Stat 430 Applied Bayesian Analysis

Final Project Proposal

Kexin Fei, Dajun Xu, Bin Xu, Kefu Zhu

1. Background

The simple bayesian model for a dataset (X,Y) is the fundamental theory for this project. The one-layer model is effective in some inference tasks when the expert's knowledge contribute significantly to the modeling procedure. However, the data in real life usually have their internal structures. It might be too easy for a one-layer model to approximate the distribution of data. It is intuitive to apply a hierarchical model to the data when they can be naturally grouped.

National Basketball Association (NBA) is a professional league which has detailed statistics for every player. Each player belongs to a certain team and the team is either in Eastern or Western Division, which makes up the entire league. Due to the existence of trading system, the managers of each team would like to know if a player can improve his performance by data analysis. It is also interesting to analyze the the performances of all players in each team or even in the whole league.

2. Data Preparation

The data are obtained from 'NBAPlayerStatistics0910' dataset of 'SportsAnalytics' library in R[1]. The dataset contains records of 25 variables for 441 different NBA players. We will use Team, Division, League as three stages

[1]Data retrived from: https://www.rdocumentation.org/packages/SportsAnalytics/versions/0.2/topics/NBAPlayerStatistics0910

in our project. The team category is in the dataset while division category is collected from external source. We also regenerate the response variable Points Per Game (PPG), which is calculated from dividing the value of total points (TotalPoints) by the number of games played (GamesPlayed) in the dataset.

3. Model Construction

Since there are three stages in our model, we need to propose models for the distributions in each stage. We made a tentative plot (Figure 1) to check the distribution of PPG in league-level and make the assumption that the data may follow gamma or beta distribution. Assume the parameters in this level are $\boldsymbol{\theta}$, and thus the division-level parameters are θ_E and θ_W . The parameters in the lowest level are denoted as θ_i for team i.Therefore, we would try two kinds of likelihood distribution. We will also make assumptions that the variances of each distribution in each layer are the same. For the priors, we will primarily try different non-informative or conjugate prior distributions**.**

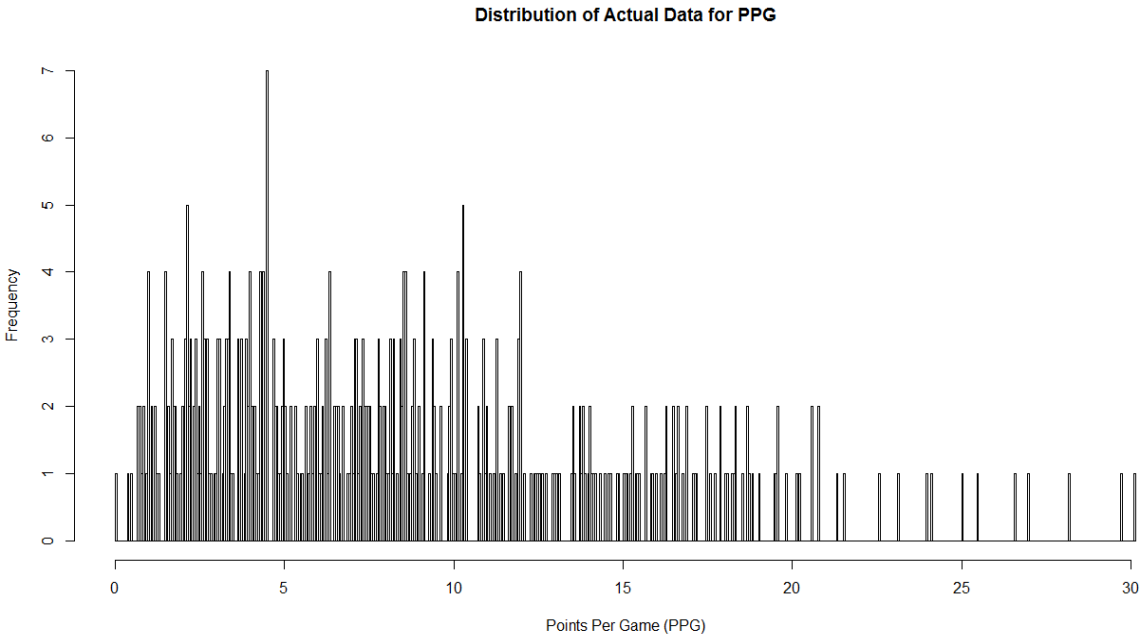

Figure 1. Distribution of Actual Data for PPG