An Analysis on the Effects of Chess Openings on Games

Judin Thomas

Chess is one of the most universal games in the world, with over 600 million players worldwide. The mechanisms of the game create a large amount of permutations that lead to one of three outcomes- a win, loss, or draw. While there has been an abundance of statistical research based on building algorithms to play chess against humans or other computers, there exists surprisingly little literature on looking into games played. This is despite the presence of online chess, which has been growing in prominence overall and especially skyrocketed over the course of the Covid pandemic. Online games, played between two humans, are easier to notate digitally and can easily include metadata about the game in manner that was much more difficult to collect with in-person games.

However, even with this data, chess games can be very difficult to analyze on a larger level because, as Dr. Dennis Holding points out in his book about statistical analysis of chess positions, 6 turns forward in a game can often generate more than 3 billion permutations. Because of this difficulty, the analysis in this paper focuses on a compares general information about the game to one of the most crucial but more easily categorical parts of the game- the opening, or the first move or two.

Openings tend to determine a lot about the game since they indicate how open or closed the board is and how aggressively or conservatively the beginning of the game is played. There are openings for players playing both black, and white, but white tends to have a larger influence because it makes the first move. Having the first move gives two clear advantages to white- they can choose the openings they are most comfortable with and they have a slightly quicker "tempo" (they are a move ahead of their opponent, so they have an easier time attacking). This is seen statistically as estimates show that white players tend to win 52-58% of the time, giving them a slight edge.

This study primarily explores those two aspects of openings by trying to analyze what role the most common openings have on the general characteristics of the chess gamel. In addition, the study tries to understand what factors play the most into the statistical edge white players benefit from using information from openings. On a general level, this analysis will help to understand how different openings affect the game, whether in terms of how many moves are played, or the type of win, or even the skill of the players. In this process, some of what affects the ability of players playing white to win with a slight edge might be revealed. This leads to the two main objectives:

How do the most common chess openings affect other aspects of the game?

How determinate are different aspects of the games when predicting whether white will win, lose, or draw?

Data

The data was taken from Kaggle, which took the data from a common chess playing website, Lichess. Within the websites, it took games from within "teams" which are smaller communities within the website. However, this did create a large skew in terms of the skill level of those playing in these communities compared to the average chess player. Lichess is already a more niche website than its large competitor Chess.com, meaning those using the site are already likely more skilled. In addition, taking from these groups subsets the users even further into the more active players. This is why the average user rating in the dataset is 1595, much higher than the average of 943 seen on chess.com.

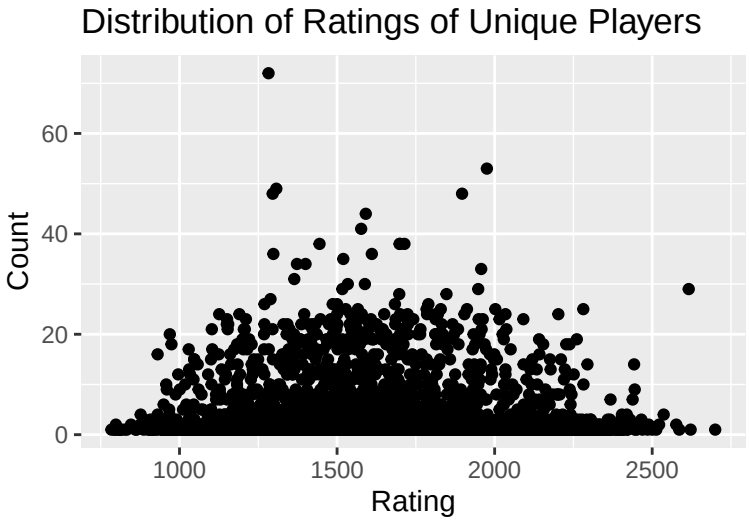


Figure 1

However, while the data comes from is a very skewed sample of games from highly rated players, this actually serves the analysis itself well because it presents a pool of games played by experienced players. This shows the game in its truer forms and takes out elements of chess that more amateur games may introduce. However, this difference is still taken into account by incorporating the ratings of the players, to see if less skilled players play differently enough to effect the outcomes being measured.

The data contains an array of variables related to the chess games it documented. It organizes observations by games and includes rating and username information about each side as well as data about the game itself, This includes what time the game was played, which side won, how they won, different classifications of openings used, the number of turns, a booleon for whether or not the player was rated, the time limit for the game, and even a full list of all the moves made.

Nonetheless, the first dependent variable used was one that was created from the dataset. This variable, Opening, was created to showcase the different openings that were played. While the data included some classifications, the number of levels was too large to be analyzed effectively. Therefore, this study created a new response variable based off of the game's move log, separating openings into three categories. These categories encompass two of the most common plays (E4 and D4), which make up a large majority of first moves, and then another category for all other openings.

Table 1: Table 1

Opening	Count	Proportion
D4	4513	0.23
E4	12569	0.63
Other	2940	0.15

Table 1 shows that E4 is played in a majority of games, with D4 being the next most common opening and all other moves only making up 15% of openings. These three categories are large enough to allow for analysis but still divide openings into categories that encompass the three major ways a player can start the game. While breaking up E4 into more openings would be more useful, the permutations quickly become too great for this to become feasible. Furthermore, each of these openings generally leads to a very different game, so making further splits might limit analysis by making categories too similar.

The second response variable was given in the data as the winner. It is divided into black, white, and draw.

Table 2: Table 2

Winner	Count	Proportion
black	9092	0.45
draw	949	0.05
white	9981	0.50

As seen in Table 2, in this specific dataset, white won 50% of the time, with Black winning 45% and a draw 5% of the time. Chess, as a board game, is generally considered to be one of the most skill-based games as opposed to chance-based, so it is interesting seeing such a high discrepancy even at such a high level. As mentioned before, two of the biggest factors can be the tempo advantage and choosing the opening.

Not all of the other variables were used for the explanatory variables. In order to have a single rating variable and reduce colinearity, the black and white ratings were averaged into a single rating. The victory status is used as it is, with the options being a draw, a resignation, a checkmate, or a time expiration. Turns includes the total number of moves each player makes. The final explanatory variable is the time increment, which is modified to exclude the fisher delay and instead only includes the main time increment. The newly created opening variable was also used as an explanatory variable for the second model.

Exploratory Data Analysis

One of the most interesting explanatory variables is the rating because it is not about the game itself but about the player. In addition, how chess is played generally evolves as the player gets more skilled. While most of the players are around the same skill level as seen above, there is a significant range of ratings within the dataset, which will likely have some effect on the model. The correlation of rating with the dependent variables is figures 2 and 3.

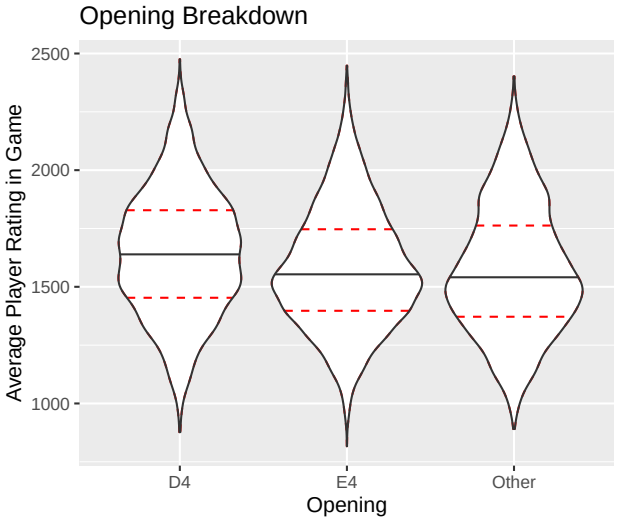


Figure 2

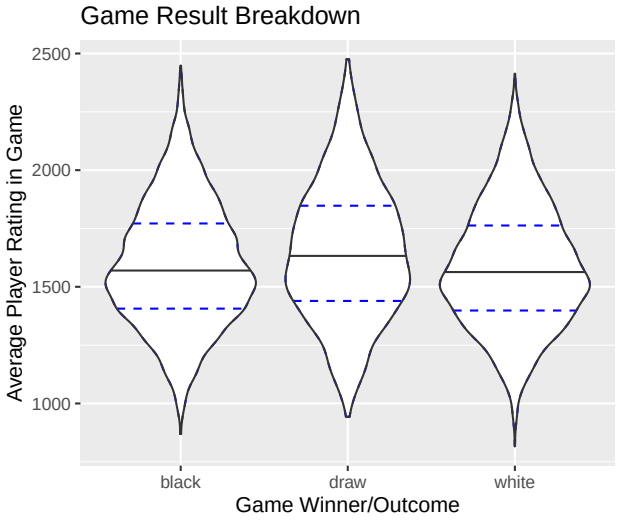


Figure 3

The opening distributions are interesting because the most common opening, E4, generally follow the same distribution of ratings as the general dataset. Since E4 is the most general and common opening, this is expected. It is interesting to see the distributions of the other openings. The other openings are used in generally the same way as E4, but with a larger tail for the most skilled players. D4 tends to be used by more skilled players overall, perhaps because advanced players may want to gain an advantage by changing the game into something they might be more comfortable with. They would be able to utilize a common but different opening in D4. However, the second graph shows this might not be the case. The distributions reveal that gaining skill does not increase the white/black winning gap. This visualization also shows an

interesting trend-draws are more common among better players. The ways a game can be drawn generally require a tighter game with no major mistakes, which is why this pattern may appear.

Two other interesting variables that tell a lot about the game itself is the number of turns and the timing for the game. As shown in figure 4, these variables do not have the relationship that an outside observer may expect.

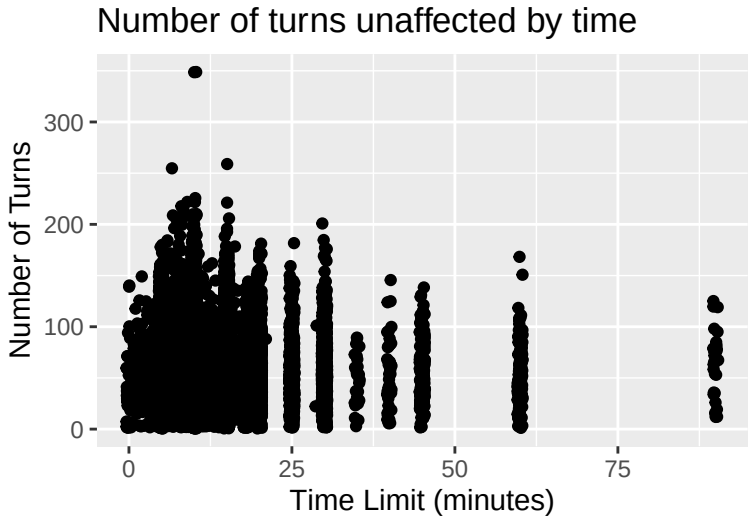


Figure 5

Contrary to expectations, games with longer time limits do not have more moves. This indicates that players tend to pace out their moves well and simply take more time for each move instead of moving more in the entire game. Luckily, the lack of correlation prevents possible issues of colinearity and allows the models to utilize both variables.

Methodology

For both models, multinomial logistic regression was used to build a model to predict either the first move or the outcome. This was appropriate since both response variables were categorical but had multiple levels. Ultimately, chess is a game of discrete moves leading to a discrete outcome, so a non-continuous outcome was likely and therefore more traditional linear regression techniques were not sufficient. Multinomial logistic regression uses a generalized linear model, which means the response variable undergoes some of transformation to allow the data to be analyzed in a more meaningful way. In this case, that function alters the response by transforming into a log-odds ratio, as written below:

$log(\frac{p_i}{1-p_i})$

This transformed response represents the log odds ratio, or the log of the probability of an event occurring over the probability of it not happening. However, in multinomial logistic models, log odds ratio is analyzed between each of the different levels within the outcome variables.

In the first model analyzing what factors impact the likelihood of a specific opening having been played, the model's output will reveal the multiplicative effect a one unit increase or binary switch in an explanatory variable will have on the likelihood of a given opening having been played. The probabilistic framework of a multinomial logistic model helps aid in the understanding of the largest factors and effects of different openings being played.

The second model is similar in its use of multinomial logistic regression, except to predict a white win versus a black win or draw. Simple logistic regression was considered for this outcome, since drawn games are far rarer than games with outcomes and a black win versus a white win is a more clear cut distinction. However, since a draw is such a unique outcome in a game and likely tied to factors being analyzed such as number of moves and opening, its inclusion may reveal more about the game of chess.

The explanatory variables of a logistic model are created into a linear equation, making the full model follow the following formula.

$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ...$

However, multinomial logistic models still have heavy assumptions that need to be met for the analysis to be interpretable. The three assumptions included in this are linearity, independence, and randomness. Independence may seem like an issue because many of the same users are playing the same games. However, no individual users played a large amount of the games, and since the players are matched with opponents of similar ratings, the games are generally independent of each other since a highly rated player will play with other highly rated players. In addition, while the group the games is selected from is does not make them a random subset of the population, the large amount of players and the randomness of the groups (ie the groups are not organized by anything related to the game characteristics), these games would tend to be good representations of general chess games played at a fairly high level. The final aspect is linearity. To check this aspect, binned residual plots were created and analyzed for general randomness. These plots are available in the appendix.

Results

Opening Analysis: D4/E4/Other Model

Opening	Term	Estimate	Statistic	Std. Error	P-Value	CI Lower	CI Upper
E4	Intercept	2.615	0.009	288.2	0.0000	2.60	2.63
E4	Turns	-0.002	0.001	-4.1	0.0000	0.00	0.00
E4	Victory Status: Checkmated	-0.016	0.026	-0.6	0.5459	-0.07	0.04
E4	Victory Status: Time Expired	-0.041	0.024	-1.7	0.0908	-0.09	0.01
E4	Victory Status: Resigned	-0.055	0.026	-2.1	0.0333	-0.11	0.00
E4	Time (min)	0.005	0.002	2.2	0.0301	0.00	0.01
E4	Rating	-0.001	0.000	-22.1	0.0000	0.00	0.00
E4	Rated: True	-0.100	0.041	-2.5	0.0142	-0.18	-0.02
Other	Intercept	1.696	0.006	269.6	0.0000	1.68	1.71
Other	Turns	-0.003	0.001	-4.3	0.0000	0.00	0.00
Other	Victory Status: Checkmated	-0.121	0.029	-4.1	0.0000	-0.18	-0.06
Other	Victory Status: Time Expired	-0.078	0.017	-4.7	0.0000	-0.11	-0.05
Other	Victory Status: Resigned	-0.268	0.030	-8.8	0.0000	-0.33	-0.21
Other	Time (min)	0.006	0.003	2.0	0.0507	0.00	0.01
Other	Rating	-0.001	0.000	-19.2	0.0000	0.00	0.00
Other	Rated: True	-0.349	0.053	-6.5	0.0000	-0.45	-0.24

Full Model: E4/D4

$$log(\frac{\hat{p_{i,E4}}}{p_{i,D4}}) = 2.615 - 0.002 \times Turns_i - 0.016 \times Checkmated_i - 0.041 \times TimeExpired_i$$
$$-0.055 \times Resigned_i + 0.005 \times Time_i - 0.001 \times Rating_i - 0.100 \times Rated_i$$

Full Model: Other Opening/D4

$$log(\frac{\hat{p_{i,Other}}}{p_{i,D4}}) = 1.696 - 0.003 \times Turns_i - 0.0121 \times Checkmated_i - 0.078 \times TimeExpired_i$$
$$-0.2.68 \times Resigned_i + 0.006 \times Time_i - 0.001 \times Rating_i - 0.349 \times Rated_i$$

Opening Analysis Output: Exponentiated

Opening	Term	Estimate	Statistic	Std. Error	P-Value	CI Lower	CI Upper
E4	Intercept	13.663	0.009	288.2	0.0000	13.42	13.91
E4	Turns	0.998	0.001	-4.1	0.0000	1.00	1.00
E4	Victory Status: Checkmated	0.984	0.026	-0.6	0.5459	0.93	1.04
E4	Victory Status: Time Expired	0.960	0.024	-1.7	0.0908	0.92	1.01
E4	Victory Status: Resigned	0.946	0.026	-2.1	0.0333	0.90	1.00
E4	Time (min)	1.005	0.002	2.2	0.0301	1.00	1.01
E4	Rating	0.999	0.000	-22.1	0.0000	1.00	1.00
E4	Rated: True	0.904	0.041	-2.5	0.0142	0.83	0.98
Other	Intercept	5.451	0.006	269.6	0.0000	5.38	5.52
Other	Turns	0.997	0.001	-4.3	0.0000	1.00	1.00
Other	Victory Status: Checkmated	0.886	0.029	-4.1	0.0000	0.84	0.94
Other	Victory Status: Time Expired	0.925	0.017	-4.7	0.0000	0.89	0.96
Other	Victory Status: Resigned	0.765	0.030	-8.8	0.0000	0.72	0.81
Other	Time (min)	1.006	0.003	2.0	0.0507	1.00	1.01
Other	Rating	0.999	0.000	-19.2	0.0000	1.00	1.00
Other	Rated: True	0.705	0.053	-6.5	0.0000	0.64	0.78

Game Outcome Analysis: White/Black/Draw Model

Opening	Term	Estimate	Statistic	Std. Error	P-Value	CI Lower	CI Upper
draw	Intercept	-4.199	0.003	-1203.0	0.0000	-4.21	-4.19
draw	Turns	0.018	0.001	21.0	0.0000	0.02	0.02
draw	Rating	0.000	0.000	7.3	0.0000	0.00	0.00
draw	Rated (True)	-0.458	0.009	-52.0	0.0000	-0.48	-0.44
draw	Time (min)	0.021	0.004	5.7	0.0000	0.01	0.03
draw	Opening: E4	0.016	0.055	0.3	0.7764	-0.09	0.12
draw	Opening: Other	0.079	0.030	2.6	0.0092	0.02	0.14
white	Intercept	0.321	0.011	30.5	0.0000	0.30	0.34
white	Turns	-0.003	0.000	-5.9	0.0000	0.00	0.00
white	Rating	0.000	0.000	-1.7	0.0896	0.00	0.00
white	Rated (True)	-0.014	0.034	-0.4	0.6886	-0.08	0.05
white	Time (min)	0.004	0.002	2.1	0.0376	0.00	0.01
white	Opening: E4	0.003	0.032	0.1	0.9258	-0.06	0.07
white	Opening: Other	-0.139	0.044	-3.1	0.0017	-0.23	-0.05

Full Model: Draw/Black Win

$$log(\frac{p_{i,\hat{Draw}}}{p_{i,BlackWin}}) = -4.199 + 0.018 \times Turns_i - 0.458 \times Rated_i + 0.021 \times Time_i + 0.016 \times E4_i + 0.079 \times OtherOpening_i$$

Full Model: White Win/Black Win

$$log(\frac{p_{i,\hat{WhiteWin}}}{p_{i,BlackWin}}) = 0.321 - 0.003 \times Turns_i - 0.014 \times Rated_i + 0.004 \times Time_i + 0.003 \times E4_i - 0.139 \times OtherOpening_i$$

Opening Model Output: Exponentiated

Opening	Term	Estimate	Statistic	Std. Error	P-Value	CI Lower	CI Upper
draw	Intercept	0.015	0.003	-1203.0	0.0000	0.01	0.02
draw	Turns	1.018	0.001	21.0	0.0000	1.02	1.02
draw	Rating	1.000	0.000	7.3	0.0000	1.00	1.00
draw	Rated (True)	0.633	0.009	-52.0	0.0000	0.62	0.64
draw	Time (min)	1.021	0.004	5.7	0.0000	1.01	1.03
draw	Opening: E4	1.016	0.055	0.3	0.7764	0.91	1.13
draw	Opening: Other	1.082	0.030	2.6	0.0092	1.02	1.15
white	Intercept	1.379	0.011	30.5	0.0000	1.35	1.41
white	Turns	0.997	0.000	-5.9	0.0000	1.00	1.00
white	Rating	1.000	0.000	-1.7	0.0896	1.00	1.00
white	Rated (True)	0.986	0.034	-0.4	0.6886	0.92	1.05
white	Time (min)	1.004	0.002	2.1	0.0376	1.00	1.01
white	Opening: E4	1.003	0.032	0.1	0.9258	0.94	1.07
white	Opening: Other	0.870	0.044	-3.1	0.0017	0.80	0.95

Discussion

Analysis of Results

The two models revealed an interesting analysis on how the very beginning of a chess game affects the rest of the game. While very few of the results were substantial in magnitude, considering the perceived fairness in chess as well as the large sample from which this analysis is derived, even small effects are relevant.

Starting with the initial model, many of the coefficients were found to be statistically significant at an alpha level of 0.05. D4 as an opening seems to lead to longer games in terms of moves compared to E4 or other openings. Since this longer game effect for D4 is true compared to both more and less popular openings, this effect is unlikely to be from changes in familiarity with the opening affecting how the game is played. Instead, a better explanation might be that the positioning of the opening might be closing off the center in a way that leads to a less aggressive game. The D4 opening is also correlated with shorter time control games (note that this variable represents the time limit chosen at the beginning of the game, not the total time taken for the game). Games starting with D4 tend to be shorter time limit games but contain more moves, a unique combination that reflects the effects of a more closed opening.

Another interesting result is that rated players tend to prefer the D4 opening over E4 or other openings. As mentioned before, D4 is the most common opening outside of the most popular opening, E4, so this may indicate more experienced players wanting to change their opening but not stray too far from what is known. Finally, the victory status tells an interesting story about playing other openings. The impacts of playing the other openings on the final outcome of the game, when compared to the baseline of playing D4, are far more significant than playing E4 compared to D4. This indicates that playing D4 more often leads to a draw. Considering the previous findings about D4, the underlying effect of playing a more closed off game with more moves may be this correlation with a larger amount of draws. However, the model also shows that games played with less unique openings are finished more often, as opposed to a resignation. On the opposite spectrum of D4, this could be from these more unique openings correlating with more aggressive games as the more unique openings introduce a larger level of uncertainty.

The second model investigates the effects playing white or black on the game by looking at what correlates with a white or black win. White won games generally look very similar to black won games, with only a few exceptions. Longer time limit games tend to help white, compared to black by a small but statistically significant amount. A longer time control may reduce mistakes, leading to an easier emergence of this natural benefit from playing white. However, the most interesting result comes from playing non-traditional openings. The advantage between playing white and black greatly diminishes when white plays a non-traditional other opening (one besides E4 or D4). This correlation may show because these openings lead to a riskier game than the alternative openings. Another surprise was that rating did not affect the models in a significant way, indicating that within this smaller subsection of skill level, openings play a similar role across ratings.

Conclusion

While few effects from the model outputs create a substantial difference, this is expected considering any major inequities would have been realized. Nonetheless, the analysis of the Lichess data does reveal some suggestions for increasing the odds of winning. According to the analysis, white should avoid D4 to prevent losing the significant advantage that comes with playing as white. This occurs because playing D4 tends to leads to longer games which are more often to end in a draw. White can also be benefited by choosing a longer time control, which the secondary analysis revealed. In addition, other non-traditional openings introduce uncertainty in the mix that reduces white's ability to capitalize on its advantage, E4 is the most correlated opening with white wins. In addition, longer time controls may help white avoid the mistakes and blunders that prevent the natural advantage from showing. Through this analysis, the findings of differences in openings were able to be combined with secondary discoveries about playing as white to create suggestions based off of these correlations in the data set.

Limitations and Future Work

There still exists a number of limitations to this analysis. Some of the more fundamental issues are the difficulties with analyzing chess as a game. Because of the complexity of the game, there are few variables generalizable to every game to analyze. In reality, the three divisions made to analyze openings are still very broad. However, given the permutations opened when black makes its first move, deeper analyses are difficult. Generally, chess is analyzed using more complex methodologies that rely on heavy computation, so more classical statistical techniques can fall short. A deeper analysis of the metadata of the game could be done utilized using nonparametric methods that might lead to a better analysis of the gameplay itself. These sorts of methods might be able to take a deeper look into the moves made in the game than some of the larger groups and generalizations required to be used in this analysis.

While the analysis in this paper uses a useful subset of games, the analysis could also be furthered with a larger and more general sample. One addressable issue with the data is the differences in ratings between players. As seen in the appendix in Figure 5, the rating differences are still generally centered, but having even smaller rating differences may eliminate some of the trends seen in the results as the games would be more fair. Having these smaller differences could lead to more valid analyses. This data could be collected easily from Lichess or Chess.com from the general population of games, which tend to be closer in rating than the ones seen in this dataset.

Sources

"Binnedplot: Binned Residual Plot." RDocumentation, Datacamp, https://www.rdocumentation.org/packages/arm/versions/1.12-2/topics/binnedplot.

J, Mitchell. "Chess Game Dataset (Lichess)." Kaggle, 4 Sept. 2017, https://www.kaggle.com/datasnaek/chess. LegendOfKass

"How to Format All Numbers in a Table in R Using Kable?" Stack Overflow, 1 Apr. 1967, https://stackoverflow.com/questions/55093178/how-to-format-all-numbers-in-a-table-in-r-using-kable.

"On Chess: Online Chess Interest Soars since the Start of the Pandemic." STLPR, St. Louis Chess Club, 1 Apr. 2021, https://news.stlpublicradio.org/2021-04-01/on-chess-online-chess-interest-soars-since-the-start-of-the-pandemic.

Tackett, Maria. Multinomial Logistic Regression, Duke University, https://www2.stat.duke.edu/courses/Spring19/sta210.001/slides/lec-slides/17-multinomial-logistic.html#27.

Y.Coch, et al. "Manual Specification of draw_quantile Argument in Violin Plot Call in R." Stack Overflow, 1 Apr. 1966, https://stackoverflow.com/questions/48936704/manual-specification-of-draw-quantile-argument-in-violin-plot-call-in-r.

https://stackoverflow.com/questions/48936704/manual-specification-of-draw-quantile-argument-in-violin-plot-call-in-r

https://www2.stat.duke.edu/courses/Spring19/sta210.001/slides/lec-slides/17-multinomial-logistic.html#27

https://www.rdocumentation.org/packages/arm/versions/1.12-2/topics/binnedplot

https://www.kaggle.com/datasnaek/chess

https://stackoverflow.com/questions/55093178/how-to-format-all-numbers-in-a-table-in-r-using-kable

https://news.stlpublicradio.org/2021-04-01/on-chess-online-chess-interest-soars-since-the-start-of-the-pandemic

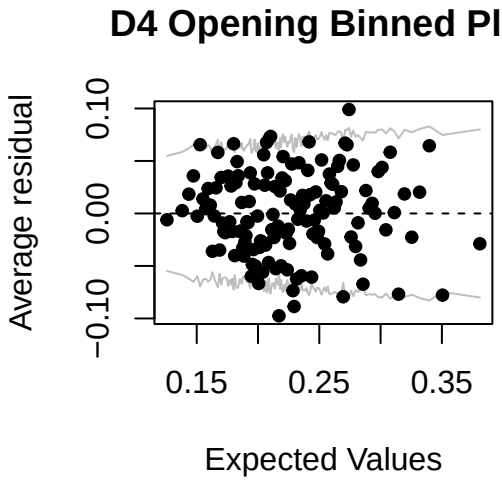
Appendix

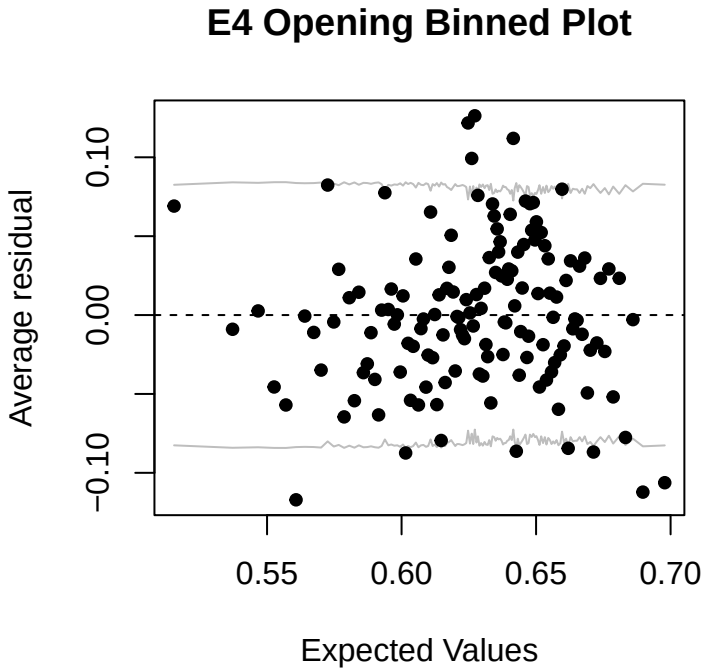
Outcome	Count	Proportion
draw	905	0.05
mate	6308	0.32
outoftime	1679	0.08
resign	11130	0.56

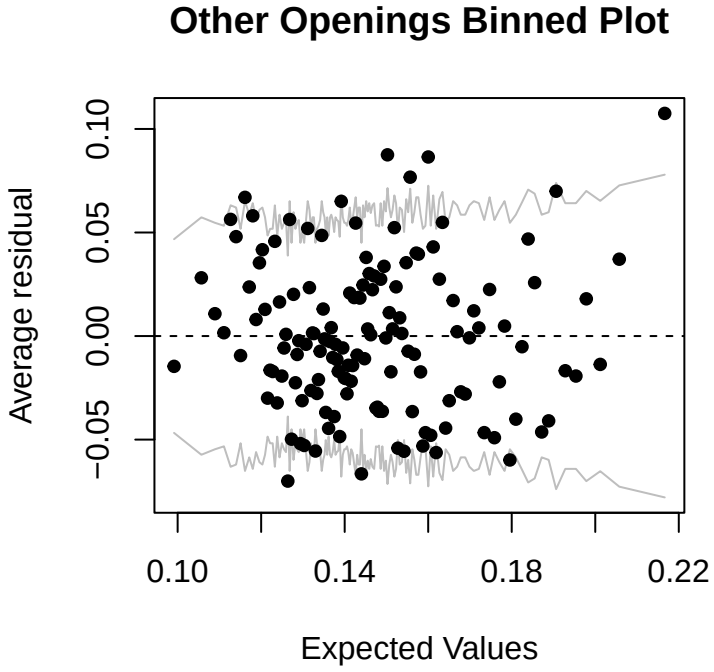
This chart shows the breakdown of game endings for the variable used in the first model on openings.

Below are the residual bin plots for the response variable levels.

Opening Response Residual Binned Plots

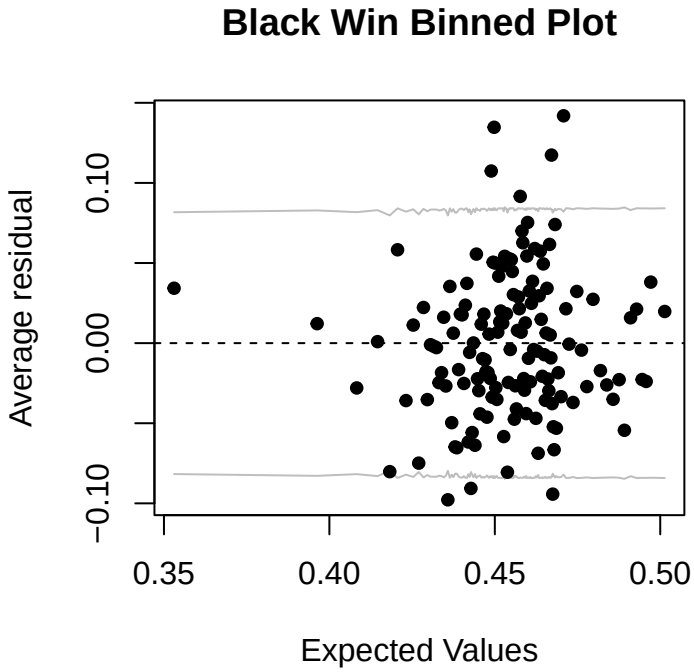


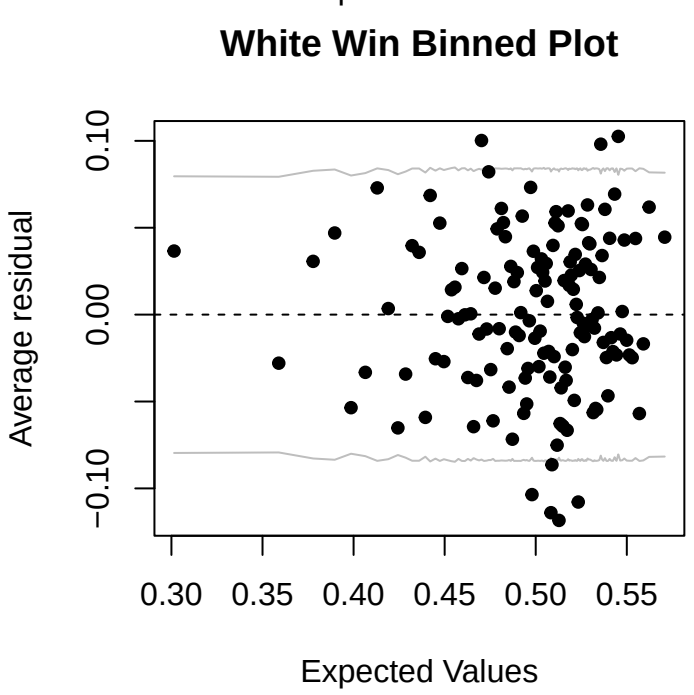


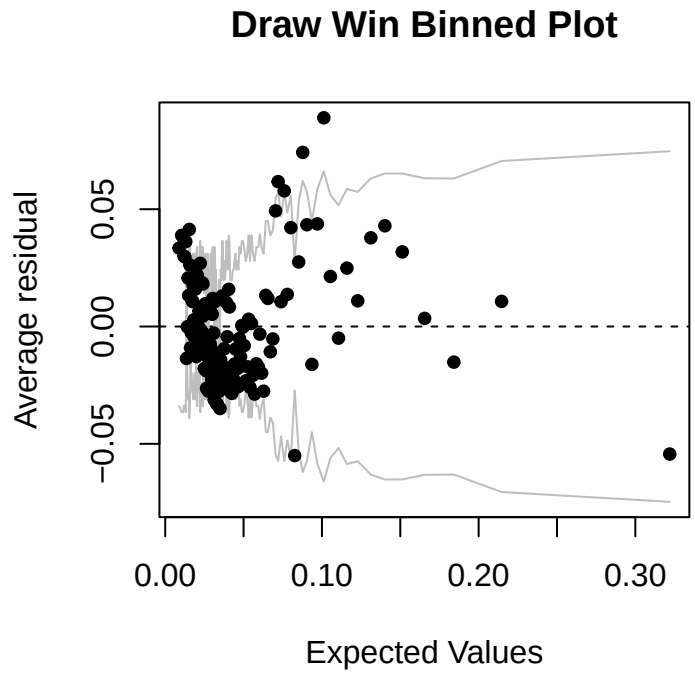


Game Outcome Residual Binned Plots

```
## Joining, by = "obs_num"
## Joining, by = "obs_num"
```


Black Win Binned Plot
Average residual
0.10
0.00
−0.10
0.35
0.40
0.45
0.50
Expected Values


White Win Binned Plot
Average residual
0.10
0.00
−0.10
0.30
0.35
0.40
0.45
0.50
0.55
Expected Values



These residual bin plots prove the linearity assumption of multinomial logistic regression.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

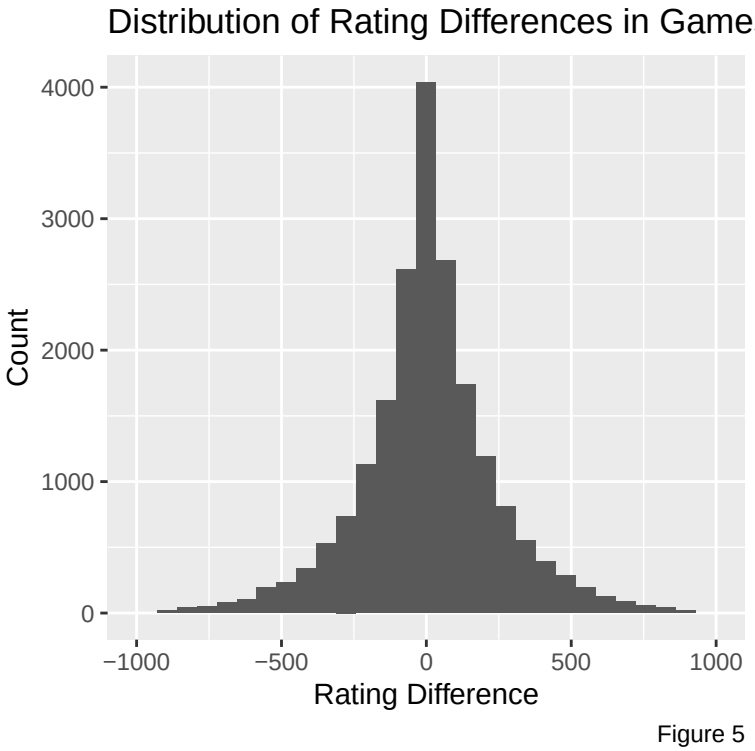


Figure 5

Figure 5 reveals the distribution of rating differences mentioned in the discussion.