Bayesian Data Analysis - Project done on Tennis Data

December 1, 2024

1 Introduction, Description of data and problem

Tennis is a globally popular sport, played on a variety of surfaces, and provides rich opportunities for data-driven analysis. Over the years, significant attention has been devoted to understanding players' performances, their progression in rankings, and the influence of playing conditions. This project leverages a comprehensive dataset of tennis matches spanning over two decades, containing 106,716 observations across 49 variables. The dataset provides detailed information on matches, players, and performance metrics, making it an ideal resource for Bayesian analysis. The dataset was sourced from Kaggle (https://www.kaggle.com/datasets/guillemservera/tennis), a well-known platform for sharing and exploring data.

1.1 Dataset

The dataset includes a rich variety of variables that capture both match-level and player-level details. Match-level variables such as `tourney_name`, `surface`, `draw_size`, `tourney_date`, and `score` provide context about the tournament and match conditions. Player-level details for both winners and losers include attributes like `name`, `age`, `height`, `handedness` (e.g., right-handed or left-handed), and `country`. Performance metrics such as the number of aces (`w_ace`, `l_ace`), double faults (`w_df`, `l_df`), first-serve success rates (`w_1stIn`, `l_1stIn`), break points saved (`w_bpSaved`, `l_bpSaved`), and many more are also available. Ranking information, including `winner_rank`, `winner_rank_points`, `loser_rank`, and `loser_rank_points`, adds a competitive context to the matches.

Upon exploring the dataset's page on Kaggle, no prior analysis was evident. This presents an exciting opportunity to explore the data in novel ways and contribute fresh insights into the factors influencing tennis players' performance and progression.

1.2 Motivation

The motivation for this project stems from personal experience and curiosity. Growing up playing tennis predominantly on clay courts sparked an interest in understanding how court types might influence player performance, especially across different regions. For instance, do players from countries with predominantly clay courts perform better on this surface compared to others? Additionally, the rapid ascent of players like Novak Djokovic in global rankings raises intriguing questions about the factors contributing to ranking progression across different regions and countries.

To illustrate the richness of the dataset and inspire our analyses, we first explored several preliminary visualizations:

- **Age Distribution of Players:** A histogram showcasing the distribution of players' ages revealed a concentration of professional players in their mid-20s, highlighting the peak years of athletic performance. (See Figure 1)
- **Average Aces per Match by Surface:** A comparison of aces across surfaces demonstrated that grass courts tend to yield higher ace counts compared to clay courts, likely due to differences in ball speed and bounce characteristics. (See Figure 2)
- **Ranking Progression of Novak Djokovic:** A time-series plot of Novak Djokovic's ranking progression highlighted his rapid rise to the top ranks within approximately three years, showcasing the career trajectories of elite players. (See Figure 3)

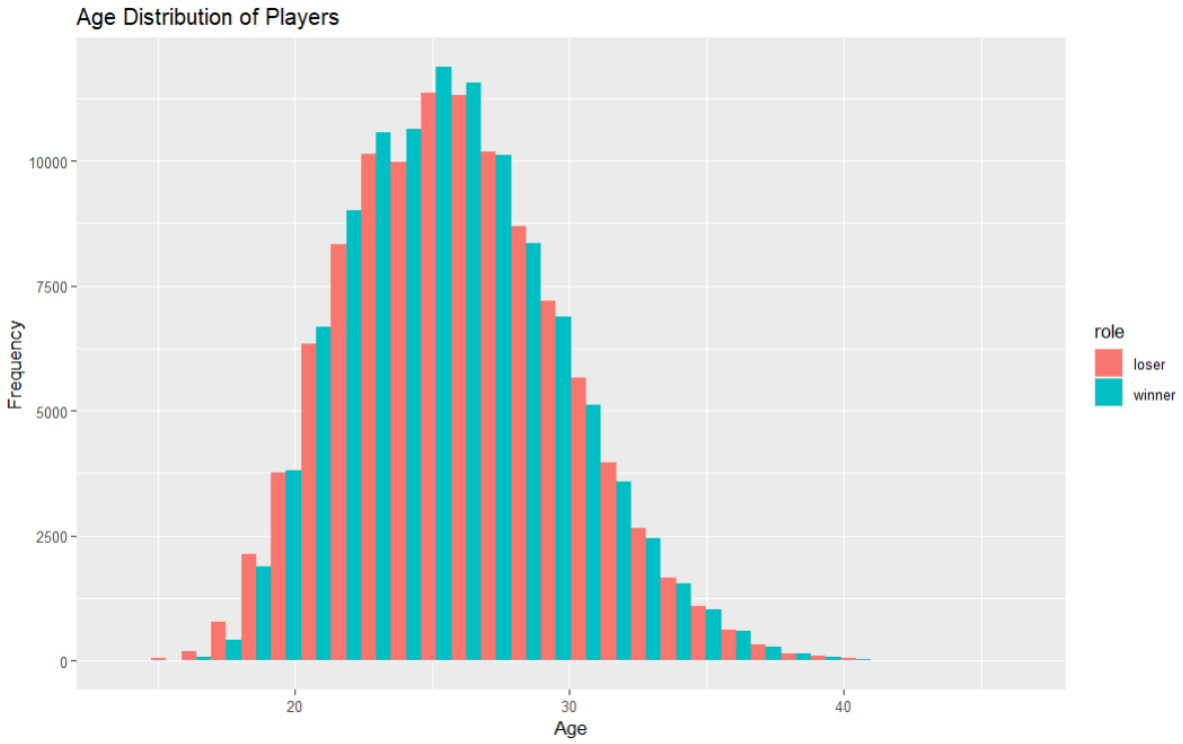

Figure 1: Age Distribution of Players

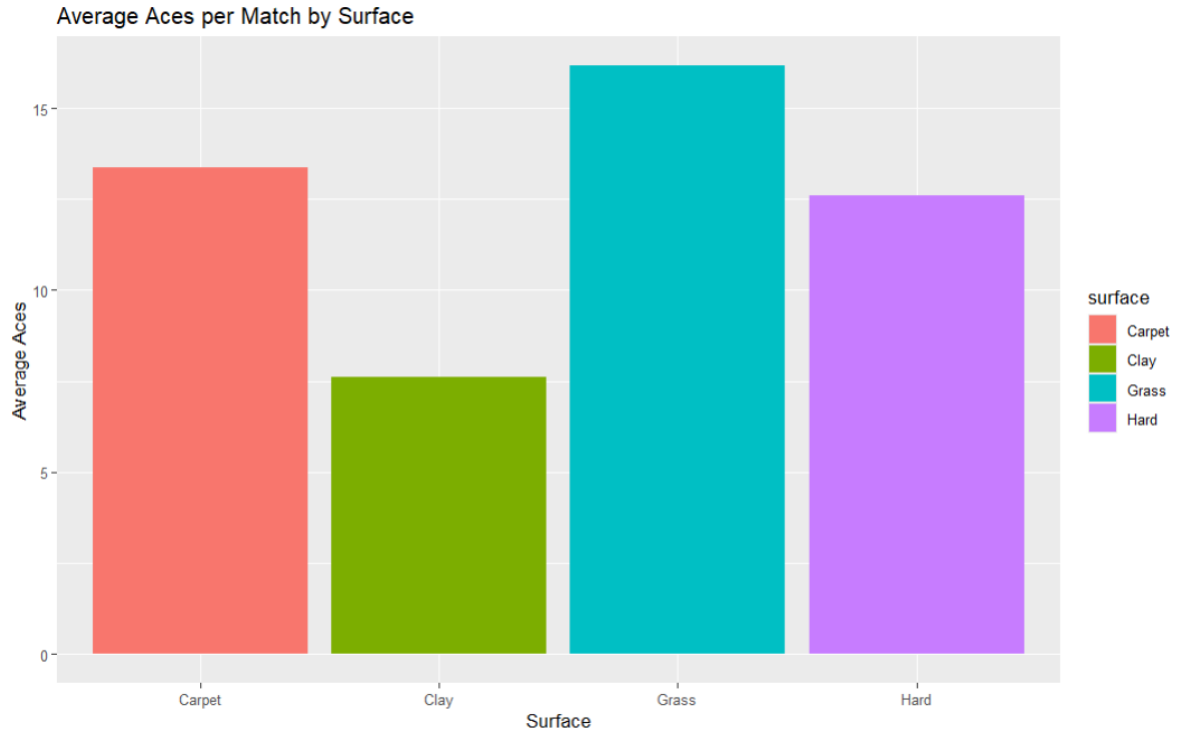

Figure 2: Average Aces per Match by Surface

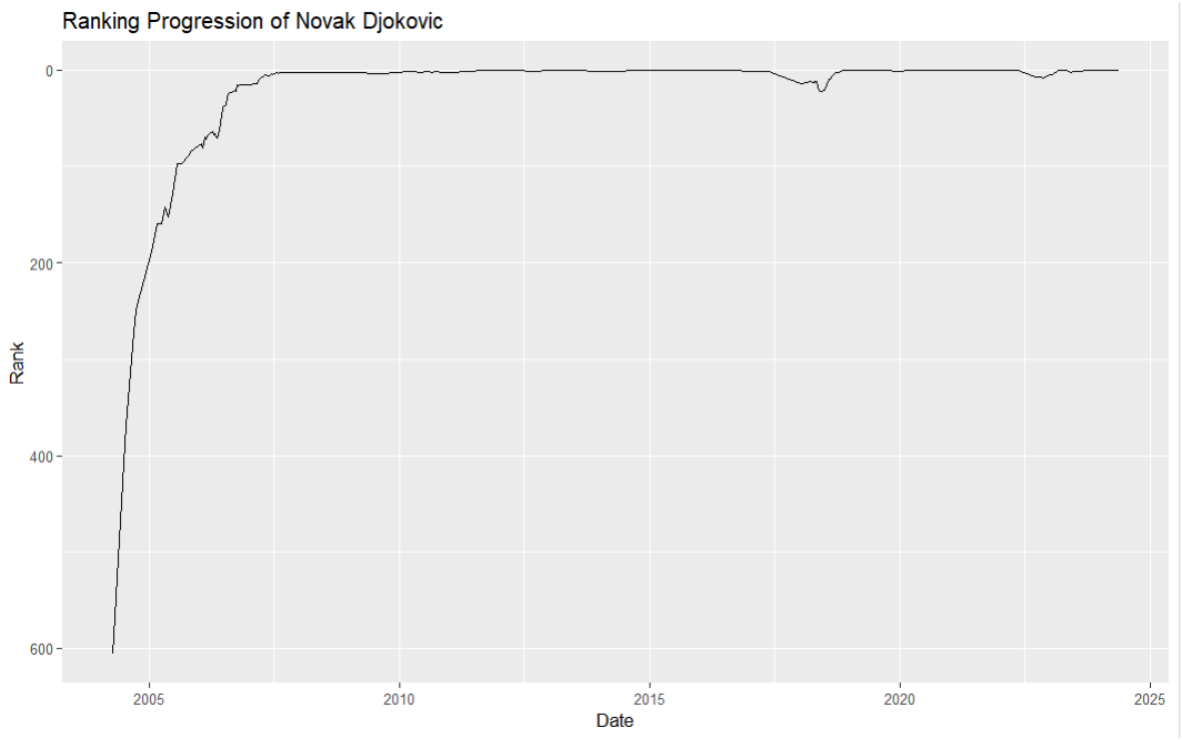

Figure 3: Ranking Progression of Novak Djokovic

1.3 Analysis Problems

Building on these observations, we propose two primary analyses:

1. Measuring the time it takes for players from different regions or countries to reach the top 20 ranks. This analysis seeks to uncover regional differences in player development and progression in professional tennis.

2. (Extra) Investigating the win rates of players from various regions (e.g., Asia, Europe, America) on different court types (e.g., clay, grass, hard). This study aims to identify whether surface-specific advantages exist for players from particular regions.

These analyses not only aim to provide insights into player development and performance but also offer practical implications for training strategies and surface-specific preparations for aspiring players.

2 Time Takes to Reach Top 20

2.1 Description of Models

In this analysis, we employ two Bayesian models to analyze the first investigation:

1. **Hierarchical Gaussian Model:** This model examines the time taken for players from different continents and countries to reach the top 20 ranks. We assume that the time is normally distributed and include random effects for continents and individual countries to capture variability across these groups.

2. **Hierarchical Gamma Model:** To account for the positive and skewed nature of the time-to-top-20 variable, we use a hierarchical Gamma model with a log-link function. Similar to the Gaussian model, this includes random effects for continents and individual countries.

The hierarchical structure allows for partial pooling, which balances individual country data with the overall regional (continental) trends. The models are expressed as follows:

$$\text{Gaussian Model: } y_i \sim \mathcal{N}(\mu_i, \sigma), \quad \mu_i = \alpha + u_{\text{region}[i]} + u_{\text{country}[i]}$$

$$\text{Gamma Model: } y_i \sim \text{Gamma}(\lambda_i, \theta), \quad \log(\lambda_i) = \alpha + u_{\text{region}[i]} + u_{\text{country}[i]}$$

where $u_{\text{region}[i]} \sim \mathcal{N}(0, \sigma_{\text{region}})$ and $u_{\text{country}[i]} \sim \mathcal{N}(0, \sigma_{\text{country}})$.

2.2 Priors

We use weakly informative priors to ensure that the models are stable while allowing the data to influence the results:

- **Intercept:**
  - For the Gaussian model, we use a prior of Normal$(500, 500)$, reflecting the expectation that time to top 20 typically spans around 500 days. The 500 as the standard deviation reflects high uncertainty and allows for a broad range of possible baseline values for how long it might take for players to reach the top 20, ranging from fast climbers (less than 500 days) to slow climbers (more than 500 days).
  - For the Gamma model, we use Normal$(7, 1)$ on the log scale. The mean of 7 corresponds to a rough estimate of 1000 days (since $\log(1000) \approx 7$). This suggests that, in the absence of other information, we expect the log of time to top 20 to be around 7, corresponding to a time around 1000 days (approximately 2-3 years). The standard deviation of 1 reflects moderate uncertainty in the log scale of this baseline value, allowing the model to adjust based on observed data without being too rigid.
- **Standard Deviations:**
  - For the Gaussian model, The standard deviations for the random effects of continents and countries are assigned Cauchy$(0, 1000)$, which is a common choice for hierarchical models to allow flexibility while avoiding overly large values. The scale 1000 gives the model flexibility to accommodate large variability in player progression. Since tennis is highly competitive, the variability between players (e.g., elite vs. mid-tier) could be large, making this broad prior reasonable.
  - For the Gamma model, we use Cauchy$(0, 2)$ on the log scale. A scale of 2 is chosen to indicate moderate variability in how players from different IOCs and regions progress to the top 20, while still allowing for substantial variation in player performance across IOCs and regions. This reflects a belief that player development across regions may vary, but not excessively.

These priors are selected based on domain knowledge and exploratory data analysis, balancing informativeness and flexibility.

2.3 Stan

The models are implemented using the `brms` package, which provides a high-level interface to `Stan`. Below is the code for the Gamma model:

```
bayesian_model_gamma <- brm(
  time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc),
  data = time_to_top_20_regional_nonzero,
  family = Gamma(link = "log"), # Gamma distribution with log-link
  prior = c(
    prior(normal(7, 1), class = "Intercept"),
    prior(cauchy(0, 2), class = "sd", group = "player_ioc"),
    prior(cauchy(0, 2), class = "sd", group = "region")
  ),
  chains = 4,
  iter = 2000,
```

```
  warmup = 1000,
  cores = 4,
  control = list(adapt_delta = 0.99, max_treedepth = 15)
)
```

Similarly, the Gaussian model was implemented with analogous syntax but using a Gaussian family:

```
bayesian_model <- brm(
  time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc),
  data = time_to_top_20_regional,
  family = gaussian(),
  prior = c(
    prior(normal(500, 500), class = "Intercept"),
    prior(cauchy(0, 1000), class="sd", group="player_ioc"),
    prior(cauchy(0, 1000), class = "sd", group="region")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  control = list(adapt_delta = 0.99, max_treedepth = 15)
)
```

2.4 MCMC Inference Options

The MCMC sampling was performed using `brms` with the following configurations:

- **Number of Chains:** 4 chains were run to ensure adequate exploration of the posterior distribution.
- **Iterations:** Each chain was run for 2000 iterations, with the first 1000 used for warm-up to allow the Markov chain to converge.
- **Control Parameters:** `adapt_delta` was set to 0.99 and `max_treedepth` to 15 to handle potential divergences and improve sampling efficiency.

The command used to fit the model:

```
bayesian_model <- brm(
  time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc),
  ...
)
```

Diagnostic plots such as posterior predictive checks (`pp_check`) and trace plots confirmed convergence and model fit for one of the models. PSIS-LOO cross-validation was performed to compare models.

2.5 Convergence Diagnostics

Convergence diagnostics are essential to ensure the reliability of the Bayesian model's results. Key metrics include R_{hat}, Bulk Effective Sample Size (Bulk_ESS), Tail Effective Sample Size (Tail_ESS), and the number of divergent transitions. Below we present the convergence diagnostics for both the Gamma and Gaussian models, with results visualized in figures 4 and 5.

2.5.1 Gamma Model Convergence Diagnostics

The convergence diagnostics for the Gamma model are summarized as follows:

$$\text{Family: Gamma}, \quad \text{Links: } \mu = \log, \quad \text{shape} = \text{identity}$$

Formula: time_to_top_20 $\sim 1 + (1|\text{region}) + (1|\text{player\_ioc})$

Data: 464 observations, 4 chains, each with iter = 2000; warmup = 1000; total post-warmup draws = 4000

The convergence diagnostics for the Gamma model, including R_{hat}, Bulk_ESS, and Tail_ESS, are visualized in figure 4. The results are generally acceptable with all R_{hat} values being 1. However, there were 9 divergent transitions after warm-up, which is highlighted in the warning message. This suggests that further tuning might be needed for optimal convergence.

```
> summary(bayesian_model_gamma)
 Family: gamma
  Links: mu = log; shape = identity
Formula: time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc)
   Data: time_to_top_20_regional_nonzero (Number of observations: 186)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~player_ioc (Number of levels: 41)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.07      0.06     0.00     0.21 1.00     2123     1672

~region (Number of levels: 5)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.15      0.17     0.00     0.63 1.00     1399     1071

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     7.40      0.13     7.17     7.69 1.00     1453      594

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     1.96      0.19     1.61     2.36 1.00     5936     2856

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
Warning message:
There were 9 divergent transitions after warmup. Increasing adapt_delta above 0.99 may help. See http://mc-sta
n.org/misc/warnings.html#divergent-transitions-after-warmup
>
```

Figure 4: Gamma Model Convergence Diagnostics

- R_{hat}: 1.00, 1.00, 1.00, 1.00: Shows good convergences of the Markov chains.
- Divergences: 9: A small number of divergences is usually okay, but it suggests that there may be regions of the posterior that are challenging for the sampler.
- Tree depth: 15: A higher value can improve sampling quality, but it also increases computation time.
- Tail ESS: 1672, 1071, 594, 2856: The chains have effectively sampled from the posterior. Larger ESS values are better because they imply that the chains have provided a wide and stable exploration of the posterior distribution.
- Bulk ESS: 2123, 1399, 1453, 5936: Similar to above.

Given the results from the Gamma model, we proceeded to attempt a Gaussian model to assess convergence more thoroughly.

2.5.2 Gaussian Model Convergence Diagnostics

The convergence diagnostics for the Gaussian model are summarized as follows:

Family: Gaussian, Links: μ = identity, σ = identity

Formula: time_to_top_20 $\sim 1 + (1|\text{region}) + (1|\text{player\_ioc})$

Data: 511 observations, 4 chains, each with iter = 2000; warmup = 1000; total post-warmup draws = 4000

The convergence diagnostics for the Gaussian model, including R_{hat}, Bulk_ESS, and Tail_ESS, are shown in figure 5. All values are satisfactory, with R_{hat} values close to 1 and adequate effective sample sizes. No divergent transitions were noted in the Gaussian model.

```
> summary(bayesian_model)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc)
   Data: time_to_top_20_regional (Number of observations: 206)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~player_ioc (Number of levels: 41)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)   169.66    110.84     8.05   419.61 1.00     1338     1941

~region (Number of levels: 5)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)   253.34    254.79     9.28   933.19 1.01     1398     1875

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept  1411.00    188.96   932.59  1718.19 1.00     2157     1498

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma  1028.72     52.68   929.15  1138.60 1.00     6146     2157

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
>
```

Figure 5: Gaussian Model Convergence Diagnostics

- R_{hat}: 1.00, 1.01, 1.00, 1.00: Shows good convergences of the Markov chains.
- Divergences: 0: Good sign. The absence of divergent transitions means the model is well-behaved.
- Tree depth: 15: A higher value can improve sampling quality, but it also increases computation time.
- Tail ESS: 1672, 1071, 594, 2856: The chains have effectively sampled from the posterior. Larger ESS values are better because they imply that the chains have provided a wide and stable exploration of the posterior distribution.
- Bulk ESS: 2123, 1399, 1453, 5936: Similar to above.

In summary, the convergence diagnostics for both models show that the Gaussian model was more stable and converged more easily than the Gamma model, which had 9 divergent transitions. Both models, however, show good convergence overall, with effective sample sizes and R_{hat} values indicating reliable results.

2.6 Posterior predictive checks

Posterior predictive checks are essential for assessing the quality of a Bayesian model's fit to the observed data. These checks involve comparing the observed data with data simulated from the posterior predictive distribution. The goal is to identify discrepancies between the observed data and the model's predictions, which could indicate model misspecification.

We conducted posterior predictive checks for both the Gamma and Gaussian models. The results for both models are visually represented in Figures 6 and 7. The visual checks suggest that while both models capture the general patterns in the data, the Gaussian model provides a better fit, especially in the peak regions of the distribution. In particular, the Gaussian model better accounts for the observed

variability in the data, especially in the left tail, where the Gamma model seems to under-predict the frequency of extremely low values.

This difference in performance can be attributed to the nature of the distributions. The Gaussian model, with its symmetric bell-shaped curve, provides a more flexible fit for data that may not be heavily skewed. In contrast, the Gamma model, being skewed by design, struggles to fit the left tail adequately, as extreme cases, such as young tennis prodigies achieving rapid success, deviate significantly from the norm. These rare cases—such as players like Carlos Alcaraz or Coco Gauff, who reached the Top 20 at remarkably young ages—represent outliers that can distort the Gamma model's predictions, as they rise quickly through exceptional talent, physical readiness, and mental toughness.

Given that the Gaussian model more accurately captures the distribution of the data, especially in terms of the left tail where rare rapid risers occur, it is considered the more appropriate model in this case. If posterior predictive checks had indicated substantial issues with either model, we would have considered further model refinements, such as exploring alternative structures, including additional covariates, or experimenting with different priors. However, as the Gaussian model performed better in posterior predictive checks, no further adjustments were necessary.

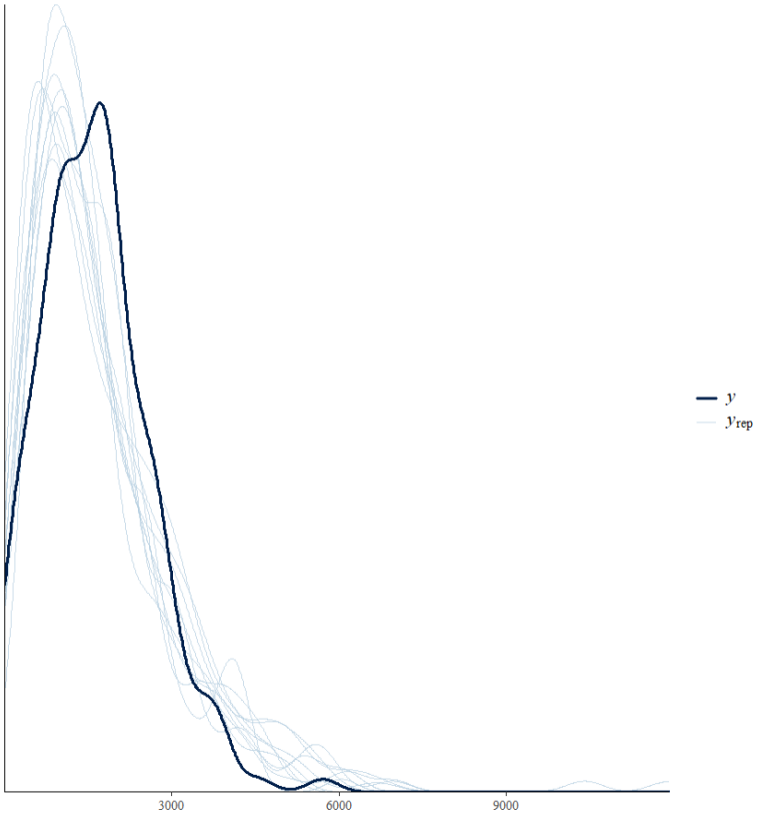

Figure 6: Posterior Predictive Check for the Gamma Model

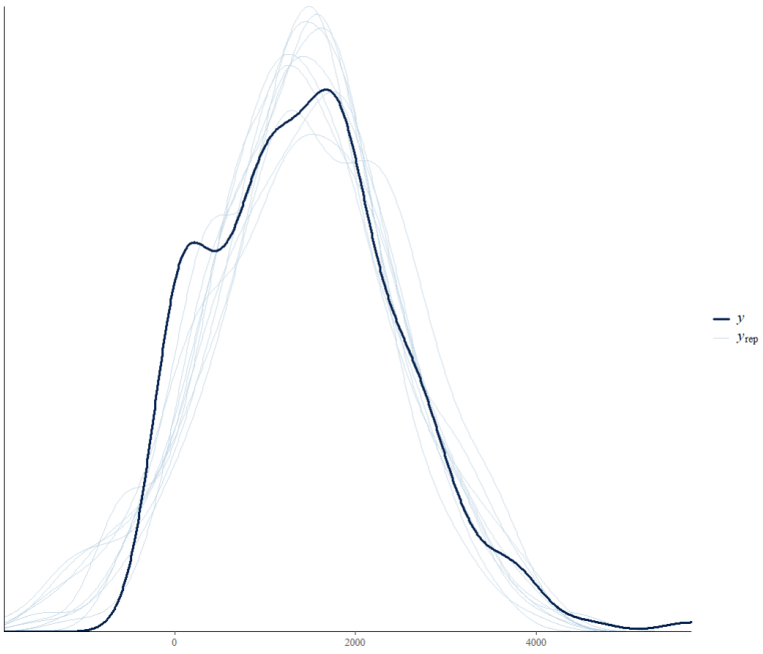

Figure 7: Posterior Predictive Check for the Gaussian Model

2.7 Sensitivity analysis and Model comparison

Sensitivity analysis is an important step in Bayesian modeling to evaluate the robustness of the results to different prior choices. In our case, we assessed the sensitivity of both the Gamma and Gaussian models to changes in prior distributions. We found that, for both models, the results remained stable when reasonable priors were used. However, when poorly chosen priors were applied, there was a significant shift in the model results, indicating the importance of prior specification in Bayesian analysis.

To quantitatively assess model performance, we conducted Leave-One-Out Cross-Validation (LOO-CV). LOO-CV provides an estimate of the model's predictive performance by computing the expected log pointwise predictive density (elpd_loo). We present the results of LOO-CV for both the Gamma and Gaussian models, which show similar elpd_loo values when reasonable priors are applied. Specifically, for the models with good priors, the elpd_loo value was approximately -1725.4, as shown in Figure 8.

However, when bad priors were used, the model's performance deteriorated, resulting in a significantly worse elpd_loo value of -3773.8, as shown in Figure 9. This highlights the importance of careful prior selection in Bayesian modeling and its effect on the predictive performance of the model.

```
> print(loo_result_)

Computed from 4000 by 206 log-likelihood matrix.

         Estimate   SE
elpd_loo  -1725.4 11.8
p_loo         8.5  1.3
looic      3450.8 23.7
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.7, 1.6]).

All Pareto k estimates are good (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

Figure 8: LOO-CV with **Good Priors** for the Gaussian Model (elpd_loo = -1725.4)

```
> print(loo_result_)

Computed from 4000 by 464 log-likelihood matrix.

         Estimate   SE
elpd_loo  -3773.8 15.6
p_loo         6.6  0.6
looic      7547.7 31.2
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.7, 1.6]).

All Pareto k estimates are good (k < 0.7).
```

Figure 9: LOO-CV with **Bad Priors** for the Gamma and Gaussian Models (elpd_loo = -3773.8)

However, interestingly, the LOO results suggest that the Gamma model performs better in terms of predictive accuracy, with an elpd_loo value of approximately -1544, compared to -1725 for the Gaussian model.

```
> print(loo_result)

Computed from 4000 by 186 log-likelihood matrix.

         Estimate   SE
elpd_loo  -1544.0  8.9
p_loo         4.9  0.7
looic      3088.1 17.8
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.7, 1.4]).

All Pareto k estimates are good (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

Figure 10: LOO-CV with **Good Priors** for the Gamma Model (elpd_loo = -1544.0)

The discrepancy in the LOO results can be attributed to the way each model handles the distribution of the data. While the Gaussian model provides a better fit to the observed data in terms of the posterior predictive checks, the Gamma model appears to generalize better. The Gamma model's skewed distribution might be less sensitive to overfitting or noise in the data, making it more robust for predictive accuracy in terms of out-of-sample data.

This outcome indicates that the Gamma model, despite not fitting the data as well in the peak and tail regions, offers a more stable prediction across different subsets of the data. Thus, even though the Gaussian model might be more accurate in modeling the data's distribution, the Gamma model performs better on the LOO-CV, suggesting that it might be more suitable for future predictions.

2.8 Discussion of Issues and Potential Improvements

While the posterior predictive checks for the Gamma and Gaussian models generally indicated a reasonable fit to the data, some issues remained that could potentially affect the precision of the model's predictions. Specifically, the plots revealed subtle discrepancies, particularly regarding out-of-the-distribution observations and the influence of young prodigies.

Out-of-the-Distribution Observations: The left tail of the distribution did not capture the observed data well. This is likely due to the presence of young prodigies, such as Carlos Alcaraz and Coco Gauff, who defy typical player development trajectories. These exceptional cases, where players

rapidly ascend to the top ranks, are outliers that significantly affect model predictions. Since these instances are rare but impactful, they present a challenge for standard distribution-based models, which struggle to represent such extreme events accurately.

Improving the Model: To address the above issues, potential improvements to the model could involve incorporating more complex distributional assumptions that can better account for outliers, such as truncated or heavy-tailed distributions. Alternatively, robust regression techniques could be explored to mitigate the impact of these extreme observations on the overall model fit. Furthermore, the inclusion of covariates that explicitly account for player age, talent, or early career milestones could help refine the model's predictions and better capture the behavior of young prodigies.

Posterior Predictive Checks Improvement: Another way to enhance the model's performance would be to improve the posterior predictive check plots by refining the model structure. A more detailed examination of individual player characteristics, such as training intensity or early career achievements, could help in adjusting the model to better fit outlier data points, leading to more accurate and reliable predictions.

In conclusion, while the models provided useful insights into player performance across different regions, addressing these issues could further enhance their accuracy and robustness, particularly when considering rare and extreme cases.

2.9 Conclusion

Based on the observations and posterior predictive checks, we can conclude that the impact of the country is more significant compared to the region (continent) when explaining the time it takes for players to reach the top 20.

From the posterior plots, we observe that countries such as the United States (USA) and Switzerland (SUI) exhibit higher frequencies of negative values, indicating that players from these countries tend to reach the top 20 faster compared to players from other countries. This is particularly evident from the figures shown for different countries in Figures **??** and 11.

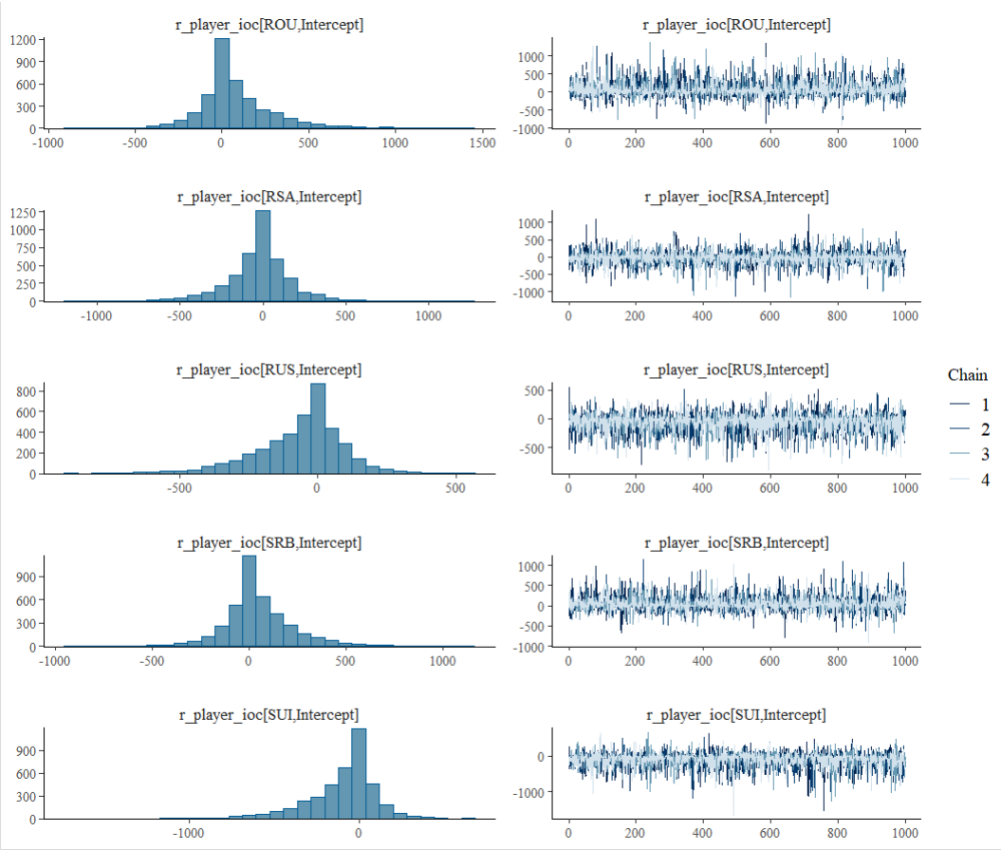

Figure 11: Player IOC (International Olympic Committee) effects on time taken to reach top 20

Regarding regions (continents), the results are more uniform. Based on the plots, we can generally state that Asia, Africa, and Europe show similar patterns in terms of the time it takes players to reach the top 20. However, players from America (North and South America combined) appear to reach the top 20 in less time on average, as indicated by the posterior distributions for players from that region.

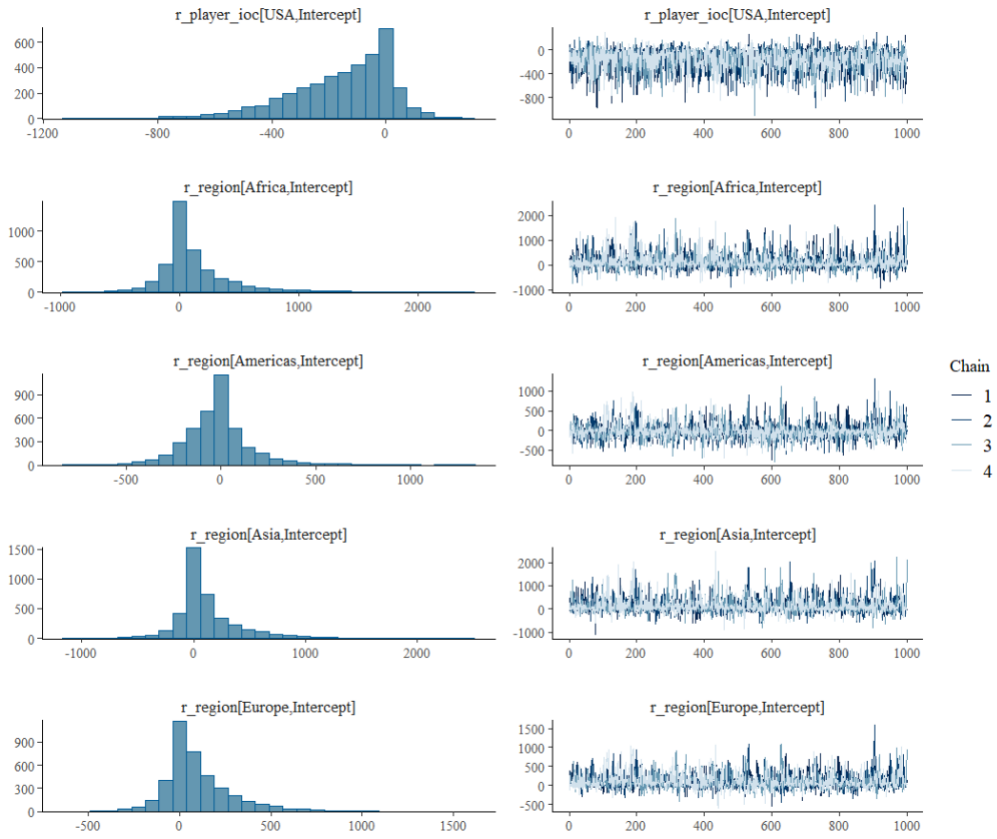

Figure 12: Player IOC (International Olympic Committee) and region effects on time taken to reach top 20

In summary, while the regional differences in the time to reach the top 20 are relatively small, significant country-level effects, particularly from the USA and SUI, are observed. These findings highlight the importance of considering the country as a more influential factor in predicting the time for players to reach top rankings in tennis.

2.10 Self-reflection

By doing analysis on an entirely new and unique project, we learned many things that span from teamworking skills - such as GitHub, conversating our own points of view and reasoning, and how to divide the tasks in a parallel way - to coding skills such as checking different priors for our model, choosing different model families, checking how well they do and overall this sense of trial and error combined with human reasoning.

3 (Extra) Win Rates of Players from Various Regions

3.1 Model Description, Priors, and MCMC Inference Options

In this section, we outline the methodology used for investigating the win rates of players from various regions on different court types. This analysis is an additional investigation and will not be as in-depth as we discussed for the first investigation. We provide this model as an extension to gain further insights into surface-specific advantages and how players from different regions perform on these surfaces.

3.2 Data Preparation

The data preparation steps are crucial for setting up the model. The data were first enriched by adding region information to the winner and loser data, using the 'region-map' which maps players' countries

to their respective regions. This was done by joining the 'region-map' with the winner and loser data on the respective player IOC codes. The resulting dataset now includes regional information for both winners and losers.

Next, we focused on the match outcomes on different surfaces. The surface data was created by combining the winner and loser data into one dataset, where each match was labeled as a win (1) or a loss (0), with corresponding surface information. After filtering out missing values, we calculated the win rate for each player on each surface.

The surface data was then merged with region information, ensuring that only valid entries with region data were included. This dataset was then grouped by region and surface, and we collected the win rates for each region-surface combination. The final dataset, referred to as 'bayesian-data', contains the win rates for players from different regions on different surfaces, ready for Bayesian modeling.

3.3 Bayesian Model

For each region-surface combination, we fitted a separate Bayesian model to estimate the win rate distribution. The model used is a simple group-level model where the win rate for each region-surface combination is modeled as a single group parameter. Specifically, the model follows the form:

$$\text{win\_rates}_{region,surface} \sim \text{Normal}(\mu, \sigma)$$

where μ is the mean win rate for each region-surface combination, and σ is the standard deviation representing the variability in win rates.

3.4 Priors

The prior distributions for the parameters of the model were chosen based on standard practice for Bayesian modeling, while ensuring reasonable flexibility in the estimates:

- The intercept (μ) was assigned a normal prior with a mean of 0.35 and a standard deviation of 0.35, reflecting a wide range of possible win rates around the assumed average win rate of 0.35.
- The standard deviation (σ) was assigned a Cauchy prior centered at 0 with a scale of 0.3, providing a heavy-tailed prior that allows for potential larger variability in win rates across regions.

3.5 MCMC Inference Options

The models were fitted using Markov Chain Monte Carlo (MCMC) with the following settings:

- Four chains were used to ensure convergence of the model.
- Each chain ran for 2000 iterations with 1000 warm-up steps, ensuring enough post-warm-up samples for inference.
- The number of CPU cores used for the parallel execution of MCMC was set to 4, allowing for faster computation.
- A random seed of 123 was set to ensure reproducibility of the results.

3.6 Model Estimation and Posterior Analysis

After fitting the models, we summarized the results for each region-surface combination. The posterior distributions for the mean win rate (μ) and the standard deviation (σ) were examined. These distributions were used to assess the surface-specific advantages and identify potential patterns in win rates for players from different regions.

Additionally, we visualized the posterior distributions of win rates across regions and surfaces to compare the effectiveness of players from different regions on various court types. The plots provide insight into the surface-specific advantages for players from different regions and can help identify trends in performance across surfaces.

In the next section, we present the results from these models and discuss the insights they offer regarding surface-specific advantages for players from different regions.

3.7 Results and Conclusion

The analysis of win rates for players from different regions on various surfaces yielded results that align with our expectations, particularly in the case of Asian players. In Asia, where the majority of tennis courts are clay courts and players tend to play predominantly on this surface, we observe that Asian players perform better on clay courts compared to other surfaces. Specifically, the mean win rate for Asian players on clay courts is 0.39, which not only surpasses their performance on other types of courts (with differences ranging from 0.05 to 0.11) but also outperforms players from other regions when playing on clay courts.

This result is consistent with the general playing conditions in Asia, where the familiarity and dominance of clay courts likely provide an advantage for Asian players. The strong performance of Asian players on clay is clearly reflected in the posterior distributions of win rates, where the clay court win rate for Asian players stands out.

To better visualize these findings, we present a plot that shows the posterior normal distributions of win rates by region and surface. This figure highlights the difference in performance across regions, with Asian players showing a higher win rate on clay compared to players from other regions, further reinforcing the impact of the type of court on player performance.

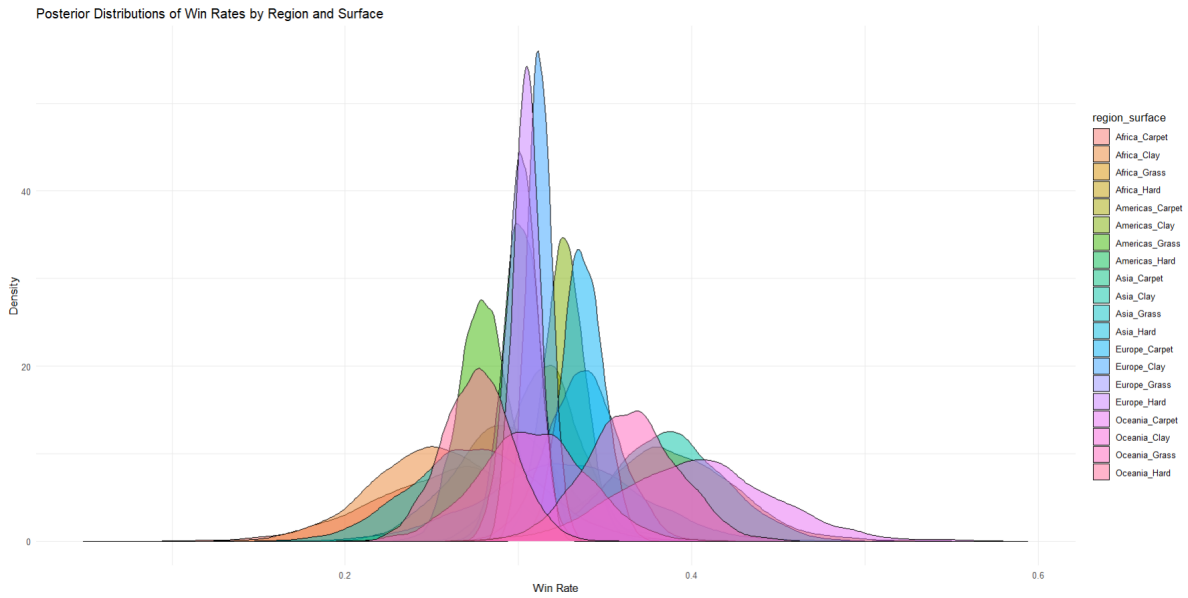

Figure 13: Posterior Distributions of Win Rates by Region and Surface. The plot highlights the performance of players from different regions on various surfaces, with Asian players showing higher win rates on clay courts.

Overall, this analysis confirms that surface-specific advantages are significant in determining player performance. The advantage of playing on familiar surfaces, like clay courts for Asian players, not only contributes to higher win rates but also illustrates the importance of surface preferences in professional tennis. These findings could have practical implications for training and tournament preparation, as players may perform better on surfaces they are most accustomed to.

In conclusion, our investigation provides useful insights into the regional and surface-specific performance of players. The observed higher win rates for Asian players on clay courts reflect the broader trend of surface familiarity influencing player success. Future studies could further explore these relationships, including how player preparation and training on specific surfaces contribute to these performance disparities.