Testing the Hypothesis that Tennis Points are Independent and Identically Distributed Using Statistical Methods

Ernesto José Ugona Santana

A Senior Thesis submitted in partial fulfillment
of the requirements for graduation
in the Honors Program
Liberty University
Spring 2024

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial fulfillment of the requirements for graduation from the Honors Program of Liberty University.

[redacted]

David E. Schweitzer, Ph.D.
Thesis Chair

[redacted]

Andrew H. Volk, M.S.
Committee Member

[redacted]

Morgan Roth, Ph.D.
Assistant Honors Director

May 8, 2024
Date

Abstract

Most research on the probability of winning a tennis match is based on the assumption that the points are independent and identically distributed, treating each point as a Bernoulli trial with fixed probability of success. This assumption, however, seems to contradict experience. Players' performance appears to fluctuate as the match progresses due to the psychological effect of past performance. To test this counterintuitive yet central assumption, previous research has attempted to test the independence hypothesis. However, there exists a research gap in evaluating the identicality-of-distribution hypothesis, a question of broader scope than that of independence. Hence, the purpose of this study is to test the hypothesis that tennis points are identically distributed throughout a server's match. This objective is accomplished by initially identifying, through appropriate homogeneity tests for sparse data, deviations from the base distribution, with the goal of developing a forecasting model that accounts for perturbations in the distribution.

Testing the Hypothesis that Tennis Points are Independent and Identically Distributed Using Statistical Methods

“We have realized that mental toughness often makes a difference in a match,” Serena Williams’s coach Patrick Mouratoglou claims (Mouratoglou Academy, n.d.). From the casual tennis enthusiast to the legend John McEnroe, everyone recognizes the unique mental toughness that sets Williams apart (ESPN, 2017). The role of the mind in athletic performance has been a topic of wide interest since the 1980s when researchers from Stanford and Cornell debunked the now-called *hot hand fallacy* in basketball (Gilovich *et al*., 1985). Since they publicized the hot hand paper in 1985, many studies have sought to detect the existence and effect of psychological momentum in virtually every sport (Avugos *et al*., 2006).

The question of momentum attempts to answer whether athletes’ performance in sports is affected by their self-assessment of previous performances due to psychological reasons, an effect known as psychological momentum. The present study aims to answer the question of momentum in tennis, determining whether momentum exists and what its effects are. The relevance of understanding momentum is twofold. On one hand, understanding the nature of psychological momentum in tennis would be crucial for tennis players and coaches, as its existence could imply mental toughness is a key component of skill in tennis. On the other hand, understanding momentum is highly relevant for the advancement of tennis analytics, as forecasting models are commonly built under the assumption that points in a match are independent and identically distributed, implying that momentum does not exist. Should momentum exist, tennis forecasting models might need revision and improvement.

Momentum in tennis: A literature review

A careful review of the existent literature on momentum in tennis reveals that, far from being a complete discussion, research on the matter is still necessary to understand the effects of psychological momentum in tennis. While there have been varied attempts to answer the question of momentum in tennis, most studies attempt to deal with two challenges. First, momentum is hard to measure. Being a psychological effect, it is complicated to know how and when momentum is triggered during a match. Second, claiming to influence a player's probability of success in winning a tennis point, momentum can be easily mistaken for other determinants of success, such as exhaustion (Klaassen and Magnus, 2001).

Despite the diversity of methodologies implemented, there seems to be a disconnect between the way players experience the momentum and the means researchers have tried to answer the question. Regarding the nature of momentum, some of the most successful studies define momentum by the effect of a point's outcome on the point immediately following it (Klaassen and Magnus, 2001; Goyal, 2020). However, psychological momentum, as experienced by players, does not vanish immediately after one point, and often it does not seem to have its effect immediately after an uncommonly negative or positive action. Regarding its impact on the probability of success, many studies implement the binomial distribution due to its simplicity and its adequacy in representing binary and discrete events, such as tennis points. However, the parameter of the binomial distribution, representing the fixed probability of success, is clearly unknown before a match starts. An estimator based on the players' rankings is not sufficiently accurate to capture the players' functional probability of success on the day of the match, as multiple factors other than a player's trajectory come into play to determine their success rate on any given day.

The first step in constructing a successful model to answer the question of momentum in tennis is to carefully explore the existing literature. Psychological momentum is defined as a psychological power gained or lost through success, or lack thereof, that results in an altered view of oneself or others (Iso-Ahola & Mobily, 1980). This change in psychological power ultimately leads to changes in the individual's future mental and physical performance and creates in the individual a psychological advantage over their opponents (1980). One of the first formal studies on the existence of momentum in sports was an attempt to determine whether the percentage of baskets by basketball players changes as a function of previous successes or failures, or, in other words, whether winning streaks are the result of psychological momentum, a phenomenon commonly known as the *hot hand* in basketball (Gilovich *et al*., 1985). Researchers have widely rejected the veracity of the *hot hand* on the grounds that streaks of success are explained primarily by the binomial probability distribution that drives events with binary outcomes and fixed probability of success (1985). Since 1985, many others have conducted similar research, quantitatively testing the existence and effect of momentum in sports other than basketball. Avugos, Bar-Eli, and Raab (2006) systematically compiled and reviewed all existing research between 1985 and 2004 on psychological momentum in sports. In their review, they concluded, on one hand, that evidence for the existence of momentum is very limited, but, on the other hand, suggested that the apparent feeling of momentum may function as an indicator of skill and thus lead to better decision-making. For this reason, the general trend has been to dismiss the effects of momentum as a key determinant of success in sports.

Specific to tennis, there has been considerable research on the topic of momentum. In one of the most prominent studies of momentum in tennis, Klaassen and Magnus (2001) approached the topic by seeking to answer the question of whether tennis points are independent and identically distributed (i.i.d.). Whether points are i.i.d. is a very relevant question when it comes to probabilistic forecasting, and thus this work was very influential for the present research. Their research, conducted on data from Wimbledon, robustly rejected the i.i.d. hypothesis, indicating that winning a tennis point increases the chances of winning the following point. Similarly, their study concluded that the server has a disadvantage at important points of increased psychological pressure. Klaassen and Magnus also argued that more skilled players are more resilient to psychological pressure and thus less affected by momentum. However, while supporting the existence of momentum, Klaassen and Magnus considered its effects to be small enough that momentum can be ignored at the time of designing forecasting models.

The work by Klaassen and Magnus (2001) is certainly not the only one when it comes to momentum in tennis. As Goyal (2020) found in his literature review on the topic, many other researchers have quantitatively analyzed momentum in tennis through indicators other than the relationship between consecutive points, including the relationship between consecutive sets and consecutive matches, and arriving at similar results (Jackson & Mosurski, 1997; O'Donoghue, 2000; Meier *et al*., 2019; Newton & Aslam, 2009; Newton and Keller, 2005; Pollard *et al*., 2006; Madurska, 2012). Following a more qualitative approach, Taylor and Demick (2006) assessed momentum through the Multidimensional Momentum Model. This model identifies precipitating events, such as attempting to win strokes too early in a point, as actions that trigger a chain of altered behaviors and ultimately affect performance.

Despite the abundance of studies on momentum in tennis, the literature review reveals that most of the quantitative studies focus on the independence among different sections of a match. Clearly, lack of dependence would be an obvious indicator of momentum, whether it is as narrow as dependence among points, or as wide as dependence among sets. Independence, however, is not the only way of characterizing momentum. In fact, when understanding momentum in terms of the i.i.d. condition, independence only accounts for half the scenario. The present study will focus on the homogeneity of distribution throughout the match, the least explored half of the i.i.d. hypothesis. While dependence would be an immediate result of momentum, it seems that the long-term effects of momentum can be perceived in a change in the probability of winning a point throughout the match. As opposed to Klaassen and Magnus (2001), I suspect this change might not take the form of an immediate reaction. Instead, this study will give consideration to the possibility of a slow and steady change in momentum throughout the match. While developing a new approach, the present study will greatly benefit from the statistical tools employed in previous studies. Ultimately, it is the aim of the present study to add valuable information to the existing literature on the topic by building on previous findings and expanding their results. The hypothesis of the present study is that statistical analysis will suggest the distribution of tennis points is not homogenous throughout a match. Should this hypothesis be supported by the experiment, it will be the aim to determine the causes for momentum shifts. Not only would the defense of this hypothesis provide much clarity into the nature of momentum in tennis, but it would also allow for the creation of forecasting models that take into account the effects of momentum.

Preliminary Work: Independence Test

The broad motivation of studying the existence and nature of psychological momentum in tennis led to various personal investigations that culminated in this paper. While the final aim of this study is to explore the homogeneity of distribution throughout a tennis match, this was not always the case. After a brief literature review, I implemented techniques from other researchers to test the hypothesis of independence for myself. It was not until I analyzed my results and conducted a more extensive literature review that my final research question was developed. The starting point of this study was the work of Klaassen and Magnus (2001), one of the most recognized existing studies addressing the independence of tennis points. Aiming to understand and apply their work, I applied their methods to test the hypothesis of independence on data from the 2020 Men's U.S. Open. My conclusions from this initial study gave rise to the research questions addressed in this paper. Similarly, the organization of the dataset became the starting point for the present work. Hence, I consider my initial study on the independence of tennis points to be the preliminary work for the subsequent study on homogeneity of distribution. The methodologies and results of this preliminary work are summarized below.

Research Question on Independence Test

As seen in the literature review, there is not a unique approach to test the role of momentum in tennis. One of the major successes in the work of Klaassen and Magnus (2001) was to narrow down a rather broad topic into a simple and specific research question to be tested. Seeing every point as a Bernoulli trial, Klaassen and Magnus narrowed the question of momentum to the well-known i.i.d. hypothesis in statistics. In particular, they sought to answer whether consecutive tennis points were statistically independent. In a similar vein, the research question for my preliminary work was to determine whether the outcome of a tennis point is a good predictor of the outcome of the point immediately following it. In short, this preliminary work consisted of testing if consecutive points are statistically independent from each other. The idea behind this research question is that psychological momentum could manifest in a sudden change in the level of play after a success or a failure. While this is certainly a simplistic view of momentum, this simple hypothesis could shed light on other questions, such as the relation between a player's quality and momentum. The characterization of the question of momentum in testing the i.i.d. hypothesis is more than a matter of convenience, but it has important implications. Most of the tennis forecasting techniques assume that tennis points are independent and identically distributed Bernoulli trials (Klaassen & Magnus, 2014). Thus, should the hypothesis be false, tennis forecasting techniques might not be as accurate as they could be.

Methodology on Independence Test

Data

One of the main obstacles to the advancement of tennis analytics is the lack of publicly available data. Some sport analysts suggest this is due to the nature of the professional tennis tours and the individual aspect of tennis (Sackmann, 2015). In major sports, most competitions are organized by a small number of big corporations, such as the NFL or NBA, which are willing to publicize data. On the other hand, the tennis tour is composed of a plethora of tennis tournaments around the globe; there has not been a wide collaborative initiative to release a public data pool. Besides, major sports are composed of teams or clubs, which have the capacity to hire professional analytics services. Tennis, however, is characterized by a significant income gap. Players ranked between 500 and 1000 average a yearly income of $7,000, whereas the top five players average an income of $8 million (Wang, 2022). Many see the potential for a revolution in tennis analytics; in particular, the *hook eye* system utilized by all major tours is said to gather significant data besides detecting the location of the ball (Annacone *et al*., 2012).

Despite the less-than-ideal condition when it comes to data availability, there has been significant progress in the field. Much of the public advancement in tennis analytics is due to the work of Jeff Sackmann and his Match Charting Project (n.d.). The Match Charting Project is an initiative to track point-by-point data in all matches of the professional tours; all data is publicly available and charted by volunteers. For my project on independence, I decided to utilize point-by-point data on the 2020 U.S. Open, one of the four major yearly tournaments. Due to computational limitations at the time, this preliminary project was limited to First Round data from the Men's Singles bracket. I downloaded all the raw data into Microsoft Excel, where I also created new variables described below. I analyzed the data in RStudio.

Variables

The general approach to the choice and classification of variables was modeled after the work of Klaassen and Magnus (2001), although I defined additional variables. Every observation in the dataset consisted of a single tennis point. I considered the binary outcome of the point as the response variable, whereas I treated the outcome of the previous point as the main regressor of interest. As did Klassen and Magnus, I divided other regressors between Quality and Dynamic regressors. The quality regressors, as the name suggests, are those related to the quality of the players. These are the variables, under the i.i.d. assumption, that should completely determine the likelihood of winning a point. The quality variables are determined before the match starts and remain constant throughout. These are composed of a variable measuring the overall quality of the match and a variable measuring the quality difference. The former serves to reveal the effect that the quality of players produces on the validity of the i.i.d. assumption when seen in interaction with dynamic regressors, while the latter seems to be the most intuitive way of determining the probability of winning a point. In some of the models discussed ahead (GLMM and FGLS), the quality variables also include a random effect, accounting for the unmeasurable quality of players.

More related to momentum, the dynamic regressors consist of both categorical and continuous variables that measure aspects that change after every point, and which might have an impact on subsequent points. The significance of these regressors in estimating success might reveal dependence among the points as well as varying probability distributions. Further, each observation includes variables that specify general information on the point. These include a match ID unique for every match, as well as a binary variable indicating the server of the point. In the testing, I defined success from the server's point of view. The organization of the data in this preliminary work became the basis for the subsequent study on homogeneity. A detailed presentation of the variables is found in Appendix A.

Statistical Analysis

I.I.D. Condition

In statistics, one of the ideal characteristics of data points is that they are independent and identically distributed; when this condition is present, the data is often said to be i.i.d. Precisely, this condition implies that all trials in a statistical experiment are statistically independent from each other, and that the probability distribution, as well as the parameter values driving each trial, are identical.

Bernoulli Trials and the Binomial Distribution

A Bernoulli trial is a statistical experiment with a mutually exclusive and exhaustive binary outcome driven by a probability parameter *p.* This can intuitively be pictured as the experiment of flipping a coin, not necessarily fairly weighted, where the probability of tails is *p* whereas the probability of heads is $1 - p$. The similarity between Bernoulli trials and tennis points is not difficult to see. The binomial distribution is a probability distribution modeling the outcome of *n* i.i.d. Bernoulli trials with probability parameter *p.* While the number of points, *n,* in a tennis match is unknown until the end of the match, it seems reasonable to picture a tennis match as a binomially distributed experiment of *n* trials. More precisely, the match can be split into two binomially distributed experiments, classifying the points by their respective servers and letting the parameter *p* be the probability of the server winning any given point. Now, there are immediate questions about this approach. First, the parameter *p* is unknown, even if at all constant. Also, assuming that points are independent seems to be an unjustifiable assumption based on apparent momentum. In a traditional forecasting setting, the parameter *p* would be estimated from various measures of quality, such as the players' rankings and recent performances. The focus of this preliminary work, as it was for Klaassen and Magnus (2001), is to specifically test the assumption of independence between points, without which tennis matches would not be binomially distributed.

Models

The project is composed of five different combinations of regressor variables. For all the models, the response variable is a binary variable indicating whether the server of the point won. Model 0 uses only the quality regressors, just to confirm the significance of the quality regressors in estimating the winner of the point. Model 1 introduces the dynamic regressors (the result of the previous point for dependence, and the importance of the point for change in probability distribution). Model 2 adds an additional dynamic regressor for dependence, indicating whether the server won the previous two points. Model 3 introduces interactions between the quality and dynamic regressors to determine whether the effects of the dynamic regressors are dependent on the quality of players. Finally, Model 4 introduces many other dynamic regressors, which are depurated by using Akaike information criterion (AIC) to determine the most significant model. A detailed mathematical description of the models is found in Appendix B.

Hypothesis Testing

The hypothesis was tested through three regression models, which increased in complexity and culminated with the feasible generalized least squares method implemented by Klaassen and Magnus (2001).

Logistic Regression

The initial model consisted of a generalized linear model in the form of logistic regression. Generalized linear models (GLM) provide a generalization of ordinary linear regression (OLR) by letting the response variable be related to the linear model through a link function and letting the magnitude of the variance of the errors be a function of their predicted values. Moreover, generalized linear models lift the normality assumption and allow the response variables to follow arbitrary distributions of the exponential family. Logistic regression is a specific form of a generalized linear model, where the response variable follows a Bernoulli distribution with probability of success dependent on the value of the regressors, and where the response variable is related to the linear model through a log-odds link. In the context of analyzing the effect of diverse conditions represented by continuous and categorical regressors, logistic regression provides an effective method, although it faces some challenges. This generalization of ordinary linear regression modifies two assumptions of OLR. First, in ordinary linear regression, the means of the observations are a linear function of some regressors, while in GLM a transformation of the mean through the link function is a linear combination of the regressors. Second, in ordinary linear regression, the variance of the observation is constant, while in GLM it is a function of the mean. The goal of the experiment is to assess the effect of previous points on the current point in order to test independence. However, the result of the previous point, which is represented in a binary regressor that indicates the winner of the previous point, contains some information about the relative quality of the players. The better player is expected to win more points, and therefore observing the winner of the previous point provides a small, yet positively correlated estimator of the outcome of the following point. In order to avoid this misleading positive correlation, it is necessary to correct for the quality of the players, absorbing most of the information conveyed by the winner of the previous point. However, yet another problem arises; while part of the quality of players is measurable in ranking systems, some of the true quality on the day of the match is unmeasurable and dependent on multiple factors, such as the form of the day, location of the match, and physical and mental form of the players. Therefore, the observed quality measures fail to approximate with high accuracy the true difference in quality. Different statistical methods provide effective solutions for this problem; these are considered below.

Generalized Linear Mixed Models

The second model consists of generalized linear mixed models. Generalized linear mixed models (GLMM) are an extension of generalized linear models, where random effects are included in the linear predictor. Similar to GLM, in GLMM observations are assumed to be conditionally independent with means dependent on the linear prediction, as defined in the link function, and with conditional variance specified on a variance function. However, GLMM includes an unobserved vector of random effects, assumed to be normally distributed with mean zero and dispersion matrix dependent on unknown variance components. In GLMM, it is only when other regressors are conditioned on these random effects that the response variable follows a distribution of the exponential family. GLMM fitting involves integrating over the random effects, and this process is commonly approximated by different methods such as numerical quadrature or Markov chain Monte Carlo. The final method is that implemented by Klaassen and Magnus (2001), commonly known as feasible generalized least squares.

Feasible Generalized Least Squares

Another problem arises while trying to fit a model where there is certain degree of correlation between residuals and the model. This is the case in dynamic panel data, data that is collected across time and that is naturally correlated with previously observed values, eliminating real independence between the observations. Trying to estimate the parameters through ordinary least squares (OLS) when such dependence is present makes the estimates for both the fixed effects and the random effects discussed in GLM biased and inconsistent due to endogeneity. Endogeneity is a measure of the correlation between the error term and the parameter. One of the goals of generalized least squares (GLS) is transforming heteroscedastic models, which are models with changing variance in their error terms, to homoscedastic models, which have a constant variance. This is accomplished by using the known variance-covariance matrix. Feasible generalized least squares, however, instead of assuming this variance-covariance matrix is known, it is initially approximated through OLS. Klaassen and Magnus (2001) claim that, contrary to OLS, FGLS is successful in finding consistent estimates for parameters of binomial data coming from dynamic data panels. Thus, FGLS provides a practical solution to the unsolved problem of fitting discrete dependent variables from dynamic data panels. Two conditions guarantee the consistency of the estimations. The first condition is the independence between the observed quality and the unobserved quality represented in the random effects. The second one is the lack of initial conditions; the first observation of regressors in every match is not correlated to a previous point as there were no previous points recorded. After applying FGLS, a linear model is used to fit the data, as suggested by Klaassen and Magnus. While there are usually some problems using linear models to fit binary response variables, they justify their choice (1) by the fact that the expected value of the response will always lie between 0 and 1 due to the binary nature of the response, and (2) by the imposition of the restriction of the second moment of the response to equal the expected value.

Results of Independence Test

Before the beginning of the study, the assumption was that points were slightly dependent on previous events during the match, and that the underlying probability distributions driving the outcomes of the match slightly varied due to different factors, mainly psychological. The study tested these hypotheses by analyzing the effect of multiple variables on the likelihood of winning points. The results reflected that the i.i.d. hypothesis is not entirely correct, but its effects are sufficiently small that they can be ignored in practical forecasting applications.

While the small deviations from independence can be overlooked from a forecasting standpoint, the effects of the different dynamic regressors should not be ignored from the perspective of the players and coaches. The study suggested that momentum is real, and, as expected, its effects are higher in weaker players. This suggests that developing the ability to remain consistent and sober throughout the match is an important area for improvement in less experienced players. The positive coefficients for the regressors measuring the importance of the point were unexpected. These could be explained by an effect on the receiver being higher than on the server. Similarly, I did not expect to see a negative correlation between winning two consecutive points and winning the following point, which might be explained by the relative similarity in the skill level of all players in this high-tier competition.

Regarding the statistical methods I used, as suggested by Klaassen and Magnus, feasible generalized least squares seem to provide the best method for approximating the parameter coefficients, as it effectively deals with the dependence between observations coming from a dynamic panel data structure (2001). However, logistic regression and the generalized linear mixed models seem to present similar results, always agreeing on the sign of the estimated coefficients, just lacking some significance. A detailed explanation of the results of every test is included in Appendix C.

Further Analysis: Designing a New Project

Initially, the project consisted of testing the independence among all the points played in a match and served by a fixed player. Applying the methods of Klaassen and Magnus (2001) to

2020 U.S. Open data, I was not able to reject the null hypothesis of independence. These results corroborate past research. The inconclusive results motivated further consideration of the initial motivation, which was testing the existence of psychological momentum in tennis. While Klaassen and Magnus propose a simple experiment based on the independence of consecutive points, it may be the case that this approach is far too simplistic. Certainly, the experience of momentum seems to go beyond consecutive points, and is not often experienced as an immediate, drastic effect. Instead, momentum is often experienced as a long-term, progressive increase, or decrease, in the level of play throughout the match. The long-term effects of momentum have not been widely explored. These thoughts turned my attention to homogeneity of distribution, the less explored aspect of the i.i.d. hypothesis.

Research Questions

The desire to test the homogeneity of distribution motivated the design of a whole new experiment. While approaching the problem of momentum from a different angle, the new study is certainly a continuation of the previous one, as it builds on its results and seeks to answer related questions. The central research question is (1) to determine whether points served by a fixed player are identically distributed within a match. I again utilize data from the 2020 U.S. Open, but now I use every point from the Men's and Women's brackets (ATP, 2020). In the case homogeneity is rejected, I propose a follow-up question of (2) determining what causes the perturbations of homogeneity. Finally, I propose the goal of (3) developing a non-i.i.d. forecasting model to be tested against the i.i.d. model through Monte Carlo simulations.

Methodology

Homogeneity of Distribution

If tennis matches can truly be broken down into two binomially distributed experiments, one for each server, not only points would be independent. The i.i.d. assumption of the binomial distribution assumes that the parameter *p* ruling the probability of success is constant. This uniqueness of parameters *p* is called homogeneity of distribution across trials, or tennis points in this case. The negation of homogeneity would imply that there exist some trials in which the probability of winning the point is not the same as in the rest of the points. In other words, a lack of homogeneity would mean that the server is not equally likely to win every point. In order to test whether data is homogenous, it is conventional to employ tests for homogeneity. These tests partition the data into multiple bins, categories that are suspected to have different probabilities of success. Thus, the null hypothesis is that the parameter *p* is the same for all the partitions, whereas the alternative hypothesis is that the parameter *p* is not unique across all the partitions. A mathematical description of the test is shown in Appendix D. Hence, in order to answer my central question on the existence of homogeneity, the first necessary step is to decide how to partition the data. In other words, multiple subsets of points need to be defined and tested for homogeneity of distribution.

Partitioning the Data

When trying to decide how to partition the data, some initial hypotheses based on experience prove helpful. These hypotheses attempt to point out situations that may cause the probability distribution of the tennis points by a fixed server on a fixed match to change. Two types of partitions seemed reasonable. The most obvious way to partition the data would be sequentially, expecting to see fluctuations in the probability parameter across time. This can be represented by partitioning the data based on games or sets, hoping to see how the probability of winning a point varies as the match progresses. However, another way of partitioning the data would be based on events, specific outcomes that are expected to represent a change in the probability parameter. Two specific events of interest are unforced errors, which are popularly believed to cause negative momentum, as well as the outcome of the previous point, which was the initial topic of interest when testing for independence.

Data

As before, the data from Tennis Abstract (n.d.) is the starting point. This time, however, all points from the Men's and Women's singles tournaments are used. Using the GROUP BY function in SQL, the data is initially partitioned by match and server. For the first analysis, the data was further partitioned by games, creating different datasets to be tested for homogeneity. Similarly, the second test consists of partitioning the data by sets. The raw data from Tennis Abstract has the convenient feature of including binary indicators for different events that often occur on tennis points, such as double faults or unforced errors. This allows the data to be partitioned by the binary indicator for unforced errors, grouping the points by a fixed server and match into two categories. The final tests followed the same approach as when testing for independence among points. As before, the lag indicator is created, which is then used to partition the data into two groups.

Now, each of these partitions provides the data with a different structure. For instance, when partitioning the data by games, most partitions will consist of a small number of data points, typically between four and seven based on the length of a game. When partitioning by unforced errors or lag, the partition is binary, creating only two categories. For this reason, it is necessary to study the literature on homogeneity tests in order to choose appropriate tests for each scenario.

Statistical Analysis

Tests of Homogeneity

Klein and Linton (2013) wrote an extensive review of the different tests of homogeneity. In their study Klein and Linton summarize the theoretical underpinnings behind nine tests for homogeneity. Furthermore, they test efficacy of these tests under different scenarios, including changing values for the size and probability parameters of the binomial distribution. Their results prove extremely valuable at the time of making a choice on the test to be used for each of my four scenarios. After a careful review, I determined the test of Nass, a modification of the standard Chi-squared test, to be appropriate for all four scenarios.

Test of Nass

When it comes to testing for homogeneity across multiple groups, Pearson's chi-squared test is the norm. This test is based on a simple chi-squared test statistic based on the number of observations and the number of expected observations under the null hypothesis of homogeneity for each category. While Pearson's test is known for its simplicity, it does not perform adequately under certain extreme conditions. Pearson's test has a high probability of Type I error under sparse data scenarios. Particularly, this test does not perform adequately when faced with a small number of partitions, k, that in turn contain a small number of observations, n_i. Unfortunately, this kind of scenario is prevalent in this study. For instance, when partitioning by games, the number of observations per game can be as low as four, some of which can contain no successes at all. Among the multiple tests that Klein and Linton (2013) evaluated, the test of Nass, a modification of Pearson's test, is characterized by low *Type I* error probabilities.

With similar concerns as Klein and Linton, Potthoff and Whittinghill (1966) published a paper evaluating distinct tests of homogeneity and their ability to model extreme conditions. Concerned with biological applications, Potthoff and Whittinghill specifically consider the scenario of having small sample sizes, n_i, in some or all of the k categories. In their study, the authors suggest the test of Nass, a modification of Pearson's test, to be particularly effective under these scenarios. Corroborating these observations, Klein and Linton (2013) determined Nass's test to possess a low probability of Type I error under all the scenarios applicable to this study, as well as a relatively high statistical power under the relatively small sample sizes dealt with in all of the scenarios considered in this project. A low probability of Type I error implies that the test will not be likely to erroneously reject the hypothesis of homogeneity, being sufficiently conservative. On the other hand, a high power implies that the test is effective at rejecting the null hypothesis when it should indeed be rejected.

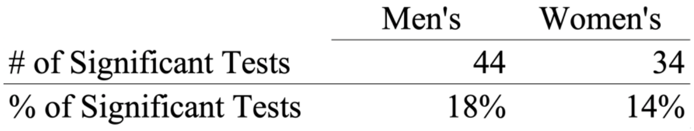
With these considerations in mind, I decided to apply Nass's test for all the scenarios under consideration. Nass's test modifies the test statistic in a standard Chi-squared test, T_p. Under the null hypothesis, the distribution of $c \times \mathrm{T}_p \sim \chi^2{}_v$, where c and v are chosen so that the conditional mean and variance of $c \times \mathrm{T}_p$ match the mean and variance of the approximating chi-squared distribution (Potthoff & Whittinghill, 1966). A more mathematical explanation of Nass's test is included in the Appendix. I calculated T_p, c, and v values in SQL, and p-values in Excel.

Test Results

For each scenario, I look at the percentage of experiments in which I reject the null hypothesis with 95% confidence. I consistently failed to reject the null hypothesis for homogeneity in three of the scenarios. With less than 7% of experiments rejecting H_0 for unforced error and lag scenarios (for both men and women), I cannot reject homogeneity and thus have no evidence of instantaneous psychological momentum. This result corroborates previous research. When partitioning by games, 18% of experiments in the male category reject H_0. When partitioning by sets, 37% of experiments in the male category reject H_0. These numbers suggest I pay attention to changes in probability parameters throughout time. In other words, an increased number of significant experiments when partitioning by sets might suggest there is evidence for long-term momentum, which was precisely the motivation of this study. Tables 1 through 4 summarize the number and percentage of significant tests for each scenario. As the name suggests, the last row in every table shows the average percent change between the lowest and highest experimental *p* parameter for every experiment.

Table 1

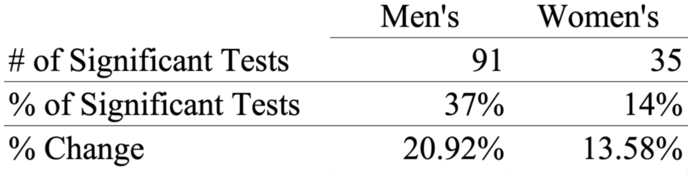
Partition by Games



	Men's	Women's
# of Significant Tests	44	34
% of Significant Tests	18%	14%

Table 2

Partition by Sets



	Men's	Women's
# of Significant Tests	91	35
% of Significant Tests	37%	14%
% Change	20.92%	13.58%

Table 3

Partition by Unforced Errors

	Men's	Women's
# of Significant Tests	14	13
% of Significant Tests	6%	5%
% Change	0.26%	0.79%

Table 4

Partition by Outcome of Previous Point

	Men's	Women's
# of Significant Tests	17	11
% of Significant Tests	7%	4%
% Change	0.76%	0.54%

Further Analysis

The results suggest there might exist evidence for a long-term change in the probability parameter. Rejecting the hypothesis of homogeneity can be an important result in and of itself, as it questions the validity of the i.i.d. hypothesis and hence the validity of tennis forecasting models. This result, however, is of minimal practical use as it does not indicate what causes the perturbations of homogeneity. In other words, in order to make significant use of these results, it is necessary to investigate what triggers the probability of winning a point to change from set to set. Identifying an underlying pattern to the change in the probability parameter could result in the creation of a forecasting model that does not assume the i.i.d. hypothesis but instead lets the parameter p become the output of a non-constant function. The question of determining what correlates to changes in the probability parameter, a question which was initially proposed as a potential follow-up question, leads to further analysis. I start with general hypotheses based on experience. Specifically, I question whether stronger players improve their level of play after the first game. As seen in the preliminary work, stronger players are thought to be more resilient and hence able to react better to a bad start. Similarly, I consider the possibility of weaker players lowering their level of play after the initial set. In a different direction, I examine whether there is a pattern in the spread of the *p* parameter throughout the matches. In other words, I consider if there is a correlation between the set number and its corresponding *p* parameter.

In this analysis, I define "strong players" as those ranked 19 or better in the ATP ranking. I define *weak players* as those ranked 126 or worse. Before performing any formal tests for these hypotheses, I graphed the data in order to visually identify any patterns. Unfortunately, there were no visible patterns, discouraging me from proceeding to any formal tests.

Conclusions

Summary of Results

I attempted to reject the hypothesis of homogeneity under four scenarios. As an attempt to find an immediate shift in the probability parameter, I partitioned the data into points followed by unforced errors, often considered to cause a negative shift in momentum. I also partitioned the data based on the success in the previous point, commonly associated with a positive shift in momentum. I hoped to construct a non-i.i.d. model that adjusts the probability of winning a point when one of these events occur. These models were to be tested against the i.i.d. model through Monte Carlo simulations. However, I was unable to consistently reject the hypothesis of homogeneity under either of these scenarios. Considering our data can be seen as a dynamic panel, I also partition data across time, grouping by games and sets. When grouping by games, I saw small evidence to reject homogeneity. However, when I partitioned into sets, 37% of matches in men's tournament did not exhibit homogeneity with a 95% confidence. While I have reasonable evidence to question the homogeneity across sets, I did not find relevant patterns in the data that would require a non-i.i.d. model.

Further Research

A driving motivation for the study, more than the lack of research, was skepticism in the methods for rejecting the i.i.d. hypothesis. Most research focuses on short-term changes based on specific events, comparable to our tests on unforced errors and lag outcomes. Based on experience, however, it seems more appropriate to look for long-term, time-dependent changes in homogeneity, such as our tests on games and sets. The reasonable amount of statistically significant tests in my project suggests that, as opposed to the general consensus, there is a place for further investigation of the matter. The central remaining question is to determine what truly dictates the probability parameters. Perhaps the simplistic approaches to momentum needed for a quantitative test do not fully capture the effects of momentum. Certainly, psychological momentum, as its name suggests, is driven by qualitative factors that might be difficult to quantify. Due to the lack of evident patterns, I suggest considering qualitative information that might have a relevant effect on psychological momentum. Thus, collaboration from experts in psychology might be a great contribution to the matter.

While the study is fundamentally inconclusive, it certainly contributed to the diverse pool of approaches to the question of momentum in sports. As suggested by most researchers on the topic, I consider the current methods of forecasting to be the best approach based on the current knowledge. With an increased comfort on the idea of accepting the i.i.d. hypothesis on tennis forecasting, I look forward to future research on the matter that might give the tennis community a more clear picture of the role momentum plays.

References

ATP Tour. (2020, August 31). ATP Rankings.

Annacone, P., Martin, T., O'Shannessy, C., & Stein, M. (2012). *2012 Sloan Sports Analytics Conference: Tennis.* [Conference session]. Massachusetts Institute of Technology Sloan Sports Analytics Conference, Cambridge, MA, United States.

Avugos, S., Bar-Eli, M., Raab, N. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sports and Exercise*, 7(6), 525-553.

Breslow, N.E., & Clayton D.G.. (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88(421), 9-25.

Crust, L., & Nesti, M. (2006). A review of psychological momentum in sports: Why qualitative research is needed. *Athletic Insight*, 8(1).

Dawson, R.B. (1954). A simplified expression for the variance of the χ2-function on a contingency table. *Biometrika*, 41, 280.

ESPN. (2017). *After Serena Williams fires back on Twitter, John McEnroe declines to apologize.*

Gilovich, T., Tversky, A., & Vallone, R. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3), 295-314.

Goyal, A. (2020). Hot Racquet or Not? An Exploration of Momentum in Grand Slam Tennis Matches. New York University.

Haldane, J.B.S. (1940).The mean and variance of χ2, when used as a test of homogeneity, when expectations are small. *Biometrika*, 31, 346-355.

Iso-Ahola, S., & Mobily, K. (1980). "Psychological momentum": A phenomenon and an empirical (unobtrusive) validation of its influence in a competitive sport tournament. *Sage Journals*, 46(2).

Jackson, D. & Mosurski, K. (1997). Heavy Defeats in Tennis: Psychological Momentum or Random Effect? *Chance*, 10(2), 27-34.

Jiang, J., & Nguyen, T. (2021). *Linear and generalized linear mixed models and their applications*. Springer.

Klaassen, F., & Magnus, J. (2001). Are tennis points independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96(June), 500-509.

Klaassen, F., & Magnus, J. (2014). *Analyzing Wimbledon: The power of statistics*. Oxford Academic.

Klein, M., & Linton, P. (2013). On a comparison of tests of homogeneity of binomial proportions. *Journal of Statistical Theory and Applications, Vol. 12, No. 3, 208-224.*

Madurska, A. (2012). A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches. 4th year Software Engineering MEng project.

Flepp, R., Rudisser, M. & Franck, E. (2019). Investigating the conditions for psychological momentum in the field: Evidence from men's professional tennis. *UZH Business Working Paper No. 383*.

Montgomery, D., Peck, E., Vining, G. Introduction to linear regression Analysis. *Wiley*.

Mouratoglou Academy. (n.d.). *Mental Training in Tennis.*

Newton, P. & Aslam, K. (2009). Monte Carlo Tennis: A Stochastic Markov Chain Model. *Journal of Quantitative Analysis in Sports*, 5(3).

Newton, P. & Keller, J. (2005). Probability of Winning at Tennis I. Theory and Data. *Studies in Applied Mathematics*, 114(3), 241-269.

O'Donoghue, P. (2000). The Most Important Points in Grand Slam Singles Tennis. *Research Quarterly for Exercise and Sport*, 72(2).

Panchanan, D. (2019). Dynamic Panel Model. In: Econometrics in Theory and Practice. Springer.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series* 5, 50, 157-175.

Pollard, G., Cross, R. & Meyer D. (2006). An Analysis of Ten Years of the Four Grand Slam Men's Singles Data for Lack of Independence of Set Outcomes. *Journal of Sports Science and Medicine, 5(4), 561-566.*

Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity: the binomial and multinomial distributions. *Biometrika , Jun., 1966, Vol. 53, No. 1/2 (Jun., 1966), pp. 167-182.*

Sackmann, J. (n.d.). *Match Charting Project*. Tennis Abstract.

Sackmann, J. (2015). *2015 Sloan Sports Analytics Conference: First service: The advent of actionable tennis analytics.* [Conference session]. Massachusetts Institute of Technology Sloan Sports Analytics Conference, Cambridge, MA, United States.

Taylor, J. & Demick, A. (1994). A multidimensional model of momentum in sports. *Journal of Applied Sport Psychology*, 6, 51-70.

Wilcox, R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics,* 6(1), 3-32.

Yamano, T. (2009, Fall). [Lecture notes on advanced econometrics].

Wang, C. (2022). *The harsh truth behind income inequality in tennis.* Medium.

Appendix A

Variables for Test on Independence

Response Variable

- y_i = ServerWon_i, in {0,1}

Quality Variables

Quality variables are determined before the match starts and remain constant throughout.

- q_1 = Rel.Quality = $(R_A - R_B) - (R_A - R_B)^*$
    - The relative quality. Gap between the two players.
    - Variable suggested by Klaassen and Magnus (2001). R_A is a transformation obtained by $R_A = 8 - \log2(\text{RANKa})$, where RANKa is the ATP or WTA ranking for the player in consideration. This provides a better estimate of the quality of the player. Still, this measure derived from the professional rankings has many limitations. A more appropriate measure for the quality of players would be the ELO ranking by Jeff Sackmann (n.d.); however, data for the time of the tournament into consideration is unavailable.
    - $(R_A - R_B)^*$ is the average for the entire tournament.
- q_2 = Overall.Qual = $(R_A + R_B) - (R_A + R_B)^*$
    - The absolute quality of the match might be important, as usually more service points are scored in a match between two strong players than in a match between two weaker players (Klaassen & Magnus, 2001).

- q_3 = match_id
    - A variable containing a unique value for every match; it is used to create the random quality effect unique to every match.

Dynamic Variables

As opposed to quality variables, dynamic variables develop and change throughout the match. For the study, these are classified into variables that affect independence of points, and variables that affect the identicality of the point distributions.

- Independence Variables:
    - d_1 = Prev.Pt.Won, in {0,1}
        - Indicates whether the server won the previous point ($\text{Prev.Pt.Won} = 1$). However, it takes the value of 0 at the first point in every game, as there is a break between the games and therefore little influence on the following point.
    - d_2 = Lag2, in {0,1,2}
        - Indicates if the server won ($\text{Lag2} = 1$) or lost ($\text{Lag2} = 2$) the previous two points, ($\text{Lag2} = 0$) otherwise. It also takes on the value of 0 in the first two points of every game.
    - d_6 = Momentum.Server
        - Measure of the momentum of a player (not defined by creator of the statistic). This variable was created by Jeff Sackmann (n.d.) and is presented in his Tennis Abstract.

        - d_7 = Prev.Ace, in {0,1}
            - Indicates whether the server served an ace in the previous point.
        - d_8 = Prev.BF, in {0,1}
            - Indicates whether the server committed a double fault in the previous point.
        - d_9 = Prev.UF, in {0,1}
            - Indicates whether the server committed an unforced error in the previous point.
    - Identicality
        - d_3 = Imp.For.Winning.Game
            - This variable accounts for the probability that the server wins the current game given that he will lose the current point, subtracted from the probability that the server wins the current game given that he will win the current point. Estimations of the probabilities assume that each point by a given server in a given match is an independent and identically distributed Bernoulli trial. While treating points as i.i.d. Bernoulli trials for this calculation seems inappropriate considering that this is the assumption trying to be tested in the study, this assumption still allows for an easy calculation of how important a point is relative to others in terms of the effect on the probability of winning the whole match.

- d_4 = Imp.of.Game
    - Attempts to estimate the importance of a game. Due to computational and time constraints, instead of following the approach from the previous regressor, it is estimated by a less accurate yet simpler formula. $\text{Imp.of.Game} = 6 - (\text{Games won in current set by player A} - \text{Games won in current set by player B})$. This implies that games are more important to win when the score of the set is tied or close to being tied.
- d_5 = PointNumber, in {1, 2, 3, …}
    - Serves as an indication of match length, which can lead to varying degrees of performance due to fatigue.
- d_{10} = GamePtServer, in {0,1}
    - Indicates whether the point is a game point for the server.

Appendix B

Models for Test on Independence

Model 0 – Model with Quality Regressors

- GLM (Logistic Model)

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1 - E(y_i)})$$

- GLLM

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + q_{3i} * \hat{\beta}_3 + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1\text{- } E(y_i)})$$

- FGLS

Model 1 – Model with no interactions between quality and dynamic regressors

- GLM (Logistic Model)

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{3i}d_{4i} * \hat{\delta}_{34} + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1\text{- } E(y_i)})$$

- GLLM

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{3i}d_{4i} * \hat{\delta}_{34} + q_{3i} * \hat{\beta}_3 + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1\text{- } E(y_i)})$$

Model 2 – Model including Lag(2)

- GLM (Logistic Model)

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1\text{- } E(y_i)})$$

- GLLM

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + q_{3i} * \hat{\beta}_3 + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1\text{- } E(y_i)})$$

Model 3 – Model with Interactions between the Quality and Dynamic Regressors

- GLM (Logistic Model)

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + q_{1i}\, d_{1i} * \widehat{\delta\beta}_{11}$$

$$+\ q_{2i}\, d_{1i} * \widehat{\delta\beta}_{21} + q_{1i}\, d_{3i}d_{4i} * \widehat{\delta\beta}_{134} + q_{2i}d_{3i}d_{4i} * \widehat{\delta\beta}_{234} + \varepsilon_i, \text{ where } \eta_i = \ln\left(\frac{E(y_i)}{1\text{- } E(y_i)}\right)$$

- GLLM

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + q_{1i}\, d_{1i} * \widehat{\delta\beta}_{11} + q_{2i}\, d_{1i} * \widehat{\delta\beta}_{21}$$

$$+ q_{1i}\, d_{3i}d_{4i} * \widehat{\delta\beta}_{134} + q_{2i}d_{3i}d_{4i} * \widehat{\delta\beta}_{234} + q_{3i} * \hat{\beta}_3 + \varepsilon_i, \text{ where } \eta_i = \ln\left(\frac{E(y_i)}{1 - E(y_i)}\right)$$

Model 4 – Model with Additional Dynamic Regressors

- GLM (Logistic Model)

$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + \sum_{k=5}^{10} (d_{ki} * \hat{\delta}_k) + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1 - E(y_i)})$$

  - $d_{3i}d_{4i}$: The multiplication of the importance of a point in a game and the importance of a game in a set works as an indicator of the overall importance of a point in a match.

- GLLM

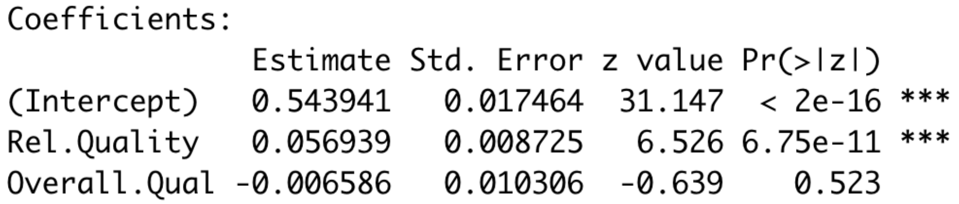
$$\eta_i = q_{1i} * \hat{\beta}_1 + q_{2i} * \hat{\beta}_2 + d_{1i} * \hat{\delta}_1 + d_{2i} * \hat{\delta}_2 + d_{3i}d_{4i} * \hat{\delta}_{34} + \sum_{k=5}^{10} (d_{ki} * \hat{\delta}_k) + q_{3i} * \hat{\beta}_3 + \varepsilon_i, \text{ where } \eta_i = \ln(\frac{E(y_i)}{1 - E(y_i)})$$

Appendix C

Results for Test on Independence: Statistical Inference and Hypothesis Testing

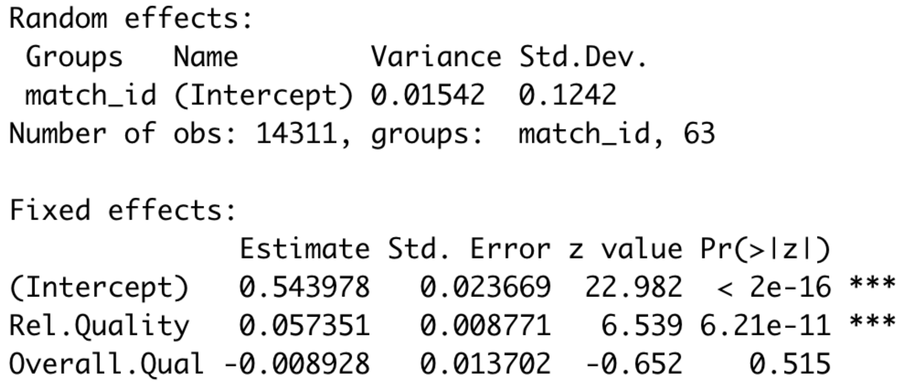
Model 0

- Logistic:



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.543941   0.017464  31.147  < 2e-16 ***
Rel.Quality   0.056939   0.008725   6.526 6.75e-11 ***
Overall.Qual -0.006586   0.010306  -0.639    0.523
```

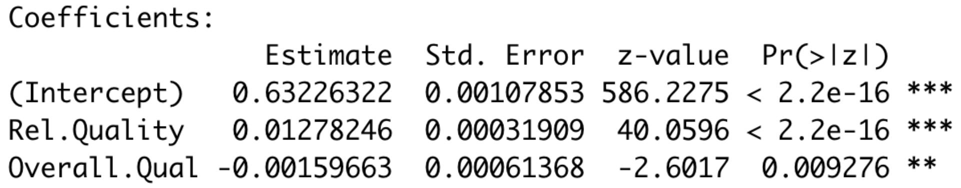
- GLMM: (AIC = 18733.6)



```
Random effects:
 Groups   Name        Variance Std.Dev.
 match_id (Intercept) 0.01542  0.1242
Number of obs: 14311, groups:  match_id, 63

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.543978   0.023669  22.982  < 2e-16 ***
Rel.Quality   0.057351   0.008771   6.539 6.21e-11 ***
Overall.Qual -0.008928   0.013702  -0.652    0.515
```

- FGLS:

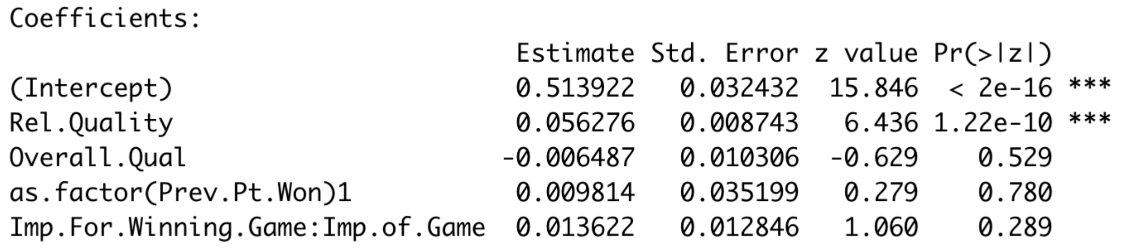


```
Coefficients:
                Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)   0.63226322  0.00107853 586.2275 < 2.2e-16 ***
Rel.Quality   0.01278246  0.00031909  40.0596 < 2.2e-16 ***
Overall.Qual -0.00159663  0.00061368  -2.6017  0.009276 **
```

- Interpretation
    - As expected, all the fitting techniques result in Rel.Qual being highly significant with positive coefficient effect, indicating the relative quality is a good predictor of the probability of winning a point. The estimated coefficient for Overall.Qual, expected to be positive as it is believed that better players tend to win more serving points, was actually negative; 0 is still in the 95% confidence interval in the GLM and GLMM.

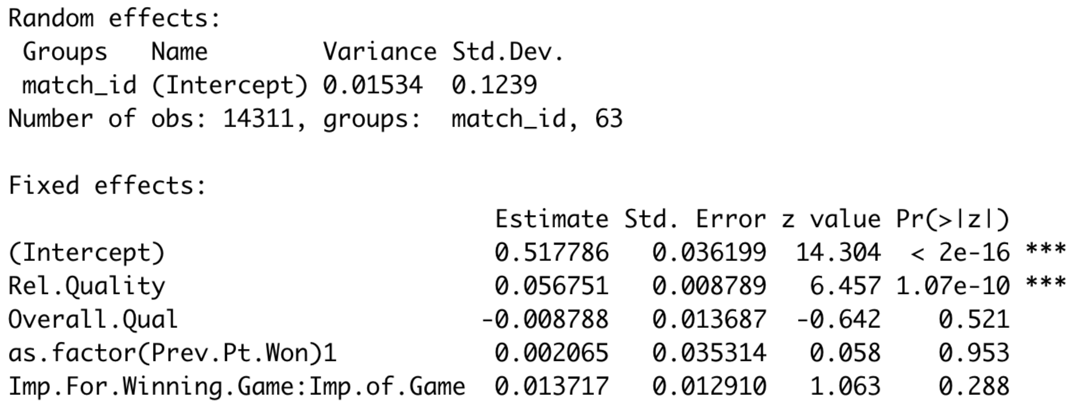
Model 1

- Logistic:



```
Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        0.513922   0.032432  15.846  < 2e-16 ***
Rel.Quality                        0.056276   0.008743   6.436 1.22e-10 ***
Overall.Qual                      -0.006487   0.010306  -0.629    0.529    
as.factor(Prev.Pt.Won)1            0.009814   0.035199   0.279    0.780    
Imp.For.Winning.Game:Imp.of.Game   0.013622   0.012846   1.060    0.289    
```

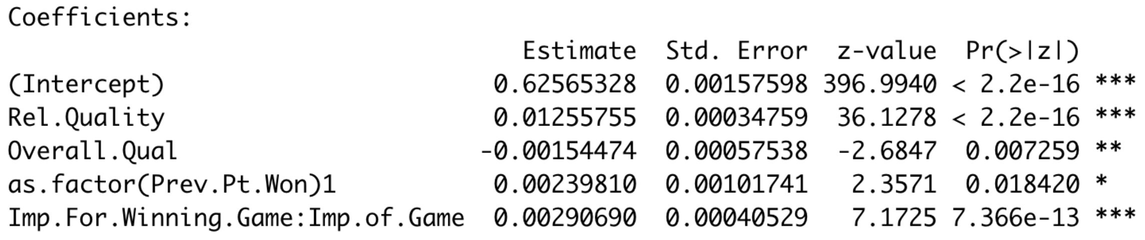
- GLMM: (AIC = 18736.4)



```
Random effects:
 Groups   Name        Variance Std.Dev.
 match_id (Intercept) 0.01534  0.1239  
Number of obs: 14311, groups:  match_id, 63

Fixed effects:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        0.517786   0.036199  14.304  < 2e-16 ***
Rel.Quality                        0.056751   0.008789   6.457 1.07e-10 ***
Overall.Qual                      -0.008788   0.013687  -0.642    0.521    
as.factor(Prev.Pt.Won)1            0.002065   0.035314   0.058    0.953    
Imp.For.Winning.Game:Imp.of.Game   0.013717   0.012910   1.063    0.288    
```

- FGLS:



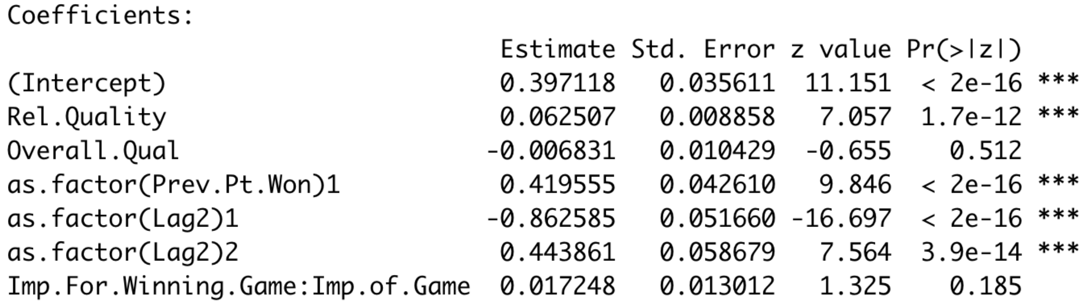
```
Coefficients:
                                    Estimate  Std. Error  z-value  Pr(>|z|)    
(Intercept)                       0.62565328  0.00157598 396.9940 < 2.2e-16 ***
Rel.Quality                       0.01255755  0.00034759  36.1278 < 2.2e-16 ***
Overall.Qual                     -0.00154474  0.00057538  -2.6847  0.007259 ** 
as.factor(Prev.Pt.Won)1           0.00239810  0.00101741   2.3571  0.018420 *  
Imp.For.Winning.Game:Imp.of.Game  0.00290690  0.00040529   7.1725 7.366e-13 ***
```

- Interpretation

- The observations in Model 0 were maintained in this model. As expected, both dynamic regressors (Prev.Pt.Won and Imp.For.Winning.Game:Imp.of.Game) resulted in a positive estimated coefficient. In the FGLS model, both are fairly significant.

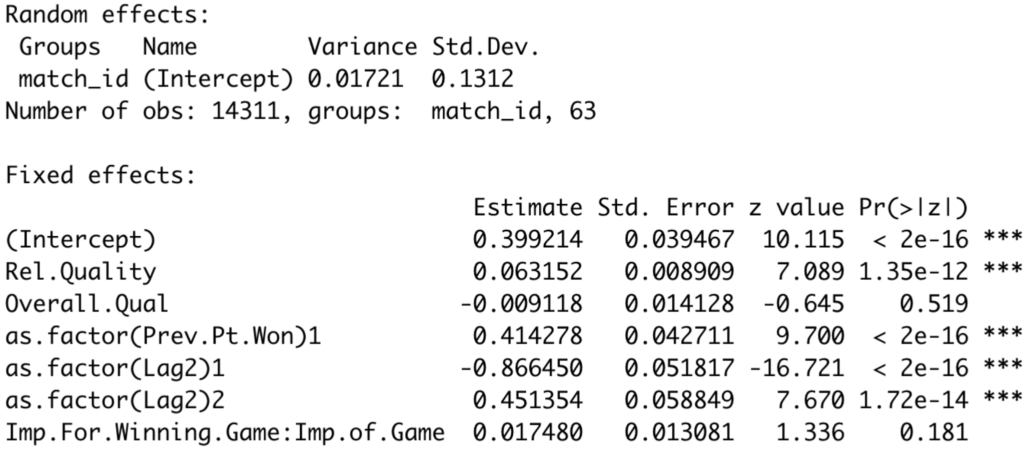
Model 2

- Logistic:



```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      0.397118   0.035611  11.151  < 2e-16 ***
Rel.Quality                      0.062507   0.008858   7.057  1.7e-12 ***
Overall.Qual                    -0.006831   0.010429  -0.655    0.512    
as.factor(Prev.Pt.Won)1          0.419555   0.042610   9.846  < 2e-16 ***
as.factor(Lag2)1                -0.862585   0.051660 -16.697  < 2e-16 ***
as.factor(Lag2)2                 0.443861   0.058679   7.564  3.9e-14 ***
Imp.For.Winning.Game:Imp.of.Game  0.017248   0.013012   1.325    0.185    
```

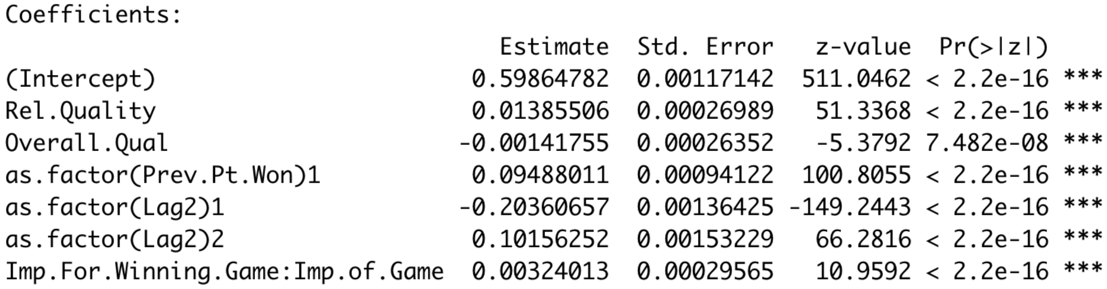
- GLMM: (AIC = 18398.7)



```
Random effects:
 Groups   Name        Variance Std.Dev.
 match_id (Intercept) 0.01721  0.1312  
Number of obs: 14311, groups:  match_id, 63

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      0.399214   0.039467  10.115  < 2e-16 ***
Rel.Quality                      0.063152   0.008909   7.089 1.35e-12 ***
Overall.Qual                    -0.009118   0.014128  -0.645    0.519    
as.factor(Prev.Pt.Won)1          0.414278   0.042711   9.700  < 2e-16 ***
as.factor(Lag2)1                -0.866450   0.051817 -16.721  < 2e-16 ***
as.factor(Lag2)2                 0.451354   0.058849   7.670 1.72e-14 ***
Imp.For.Winning.Game:Imp.of.Game  0.017480   0.013081   1.336    0.181    
```

- FLGS



```
Coefficients:
                                  Estimate  Std. Error   z-value  Pr(>|z|)
(Intercept)                     0.59864782  0.00117142  511.0462 < 2.2e-16 ***
Rel.Quality                     0.01385506  0.00026989   51.3368 < 2.2e-16 ***
Overall.Qual                   -0.00141755  0.00026352   -5.3792 7.482e-08 ***
as.factor(Prev.Pt.Won)1         0.09488011  0.00094122  100.8055 < 2.2e-16 ***
as.factor(Lag2)1               -0.20360657  0.00136425 -149.2443 < 2.2e-16 ***
as.factor(Lag2)2                0.10156252  0.00153229   66.2816 < 2.2e-16 ***
Imp.For.Winning.Game:Imp.of.Game  0.00324013  0.00029565   10.9592 < 2.2e-16 ***
```

- Interpretation
    - The previous observations on the Quality regressors were maintained. A very important result from this model is that in the three fitting methods used, Prev.Pt.Won is highly significant and has positive estimated coefficient. This seems to suggest that there is some dependence between the outcome of the previous point and the current point. The Lag2 variables are highly significant with negative coefficient estimates. While this result is contrary to that observed in other studies, it seems reasonably intuitive considering that all players are relatively good in quality. As all the sampled points come from one of the most prestigious tournaments in the Tour, it is not surprising that winning three consecutive points is not a frequent occurrence. This model shows Imp.For.Winning.Game:Imp.of.Game to have positive coefficient estimate, and it is significant under FGLS. These positive values are unexpected, as the initial assumption was that points of increased importance should result in decreased rate of success. However, it is possible that "playing under pressure" affects the receiver more than the server, resulting in positive coefficients.

Model 3

- Logistic:

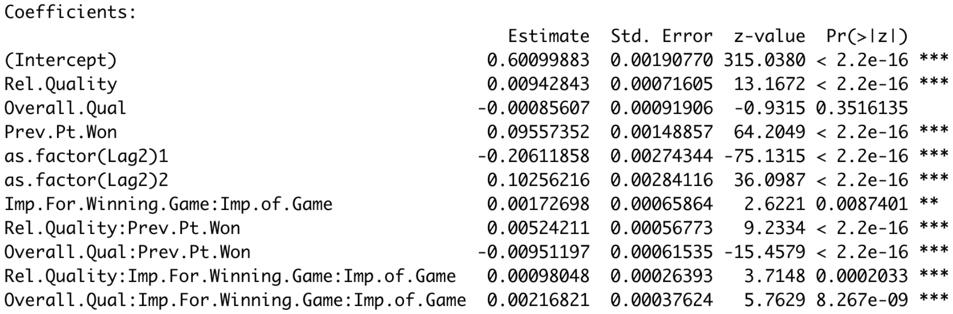
```
Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                   0.517843   0.032665  15.853  < 2e-16 ***
Rel.Quality                                   0.042132   0.016340   2.578  0.00992 ** 
Overall.Qual                                 -0.004969   0.019135  -0.260  0.79510    
Prev.Pt.Won                                   0.011180   0.035364   0.316  0.75190    
Imp.For.Winning.Game:Imp.of.Game              0.010753   0.013016   0.826  0.40875    
Rel.Quality:Prev.Pt.Won                       0.021654   0.017703   1.223  0.22126    
Overall.Qual:Prev.Pt.Won                     -0.042566   0.020863  -2.040  0.04132 *  
Rel.Quality:Imp.For.Winning.Game:Imp.of.Game  0.002264   0.006463   0.350  0.72615    
Overall.Qual:Imp.For.Winning.Game:Imp.of.Game 0.010929   0.007660   1.427  0.15363    
```

GLMM: (AIC = 18397.4)

```
Random effects:
 Groups   Name        Variance Std.Dev.
 match_id (Intercept) 0.01738  0.1318  
Number of obs: 14311, groups:  match_id, 63

Fixed effects:
                                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                   0.405113   0.039684  10.208  < 2e-16 ***
Rel.Quality                                   0.043198   0.016691   2.588  0.00965 ** 
Overall.Qual                                 -0.008120   0.021542  -0.377  0.70622    
Prev.Pt.Won                                   0.416039   0.042837   9.712  < 2e-16 ***
as.factor(Lag2)1                             -0.871420   0.051957 -16.772  < 2e-16 ***
as.factor(Lag2)2                              0.449107   0.058837   7.633 2.29e-14 ***
Imp.For.Winning.Game:Imp.of.Game              0.013835   0.013246   1.044  0.29626    
Rel.Quality:Prev.Pt.Won                       0.027767   0.018019   1.541  0.12332    
Overall.Qual:Prev.Pt.Won                     -0.045757   0.021164  -2.162  0.03062 *  
Rel.Quality:Imp.For.Winning.Game:Imp.of.Game  0.003970   0.006718   0.591  0.55452    
Overall.Qual:Imp.For.Winning.Game:Imp.of.Game 0.011959   0.007822   1.529  0.12630    
```

- FGLS



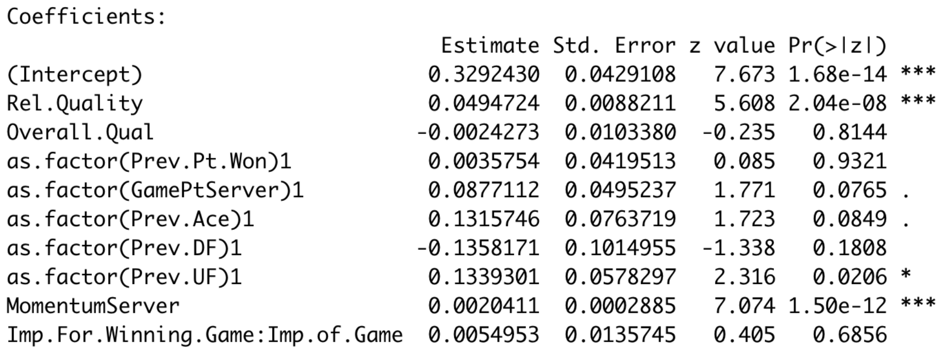
```
Coefficients:
                                               Estimate  Std. Error  z-value  Pr(>|z|)    
(Intercept)                                  0.60099883  0.00190770 315.0380 < 2.2e-16 ***
Rel.Quality                                  0.00942843  0.00071605  13.1672 < 2.2e-16 ***
Overall.Qual                                -0.00085607  0.00091906  -0.9315 0.3516135    
Prev.Pt.Won                                  0.09557352  0.00148857  64.2049 < 2.2e-16 ***
as.factor(Lag2)1                            -0.20611858  0.00274344 -75.1315 < 2.2e-16 ***
as.factor(Lag2)2                             0.10256216  0.00284116  36.0987 < 2.2e-16 ***
Imp.For.Winning.Game:Imp.of.Game             0.00172698  0.00065864   2.6221 0.0087401 ** 
Rel.Quality:Prev.Pt.Won                      0.00524211  0.00056773   9.2334 < 2.2e-16 ***
Overall.Qual:Prev.Pt.Won                    -0.00951197  0.00061535 -15.4579 < 2.2e-16 ***
Rel.Quality:Imp.For.Winning.Game:Imp.of.Game  0.00098048  0.00026393   3.7148 0.0002033 ***
Overall.Qual:Imp.For.Winning.Game:Imp.of.Game  0.00216821  0.00037624   5.7629 8.267e-09 ***
```

- Interpretation
  - This model preserves all the considerations in the previous model. Regarding the newly introduced interactions, the following observations can be made:
    - Rel.Quality:Prev.Pt.Won: Positive coefficients imply that dependence is stronger for players who are better than their opponents. This result can be misleading as it can be caused by a negative momentum by the weaker receiver. The following result gives a better idea of the relationship between dependence and quality.
    - Overall.Qual:Prev.Pt.Won: Negative coefficients imply that dependence is weaker in matches where both players are stronger. This result agrees with the expectation that most skilled players better approximate the independence assumption.
    - Rel.Quality:Imp.For.Winning.Game:Imp.of.Game: Positive coefficients imply that better players are more capable of neutralizing the effect of “playing under pressure” of important points. This result was expected, as most skilled and experienced players tend to handle pressure better, maintaining a more stable level of play throughout the match.

- Overall.Qual:Imp.For.Winning.Game:Imp.of.Game: Positive coefficients imply that the effect of "playing under pressure" is smaller when the overall quality of the match increases. This is also an expected result.

Model 4

- Logistic



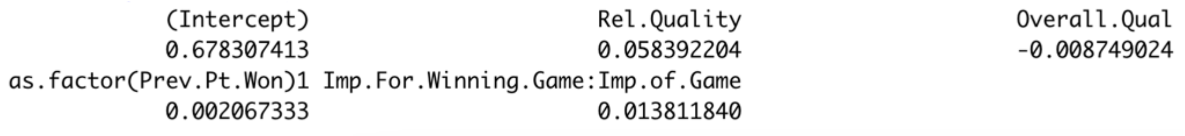
```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      0.3292430  0.0429108   7.673 1.68e-14 ***
Rel.Quality                      0.0494724  0.0088211   5.608 2.04e-08 ***
Overall.Qual                    -0.0024273  0.0103380  -0.235   0.8144    
as.factor(Prev.Pt.Won)1          0.0035754  0.0419513   0.085   0.9321    
as.factor(GamePtServer)1         0.0877112  0.0495237   1.771   0.0765 .  
as.factor(Prev.Ace)1             0.1315746  0.0763719   1.723   0.0849 .  
as.factor(Prev.DF)1             -0.1358171  0.1014955  -1.338   0.1808    
as.factor(Prev.UF)1              0.1339301  0.0578297   2.316   0.0206 *  
MomentumServer                   0.0020411  0.0002885   7.074 1.50e-12 ***
Imp.For.Winning.Game:Imp.of.Game 0.0054953  0.0135745   0.405   0.6856    
```

- Interpretation
  - Introducing all these regressors yields some unintuitive results. When using backward and stepwise AIC to find reduced models, both maintain Rel.Quality and MomentumServer as significant, but the rest of the regressors seem to add no meaningful contributions to the model. The use of a simple model, with one or two regressors per category, is also the approach followed by the other research in the field mentioned throughout the paper.

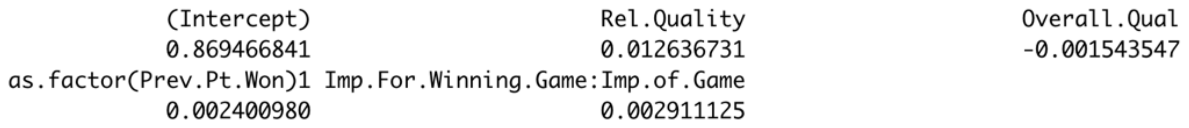
Odd Ratio Analysis

- The fitted probability of success divided by its complement, the fitted probability of failure, is called odds. Dividing the odds at x_{i+1}, where x is certain arbitrary regressor, by the Odds at x_i results in the odds ratio. The odds ratio of a particular regressor is equivalent to the estimated increase in the probability of success associated with a unit increase in the value of the regressor (Montgomery *et al.*, 2012).

- Below are the odds ratios for Model 1 (the model with the highest AIC value) using GLMM and FGLS.



```
           (Intercept)                     Rel.Quality                Overall.Qual
           0.678307413                     0.058392204                -0.008749024
as.factor(Prev.Pt.Won)1 Imp.For.Winning.Game:Imp.of.Game
           0.002067333                     0.013811840
```



```
           (Intercept)                     Rel.Quality                Overall.Qual
           0.869466841                     0.012636731                -0.001543547
as.factor(Prev.Pt.Won)1 Imp.For.Winning.Game:Imp.of.Game
           0.002400980                     0.002911125
```

- From the results obtained, there is some increase in the likelihood of winning a point by having won the previous point, as well as by playing a more important point. While the regressors are fairly significant, the change in probability is small enough that for most practical purposes it is appropriate to assume that points are independent and identically distributed.

Appendix D

Tests of Homogeneity

Formally, the test of homogeneity consists in partitioning a dataset into k samples, each of which represents a random variable $X_i \sim \text{Binomial}(n_i, \pi_i)$, where $n_1, \dots, n_k$, are known, and $\pi_1, \dots, \pi_k$ are unknown (Klein and Linton, 2013). The null hypothesis for the homogeneity claims that all the probability parameters are equal, while the negation of such a claim represents the alternative hypothesis. Formally, H_0: $\pi_1 = \ldots = \pi_k$, and H_1: $\pi_1 \neq \pi_j$ for some $i \neq j$.

Appendix E

Test of Nass

The Test of Nass consists of a modification to the standard chi-squared test that improves approximations on sparse data (Klein and Linton, 2013). In particular, the test statistic is defined by $c \times T_p|X_+ \sim \chi^2_{v,}$ where T_p is the Pearson's (1900) test statistic for the standard chi-squared test as defined below, χ_v is a chi-square random variable with v degrees of freedom, X_+ represents the total number of successes, and c and v are chosen so that the conditional mean and variance of $c \times T_p$ would equal the respective mean and variance of the approximating chi-square distribution. Mathematically,

$$E(c \times T_p|X_+) = v,\ \text{Var}(c \times T_p|X_+) = 2\text{v},$$

or equivalently,

$$c = 2E(T_p|X_+) \,/\, \text{Var}(T_p|X_+),\ v = cE(T_p|X_+).$$

Dawson (1954) presented a simplified version of Haldane's (1940) equations for the mean and variance of T_p conditioned on X_+ under the null hypothesis.

Here, k represents the total number of partitions, n_i is the sample size for the ith group, and X_i is the number of successes in a group. Similarly, $n_+ = \sum_{i=1}^{k} n_i$ and $X_+ = \sum_{i=1}^{k} X_i$. Hence, the p-value for the test can be computed as the probability $\Pr\{\, \chi^2_v > ct_p\}$. Pearson's (1900) test statistic, as displayed by Klein and Linton (2013), is defined by $T_p = T_p(X_1, \ldots, X_k) = \sum_{i=1}^{k} \frac{n_i\,(\hat{\pi}_i - \hat{\pi})}{\hat{\pi}\,(1 - \hat{\pi})}$), where, $\pi_i = x_i/n_i$ is the ratio of successes over occurrences in a group. $\hat{\pi} = x/n$ is the overall sample rate of success, which is the maximum likelihood estimator of the population probability of success given the null hypothesis.