Bayesian Analysis of NBA salaries using Generalized Linear Models

Pranay Raj Kapoor

December 21, 2024

1 Introduction

The NBA is the top league for men's basketball in the US. Particularly, the average salaries for NBA players for the 2023-2024 season was upwards of 12 million USD[1]. Consequently, given the size of the national basketball league in the US, there has been a growing interest in using advanced analytics and statistics for NBA team and player performance analysis. Bayesian statistics and analysis have also been extensively employed for a variety of analysis, including studying home team advantage in games (Higgs & Stavness (2021)) and player performance (Deshpande & Jensen (2016)).

In this project, we employ Bayesian Generalized Linear Models (GLMs) and Bayesian Heirachichal Models for the purposes of predicting NBA salaries using a variety of covariates including mean statistics of player performance over the past season such as the including mean minutes per game played, mean points scored per minute played and so forth. We compare and contrast the results of three models: 1) A Bayesian Fixed Effects Linear Regression with weakly informative priors, 2) A Bayesian Random Effects Linear Regression with weakly informative priors and 3) A Bayesian Hierarchichal Model that accounts for predictive information in covariates such as age and team.

The advantage of using Generalized Linear Models over regular regression methods such as OLS is that with Generalized Linear Models, we are able to model the outcome variables based on their specific distributions. For instance, we can model salary data to follow a Gamma Distribution to account for positive-skew as well as the log-salary data to follow a normal distribution. This coupled with informative priors on other covariates based on our understanding of NBA salary structures in our Bayesian analysis will allow us to explain higher variability in our salary data and better predict salaries.

2 The Data

We obtain data on NBA player salaries from Kaggle. Particularly, we obtain salary data for NBA players across seasons from 2000-2019[2] as well as player performance across seasons for the preceding years[3] from the Kaggle sources outlined in the footnote below.

2.1 NBA Salary Data

For the salary data, there is data for 37,000 players. However, the salary data is missing for 28,000 rows. Given the large number of rows of missing data, we drop the rows with missing data as imputing them using missing data methods would not be a better choice. Doing, so we have 9347 rows of data with roughly equal number of rows for each year. The distribution of salaries by year is shown in the picture below.

[1] https://www.statista.com/statistics/1120680/annual-salaries-nba-wnba

[2] https://www.kaggle.com/datasets/hrfang1995/nba-salaries-by-players-of-season-2000-to-2019

[3] https://www.kaggle.com/datasets/jacobbaruch/basketball-players-stats-per-season-49-leagues

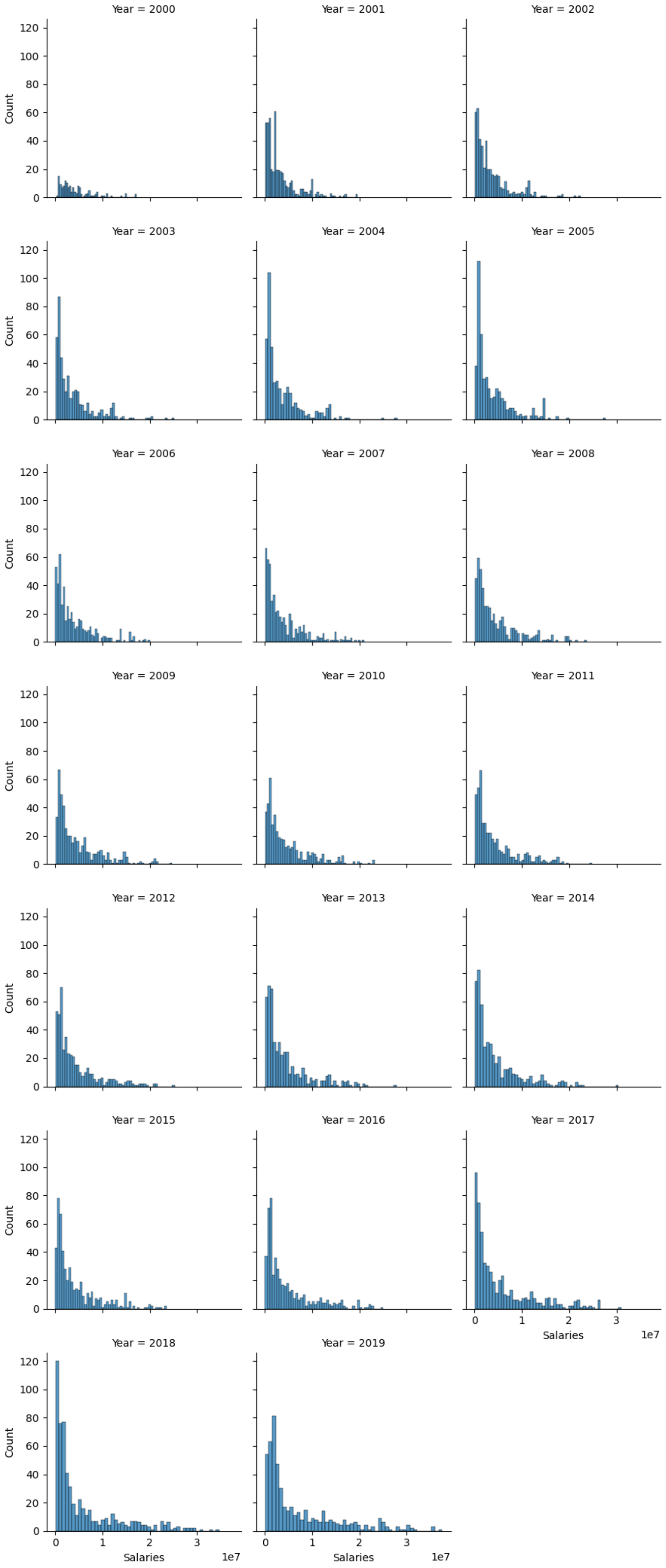

Year = 2000
Year = 2001
Year = 2002
Year = 2003
Year = 2004
Year = 2005
Year = 2006
Year = 2007
Year = 2008
Year = 2009
Year = 2010
Year = 2011
Year = 2012
Year = 2013
Year = 2014
Year = 2015
Year = 2016
Year = 2017
Year = 2018
Year = 2019
Count
0
20
40
60
80
100
120
0
1
2
3
Salaries
1e7

Table 1: Summary Statistics by Year

Year	min	max	mean	count
2000	301000.00	17142000.00	4232652.78	144
2001	40000.00	19610000.00	3304928.99	455
2002	26190.00	22400000.00	3463517.48	450
2003	29832.00	25200000.00	3702467.96	451
2004	32375.00	28000000.00	3684159.76	454
2005	34118.00	27696430.00	3773687.55	470
2006	24315.00	20000000.00	3961305.05	479
2007	20133.00	21000000.00	3917887.23	495
2008	42203.00	23750000.00	4400600.82	469
2009	46812.00	24751934.00	4692347.60	460
2010	26917.00	23239562.00	4633963.24	456
2011	55718.00	24806250.00	4413547.23	459
2012	42009.00	25244493.00	4354507.71	463
2013	25073.00	27849000.00	4274055.67	494
2014	28834.00	30453000.00	4322503.51	490
2015	20000.00	23500000.00	4293242.99	513
2016	9266.00	25000000.00	4708117.08	500
2017	24022.00	30963450.00	5442310.90	545
2018	46079.00	34682550.00	5729243.15	586
2019	20000.00	37457154.00	7010579.83	513

Based on the distribution of salaries shown above, the salary data follows a Gamma distribution given the positive-skew we observe. This makes the salary data suitable to be modeled by a Generalized Linear Model with either a Gamma Distribution for raw salaries or a Normal Distribution for the salary on the log scale. We can also observe the distribution of salaries by year in the table below. Annual salaries range from about about 30 thousand USD to upwards 30 million USD in the data with mean salaries ranging from 3.5 to 7 million USD by year. Given the low count for the year 2000 and the different skew of the salary data, we would exclude that year from our analysis given the salary data is likely not complete. Thus, we consider the years 2001-2019 for our analysis.

2.2 NBA Statistics Dataset

We now consider the NBA statistics dataset and clean it. The dataset contains various metrics for player performance and we create normalized variables for these before running our Bayesian Generalized Linear Model. The variables include:

- FG_{pct} : Percentage of Field Goals made
- $Three_{pct}$: Percentage of Three Pointers made
- FT_{pct} : Percentage of Free Throws
- MIN_{pergp}: Average minutes per game played
- PF_{permin}: Personal Fouls per minute
- TOV_{permin}: Turnover per minute
- REB_{permin}: Rebounds per minute
- AST_{permin}: Assists per minute
- BLK_{permin}: Blocks per minute
- PTS_{permin}: Points per minute

The summary statistics for these normalized variables are shown in the table above. The distribuition for all the numerical variables appears logical. Particularly, all our normalized covariates have values less than 1.0 suggesting that the data doesn't have any anomalies. The games played and minutes played also seem consistent with what we would expect from the data. In terms of the dis-balance in the count for our percent variables, this is because those players would have not made any field goals or three pointers and thus the missing percent variables will be populated with a value

	GP	MIN	FGpct	THREESpct	FTpct	Age	MINpergp	PFpermin	TOVpermin	REBpermin	ASTpermin	STLpermin	BLKpermin	PTSpermin
mean	48.67	1295.69	0.45	0.30	0.75	28.07	26.51	0.09	0.05	0.18	0.08	0.03	0.02	0.40
std	31.62	966.23	0.07	0.15	0.13	4.24	8.66	0.03	0.02	0.08	0.06	0.01	0.02	0.13
min	1.00	0.70	0.00	0.00	0.00	20.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	11.00	266.80	0.41	0.25	0.69	25.00	19.78	0.07	0.04	0.11	0.04	0.02	0.01	0.31
50%	66.00	1312.30	0.45	0.34	0.77	28.00	26.93	0.09	0.05	0.16	0.07	0.03	0.01	0.38
75%	77.00	2120.10	0.49	0.38	0.83	31.00	33.52	0.11	0.07	0.23	0.11	0.04	0.03	0.47
max	85.00	3485.00	0.83	1.00	1.00	43.00	47.60	0.40	0.26	0.49	0.36	0.26	0.13	0.98
count	7633.00	7633.00	7631.00	6789.00	7553.00	7633.00	7633.00	7633.00	7633.00	7633.00	7633.00	7633.00	7633.00	7633.00

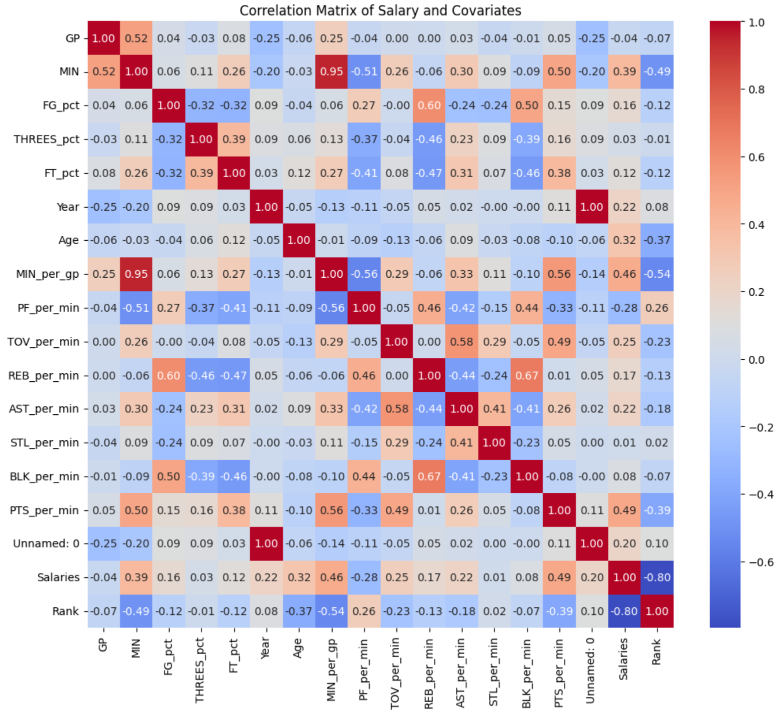
of 0 for the Field Goals percent, Three Pointers percent and Free Throws percent with 0 for the missing values. This seems a natural step given the context of the data and thus no missing data imputation methods are necessary.

Overall, among the metrics above, we would expect positive and potentially significant coefficients for all covariates with the exception of Personal Fouls per minute and Turnover per minute for which we would expect coefficients to be negative. We can test this by looking at the correlation between each covariate and the salary variable and seeing what kind of correlation we observe in the data.

2.3 Merging the Two Datasets

Given we have pre-processed relevant variables across the NBA Salaries and NBA Statistics dataset, we merge the two datasets on player name and season. As discussed earlier, we exclude the year = 2000 for our analysis. After restricting the NBA stats dataset to only the Regular Season statistics and NBA as League, we have 4658 rows. We do an inner merge with our Salaries data on 'Player Name' and 'Year'. Doing so, we get a new panel dataset with about 4000 rows of data at the Player-Year level. Overall, there are not many missing values except the percent variables created in the Statistics dataset, which we impute missing values for with 0 as discussed above. The panel also seems to be balanced with about the same number of observations per year and more or less per team with about 100 observations per team. We drop those teams with fewer than 50 observations for our analysis.

After doing the merge, we can also analyze correlation between the covariates. The heatmap below shows the correlation between the covariates. We are particularly interested in the correlation between heatmap and salary. We can note that except for Personal Fouls per minute, there is a positive correlation between all of our covariates of interest and salary. This seems logical and in line with what we were expecting particularly the most significant variables being minuter per game played and points per minute followed by age, year, assists per minute and rebounds per minute.

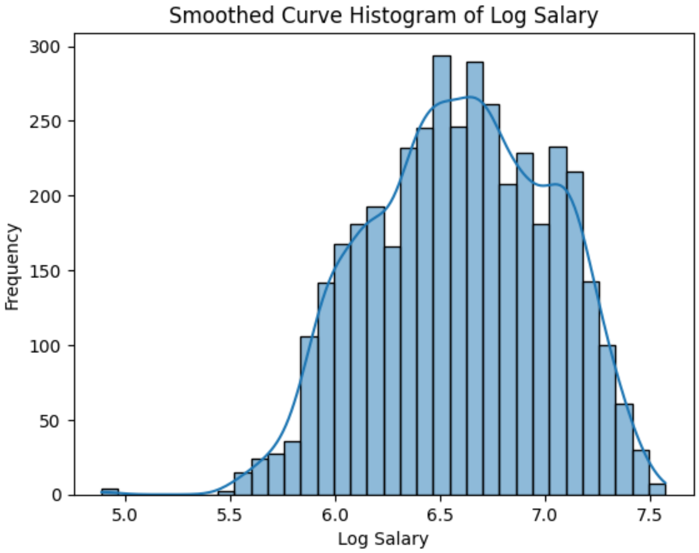


	GP	MIN	FG_pct	THREES_pct	FT_pct	Year	Age	MIN_per_gp	PF_per_min	TOV_per_min	REB_per_min	AST_per_min	STL_per_min	BLK_per_min	PTS_per_min	Unnamed: 0	Salaries	Rank
GP	1.00	0.52	0.04	-0.03	0.08	-0.25	-0.06	0.25	-0.04	0.00	0.00	0.03	-0.04	-0.01	0.05	-0.25	-0.04	-0.07
MIN	0.52	1.00	0.06	0.11	0.26	-0.20	-0.03	0.95	-0.51	0.26	-0.06	0.30	0.09	-0.09	0.50	-0.20	0.39	-0.49
FG_pct	0.04	0.06	1.00	-0.32	-0.32	0.09	-0.04	0.06	0.27	-0.00	0.60	-0.24	-0.24	0.50	0.15	0.09	0.16	-0.12
THREES_pct	-0.03	0.11	-0.32	1.00	0.39	0.09	0.06	0.13	-0.37	-0.04	-0.46	0.23	0.09	-0.39	0.16	0.09	0.03	-0.01
FT_pct	0.08	0.26	-0.32	0.39	1.00	0.03	0.12	0.27	-0.41	0.08	-0.47	0.31	0.07	-0.46	0.38	0.03	0.12	-0.12
Year	-0.25	-0.20	0.09	0.09	0.03	1.00	-0.05	-0.13	-0.11	-0.05	0.05	0.02	-0.00	-0.00	0.11	1.00	0.22	0.08
Age	-0.06	-0.03	-0.04	0.06	0.12	-0.05	1.00	-0.01	-0.09	-0.13	-0.06	0.09	-0.03	-0.08	-0.10	-0.06	0.32	-0.37
MIN_per_gp	0.25	0.95	0.06	0.13	0.27	-0.13	-0.01	1.00	-0.56	0.29	-0.06	0.33	0.11	-0.10	0.56	-0.14	0.46	-0.54
PF_per_min	-0.04	-0.51	0.27	-0.37	-0.41	-0.11	-0.09	-0.56	1.00	-0.05	0.46	-0.42	-0.15	0.44	-0.33	-0.11	-0.28	0.26
TOV_per_min	0.00	0.26	-0.00	-0.04	0.08	-0.05	-0.13	0.29	-0.05	1.00	0.00	0.58	0.29	-0.05	0.49	-0.05	0.25	-0.23
REB_per_min	0.00	-0.06	0.60	-0.46	-0.47	0.05	-0.06	-0.06	0.46	0.00	1.00	-0.44	-0.24	0.67	0.01	0.05	0.17	-0.13
AST_per_min	0.03	0.30	-0.24	0.23	0.31	0.02	0.09	0.33	-0.42	0.58	-0.44	1.00	0.41	-0.41	0.26	0.02	0.22	-0.18
STL_per_min	-0.04	0.09	-0.24	0.09	0.07	-0.00	-0.03	0.11	-0.15	0.29	-0.24	0.41	1.00	-0.23	0.05	0.00	0.01	0.02
BLK_per_min	-0.01	-0.09	0.50	-0.39	-0.46	-0.00	-0.08	-0.10	0.44	-0.05	0.67	-0.41	-0.23	1.00	-0.08	-0.00	0.08	-0.07
PTS_per_min	0.05	0.50	0.15	0.16	0.38	0.11	-0.10	0.56	-0.33	0.49	0.01	0.26	0.05	-0.08	1.00	0.11	0.49	-0.39
Unnamed: 0	-0.25	-0.20	0.09	0.09	0.03	1.00	-0.06	-0.14	-0.11	-0.05	0.05	0.02	0.00	-0.00	0.11	1.00	0.20	0.10
Salaries	-0.04	0.39	0.16	0.03	0.12	0.22	0.32	0.46	-0.28	0.25	0.17	0.22	0.01	0.08	0.49	0.20	1.00	-0.80
Rank	-0.07	-0.49	-0.12	-0.01	-0.12	0.08	-0.37	-0.54	0.26	-0.23	-0.13	-0.18	0.02	-0.07	-0.39	0.10	-0.80	1.00

3 Statistical Models

After merging the two datasets above, we decide to regress the logarithm of salary on our covariates including standardized variables for age and minutes per game played as well as the field goals percentage, three pointer percentage, and free throw percentage. Other performance metrics we include at the per minute level include turnover per minute, rebounds per minute, points per minutes, assists per minute, blocks per minute, steals per minute, and personal fouls per minute. We expect all these covariates except turnover and personal fouls to have positive coefficients. In our bayesian and frequentist models, we also consider fixed effects at the Year, Team and Player level. In addition to a bayesian fixed effects model, we consider a random effects bayesian model and a hierarchical bayesian model

Scaling the salary on log base 10 scale and scaling the parameters such as age and minutes per game makes our data suitable for an Ordinary Least Squares regression as our residuals would be normally distributed and make the distribution of residuals homoskedastic. The distribution for the salary is still slightly left-skewed but more or less follows a normal distribution making it suitable for both OLS and Normal Generalized Linear Model.



3.1 Frequentist Fixed Effects Regression

The OLS regression we consider is a fixed-effects regression that allows us to account for any variation across salaries at the team and year level. The model can be specified as follows, where y_{ij} represents the salary for each year-team combination, X_{ij} are the performance and demographic statistics discussed above and γ_i, η_j and are the team and year level fixed effects:

$$y_{ij} = \alpha + \beta \cdot \mathbf{X_{ij}} + \gamma_i + \eta_j + \epsilon_{ij} \tag{1}$$

The results of the OLS fixed effects regression are shown on the next page. Each coefficient β_l can be interpreted as that a 1 unit increase in the variable leads to a $100 * \beta_l\%$ increase in the salary levels. For instance, from the results below it is evident that assists per minute and points per minute have significant and positive effect on wages with a 1 unit increase in those variables leading to a 44 % increase and 72 % increase in salary. In addition, other performance metrics that have a positive and significant effect on wages include rebounds per minute and blocks per minute with an increase in wages by 79 % and 179 % respectively. In terms of negative effect, steals per minute has a negative and statistically significant effect with a decrease in salaries by 200 % for a 1 unit increase in steals per minute. Again, while these coefficients seem large, they are logical given an increase in steals per minute or assists per minute by a unit or so would be impractical and thus the coefficients are logical.

Other statistically significant coefficients include coefficients for years, where we are seeing an increase from the base year 2001 in salaries ranging from 15% to 35% as well as age and minutes per game played, where a 1 standard deviation increase in age or minutes played over the mean yields a

17% and 12.77% increase in salaries. Turnover and personal fouls per minute do not have a statistically significant effect on wages. We also note a counter-intuitiuve sign for the field goals percent, threes percent and free throws percent but this may be driven due to multi-collinearity with other covariates such as points scored and maybe perhaps should be excluded from analysis. In terms of team-fixed effects, the effect for any one team isn't statistically significant or major in terms of magnitude to be considered. Overall, the regression results has an $R^2 = 56.2\%$ meaning it explains 56% of the variance in the log-salary data using the covariates which is good.

Overall, the frequentist regression results are encouraging. However, we maybe able to achieve better results with a Bayesian GLM with weakly informative priors. We next run a Bayesian Linear Regression with team and year fixed effects and team and year random effects.

Dep. Variable:	log_salary	**R-squared:**	0.562
Model:	OLS	**Adj. R-squared:**	0.555
Method:	Least Squares	**F-statistic:**	80.08
Date:	Fri, 20 Dec 2024	**Prob (F-statistic):**	0.00
Time:	16:08:37	**Log-Likelihood:**	-609.32
No. Observations:	3936	**AIC:**	1345.
Df Residuals:	3873	**BIC:**	1740.
Df Model:	62		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	6.2277	0.080	78.270	0.000	6.072	6.384
C(Year)[T.2002]	0.0393	0.030	1.312	0.190	-0.019	0.098
C(Year)[T.2003]	0.0418	0.029	1.435	0.152	-0.015	0.099
C(Year)[T.2004]	0.0367	0.030	1.225	0.221	-0.022	0.095
C(Year)[T.2005]	0.0883	0.029	3.022	0.003	0.031	0.146
C(Year)[T.2006]	0.1439	0.030	4.848	0.000	0.086	0.202
C(Year)[T.2007]	0.1365	0.030	4.589	0.000	0.078	0.195
C(Year)[T.2008]	0.1517	0.029	5.194	0.000	0.094	0.209
C(Year)[T.2009]	0.1599	0.030	5.371	0.000	0.102	0.218
C(Year)[T.2010]	0.1810	0.030	6.064	0.000	0.122	0.240
C(Year)[T.2011]	0.1972	0.030	6.496	0.000	0.138	0.257
C(Year)[T.2012]	0.1834	0.031	5.839	0.000	0.122	0.245
C(Year)[T.2013]	0.1927	0.029	6.574	0.000	0.135	0.250
C(Year)[T.2014]	0.1708	0.030	5.737	0.000	0.112	0.229
C(Year)[T.2015]	0.1994	0.030	6.606	0.000	0.140	0.259
C(Year)[T.2016]	0.2569	0.031	8.419	0.000	0.197	0.317
C(Year)[T.2017]	0.3401	0.029	11.877	0.000	0.284	0.396
C(Year)[T.2018]	0.3496	0.029	12.128	0.000	0.293	0.406
C(Year)[T.2019]	0.3509	0.029	12.034	0.000	0.294	0.408
Age_std	0.1777	0.005	36.285	0.000	0.168	0.187
GP_std	-0.0114	0.006	-2.040	0.041	-0.022	-0.000
MIN_per_gp_std	0.1950	0.007	26.705	0.000	0.181	0.209
FG_pct	-0.3768	0.113	-3.325	0.001	-0.599	-0.155
THREES_pct	-0.0806	0.039	-2.090	0.037	-0.156	-0.005
FT_pct	-0.1259	0.062	-2.019	0.044	-0.248	-0.004
PF_per_min	-0.0353	0.258	-0.137	0.891	-0.540	0.470
TOV_per_min	0.6508	0.405	1.606	0.108	-0.143	1.445
REB_per_min	0.7940	0.098	8.133	0.000	0.603	0.985
AST_per_min	0.4472	0.137	3.269	0.001	0.179	0.715
STL_per_min	-2.1402	0.427	-5.009	0.000	-2.978	-1.303
BLK_per_min	1.7932	0.332	5.396	0.000	1.142	2.445
PTS_per_min	0.7205	0.061	11.720	0.000	0.600	0.841

Omnibus:	131.346	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	144.275
Skew:	-0.453	**Prob(JB):**	4.69e-32
Kurtosis:	3.239	**Cond. No.**	143.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

3.2 Bayesian Fixed Effects Regression

We now run a Bayesian Generalized Linear Model with weakly informative priors. The model is specified as below based on our summary statistics and frequentist regression results. The advantage of using Bayesian Linear Regression is that Generalized Linear Models in general allow for varied non-linear distributions for the output to be modeled while the priors when informative help provide direction on the magnitude and sign of the coefficients. We include fixed effects for team and year to account for variation in salaries at the team and year level.

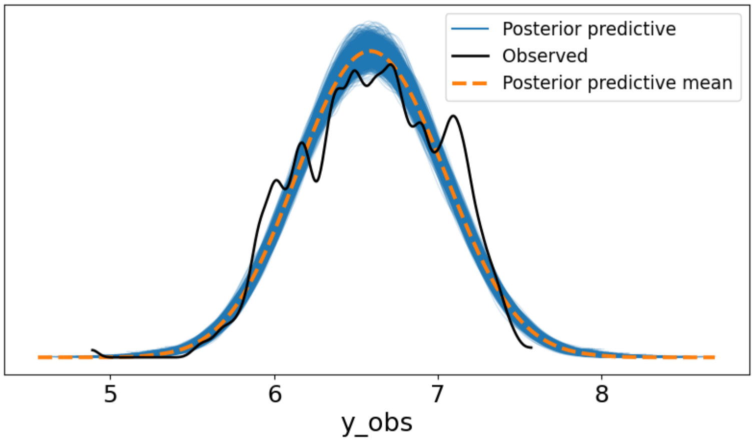
$$\begin{gathered} y_{ij} \sim N(\alpha + \beta_{ij}\mathbf{X_{ij}} + \gamma_i\mathbf{Year_{ij}} + \eta_j\mathbf{Team_{ij}} + \epsilon_{ij}, \sigma^2) \\ \sigma \sim Exp(1) \\ \beta_{ij} \sim N(0,3) \\ \gamma_i, \eta_j \sim N(0,2) \end{gathered}$$

Based on the model above, we run the Bayesian Generalized Linear Model with fixed effects for year and team . The fixed effects help account and control for any variation in salary due to the year or team and thus allow for better measures of the performance coefficients and their impact on salaries for the subsequent season.

From the results again we find that Points per minute and Assists per minute have the most significant positive impact on log(salaries) and the 95 % HDI interval is significant. In contrast, in addition to personal fouls being a large negative coefficient, we also find that salaries decrease significantly with an increase in steals per minute and blocks per minute. This appears to be counter-intuitive. This could be a result of aggressive players being penalized salaries or other positions such as shooters and forwards having more value than point guards. Again, the coefficients seem to be in line with what we had observed with the OLS fixed effects regression with an increase in points per minute by 1 unit leading to a 72 % increase in salaries, 177 % for blocks per minutes and 45 % for assists per minute and 80% for rebounds per minute. We see non-significant coefficients for personal fouls per minute and turnover suggesting these are not significant coefficients affecting wages. However, in the Bayesian setting, it would also be interesting to see a comparison with a random effects model to account for heterogeneity at the team and year level that occurs due to random shocks.

	mean	sd	hdi$_{2.5\%}$	hdi$_{97.5\%}$	mcse$_{mean}$	mcse$_{sd}$	ess$_{bulk}$	ess$_{tail}$	r$_{hat}$
Intercept	6.22	0.08	6.05	6.37	0.00	0.00	654.00	1282.00	1.00
beta$_{age}$	0.18	0.01	0.17	0.19	0.00	0.00	2539.00	1160.00	1.00
beta$_{gp}$	-0.01	0.01	-0.02	-0.00	0.00	0.00	2441.00	1511.00	1.00
beta$_{min}$	0.20	0.01	0.18	0.21	0.00	0.00	1582.00	1470.00	1.00
beta$_{fg}$	-0.37	0.11	-0.59	-0.15	0.00	0.00	1830.00	1375.00	1.00
beta$_{threes}$	-0.08	0.04	-0.15	-0.01	0.00	0.00	2313.00	1585.00	1.00
beta$_{ft}$	-0.13	0.06	-0.25	-0.01	0.00	0.00	1552.00	1265.00	1.00
beta$_{pf}$	-0.04	0.25	-0.49	0.48	0.01	0.01	1468.00	1517.00	1.00
beta$_{tov}$	0.63	0.41	-0.20	1.39	0.01	0.01	1426.00	1313.00	1.00
beta$_{reb}$	0.80	0.10	0.60	1.00	0.00	0.00	1767.00	1474.00	1.00
beta$_{ast}$	0.45	0.14	0.17	0.72	0.00	0.00	1574.00	1522.00	1.00
beta$_{stl}$	-2.09	0.42	-2.95	-1.29	0.01	0.01	2409.00	1632.00	1.00
beta$_{blk}$	1.77	0.32	1.15	2.42	0.01	0.00	2759.00	1381.00	1.00
beta$_{pts}$	0.72	0.06	0.61	0.84	0.00	0.00	1501.00	1408.00	1.00

We also run a posterior predictive check on the model. While our model had 0 divergences, the PPC shows that the salary does deviate from our sampled posterior predictive distribution. Again, this is because of unsmoothed observations of the raw salary data and so overall the PPC looks fine and in line with what we should be expecting. Perhaps increasing the variance of the model would allow for the entire observed set of salaries to lie within the posterior predictive distribution area.



3.3 Bayesian Random Effects Regression

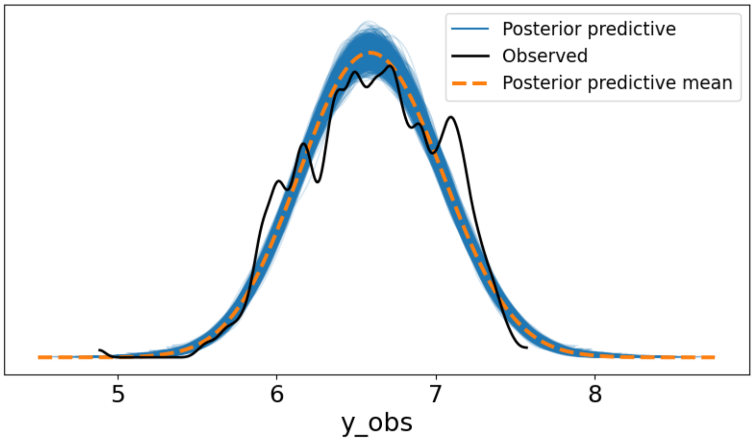
Finally, we also run a Bayesian Generalized Linear Models regression with random effects for player, teams, and year. Random effects as the name suggests attempts to incorporate random effects that affect salaries due to the year or team level. For instance, if due to an economic shock salaries are affected for a particular year then it is a good idea to include random effects for years or if salaries for a team to account for player-level heterogeneity or team-level heterogeneity, it is wise to include team and player level effects. Overall, in the context of NBA, it is ideal to include random effects at the team level given team budgets dictate salaries as do player performance but given economic shocks are rare and unlikely to affect salaries significantly in the context of NBA, it may be better to have year fixed effects. In any case for this regression we include both year and team random effects as shown in the model below.

$$y_{ij} \sim N(\alpha + \beta_{ij}\mathbf{X_{ij}} + \gamma_i + \eta_j + \epsilon_{ij}, \sigma^2)$$
$$\sigma \sim Exp(1)$$
$$\beta_{ij} \sim N(0,3)$$
$$\gamma_i, \eta_j \sim N(0,2)$$

The model results are shown below and again consistent with the fixed effects frequentist and bayesian regression results. Again rebounds, assists, blocks and points scored on a minute basis have a positive and statistically significant effect on salaries. While steals has a negative and statistically significant effect on salaries. Other covariates of interest are minutes per games played and age where a 1 sd increase in age or minutes per game played yields an 18 % and 20% increase in salaries respectively.

	mean	sd	hdi 2.5%	hdi 97.5%	mcse mean	mcse sd	ess bulk	ess tail	r hat
Intercept	6.39	0.57	5.36	7.57	0.03	0.02	299.00	407.00	1.00
beta age	0.18	0.01	0.17	0.19	0.00	0.00	2191.00	1160.00	1.00
beta gp	-0.01	0.01	-0.02	-0.00	0.00	0.00	1967.00	1419.00	1.00
beta min	0.20	0.01	0.18	0.21	0.00	0.00	1543.00	1230.00	1.00
beta fg	-0.37	0.11	-0.58	-0.15	0.00	0.00	1761.00	1188.00	1.00
beta threes	-0.08	0.04	-0.16	-0.01	0.00	0.00	1801.00	1255.00	1.00
beta ft	-0.12	0.06	-0.24	-0.01	0.00	0.00	1942.00	1232.00	1.00
beta pf	-0.04	0.25	-0.52	0.45	0.01	0.01	1783.00	1534.00	1.00
beta tov	0.64	0.39	-0.16	1.39	0.01	0.01	1631.00	1344.00	1.00
beta reb	0.80	0.10	0.61	1.01	0.00	0.00	1797.00	1287.00	1.00
beta ast	0.45	0.14	0.19	0.75	0.00	0.00	1239.00	908.00	1.00
beta stl	-2.11	0.42	-2.90	-1.33	0.01	0.01	1717.00	1390.00	1.00
beta blk	1.77	0.34	1.15	2.48	0.01	0.01	1558.00	1060.00	1.00
beta pts	0.72	0.06	0.60	0.84	0.00	0.00	1738.00	1404.00	1.00

Again, the random effects model performs similarly to the Bayesian fixed effects and frequentist models with the posterior predictive check looking almost identical. Again, given the unsmoothed distribution of the salary data it would be meaningful to look at a hierarchical model that incorporates contract information or other information that drives the multi-modal salary distribution we observe.



However, we compare our results of our random effects and fixed effects models and find that through WAIC the random effects model performs better. This again seems to be consistent with what we were expecting as given the context of NBA where salaries are governed by team budgets, the random effects model would be a better fit but again the difference in deviance is not very significant as can be seen in the table below.

	rank	elpd$_{waic}$	p$_{waic}$	elpd$_{diff}$	weight	se	dse	warning	scale
random effects	0	1351.22	66.83	0.00	1.00	94.62	0.00	True	deviance
fixed effects	1	1353.03	67.81	1.81	0.00	94.59	0.50	True	deviance

3.4 Bayesian Hierarchical Model

We next fit a Bayesian Hierarchical Model with clusters formed based on Player's Age. Given salary structures are determined based on contract basis and we don't have contract information, we use Age as a means to cluster whether the contract is for a Rookie, Novice or Veteran. We use K-Means to determine the number of clusters. Based on the elbow-method, we decide on 4 clusters as shown in the picture below and cluster our data into the clusters shown in the image below based on Age.

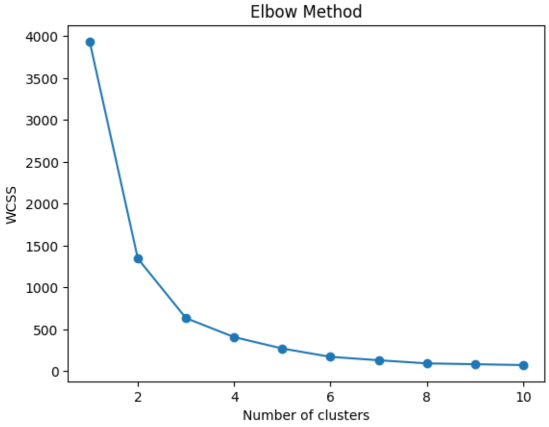


Figure 1: Elbow Method for K-Means Clustering

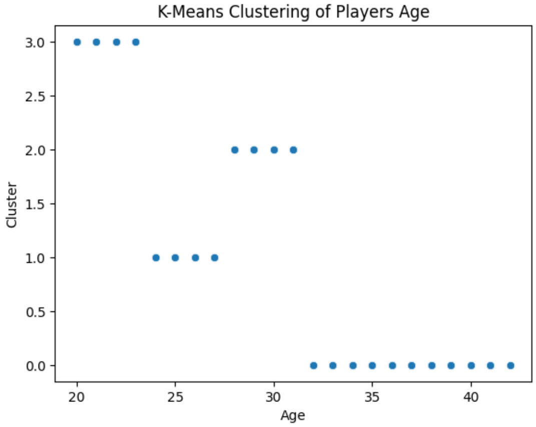


Figure 2: K-Means Clustering of Players Age

We then test our hypothesis whether Salary significantly differs by Age and run a one-way ANOVA using weakly informative priors on log-salary and Age clusters that we derived. The model we test through our ANOVA shown below.

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{One of the means for the clusters for salary is different}$$
$$y_i \sim N(\mu_i, \sigma^2)$$
$$\sigma \sim Exp(1)$$
$$\mu_i = \mu + \alpha_i, \text{ where } \mu \sim N(5,5) \text{ and } \alpha_i \sim N(0,3)$$
$$\text{Subject to the constraint that } \sum \alpha_i = 0$$

We run the ANOVA using a PPL and get the results shown below. Particularly, we find that Cluster 2 (which corresponds to ages 28-32) has the highest treatment effect with an increase in log-salary by 0.24 points over Cluster 0 (which corresponds to ages 18-24). In comparison, Cluster 1 (Ages 24-28) has a slight decrease in salary over Cluster 0 in log-salary of -0.04 points and then Cluster 3 (Ages 28 and above) has a signficant decrease in salary by -0.20 points on the log-salary scale. These results seem logical as we would expect our experienced performing players to earn the most followed by fresh talent followed by the novices followed by the oldest players in the league.

	mean	sd	hdi 2.5%	hdi 97.5%
alpha1	-0.04	0.01	-0.06	-0.02
alpha1-2	-0.28	0.02	-0.31	-0.24
alpha1-alpha3	0.16	0.02	0.12	0.20
alpha2	0.24	0.01	0.22	0.26
alpha2-alpha3	0.43	0.02	0.39	0.47
alpha3	-0.20	0.01	-0.22	-0.17

We next run a Bayesian Heirarchichal Model using the 4 clusters created based on Age. Our model is specified as below. We include random effects for year and team in our model and run the model shown below using PyMC.

$$y_{ij} \sim N(\alpha + \beta_{cluster[i]}\mathbf{X_{ij}} + \gamma_i + \eta_j + \epsilon_{ij}, \sigma^2)$$
$$\sigma \sim Exp(1)$$
$$\beta_{cluster[i]} \sim N(\beta_j, \sigma_{cluster[i]})), \text{ where } \beta_j \sim N(0,5) \text{ and } \sigma_{cluster[i]} \sim Exp(1)$$
$$\gamma_i, \eta_j \sim N(0,2)$$

The coefficients for the hierarchical model are shown on the next page. As can be observed that the coefficients for the 12 performance covariates do differ across the age clusters we found and are significant, suggesting the salary distribution is well fit by a hierarchical model. This maybe due to the fact that Age potentially captures the experience of the player and thus helps determine contract type of the player. In terms of the other coefficients, the intercept is 6.07 suggesting base salaries in millions and in line with our ANOVA results, cluster 1 performance metrics coefficients are largely positive. What is also interesting employing the hierarchical model is that the percent variables particularly Field Goals percent, Threes percent and Free Throws percent are positive and significant for certain clusters, which is a promising sign on the fit of the model. Overall, again rebounds, assists, blocks and points scored per minute have significant and positive impact on base salaries across most clusters. The trend seen in the coefficients of the hierarchical model is encouraging. We also evaluate and look at the PPC distribution and the WAIC score for the model fit. Overall, the WAIC value is lower than in the models above suggesting that the model is better fit relative to the random effects and fixed effects Bayesian models we considered. The distribution of the PPC however looks consistent with that of the random effects and fixed effects model with the posterior area not fully covering the observed data due to the non-normality and multimodal distribution of the salary data. Overall, the model fits well relative to the random effects and fixed effects model and improves on its deficiencies with the WAIC score being significantly lower.

	mean	sd	hdi 2.5%	hdi 97.5%	mcse mean	mcse sd	ess bulk	ess tail	r hat
Intercept	6.17	0.56	5.04	7.25	0.03	0.02	265.00	494.00	1.01
beta clusters[0, 0]	-0.01	0.01	-0.03	0.01	0.00	0.00	3106.00	1725.00	1.00
beta clusters[0, 1]	0.20	0.01	0.18	0.21	0.00	0.00	2631.00	1752.00	1.00
beta clusters[0, 2]	-0.18	0.13	-0.42	0.07	0.00	0.00	2913.00	1368.00	1.00
beta clusters[0, 3]	-0.06	0.06	-0.18	0.05	0.00	0.00	3066.00	1442.00	1.00
beta clusters[0, 4]	-0.24	0.08	-0.39	-0.08	0.00	0.00	2520.00	1524.00	1.00
beta clusters[0, 5]	-0.12	0.24	-0.58	0.34	0.00	0.01	2930.00	1346.00	1.01
beta clusters[0, 6]	0.03	0.24	-0.44	0.50	0.00	0.01	3169.00	1521.00	1.00
beta clusters[0, 7]	0.87	0.13	0.62	1.13	0.00	0.00	2598.00	1211.00	1.00
beta clusters[0, 8]	0.41	0.15	0.13	0.71	0.00	0.00	2952.00	1553.00	1.00
beta clusters[0, 9]	-0.26	0.26	-0.76	0.24	0.01	0.00	2943.00	1574.00	1.00
beta clusters[0, 10]	0.63	0.26	0.11	1.15	0.00	0.00	3474.00	1426.00	1.00
beta clusters[0, 11]	0.88	0.08	0.73	1.03	0.00	0.00	2870.00	1669.00	1.00
beta clusters[1, 0]	-0.01	0.01	-0.03	0.01	0.00	0.00	2875.00	1437.00	1.00
beta clusters[1, 1]	0.23	0.01	0.20	0.25	0.00	0.00	2500.00	1225.00	1.00
beta clusters[1, 2]	0.18	0.14	-0.08	0.45	0.00	0.00	2914.00	1562.00	1.00
beta clusters[1, 3]	-0.13	0.06	-0.26	-0.01	0.00	0.00	3122.00	1361.00	1.00
beta clusters[1, 4]	0.02	0.08	-0.13	0.19	0.00	0.00	2538.00	1412.00	1.00
beta clusters[1, 5]	0.32	0.23	-0.11	0.78	0.00	0.00	3238.00	1669.00	1.00
beta clusters[1, 6]	0.46	0.25	-0.02	0.96	0.00	0.00	3689.00	1584.00	1.00
beta clusters[1, 7]	0.95	0.14	0.68	1.22	0.00	0.00	2791.00	1370.00	1.00
beta clusters[1, 8]	0.34	0.15	0.05	0.65	0.00	0.00	3246.00	1481.00	1.00
beta clusters[1, 9]	0.08	0.26	-0.44	0.57	0.01	0.01	3369.00	1525.00	1.00
beta clusters[1, 10]	0.61	0.25	0.10	1.09	0.01	0.00	2830.00	1338.00	1.00
beta clusters[1, 11]	0.52	0.09	0.36	0.69	0.00	0.00	2662.00	1609.00	1.00
beta clusters[2, 0]	0.01	0.01	-0.01	0.03	0.00	0.00	2394.00	1593.00	1.00
beta clusters[2, 1]	0.12	0.01	0.10	0.15	0.00	0.00	2659.00	1705.00	1.00
beta clusters[2, 2]	-0.28	0.15	-0.56	0.02	0.00	0.00	2660.00	1859.00	1.00
beta clusters[2, 3]	-0.06	0.09	-0.23	0.12	0.00	0.00	2884.00	1330.00	1.00
beta clusters[2, 4]	-0.02	0.11	-0.23	0.20	0.00	0.00	2390.00	1395.00	1.00
beta clusters[2, 5]	0.31	0.25	-0.18	0.82	0.00	0.00	3741.00	1199.00	1.00
beta clusters[2, 6]	0.42	0.27	-0.09	0.94	0.00	0.00	3888.00	1454.00	1.00
beta clusters[2, 7]	0.31	0.16	0.00	0.63	0.00	0.00	2164.00	1580.00	1.00
beta clusters[2, 8]	-0.14	0.20	-0.52	0.23	0.00	0.00	3132.00	1307.00	1.00
beta clusters[2, 9]	0.09	0.28	-0.42	0.62	0.01	0.01	3480.00	1496.00	1.00
beta clusters[2, 10]	0.37	0.27	-0.17	0.86	0.01	0.00	3266.00	1368.00	1.00
beta clusters[2, 11]	0.42	0.13	0.17	0.68	0.00	0.00	2340.00	1383.00	1.00
beta clusters[3, 0]	-0.03	0.01	-0.04	-0.00	0.00	0.00	2961.00	1317.00	1.00
beta clusters[3, 1]	0.21	0.01	0.18	0.24	0.00	0.00	2827.00	1447.00	1.00
beta clusters[3, 2]	0.34	0.15	0.08	0.67	0.00	0.00	2583.00	1424.00	1.00
beta clusters[3, 3]	-0.17	0.07	-0.32	-0.04	0.00	0.00	3087.00	1471.00	1.00
beta clusters[3, 4]	0.08	0.09	-0.09	0.26	0.00	0.00	2643.00	1424.00	1.00
beta clusters[3, 5]	0.22	0.24	-0.30	0.67	0.00	0.01	3277.00	1324.00	1.00
beta clusters[3, 6]	0.28	0.27	-0.25	0.78	0.01	0.00	3200.00	1582.00	1.00
beta clusters[3, 7]	0.73	0.15	0.42	1.01	0.00	0.00	2967.00	1176.00	1.00
beta clusters[3, 8]	0.34	0.16	0.01	0.64	0.00	0.00	3115.00	1285.00	1.00
beta clusters[3, 9]	0.09	0.28	-0.43	0.62	0.01	0.01	3200.00	1366.00	1.01
beta clusters[3, 10]	0.52	0.28	-0.04	1.04	0.01	0.00	3484.00	1453.00	1.00
beta clusters[3, 11]	0.60	0.11	0.40	0.82	0.00	0.00	2725.00	1373.00	1.00

Index for interpreting coefficients: 0 - Games Played standardized 1 - Minutes per Games Player standardized 2 - Field Goals percent 3 - Threes percent 4 - Free Throws percent 5 - Personal Fouls per minute 6 - Turnover per minute 7 - Rebounds per minute 8 - Assists per minute 9 - Steals per minute 10 - Blocks per minute 11 - Points per minute

	Results
elpd waic	1171.97
se	95.76
p waic	83.36
n samples	2000
n data points	3936
warning	True

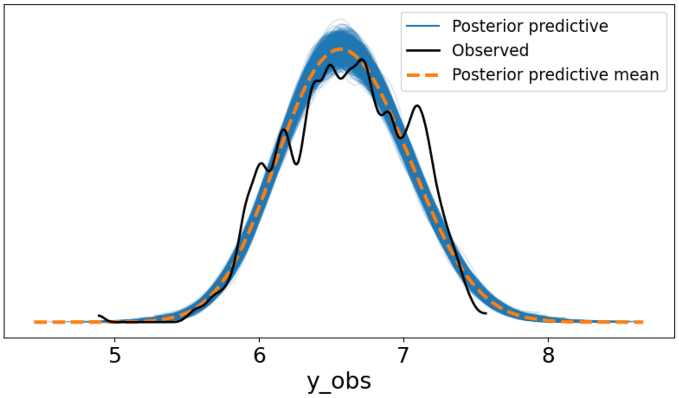


Figure 3: PPC for Heirarchichal Model

4 Conclusion

In summary, we fit three Bayesian GLM models with weakly informative priors including a fixed effects model, a random effects model and a hierarchical model based on clusters obtained from players age. Overall, the results of the Bayesian fixed effects and random effects model do not differ significantly from the results of the frequentist fixed effects model. However, the random effects model performs better than the fixed effects model. The Bayesian Hierarchical model developed based on K-means clustering on the age variable allows us to capture salary structures better with most clusters having significant variation on the performance metrics. It also performs best in comparison to the random effects and fixed effects model on the WAIC comparison scale. In terms of the most significant and positive predictors of salary, we find that assist per minute, points per minute and rebounds per minute have a positive and significant effect. Though, the Bayesian Hierarchical model allows us to capture granular differences in salary structure across the difference age brackets. For future work, we could get more granular data and attempt to smooth the log salaries data further to better fit the GLM and have a better posterior predictive check unlike seen for the three models observed. However, still as a first step we investigate how GLMs can be used to predict and analyze NBA salaries.

References

S. K. Deshpande and S. T. Jensen, “Estimating an NBA player’s impact on his team’s chances of winning,” Journal of Quantitative Analysis in Sports, vol. 12, no. 2, Jan. 2016, doi: https://doi.org/10.1515/jqas-2015-0027.

N. Higgs and I. Stavness, “Bayesian analysis of home advantage in North American professional sports before and during COVID-19,” Scientific Reports, vol. 11, no. 1, Jul. 2021, doi: https://doi.org/10.1038/s41598-021-93533-w.

Brani Vidakovic, Engineering biostatistics : an introduction using MATLAB and WinBUGS. Hoboken, Nj: Wiley, 2017.

“Annual wages in the NBA & WNBA 2019/20,” Statista. https://www.statista.com/statistics/1120680/annual-salaries-nba-wnba

hrfang, “NBA Salaries By Players of Season 2000 to 2019,” Kaggle.com, 2019. https://www.kaggle.com/datasets/hrfang salaries-by-players-of-season-2000-to-2019 (accessed Dec. 21, 2024).

“Basketball Players Stats per Season - 49 Leagues,” www.kaggle.com. https://www.kaggle.com/datasets/jacobbaruch/b players-stats-per-season-49-leagues