Predicting NBA Players Salaries with Bayesian Linear Regression

https://www.flickr.com/photos/rmtip21/9163118621

Qihang Zhang

Project Goals

1. Get hands on experience with Markov Chain Monte Carlo methods
2. Understand the difference between Bayesian Approach vs. Frequentist Approach
3. Test on No U-Turns Sampler (NUTS) method (Hamiltonian Monte Carlo method)

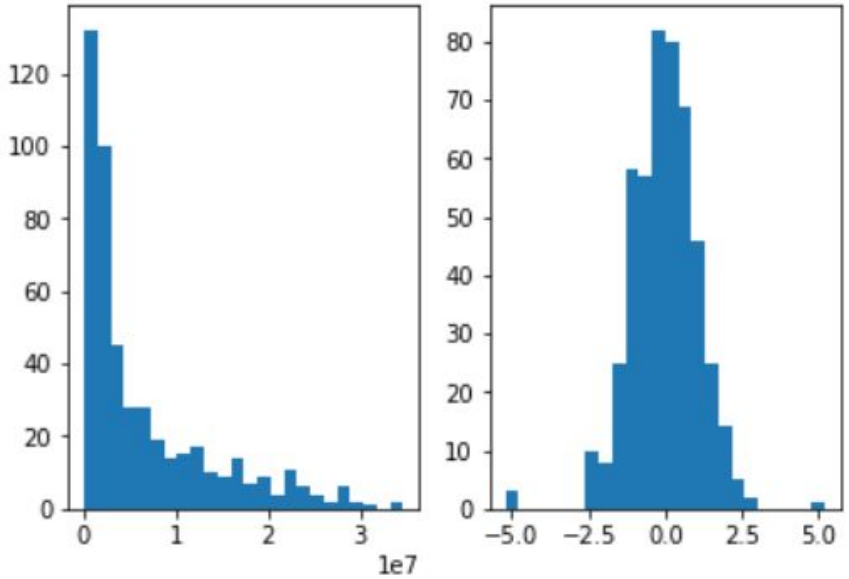
Qihang Zhang

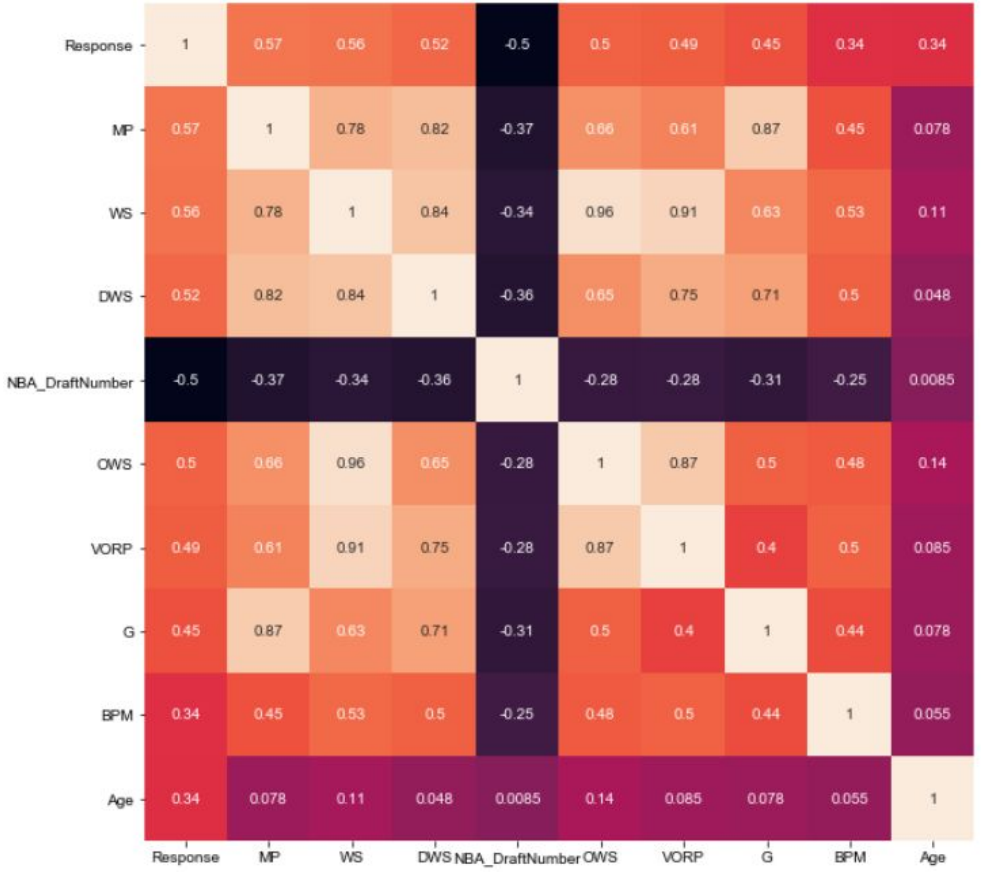
Data Preview & Preprocessing

- 485 NBA Players
- 28 Features in total
  - Independent Variable (Salary)
  - Personal Information (Name, Team, Age, Nationality, Draft Number)
  - Game Statistics (Games played, 3-point attempt rate, value over replaced player, etc.)
- Imputed missing values with median (3 Columns with 2 missing values each)
- Transformed the response to a normal distribution

https://www.kaggle.com/aishjun/nba-salaries-prediction-in-20172018-season

Data Visualization





Qihang Zhang

Feature Selection

Selected Features (VIF score < 10):

- Age
- VORP (Value over replaced player)
- BPM (Box Plus/Minus)
- Draft Number
- MP (Minutes played)

Problem Setup

$$y|\beta,\sigma^2 \sim N(\beta^T X, \sigma^2 I)$$

$$P(\beta|\sigma^2) \propto 1$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$\epsilon \sim N(0,\sigma^2)$$

$$P(y|X,\beta,\sigma^2) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(y-\beta^T X)^2}{2\sigma^2}$$

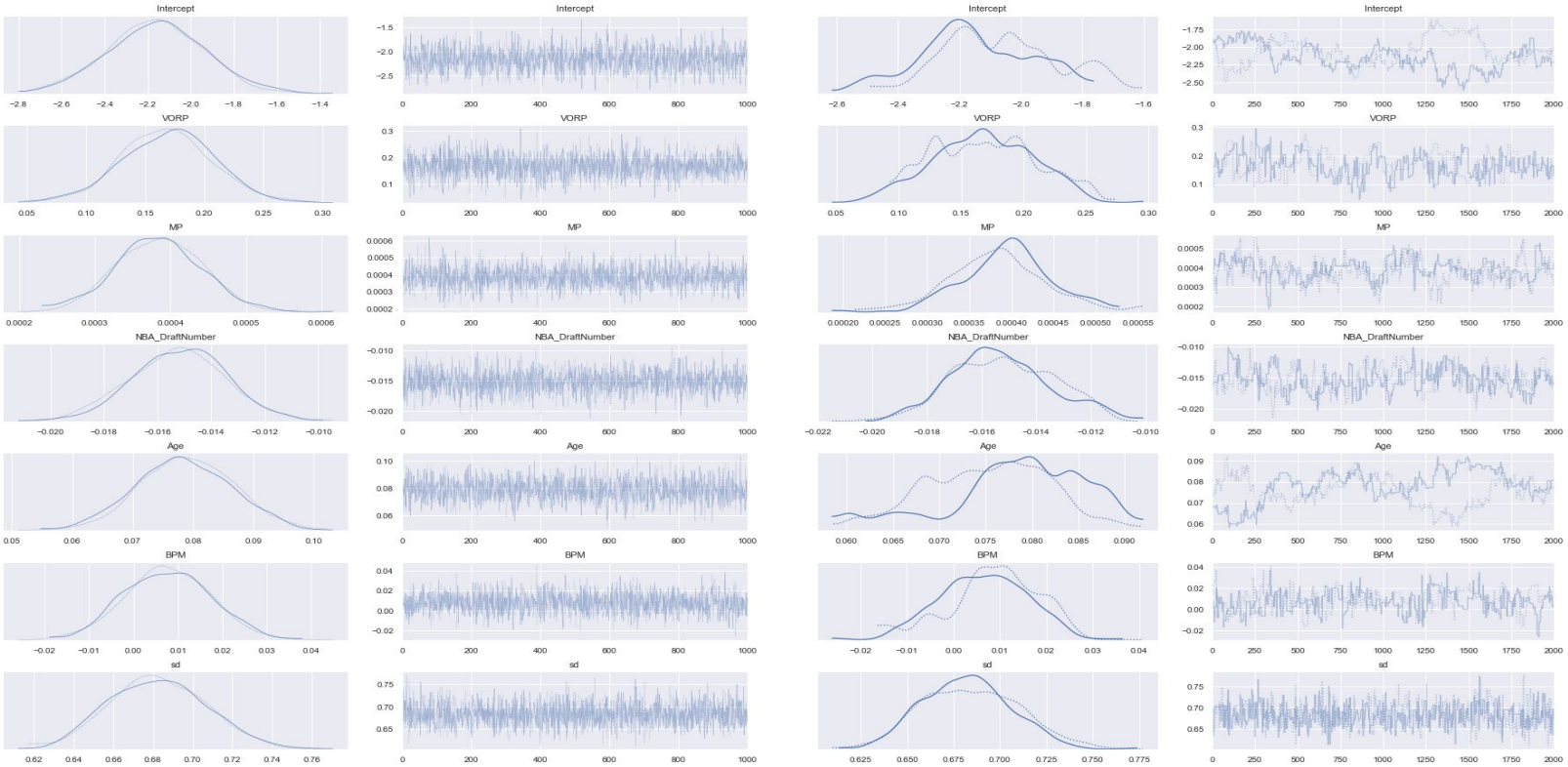
Two Approaches

- Bayesian
- Frequentist

Two Sampling Methods

- Metropolis-Hastings
- No U-Turns (NUTS)

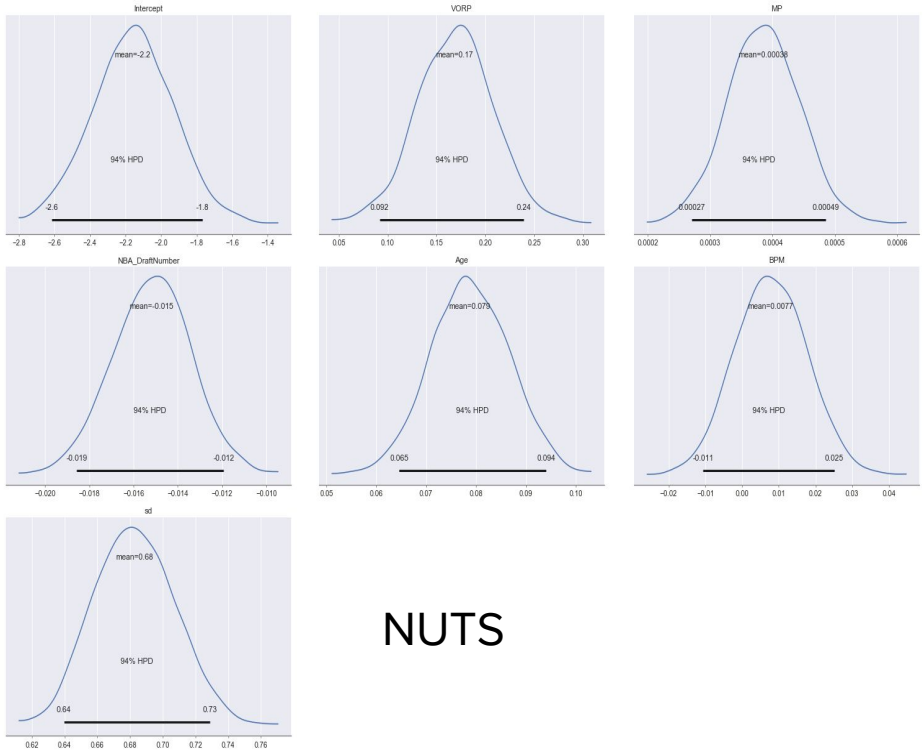
Result & Discussion



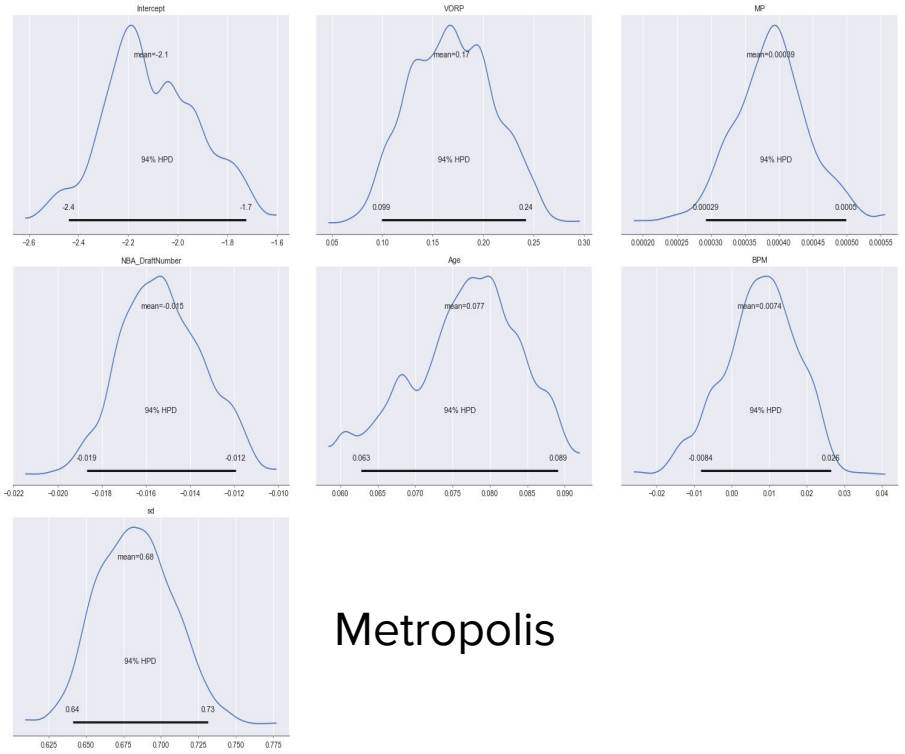
NUTS

Metropolis-Hastings

Result & Discussion



NUTS



Metropolis

Result & Discussion

	Mean Abs. Error	Root Mean Square Error
Frequentist	0.611399	0.930005
NUTS_1000	1.134446	1.494323
NUTS_2000	1.134190	1.494059
Metro_2000	1.133371	1.493248

	NUTS_1000	NUTS_2000	Metropolis_2000	Frequentist
Salary	$11984473	$11984473	$11939148	$11964463

Result & Discussion

Bayesian Approach:

- More interpretable (Parameters are estimated as probability distributions)
- Can incorporate with prior information (more suitable for industrial applications)

Future Steps

- Reduce the training size and set appropriate priors to see the result

Thanks!