topics were the generalized linear model and the utilization of link functions for its connection with a linear model and Longitudinal Analysis where time is not the primary interest.

Then our next topic was product limit estimator in incomplete observations over time like death and loss, like calculating the probability of survival after a certain amount of time. We discussed the difference between the Kaplan-Meier estimator, which includes all the data in the estimation, and the naïve method, which simply excludes losses from the data. The information that they were alive before the loss can be valuable, so using the entire data can be beneficial for the analysis.

Our last meeting was about False Discovery Rate (FDR) controlling. We focused on the definition of the FDR, its difference between Bonferroni correction, and the formulation of the FDR using an example in the paper and its expected value. We also briefly discussed the implication that FDR control can control the Family-Wise Error Rate (FWER) in a weak sense.

For me, it was a great opportunity to track back and review the origin of some fundamental concepts in statistics like MLE, Type I error control, and others. Even though it was difficult to digest all of them (especially complicated mathematical proofs), it was meaningful to get familiar with statistics papers and become less intimidated. Ethan's explanation makes papers and concepts easier and more relevant, making a huge difference from my own reading. During the reading, I felt a bit frustrated by the fact that there are numerous concepts and fields, but I only know a tiny portion of them. However, he encouraged me by saying that knowledge stacks up, so if I don't have an understanding of basic concepts I would not understand the paper. I also realized that reviewing and understanding basic concepts, and knowing how they work and what they do, is imperative for application. Lastly, it was just fun to learn new things.