Assessing the Impact of Relative Age on Professional Tennis Players' Rankings

Laurio Johannes, Musio Cristian, Slam Yeralkhan

Spring, 2024

1 Introduction

1.1 Background

The relative age effect (RAE), also known as birthdate effect, is a phenomenon where individuals born earlier in the year are thought to have an advantage over those born later in that same year. This effect has been observed in many fields, from sports to academia. The concept behind it is that, during childhood, individuals born earlier are on average more physically and cognitively developed than those born later in the year, and this advantage would compound over time leading to effects observable in adulthood.

An example of this is in sports where the cohort cutoff is January 1st, such as hockey. Here, kids born earlier in the year typically exhibit better performance thanks to simply being a few months older than their younger peers, which at a young age can be a substantial strength difference. In fact, since at the very early stages of kids' sports there's not much skill involved yet, this initial greater physical strength and performance earns them more time in the field as coaches deem them to have higher potential, hence accumulating more practice. As they grow into their teenage years and early adulthood, the compound effect of the extra practice time allows them to outpace the progress of their peers, resulting in a skewed distribution of the birth months of professional players. In the hockey field, this has been suggested by Barnsley and Thompson (1985)[1], who found that 40% of NHL hockey players were born in the first three months of the year.

However, hockey is a team-based sport, so we suggest that a better case study would be an individual-based sport because we believe it would in theory reduce confounding variables and noise, thus more clearly isolating the RAE. This is why we chose tennis.

Moreover, given that the RAE has already been observed in terms of the skewed odds of becoming professional, we are interested in seeking whether the RAE also affects performance once they reach professional level. In other words, we will discuss both the RAE on becoming professional, and the RAE on performance of the professionals. The latter has been source of less research in the literature, hence our interest.

To address this, we aim to answer our research question: "does the month of birth have any impact on the ranking of professional tennis players?". Our initial hypothesis is that, consistent with Barnsley and Thompson (1985), the distribution of professional tennis players should be skewed towards the beginning of the year and that the ranking of players born earlier in the year should be lower (i.e., better - rank number 1 being the best in the world) than those born later, indicating better performance.

1.2 Motivations

Why bother? The relative age effect can be interpreted as one of the ways luck plays a role in success, and we are interested in quantifying this phenomenon and its effects in the tennis sport because it can help shed more light on instances of survivorship bias. In fact, we believe successful players might sometimes have a biased view of the impact luck played on their opportunity to show their hard work. It's safe to assume, most of them probably do not attribute much value to the month they were born in, let alone attribute it even the smallest percentage of impact on their professional careers. Of course, the compound effect of hard work cannot be replaced, but that's an already well known fact. What is less explored however, is the compound effect that luck can have. The birth month of an individual, from their perspective, can be seen as complete luck (or unluck) and random, given that they had no involvement in this decision at all. Therefore, it serves as a good measurement of something outside of their control. How important is luck in success? Can it be quantified? This is obviously a question that can't be fully answered by a single study because luck shows in many different ways (country of birth, parents, decade of birth, etc), and it's not the goal of this project. Our scope is simply to have a glimpse of one way this question could be addressed in one specific limited area, to set the ground for future research to progressively cover more and more areas of this difficult question.

2 Literature Review

2.1 Evidence of RAE

Barnsley and Thompson (1985) conducted a study on NHL players, one of the first relative age effect studies that we were able to find. They've found that 40% of professional hockey players were born in the first 3 months of the year.

Consistent with those results were the findings of Edgar and O'Donoghue[7], who conducted a study on elite tennis players' season of birth and found a skewed distribution towards the beginning of the year. Furthermore, research has been done specifically on junior tennis by Agricola, Zhanel, Hubacek at King's College London[6], who studied this effect on schoolchildren as well.

A study on the relative age effect in the Finnish parliament was also conducted, and it was published in the European Journal of Political Economy. This study did find that for males specifically, being born early in the year was correlated with higher likelihood to be selected into the parliament.

These are just a few references, as since the Barnsley and Thompson study, many others have researched the topic in a variety of fields.

3 Data

3.1 Source

To measure the performance of professional tennis players, we use their Association of Tennis Professionals (ATP) ranking. The data was mined from the official ATP website by GitHub user Jeff Sackmann[5].

3.2 Why this data

Performance can be a tricky variable to measure, so it was critical to find data that accurately represents it. ATP rankings are widely accepted by tennis federations around the world as a good indication of players' performance, so we decided to use it as our variable of interest. In the ATP rankings, points are gained or loss proportionally to the importance of the game (world cup final, friendly game, etc.) and on the ranking of the opponent, similar to the Elo ranking system in chess or the FIFA ranking in football. The more a game is important, the more points are gained by winning, and the bigger the difference in rankings, the more points are gained by the lower-ranked player if he wins. Note that we're using ATP rankings, rather than points, and therefore a lower ranking (i.e., higher points) is indicative of better performance. The reason we're using APT rankings as outcome variable and not the win rate of each player, is because the win rate is not as good of an indicator of performance. For instance, a player could have a high win rate and still be a worse performer than another player simply by having played less games overall making it easier to have a higher win rate, or by luck (resulting from a small sample size for example).

3.3 The data processing

The match data we used ranged from 2000-2019. First, ATP games datasets for each year was merged into master_data. Then, dataset of tennis players consisting information about player id, hand, birthdate was merged with "master_data". It is called "seems_ok_2.dta". It includes information about every game, its winner, loser, each player's birthdate. From that we found winner_id_freq(number of wins for each player), loser_count(number of losses for each player), num_of_games, win_r(win rate of each player). The measured variable, ATP ranking, was denoted by the lowest (i.e., best) rank achieved by a player during their career. This allowed us to compare the optimal, peak performance of players and control for injuries or other time-variant effects. Essentially, this means that the interpretation of the graph below is "what is the maximum potential that this specific player was able to reach during their career?". Subsequently, for the sake of simplicity, all unnecessary observations were dropped so that there is only one observation for each player. Thus, overall 1315 observations, meaning 1315 players.

4 Methods

4.1 Empirical model

To address our research question and test our hypothesis, we opted for a Regression Discontinuity Design (RDD) approach because in tennis junior leagues there's a cohort cutoff on January 1st[2]. This cutoff date is used by the International Tennis Federation[3]. This would be a sharp RDD, given that the assignment to the treatment group (being born earlier in the year) is deterministic. Our running variable, or score, is the normalized date of birth such that January 1st has score 1 and December 31st has score -1. The rest of the dates are scaled accordingly. This assures discontinuity, as the cutoff score of 0 is between those two dates.

Therefore, our model is as follows:

$$D_i = \begin{cases} 1 & \text{if } x_i - c \geq 0 \\ 0 & \text{if } x_i - c < 0 \end{cases}$$

$$y_i = \beta_i + \tau D_i + \beta_1 D_i(x_i - c) + \beta_2(1 - D_i)(x_i - c) + \epsilon_i$$

y_i is our outcome variable (i.e., the variable of interest: the lowest ATP ranking achieved by each player)

x_i is the score (i.e., standardized date of birth)

D_i is the treatment status of individual i (i.e., being born earlier in the year)

β_i is the constant term

β_1, β_2 are the parameters of the linear control of the score

τ is the causal effect of the treatment on the outcome variable

c is the cutoff (January 1st)

ϵ is the error term

4.2 Assumptions

Our approach fulfills the key points and assumptions of RDD. For this information, we used the lecture slides from the Principles of Empirical Analysis course, by Matti Sarvimäki:
"Units that receive very similar score values on opposite sides of the cut-off are comparable to each other in all relevant aspects, except for their treatment status."
In our study:
1. For tennis players around the cut-off we obtain conditions that mimic a randomized experiment, meaning that units on each side of the cut-off are as good as randomly assigned to either receive treatment or not. For instance, players born December 31 in year n, and players born January 1st of the year n+1, have essentially one day difference, so they are extremely similar in their potential outcomes, but they are assigned to different groups. Their assignment to treatment can be considered as good as random, since their birth could've been delayed or postponed by one day by randomness.
2. There is a continuity of average potential outcomes near the cut-off. In fact, there's no reason to believe that a one day difference would in of itself cause a discontinuity in the potential outcome (ATP ranking) of a player.
3. The method assumes that entities cannot manipulate their assignment status around the cutoff to receive or avoid the treatment. If units could precisely determine their scores and thus choose to

receive treatment or not, then the running variable around the cutoff would not be like a randomizer anymore. In our case, birthdate cannot be manipulated neither by the player, nor by external factors.

4.3 Controls

We're using the lowest (as in the best) ATP ranking achieved by each player over their entire career to control for time specific events, such as injuries, declining performance due to age, or simply a player having a bad year. We also tried controlling for some other variables, such as hand and surface. However, the results were statistically insignificant. Thus, in the study's empirical model and results those RDD outputs were not included as we do not intend to venture into p-hacking territory.

Something we also considered accounting for was the disruption caused by leap years. As in our dataset, spanning 19 years, there would have been 5 leap years, which means the dates after February 28th are essentially "pushed" one unit further. The way that the Finnish parliament study from 2019 controlled for these leap years was through averaging out their effect, by making each year 365.25 days long. We tried this, and the effect on the results was negligible, so we decided to omit this change as it could be considered manipulation of the results, and a more advanced method of controlling for the leap years might extend the scope of our study.

5 Results

5.1 Economic and Statistical Interpretation

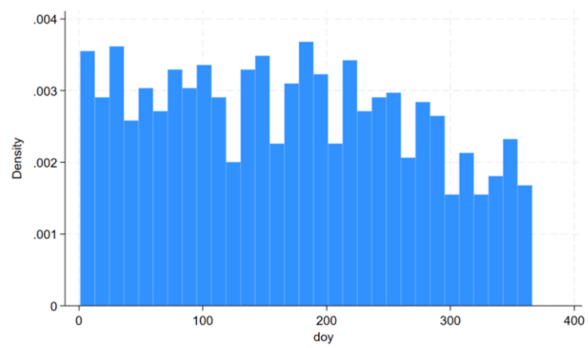

Figure 1: Birthdate distribution

Figure 1 depicts the distribution of the dates of birth, which allows us to visually confirm a clear trend indicating that the frequency of professional players being born in the late months (days ≥ 300) is clearly lower, out of 1315 players.

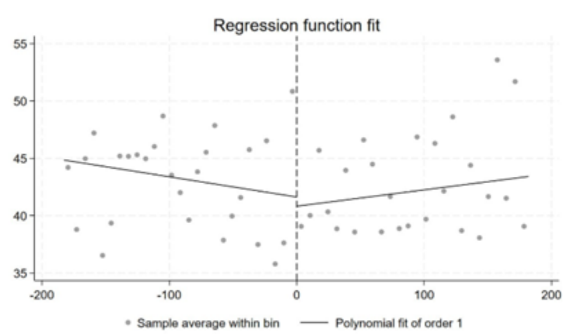

Figure 2: Win Rate RDplot

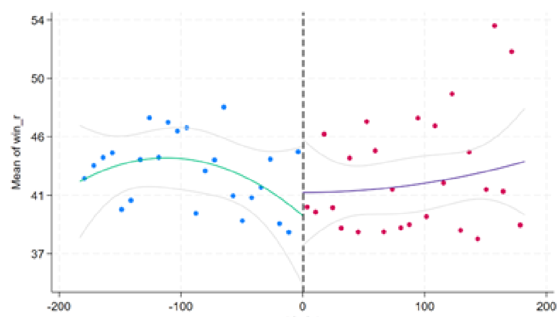

Figure 3: Win Rate RDplot Quadratic fit

From figures 2 and 3 above, we can observe that there is no significant effect on win rate based on the date of birth. Actually, it appears that players born both early and late in the year, on either side of the cutoff have comparable win rates. As previously mentioned in section 3.2, win rate is not a good variable to use, as the p-value of the coefficients of regression are higher than 0.05. (see Appendix 1).

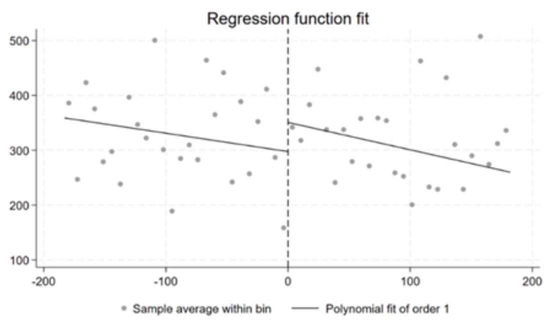

Figure 4: Winner Rank RDplot

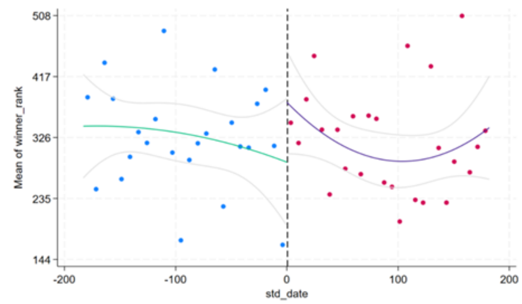

Figure 5: Winner Rank RDplot Quadratic fit

With our sample size of 1315 observations, we used a triangular kernel in our regression for higher efficiency. Figure 4 shows a polynomial fit of order 1, and figure 5 a second order. The average bin length is 7 days on either side, and the bandwidth is chosen by MSE-optimal bandwidth method. These optimal values are automatically chosen by the rdrobust Stata command. We notice that our results suggest an opposite effect to that of our hypothesis. Players born later in the year (left of the cutoff) seem to have on average better performance (lower ranking), than those born earlier in the year (right of the cutoff). This is shown by the line bump at score 0, indicating a jump in rankings as we cross from December to January. At the cutoff, Average Treatment Effect is 215 player rankings (i.e. going down 215 spots in rankings). The decreasing line also shows that the ranks decrease as the birth months approach the end of the year. The results are statistically significant, given that the coefficients have a p-value less than 0.05 (see Appendix 2).

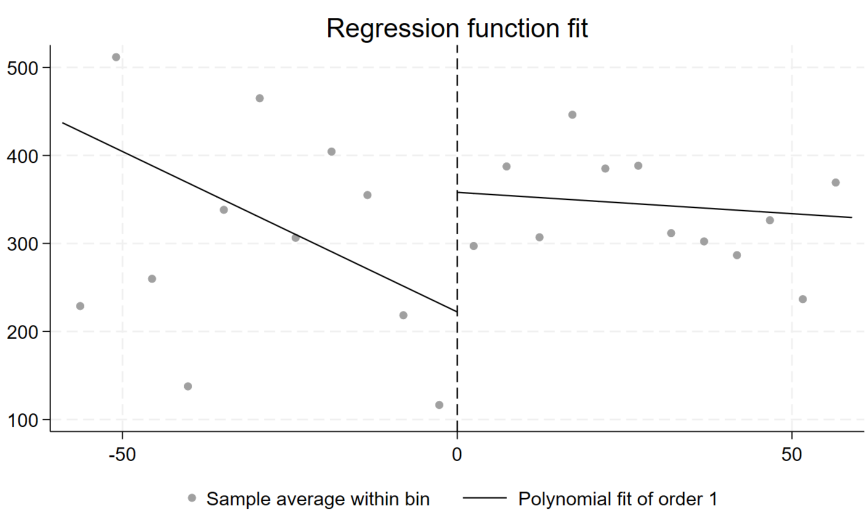

Figure 6: RD Plot 60 day interval at the cutoff

As this graph with a narrower +/- 60-day interval suggests (figure 6), the jump at the cutoff is actually even more prominent. This further supports the significance of our results, indicating that when we isolate the effect to be only around the beginning and end of the year, it appears less noisy, which goes along with intuition.

5.2 Career length

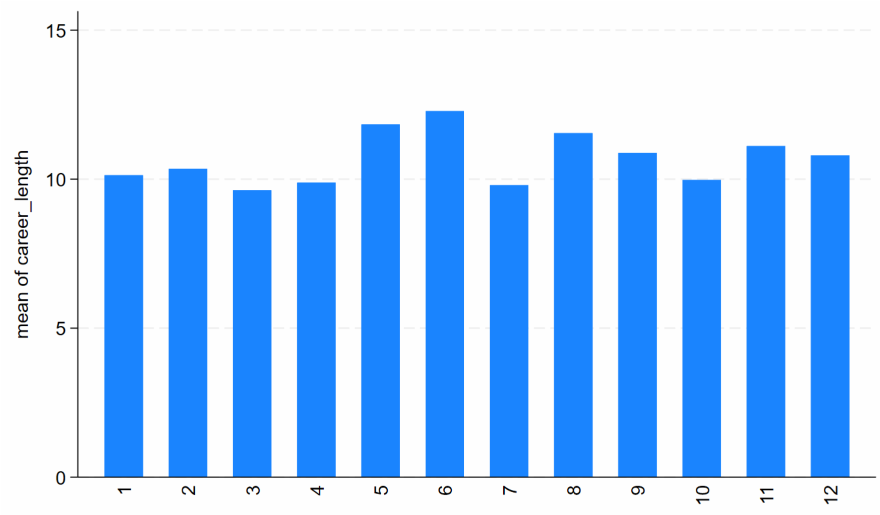

Figure 7: Career Length in Tennis

In our study, we did not find differences in the distribution of mean career length (in years) by birth month. This suggests that being born earlier or later in the year doesn't seem to affect the career length (figure 7). However, in hockey, Gibbs, Jarvis & Dufur (2011)[8], found that players born in the last quarter of the year have on average longer careers, and those born in the first quarter have the shortest average career lengths. Their findings are depicted in figure 8 below. This phenomenon might be sport-specific and warrants further research.

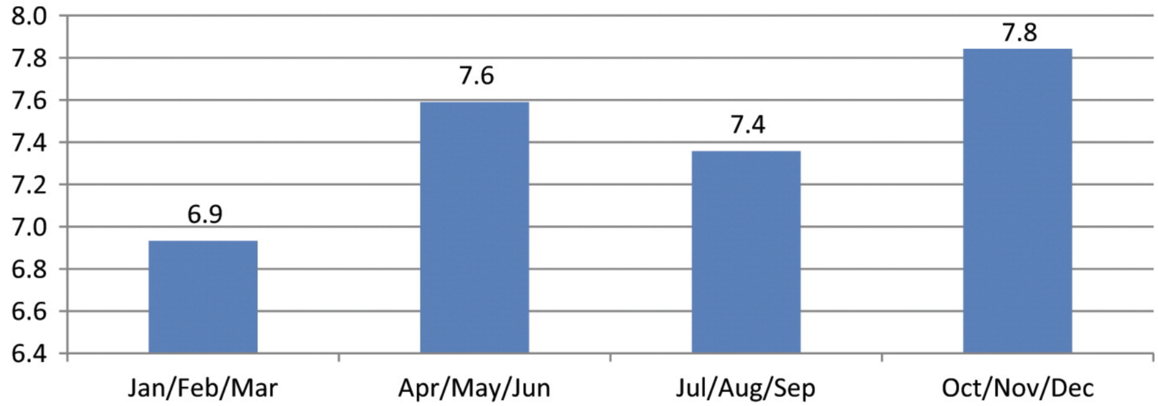

Figure 8: Career Length in Hockey, findings by Gibbs et al.

5.3 Socio-economic interpretation

In terms of the socio-economic interpretation, The relative age effect has important implications for athlete development and performance. Socio-economic factors play a role in shaping the RAE, as children born early in the year may have better access to coaching, training facilities, and competitive opportunities, which can lead to early success thanks to their early physical and cognitive advantages over their peers. These advantages can compound over time, reinforcing the RAE. Furthermore, coaches and talent scouts often select players based on age group eligibility, favoring those born earlier in the year. This can lead to athletes born later in the year having to face challenges in "catching up", leading to potential dropout rates.

6 Discussion – criticizing the model

In some countries, such as the UK, the junior cohorts span two years. So, an individual born in year n can choose to play in cohort (n;n+1) or in cohort (n-1;n). Our model assumes that each individual is rational and would choose to play in the cohort where they are more likely to be older and would therefore only be indifferent between the two cohorts if they were born exactly in the middle. This could cast doubts on our design's validity, as we assume the relative age effect is present year-by-year. However, in the UK example, a child that is physically more developed than their peers may be put into a higher age group, and vice versa. This diminishes some of the natural "hierarchy" that the relative age effect is trying to capture.

	January – December 2024
Age Group	For players born in...
8U (Red)	2016 or later
9U (Orange)	2015 or 2016
10U (Green)	2014 or 2015
11U	2013 or 2014
12U	2012 – 2013
14U	2010 – 2013
16U	2008 – 2013
18U	2006 – 2013

Figure 9: Age Group allocation in the UK

7 Conclusions – elements for further research

To answer our research question, our findings suggest that professional tennis players are more likely to be born towards the beginning of the year (as shown by the birthdate distribution in figure 1), but that, opposite to our hypothesis, they are not likely to have better rankings compared to players born later in the year. In fact, the opposite seems to be the case: those born later in the year seem to have better rankings instead. This effect is referred to as the "underdog effect" by Gibbs et al[8], who found similar results to us in hockey.

This effect could arise for a multitude of reasons. Firstly, it makes sense that if our background hypothesis is right in regards to the relative age effect being prevalent amongst youngsters, this would mean that children being born later on in the year have a tougher time getting to a high level of play, aka playing professionally. It would then somewhat logically follow, that the ones who DO end up in the highest echelon of competition end up "going the distance" and finding successful careers, as they are already used to fighting as underdogs and working hard to stay relevant their entire life. In other words, nothing was "handed to them" when they were younger, so they are able to utilize their skills and perhaps are more motivated to push themselves to the limit when older.

To take us back to the literature review section in regards to ice hockey, a similar trend has been suggested: "Despite the evidence of the RAE in ice hockey, there are reasons to question the long-term advantages of being born earlier in the competition year. In that respect, recent research suggests that the benefits of the RAE may be mostly short term, and late-born athletes could benefit from a reversal effect, in which it becomes possible for them to increase their potential to progress at a higher level." (Gibbs et al., 2012). This is in line with the results we showed in section 5.1, and supports this suggested "underdog hypothesis", where the relatively younger players actually succeed in the long-term due to struggling with disadvantages in comparison to their older counterparts at a young age.

To validate these results, further research in the category of individual sports would be preferred. Additionally, if we want to generalize even further, we could look at other fields such as academia or politics, although quantifying everything in a comparable manner would pose a great challenge. For sports, it would be interesting to look at something like MMA or boxing, which is arguably even more dependent on physical traits especially at a young age, than tennis.

8 Appendix

```
. rdrobust win_r std_date, kernel(triangular) c(0) p(1) masspoints(adjust) all
Mass points detected in the running variable.

Sharp RD estimates using local polynomial regression.

     Cutoff c = 0 | Left of c  Right of c            Number of obs =       1315
------------------+-----------------------            BW type       =      mserd
    Number of obs |       589         726             Kernel        = Triangular
Eff. Number of obs|       121         201             VCE method    =         NN
   Order est. (p) |         1           1
   Order bias (q) |         2           2
      BW est. (h) |    49.993      49.993
     BW bias (b)  |    84.225      84.225
       rho (h/b)  |     0.594       0.594
       Unique obs |       171         178

Outcome: win_r. Running variable: std_date.
--------------------------------------------------------------------------------
           Method |   Coef.    Std. Err.     z     P>|z|    [95% Conf. Interval]
------------------+-------------------------------------------------------------
     Conventional | -4.5238     4.5625   -0.9915  0.321   -13.4661      4.41861
   Bias-corrected | -6.0199     4.5625   -1.3194  0.187   -14.9622      2.92251
           Robust | -6.0199     5.2944   -1.1370  0.256   -16.3967      4.35692
--------------------------------------------------------------------------------
Estimates adjusted for mass points in the running variable.
```

Figure 10: Appendix 1

```
. rdrobust winner_rank std_date, kernel(triangular)  bwselect(mserd) c(0) p(2) masspoints(adj
> ust) all
Mass points detected in the running variable.

Sharp RD estimates using local polynomial regression.

     Cutoff c = 0 | Left of c  Right of c            Number of obs =       1315
------------------+-----------------------           BW type       =      mserd
    Number of obs |       589         726            Kernel        = Triangular
Eff. Number of obs|       148         251            VCE method    =         NN
   Order est. (p) |         2           2
   Order bias (q) |         3           3
      BW est. (h) |    60.645      60.645
      BW bias (b) |    87.555      87.555
        rho (h/b) |     0.693       0.693
       Unique obs |       171         178

Outcome: winner_rank. Running variable: std_date.
--------------------------------------------------------------------------------
           Method |   Coef.    Std. Err.    z     P>|z|    [95% Conf. Interval]
------------------+-------------------------------------------------------------
     Conventional |  215.34     87.571   2.4590   0.014    43.6989      386.972
   Bias-corrected |  225.24     87.571   2.5721   0.010    53.6055      396.879
           Robust |  225.24     98.716   2.2817   0.023    31.7629      418.721
--------------------------------------------------------------------------------
Estimates adjusted for mass points in the running variable.

.
```

Figure 11: Appendix 2

References

[1] Barnsley, Roger & Thompson, Angus & Barnsley, P.E.. (1985). Hockey success and birthdate: The relative age effect. Canadian Association for Health, Physical Education, and Recreation. 51. 23-28.
https://www.researchgate.net/publication/284328248_Hockey_success_and_birthdate_The_relative_age_effect

[2] UK Junior Competition Age Groups. LTA.
https://www.lta.org.uk/compete/junior/competition-age-groups/

[3] "How old do I have to be to compete on the 2023 ITF World Tennis Juniors?". International Tennis Federation.
https://www.itftennis.com/en/about-us/organisation/faqs/?category=world-tennis-tour-juniors&type=itf-tours

[4] Janne Tukiainen, Tuomas Takalo, Topi Hulkkonen, Relative age effects in political selection, European Journal of Political Economy, Volume 58, 2019, Pages 50-63, ISSN 0176-2680,
https://doi.org/10.1016/j.ejpoleco.2018.10.003

[5] Sackmann, J. (2024). Tennis ATP [Source code]. GitHub.
https://github.com/JeffSackmann/tennis_atp

[6] Agricola, Adrián & Zhanel, Jiri & Hubáček, Ondřej. (2013). Relative age effect in junior tennis (male). Acta Gymnica. 43. 27-33. 10.5507/ag.2013.003.
https://www.researchgate.net/publication/277936019_Relative_age_effect_in_junior_tennis_male

[7] Edgar, S., & O'Donoghue, P. (2005). Season of birth distribution of elite tennis players. Journal of sports sciences, 23(10), 1013–1020.
https://doi.org/10.1080/02640410400021468

[8] Gibbs, B. G., Jarvis, J. A., & Dufur, M. J. (2012). The rise of the underdog? The relative age effect reversal among Canadian-born NHL hockey players: A reply to Nolan and Howell. International Review for the Sociology of Sport, 47(5), 644-649.
https://doi.org/10.1177/1012690211414343