Bayesian Statistics Capstone Project

Francis Lai, December 3, 2018

Executive Summary

An NHL player who scored on average 1 point per game playing all 82 games in a season has over 77% probability of scoring 82 or more points the next season. The Poisson regression model from which this posterior probability was obtained included games played (GP) and points per game (PPG) played as covariates - the latter a much more influential effect than the former on points scored in the next season.

1. Introduction

Every year, millions of fantasy hockey (FH) participants predict individual National Hockey League (NHL) player performance based on simple, easily accessible information. In many FH leagues, the winner is determined by drafting a pool of NHL players with the highest number of points (sum of goals and assists) scored in aggregate during the regular season.

Nowadays an NHL player averaging 1 PPG is considered a superstar - the type of players FH participants must identify to maximize their chances of winning. However, most casual participants rely on "gut feeling" to pick their players. This project attempts to provide a more objective answer to the question, "How likely is a player who played in all 82 games and scored 1 PPG on average in the same season to perform at least as well next season?" using methods covered in the "Bayesian Statistics: Techniques and Models" course.

2. Data

Freely distributed NHL data were downloaded from http://hockeyabstract.com/home, which are no longer available at the time of this project's start. The latest complete data previously obtained were those from the 2015-16 season (a season starts in the fall and ends in the following year). As a result, we used player points in the 2015-16 season as the outcome variable in our models. Being mindful of this course's focus, we used only 2 predictor variables chosen for their high predictive power as traditionally perceived by FH leagues and their being commonly available in the public domain.

Since FH predictions typically need to be made **before** the hockey season begins, the predictor variables can only be obtained from the prior season. As a result, we used the following 2 official player statistics during 2014-15 to predict player points scored in the 2015-16 season.

Predictor Variables:

- **GP**: *Games Played* during the 2014-15 season - integer. Durable players are assumed to be more likely to play in many games in the following season, therefore increasing their chances to score more points.
- **PPG**: *Points Per Game* played [(goals + assists)/games played] during the 2014-15 season - numeric. A measure of productiviy from the previous season, taking into account the number of games a player actually played, is often a performance predictor used by FH enthusiasts.

Outcome Variable:

- **POINTS16**: *Points* scored (goals + assists) during the 2015-16 season - integer. By far the most common metric used to determine the winner in FH.

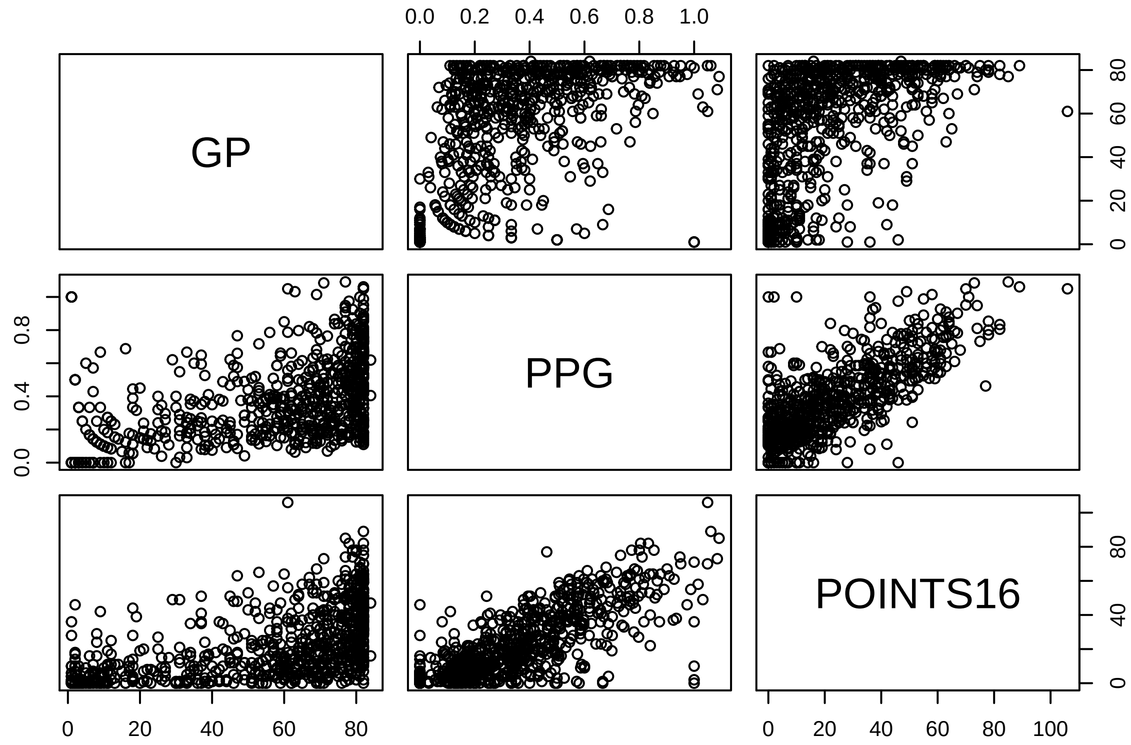
Not all players exist in both the 2014-15 and 2015-16 datasets, so after merging it was necessary to remove players with missing data in any of the above predictor and outcome variables in the final dataset for analysis, which contains 715 players with complete data.

2.1 Preliminary Data Exploration

Here is the first row of the dataset:

```
##     GP       PPG POINTS16
## 715 53 0.3584906        0
```

Here are the scatterplots of the variables:



As expected, GP and PPG both seem to be positively correlated with points scored in the next season.

3. Model

Let y_i = points scored for the i^{th} player during the 2015-16 season, independently distributed under the following conditional distribution:

$$y_i|\lambda_i \sim Poisson(\lambda_i), \quad i = 1, ..., n$$

As count data, points scored obviously cannot be negative, so we will use an appropriate Poisson regression model in which we will model the log of the rate λ_i, to avoid the few cases in which a player scoring 0 points in a season. Model 1 will include GP and PPG as the covariates:

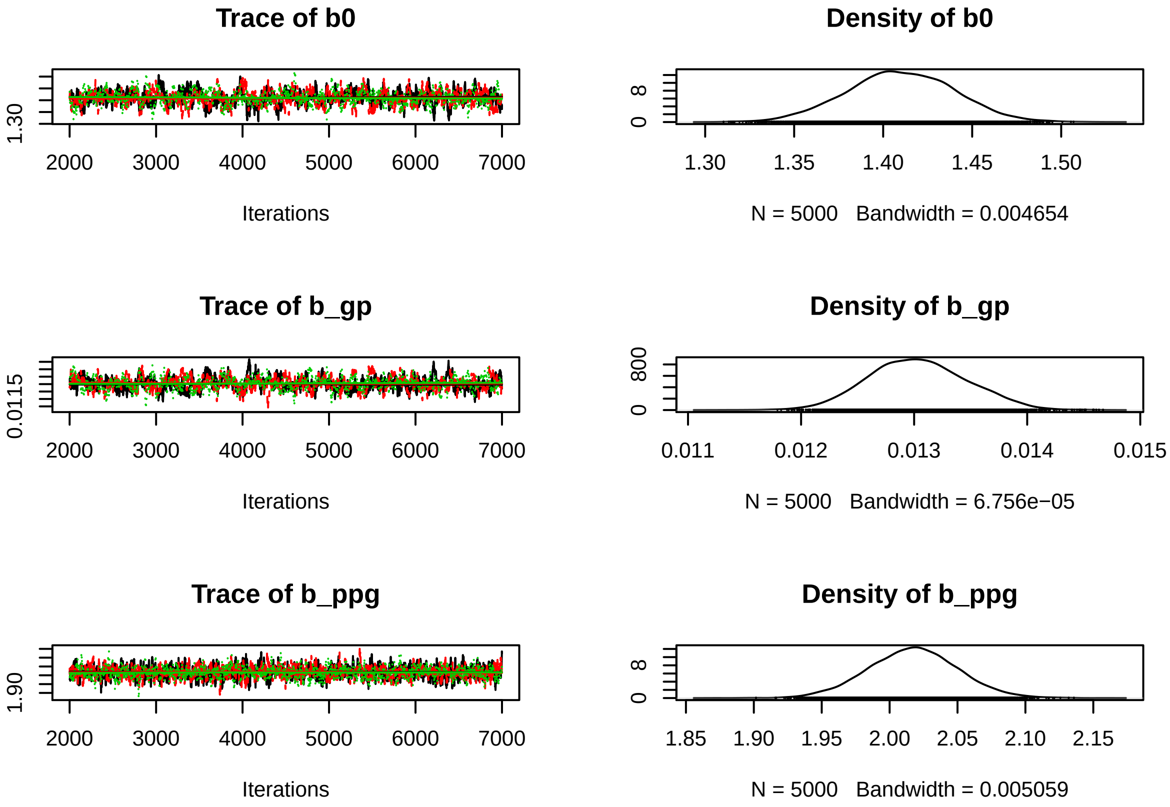
$$log(\lambda_i) = log(E(y_i)) = \beta_0 + \beta_1 x_{GP,i}^2 + \beta_2 x_{PPG,i},\ i = 1, ..., n$$

Non-informative priors were used since we had no prior knowledge of the effects of GP and PPG on points scored in the next season. The prior distribution of the $\beta_j\ 's$ are:

$$\beta_j \sim N(0, 10^{-4}),\ j = 0, 1, 2$$

3.1 Model Diagnostics

Three Markov chains of 5000 iterations each were run. The first 1000 burn-in iterations of the estimates were discarded in each. No serious convergence problems can be found in the trace plots for Model 1, indicating the Markov chain has mostly converged for each parameter.



Not shown are the following diagnostic results. Gelman-Rubin test estimates for each parameter is close to 1, indicating that the 3 chains coverge to the same posterior distribution. Autocorrelations are low in general, especially for PPG's parameter estimates. Effective sample sizes are decent, and residuals are unremarkable.

4. Results

Parameter estimates for Model 1 are as follows:

```
##
## Iterations = 2001:7000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean        SD  Naive SE Time-series SE
## b0    1.40984 0.0300436 2.453e-04      1.475e-03
## b_gp  0.01303 0.0004361 3.561e-06      2.241e-05
## b_ppg 2.01722 0.0329661 2.692e-04      1.025e-03
##
## 2. Quantiles for each variable:
##
##          2.5%     25%     50%     75%  97.5%
## b0    1.35052 1.38961 1.40964 1.43073 1.4682
## b_gp  0.01222 0.01273 0.01302 0.01333 0.0139
## b_ppg 1.95191 1.99510 2.01750 2.03886 2.0821
```

4.1 Model Comparison

Even though the GP parameter estimate's 95% credible interval excludes 0 (0.012, 0.013), its magnitude relative to PPG's is much smaller, so we created Model 2 without GP. Similar diagnostics results were obtained. In the end, Model 1 is superior since it has a lower Deviation Information Criterion (DIC) than Model 2 (8691 vs 9724), so we will continue with this model.

4.2 Prediction

To answer our original question, we need to first create a vector $x = [82, 1]')$ to represent a point-per-game-player in the 2014-15 season who also played in all 82 games, along with Model 1's posterior samples of the $\hat{\beta}_j$ *'s* to get samples for $log(\hat{\lambda}_i)$. After applying exponentiation, we can obtain the posterior predictive distribution of the points scored by such a player ($\hat{y_i}$) by drawing random samples using our $\hat{\lambda}_i$ as follows:

```
set.seed(10)
y <- rpois(n=length(loglam), lambda = exp(loglam))
mean(y > 82)
```

```
## [1] 0.7717333
```

5. Conclusions

Based on our model, there is a 77.2% posterior probability that a player who played every game and scored on average 1 point per game in the 2014-15 season will be able to maintain this scoring pace by scoring at least 82 points in the following season. This is a much better than a purely random *a priori* guess and allows one to pick players using the GP and PPG criteria more confidently.

Even though the original model with GP is a better model than the one without based on DIC comparison, the GP parameter estimate is very small. Perhaps this is an indication that GP is a weak indirect predictor for points scored through its perceived association with player durability. It is generally very difficult to predict injuries as they could happen to any player in any number of ways.

The model also suffered from being based on outdated data. NHL rules and regulations change every year and tend to have profound impact on player performance. For instance, last season's addition of the expansion franchise in Las Vegas diluted talent league wide.