National Hockey League (NHL) Player Evaluation by Estimating of Players Salary Cap Hit*

Regression analysis of NHL players based on the 2023-24 season to estimate players Cap hit resulting in an undervaluation of star players

Gavin Crooks

April 18, 2024

This paper employs regression analysis to estimate NHL player cap hits based on performance metrics from the 2023-24 season. Using data from 448 NHL players, including games played, points, plus-minus, and salary information. The analysis suggests potential undervaluation for star players like Connor McDavid and Auston Matthews, highlighting the importance of considering various performance indicators in determining player salaries. Overall, this research contributes valuable insights into NHL player valuation and the factors influencing their cap hits.

1 Introduction

In major sports leagues including the National Hockey League(NHL), teams look to sign players to contracts within the salary cap limits. Teams aim to build the best team possible, with the goal of winning the Stanley Cup. Histrionically, teams used scouts and basic player statistics such as goals and assist to analyze players, their abilities and ultimately how much to pay them. In the past few decades, professional sports organizations have began using statistical models to better evaluate players and their values. A notable example of this is from Major League Baseball (MLB) and the movie "Moneyball," directed by Bennett Miller, which portrays the analytical approach adopted by the 2002 Oakland Athletics to build their baseball team (IMDb (2011)). Now all major sports league teams use statistical approaches to build their teams in order to build the best team possible.

Of the major sports leagues the NHL has the smallest salary cap, in the 2023-24 season the NHL salary cap is 83.5 million dollars (NHL.com (2024)). This means in order for teams to

*Code and data are available at: https://github.com/Crooksyyy/NHL-Player-Evaluation

win they must value and pay players appropriately. It is important to distinguish the difference between a player's salary and cap hit. A player's salary is the total compensation for a player at the NHL level that includes signing bonuses, but not performance bonuses (Sportsnet (2024)). A player's cap hit is determined as the average annual value of their current contract. Cap hit is calculated by dividing the total salary plus signing bonuses of a contract by the contract's length (Sportsnet (2024)).

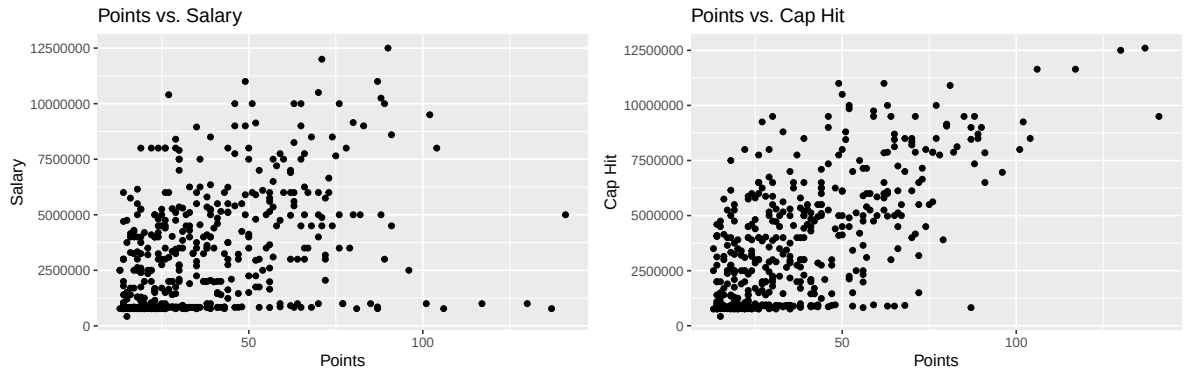
In this paper I aim to estimate players true value by using a multiple regression approach to determine a players worth based on their games played, points and plus/minus. This method yielded coefficients in which we can estimate how much a player should be paid. To complete this analysis I use R, R Core Team (2023) and a number of packages available including Wickham et al. (2019), Richardson et al. (2023), Bob Goodrich et al. (2024) and Müller (2020). This paper is inspired by Telling Stories with Data With Applications in R and Python(Alexander (July 27, 2023)).

The remainder of this paper is structured as follows. Section 2 which outlines data collection, data cleaning and showcases the data. Section 3 explains the model set-up and model justification. Section 4 summarizes the results and predicts several player's cap hit.

2 Data

The data in this paper was obtained from two resources, the player statistics were obtained from the NHL official website National Hockey League (ongoing). The salary data was obtained from Hockey Reference (ongoing). It is important to note the player statistics were pulled on April 12, 2024, this means the season was not over and they would change that night. These data sets were merged by players' names to combine player statistics and their salary information. The data cleaning procedures for this data set were extremely straightforward as the dataset is extremely well formatted and easy to work with. The data was cleaned to only include the data that was to be used in the analysis removing statics such as game-winning goals or shots per game. The last aspect of the data cleaning was to remove players that had incomplete data, which would have occurred when merging the two data sets. The resulting dataset consists of 448 NHL players, their position, games played, goals, assists, points, plus/minus, points per game, time on ice per game, salary and their cap hit. These variables were chosen as they are generally easy to understand as a measure of how good a player is. With the goal of this paper in mind, there needs to be a measure of how productive the player is and how much they are paid. In the NHL there are two different measures of how much a player is paid, salary and cap hit. Both were included in the salary data obtained from Hockey Reference (ongoing) and are included in the analysis.

Figure 1a and Figure 1b illustrate a naive approach to analyzeing how much a player should or should not be paid. Based on this approach players in the bottom right corner are producing lots of points and are relatively underpaid. The players in the top right are highly paid and producing like they should be. Players in the top left would be considered overpaid as they



(a) Scatter plot for Points vs. Salary

(b) Scatter plot for Points vs. Cap Hit

Figure 1: Scatter plots of players salary or cap hit and points during the 2023-24 NHL season as of April 12, 2024.

are not scoring but are highly paid. Scoring is not the only thing that describes how good a player is. Specifically, this approach is flawed for measuring how valuable defenseman are. Because of this other measures can be used, one of the simplest is a player's plus/minus. A plus is obtained if you are on the ice when your team scores and a minus when the other team scores.

Figure 2 uses the plus/minus and salary to illustrate aspects besides scoring. Plus/minus may be a better measure than goals, assists or points as it includes defence, which the other statistics do not. However, can still measure if your team produces offense when they are on the ice. Plus/minus is also a better measure of deference man as preventing goals is more important for defence, than scoring goals. This graph emphasizes players with more than 75 points. We can see that the majority of the players with more than 75 points also have a positive plus/minus. It also shows several with high plus/minus are not top scorers as the top 3 players in plus/minus, have less than 75 points.

3 Model

The goal of our modelling strategy is to estimate the appropriate salary for NHL players based on their performance metrics using regression analysis. This approach is not novel, as sports teams have long employed data analysis to determine player salaries. However, it remains a contentious issue, with debates often centred around intangible factors such as leadership and camaraderie. Despite these challenges, various models have been developed across sports with the shared objective of identifying undervalued players to enhance team performance. One prominent example is depicted in the movie "Moneyball," directed by Bennett Miller, which portrays the analytical approach adopted by the 2002 Oakland Athletics to build their baseball team.

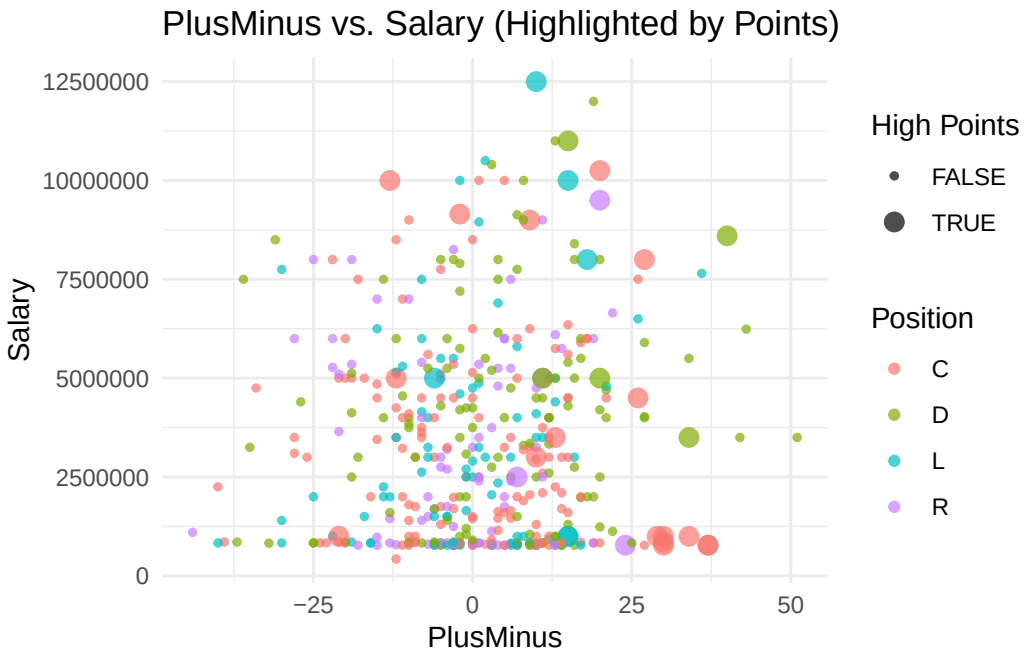


Figure 2: Scatter plot of NHL players plus/minus and salary highlighting positision and players with more than 75 points.

3.1 Model set-up

In this following sections, we present the hierarchical Bayesian model to predict player salaries in professional sports leagues, utilizing performance metrics such as points, plus-minus, and games played as predictors.

$$\text{CapHit}_i | \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_{\text{Points}} \times \text{Points}_i + \beta_{\text{PlusMinus}} \times \text{PlusMinus}_i + \beta_{\text{GamesPlayed}} \times \text{GamesPlayed}_i$$

$$\alpha \sim \text{Normal}(0, 2.5)$$

$$\beta_{\text{Points}}, \beta_{\text{PlusMinus}}, \beta_{\text{GamesPlayed}} \sim \text{Normal}(0, 2.5)$$

$$\sigma \sim \text{Exponential}(1)$$

In this representation, (CapHit_i) represents the cap hit of player (i), (Points*i*), (*PlusMinus*i), and (GamesPlayed*i*) *represent the points, plus-minus, and points per game of player (i) respectively, and* ( ), ({Points}), (*{PlusMinus}), and* ({GamesPLayed}) represent the intercept and coefficients for these variables, respectively. We run the model in R (R Core Team 2023) using the `rstanarm` package of Ben Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

Our model focuses on key performance indicators such as points, plus-minus, and games played, which are all recognized as significant contributors to player value in the NHL. Points reflect a player's offensive contribution, while plus-minus provides insight into their defensive impact. Games played serve as a measure of a player's availability and durability throughout the season, contributing to their overall value to the team. It also controls for ultimately accounts for points per game.

By incorporating these metrics into our model, we aim to capture the multifaceted nature of player performance and its influence on salary and cap hit determination. While acknowledging the inherent complexity of salary negotiations in professional sports, our model provides a systematic framework for evaluating player worth based on empirical data.

We expect a positive relationship between plus-minus and salary, with a greater magnitude than the positive relationship between points and cap hit. Plus-minus reflects a player's overall impact on the game, taking into account both offensive and defensive contributions, whereas points primarily capture offensive prowess. Therefore, players with a high plus-minus rating are likely to be valued more highly by teams due to their ability to positively influence the outcome of games through effective play in both offensive and defensive situations.

In particular, we anticipate that players with a positive plus-minus rating, indicating that their team scores more goals than it concedes while they are on the ice, will command higher salaries compared to players with a negative or neutral plus-minus. This expectation is rooted in the fundamental principle of team success in hockey, where preventing goals is as crucial as scoring goals.

Furthermore, our model accounts for the interplay between points and plus-minus, recognizing that players who contribute offensively while maintaining a strong defensive presence are likely to be perceived as more valuable assets to their teams. As such, the model seeks to identify players who excel in both offensive production and defensive responsibility, thereby providing teams with valuable insights into potential salary and cap hit allocations.

4 Results

Our results are summarized in Table 1 and Table 2. Table 1 illustrates the accuracy of each method to measure a player's pay, overall salary or cap hit against points. This simple regression analysis provides insights into the relationship between player performance (measured by points) and their corresponding salary or cap hit.

The coefficients presented in the regression models indicate the expected change in salary or cap hit for each unit increase in points. For example, in the Salary and Points regression model, the coefficient for Points is $43,107.99$, indicating that, on average, each additional point scored by a player is associated with an increase in salary of approximately $43,107.99$. Similarly, in

Table 1: Model results for compairing Salary and Cap hit as the output variable

	Salary and Points	Caphit and Points
(Intercept)	1 713 785.76	691 440.75
	(233 528.97)	(203 924.74)
Points	43 107.99	81 599.22
	(5067.94)	(4562.16)
Num.Obs.	448	448
R2	0.134	0.415
R2 Adj.	0.127	0.413
Log.Lik.	−7223.539	−7167.409
ELPD	−7226.3	−7169.8
ELPD s.e.	15.8	15.6
LOOIC	14 452.7	14 339.6
LOOIC s.e.	31.6	31.3
WAIC	14 452.7	14 339.6
RMSE	2 430 113.51	2 144 385.23

the Caphit and Points regression model, the coefficient for Points is 81, 599.22, suggesting a larger increase in cap hit for each additional point scored.

The R-squared values provide a measure of the proportion of variance in salary or cap hit explained by the independent variable (points) in each model. In both cases, the R-squared values are relatively low, indicating that points alone may not fully account for the variation in salary or cap hit. Ultimately the R-squared value is better within the cap hit model and therefore used as our estimand throughout this paper although salary and cap hit are often used interchangeably.

Table 2 presents the complete model and multiple simple model results. The first three models are simple regressions between one of Points, Plus/Minus, or Games Played on a player's Cap Hit.

Similarly to the points models from @Table 1, the Cap Hit and Plus/Minus model shows that each additional unit of Plus/Minus is associated with a decrease in salary of approximately 27286.19. For the Cap Hit and Games Played model, each additional game played is associated with an increase in salary of approximately 60955.56. It is important to note the games played model is flawed as a number of factors like injuries, greatly impact games played and why it is controlled for in the final model.

The R-squared values for these simple models range from 0.020 to 0.134, indicating that the proportion of variance in salary explained by these individual predictors is relatively low.

In contrast, the Complete Model includes all three predictors (Points, Plus/Minus, and Games

Table 2: Model results for points, Games Played, plus/minus and the complete model

	Points	Plus/Minus	Games Played	Complete Model
(Intercept)	691 440.75	3 849 466.51	−446 090.40	1 389 708.24
	(203 924.74)	(131 388.48)	(951 322.20)	(718 354.15)
Points	81 599.22			84 542.71
	(4562.16)			(5075.21)
PlusMinus		27 286.19		−5206.62
		(9209.57)		(7457.26)
GamesPlayed			60 955.56	−11 239.05
			(13 279.17)	(10 619.42)
Num.Obs.	448	448	448	448
R2	0.415	0.020	0.046	0.419
R2 Adj.	0.413	0.012	0.041	0.409
Log.Lik.	−7167.409	−7283.650	−7277.425	−7166.943
ELPD	−7169.8	−7286.0	−7279.6	−7171.2
ELPD s.e.	15.6	13.1	13.7	15.7
LOOIC	14 339.6	14 572.1	14 559.3	14 342.4
LOOIC s.e.	31.3	26.2	27.4	31.3
WAIC	14 339.6	14 572.1	14 559.3	14 342.4
RMSE	2 144 385.23	2 779 780.09	2 741 460.54	2 140 821.22

Played) simultaneously. It shows a higher R-squared value of 0.419, indicating that this model explains a larger proportion of the variance in salary. The coefficients for each predictor in the Complete Model provide insights into how each performance metric contributes to player salary when considered together. This model results in a negative coefficient for both plus/minus and games played. This means our model likely overestimates the intercept or the value of points. However, this allows us to predict players' salaries.

4.1 Prediction

In this section of the paper, we are going to use our model to estimate how much a number of players should be paid and compare it to their actual contract. To predict Connor McDavid's cap hit, we utilize the coefficients obtained from our regression model. Specifically, we input the values of players' performance metrics (points, plus/minus, and games played) into the model equation:

$$\text{CapHit}_i = \beta_0 + \beta_{\text{Points}} \times \text{Points}_i + \beta_{\text{PlusMinus}} \times \text{PlusMinus}_i + \beta_{\text{GamesPlayed}} \times \text{GamesPlayed}_i$$

By substituting the values of players' performance metrics into this equation along with the corresponding coefficients from the model, we can calculate the predicted cap hit.

The first player we will estimate is Connor McDavid as he is unanimously agreed upon as the best hockey player in the world. In the 2023-24 NHL season on April 12, 2024, he had 72 games played, 130 points and a plus/minus of 34. Our model estimates his cap hit should be $11,394,024.86$. His actual cap hit is $12,500,000$ meaning according to our model he is overpaid.

Auston Matthews is the leading goal scorer with 68 goals as of April 12, 2024. His complete statistics required for our model are 78 games played, 106 points and 37 plus/minus. Based on the complete model and Auston Matthews' performance statistics, his estimated cap hit would be approximately $9,357,267.66$. His actual cap hit is $11,640,250$, again stating he is overpaid.

The last player we will predict is Tyson Barrie, a player who is not special unlike the ones above. In our data, he has 41 games played, 15 points and -10 plus/minus. The model's resulting estimate is $2,153,176.04$. His actual Cap hit is $4,500,000$ again showing he is overpaid.

5 Conclusion

As sports teams continue to increase the use of statistical models to help build their teams, new approaches will continue to be developed. This paper uses regression to predict players' true value against teams' cap hits. Our model appears specifically to overweight points and

undervalue both plus/minus and games played. These yielded a negative coefficient, meaning playing fewer games and having a negative plus/minus is more valuable according to our model. This is flawed and inaccurate. However, it is a good base to continue regression analysis to determine player worth in the NHL.

6 Discussion

6.1 First discussion point {#sec-first-point} - Salary vs Caphit

The decision to utilize cap hit as the outcome variable in our regression model stemmed from the comparative analysis of simple regressions, particularly the one between points and salary versus points and cap hit. Notably, the caphit model exhibited a substantially larger R-squared value, suggesting a better fit to the data. However, upon reviewing the results, we encountered counter intuitive findings such as negative coefficients for plus/minus and games played. This unexpected outcome prompted a critical reflection on the underlying assumptions and model specifications.

In future analyses, we aim to delve deeper into the dynamics of player compensation by conducting similar regressions with salary as the outcome variable. By directly comparing the two measures of compensation—salary and cap hit—we anticipate gaining deeper insights into the factors influencing player worth in the NHL. This comparative approach will allow for a more nuanced understanding of the relationship between player performance metrics and their corresponding compensation. Additionally, it opens up avenues for exploring the discrepancies between the two measures

6.2 Second discussion point - Goalies

While our analysis focused solely on skaters, it's crucial to acknowledge that goaltenders represent a distinct subset within the realm of hockey players. Their role and performance metrics differ significantly from those of skaters, necessitating a separate analytical approach. Goalies play a pivotal role in determining team success, often serving as the last line of defence, which inherently alters the evaluation criteria for their worth.

Due to these fundamental differences, we made the deliberate decision to exclude goalies from the scope of this paper. However, it's essential to recognize the importance of goaltending in the overall dynamics of hockey. Some argue that goaltenders wield the most significant influence on game outcomes, further underscoring the necessity of conducting a dedicated analysis of their value.

Moving forward, future research endeavours should prioritize the development of specialized models tailored specifically to assess the value of goaltenders. These models would incorporate performance metrics unique to goaltending, such as save percentage, goals-against average, and

shutouts, among others. By doing so, we can gain a comprehensive understanding of the factors driving the value of goaltenders and their impact on team success.

6.3 Third discussion point - Underestimation

Our model yielded unexpected outcomes, particularly in the assessment of player value, where prominent figures like Auston Matthews and Connor McDavid appeared to be overpaid according to the model's estimations. This discrepancy raised significant questions regarding the accuracy and validity of our model's predictions. Upon closer examination, it became evident that the model's undervaluation of certain performance metrics, notably plus/minus and games played, contributed to the observed discrepancies.

The inclusion of games played as a predictor variable aimed to address scenarios where players with high salaries might have lower point totals due to injury-induced absences. However, this approach may have inadvertently skewed the model's estimations, particularly when considering players' overall contributions throughout a season. To mitigate this issue in future iterations, a shift towards using statistics such as points per game could provide a more nuanced understanding of player performance, thereby eliminating the need for games played as a predictor variable.

Moreover, the puzzling negative coefficient associated with plus/minus in our final model warrants further investigation. While a simple regression of plus/minus on cap hit produced a positive coefficient, the unexpected reversal in our comprehensive model suggests potential complexities or inconsistencies in the underlying data or model structure. This discrepancy underscores the importance of rigorous model validation and sensitivity analysis to identify and address such anomalies effectively.

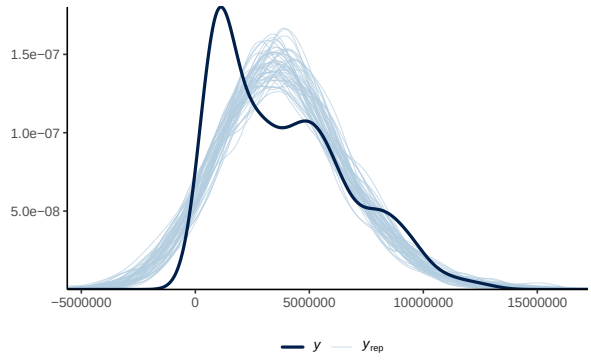
Additionally, the disproportionately large intercept in our model, nearly 1.4 million, poses significant concerns regarding its practical implications. Such a substantial intercept implies that a player with zero points, games played, and plus/minus would still command a considerable salary—a scenario that seems implausible in reality. This overestimation of the intercept likely contributed to the undervaluation of plus/minus and games played, ultimately resulting in underestimated player cap hits.

Appendix

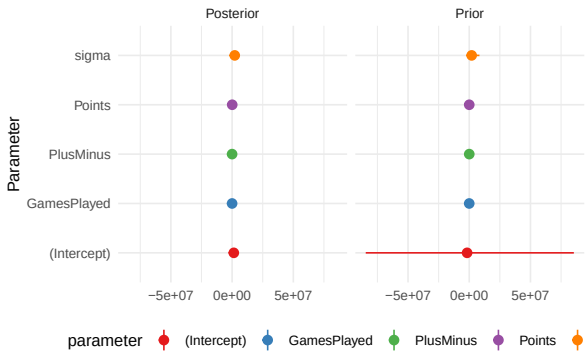
.1 Posterior predictive check

In Figure 3a we implement a posterior predictive check. This shows the posterior's distributions ability to simulate the data used to create the model. We can see some difference between the simulated data and reality.

In Figure 3b we compare the posterior with the prior. This shows the small amount of change between the posterior and priors. We see little change between them except in the intercept.



(a) Posterior prediction check

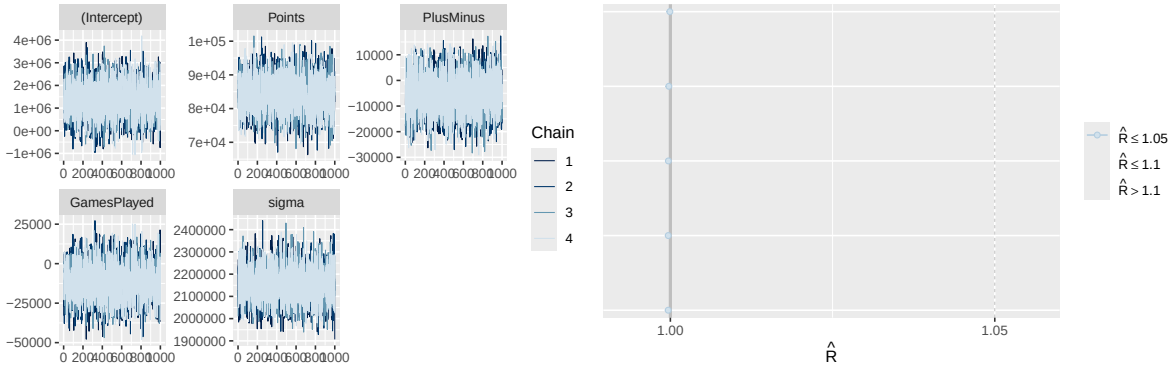


(b) Comparing the posterior with the prior

Figure 3: Examining how the model fits, and is affected by, the data

.2 Diagnostics

Figure 4 is a trace and rhat plot. The trace plot is as expected showing large amount of scatter and nothing out of the ordinary. In the rhat plot we want the values as close to 1 as possible which it is clear our model achieves to an extremely high level.



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. July 27, 2023. "Telling Stories with Data." https://tellingstorieswithdata.com/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Goodrich, Bob, Jonah Gabry, Imran Ali, and Ben Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm.

Hockey Reference. ongoing. "Hockey Reference: Current NHL Salaries." https://www.hockey-reference.com/friv/current_nhl_salaries.cgi.

IMDb. 2011. "Moneyball." https://www.imdb.com/title/tt1210166/.

Müller, Kirill. 2020. "Here: A Simpler Way to Find Your Files." https://CRAN.R-project.org/package=here.

National Hockey League. ongoing. "NHL Player Statistics." https://www.nhl.com/stats/skaters.

NHL.com. 2024. "NHL Salary Cap Expected to Rise for 2024-25 Season." https://www.nhl.com/news/nhl-salary-cap-expected-to-rise-for-2024-25-season#:~:text=The%20salary%20cap%20is%20%2483.5,of%20the%20COVID%2D19%20pandemic.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2023. "Arrow: Integration to 'Apache' 'Arrow'." https://CRAN.R-project.org/package=arrow.

Sportsnet. 2024. "NHL Salary Contract Glossary." https://www.sportsnet.ca/hockey/nhl/salary-contract-glossary/#:~:text=Cap%20hit%20is%20calculated%20by,the%20performance%20bonuses%20are%20achieved.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software*. https://doi.org/10.21105/joss.01686.