# Linear Regression on NCAA Women's Lacrosse Stats by Bayesian Model Selection and Averaging

**Stella I. Ndahiro**
Department of Applied Maths and Stats
Johns Hopkins University
Baltimore, MD 21218
sinezan1@jh.edu

## Abstract

After a full semester of learning Bayesian statistics, my goal is to use Bayesian linear regression to make a predictive model for the winning percentage of an NCAA Women's Lacrosse DI team given various game stats like assists, shots, turnovers, etc. Concurrently, I will be investigating whether or not the Bayesian approach to linear regression out-competes the frequentists' Ordinary Least Squared regression (OLS) by checking which of the two minimizes the Mean Squared Error (MSE).

## 1 Literature Review

As of now, no other studies have investigated Women's Lacrosse. Lacrosse on its own is a newly emerging sport that hasn't gotten that much attention. I found similar studies in other very well-developed sports like basketball and soccer where linear regression is used to predict the performance of a team. As an example, one study investigated the different factors that influence the winning percentage of an NBA basketball team using multiple linear regression Yao et al. (2018). The results showed that the three-point percentage, turnovers, and points per game are critical in offensive efficiency. My analysis will take a similar approach but implore Bayesian model selection in determining the most significant stats.

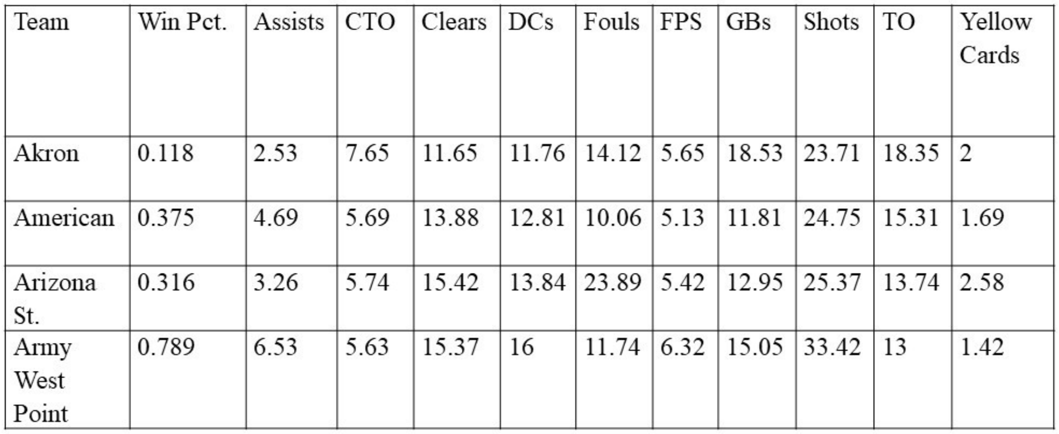
## 2 Proposed Methods

In total, there are 121 data points with 10 regressors that I will be using. These observations are from the NCAA DI Women's Lacrosse 2022-2023 season. Each data point represents a team with its winning percentage and average stats (per game) for the entire year.

**What would a Frequentist do?** One of the most popular methods of implementing linear regression is to calculate the OLS, which involves calculating the coefficients of the regressors that minimize the sum of the squared residues. To determine which coefficients are significant or not, a forward or backward selection is done.

**Bayesian Analysis** Unlike the frequentist technique that aims at minimizing the sum of squared residue, Bayesian analysis calculates the probability of observing data given a certain combination of regressors (a model). The coefficients are calculated by sampling from the marginal posterior distribution of all of the models. When we have a lot of regressors like in our case, computation techniques like Gibbs sampling and Monte Carlo simulations are used to estimate the probability distribution of models and coefficients of regressors.

NeurIPS 2023 Generative AI and Biology Workshop.



| Team | Win Pct. | Assists | CTO | Clears | DCs | Fouls | FPS | GBs | Shots | TO | Yellow Cards |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Akron | 0.118 | 2.53 | 7.65 | 11.65 | 11.76 | 14.12 | 5.65 | 18.53 | 23.71 | 18.35 | 2 |
| American | 0.375 | 4.69 | 5.69 | 13.88 | 12.81 | 10.06 | 5.13 | 11.81 | 24.75 | 15.31 | 1.69 |
| Arizona St. | 0.316 | 3.26 | 5.74 | 15.42 | 13.84 | 23.89 | 5.42 | 12.95 | 25.37 | 13.74 | 2.58 |
| Army West Point | 0.789 | 6.53 | 5.63 | 15.37 | 16 | 11.74 | 6.32 | 15.05 | 33.42 | 13 | 1.42 |

Figure 1: Summary of stats for four teams that participated in the NCAA Women's Lacrosse DI season 2022-2023.

In total, we have $2^{10}$ models to choose from given that each regressor can either be turned on or off. We will be using a G-prior that is equivalent to the number of samples ( $g$ = 121) and a uniform prior distribution for each model ( $p(z) = 1$).

## 3 Data Analysis and Result Interpretation

### 3.1 Data Extraction

The NCAA website provides all stats for each team that participates in the league. As previously mentioned, there are 121 teams. Figure 1 shows the summary of four teams on per game basis[1]:

The data was scrapped from the NCAA websiteNCA ([n. d.]). Python libraries like pandas and numpy were used to make the data more visualizable and easily manipulatable by saving it as a csv file.

### 3.2 Exploratory Data Analysis

The first step before carrying out linear regression is to do exploratory data analysis through pair-wise plots. Visualizing the data through histograms and scatter plots shows the different correlations between regressors and the predicted variable. This also gives us an expectation of what the coefficients should look like, if there is a positive or negative correlation. For example, from Figure 2, we can see that the winning percentage has a positive correlation with assists and shots.

### 3.3 Frequentists' OLS

The data was first standardized by subtracting the mean and dividing it by the standard deviation. The following are the coefficients of the regressors and Mean Squared Error of the OLS regression model:

Though the model was able to fit well the data, some coefficients don't make sense. For example, we expect to see a strong positive correlation between winning percentage and assists but the coefficient from the model is negative. The causes of such deviations will be discussed later on in the conclusion section.

### 3.4 Bayesian Model Selection and Averaging

A 10000 iteration of Gibbs sampling was run to estimate the marginal posterior distribution of the models. The coefficients of the regressors are derived from averaging Monte Carlo samples from the marginal posterior probability distribution of the models. Unlike the frequentists' OLS, the signs of

[1] CTO stands for Caused Turnovers, DCs is for Draw Controls, FPS is Free Position Shots, GBs is Ground Balls and TO stands for Turnovers

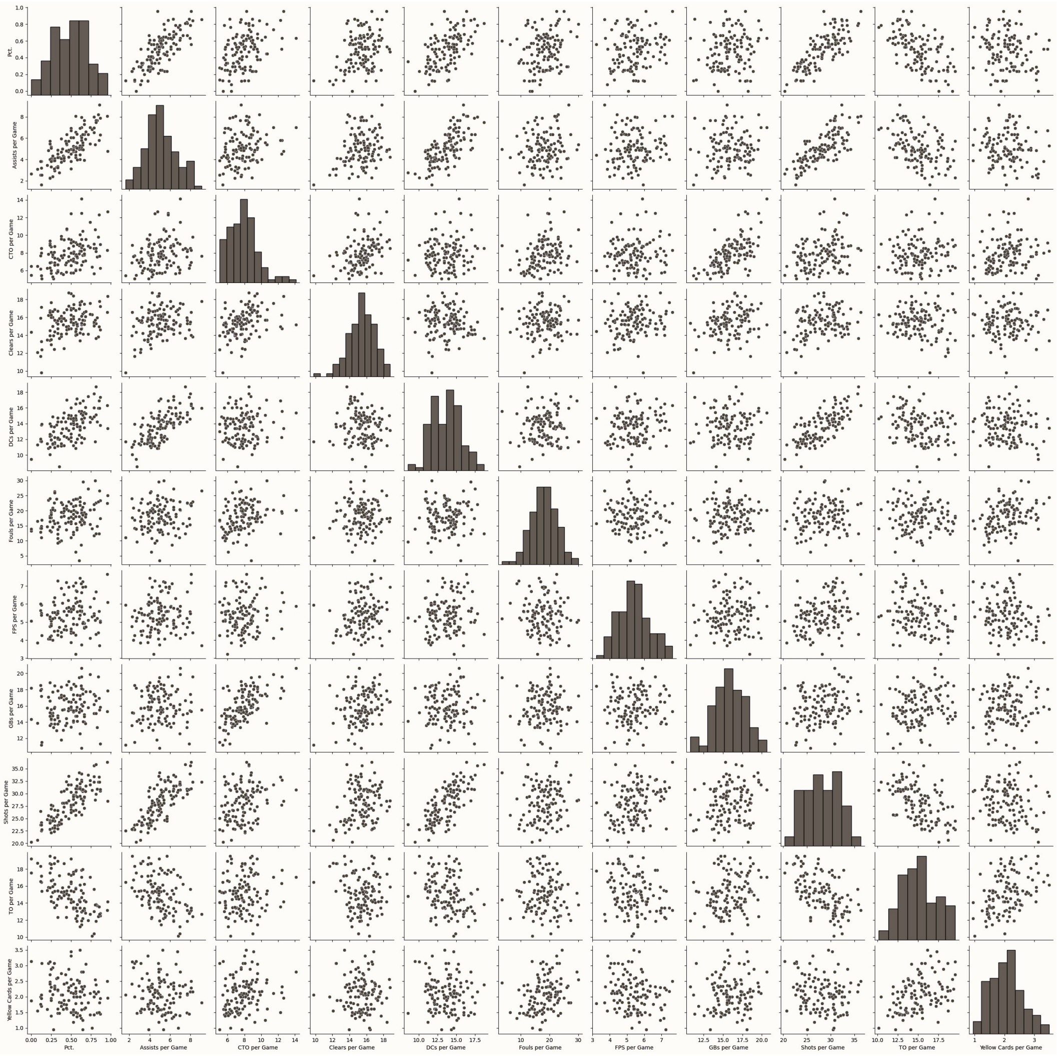


Figure 2: Pair-wise plots of the regressors and predicted variable. The first row shows the correlation of each regressor (stats) to the predicted variable (winning percentage). We can clearly see that there is a positive correlation with assists and shots, and a negative correlation with turnovers.

Table 1: Coefficients of regressors by using Ordinary Least Squares linear regression. The Mean Squared Error is 0.722

| Regressor | Coefficient |
|---|---|
| Assists | -0.161 |
| CTO | 0.358 |
| Clears | -0.11 |
| DCs | 0.296 |
| Fouls | 0.202 |
| FPS | -0.116 |
| GBs | -0.199 |
| Shots | 0.767 |
| TO | -0.374 |
| Yellow Cards | 0.33 |
| **Mean Squared Error** | 0.722 |

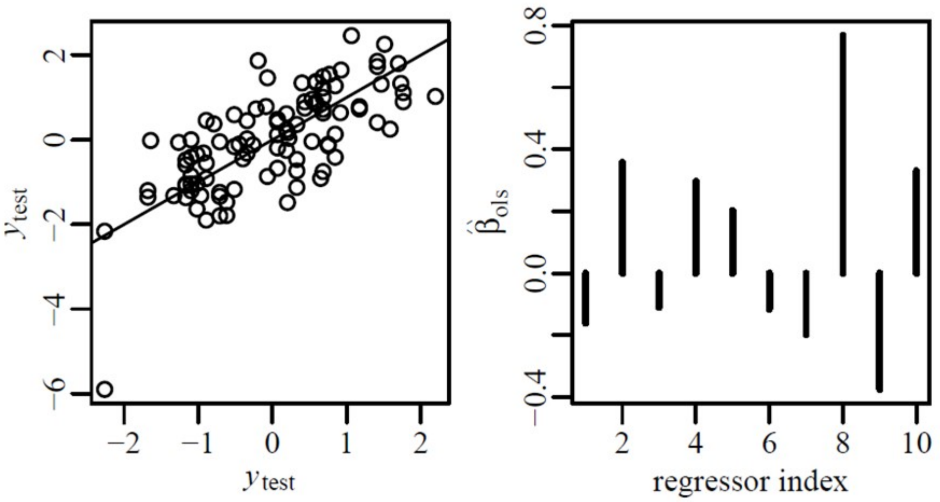


Figure 3: The left plot is a fit of the testing data using OLS linear regression. On the right are the OLS coefficients. Calculations, numbers and figures were generated by using Peter Hoff's Bayesian Statistics book Hoff (2009).

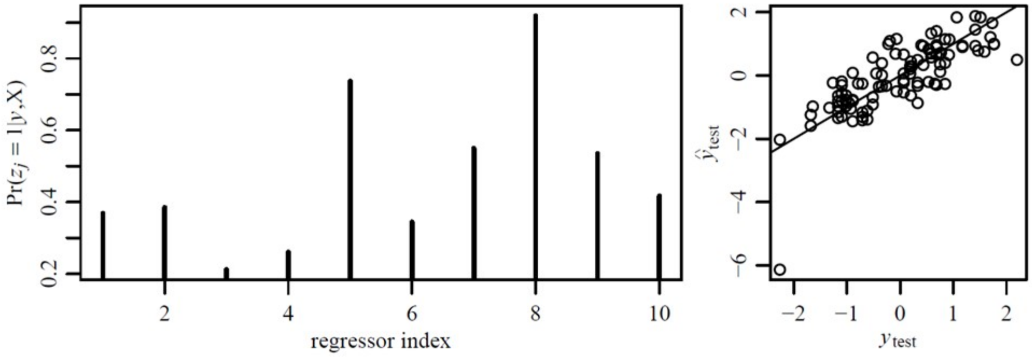


Figure 4: On the left are the probabilities of the regressor coefficients being non-zero. The right plot is a fit of the testing data using Bayesian linear regression.

the coefficient make sense for the positively correlated regressors like assists and shots. Moreover, the Mean Squared Error is as low as 0.478. Table 2 summarizes well the results.

### 3.5 MCMC Diagnostics

**Trace Plots and Burnout Period** The Gibbs sampling Markov Chain achieved stability fast. Figure 4 is a trace plot of the probabilities of the regressors coefficients being non-zero. One can clearly see that the burnout period ended after about 2000 iterations.

**Effective Sample Size** The effective sample size for the 10 regressors rounded off to the nearest tenth is 7217, 7343, 10000, 8151, 8211, 8818, 6923, 6400, and 6208. The effective sample size is high which means our Markov Chain is a good estimate of the marginal posterior distribution of the models. Moreover, we can conclude that the autocorrelation wasn't that high because the effective sample size is as big as the number of Gibbs sampling iterations.

Table 2: Coefficients of regressors by using Bayesian model selection and averaging neglecting the first 2000 iteration (burnout period). The Mean Squared Error is 0.478

| Regressor | Coefficient |
|---|---|
| Assists | 0.115 |
| CTO | 0.09 |
| Clears | -0.008 |
| DCs | 0.065 |
| Fouls | 0.155 |
| FPS | -0.052 |
| GBs | -0.078 |
| Shots | 0.819 |
| TO | -0.148 |
| Yellow Cards | 0.085 |
| **Mean Squared Error** | 0.478 |

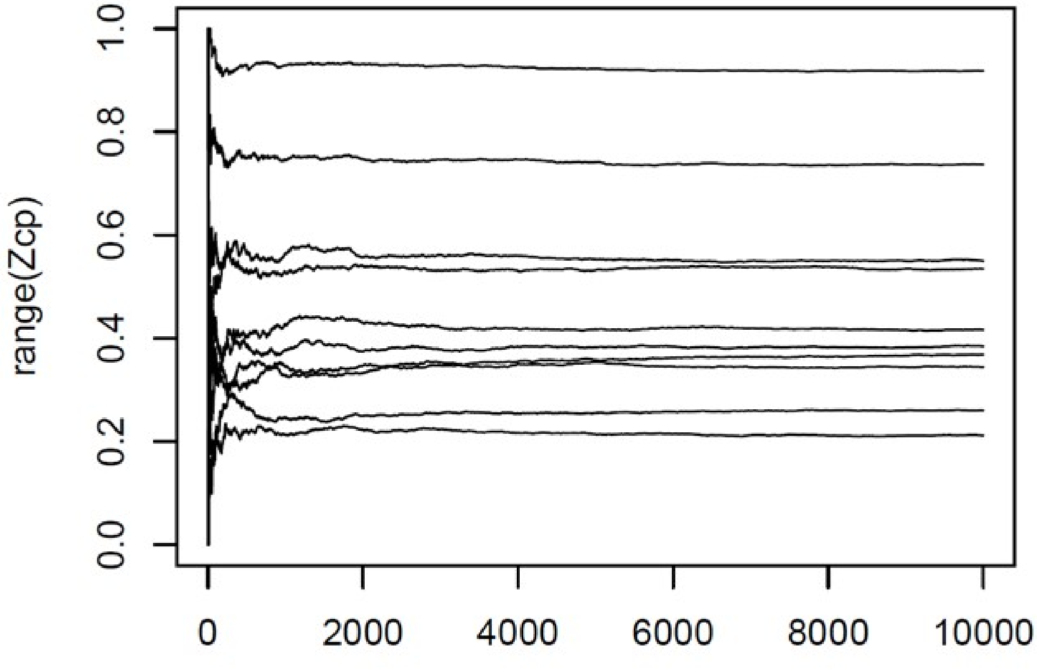


Figure 5: Posterior probability of a regressor coefficient being non-zero throughout the 10000 iteration of the Gibbs sampling. The Markov Chain converges approximately after 2000 iterations

**Thining** The sample space ($2^{10}$ models) was too large for thinning to be useful especially when there isn't that much autocorrelation between the Markov Chain.

## 4 Conclusion

Overall, the Bayesian Model averaging did better than the frequentists' OLS. The Mean Squared Error is only 0.478 for the Bayesian model regression compared to 0.722 for OLS. Therefore, Bayesian model averaging is superior in terms of predicting the winning percentage.

Moreover, it has to be noted that OLS linear regression model failed to give coefficients that are interpretable despite providing a good fit to the data. For example, according to pair-wise plots, we expect to see a positive correlation between assists and winning percentage but observed the opposite for the OLS model. These discrepancies can be explained by the multicollinearity of regressors, where one variable can be written as a linear combination of others. Therefore, Bayesian model selection might be a good option when dealing with data that have multicollinearity.

## References

[n. d.]. NCAA College Women's Lacrosse DI Stats.

Peter D Hoff. 2009. *A first course in Bayesian statistical methods*. Vol. 580. Springer.

Alan Yao, Mason Chen, Sean Yao, and Charles Chen. 2018. Applying Statistical Modeling to Predict Basketball Winning Percentage. In *Fuzzy Systems and Data Mining IV*. IOS Press, 44–52.