Probability of Models in Practice: Poisson Vs Binomial Distribution

Zuriahn Yun

Abstract

This study examines the accuracy of the Poisson and Binomial distributions when applied to a dataset containing only the maximum values (Max-Kp) of geomagnetic activity. The goal was to determine which distribution provides a better fit for this dataset. The analysis was conducted by applying both distributions to the Max-Kp dataset for four different years, each spaced seven years apart, to capture a range of solar activity levels. The model fits were evaluated using linear regression, and the R^2 scores were calculated to numerically assess the accuracy of each distribution both of which I learned in Data 311. The results show that the Binomial distribution consistently outperforms the Poisson distribution in terms of R^2 scores, indicating a better fit for the Max-Kp data. This finding has implications for the modeling of geomagnetic activity and can inform future research in space weather prediction.

Background and Significance

The K-index is a metric used to categorize the magnitude of geomagnetic storms on a 0 to 9 scale, with 9 representing the most extreme storm and 0 indicating minimal geomagnetic activity. Geomagnetic storms have significant impacts on the electrical power grid, spacecraft communications, and other technologies sensitive to space weather. Understanding and predicting storms is critical for mitigating potential disruptions to infrastructure. This study aims to determine which statistical distribution, Poisson or Binomial better fits the Kp-max dataset, which represents the maximum observed Kp values for a given day. By exploring whether there is a significant difference between the two contributions, this research contributes to improving the accuracy of geomagnetic storm predictions which can benefit fields ranging from space exploration to energy grid management. When looking at Kp-Max, is there a significant difference between the Poisson and Binomial distributions?

This research investigates whether the Binomial distribution provides a better fit than the Poisson distribution when applied to the maximum values of the Kp-index (Max-Kp). Our null hypothesis states that there will be no difference between the mean R^2 scores of the linear regression of the Binomial and Poisson distributions or Ho: $\mu1 - \mu2 = 0$. Our alternative hypothesis states that there will be a positive difference between the mean R^2 scores, or HA: $\mu1 - \mu2 > 0$.

Method

Dataset

For this study, we analyzed the maximum daily values of the Kp-index (Kp max) for the years 2001,2008,2015 and 2022. These years were chosen to capture a diverse range of solar activity levels, to capture the ambiguity in geomagnetic disturbances. The kp max reflects the highest level of geomagnetic activity observed for 24 hours, making it a critical measure for understanding extreme space weather conditions. By focusing on Kp max, we can evaluate periods of peak geomagnetic disturbances, which are particularly relevant for auroras and for assessing the potential impacts on technology and infrastructure.

Data Analysis

The data was input in R-Studio where from there each of the 4 years was mapped to the Binomial and Poisson distribution. From there they were each mapped to a linear regression to see how well each distribution fit the dataset, to measure how well the dataset fit I took the R^2 scores from the linear regression and compared the R^2 scores for both distributions using a T-test and confidence intervals.

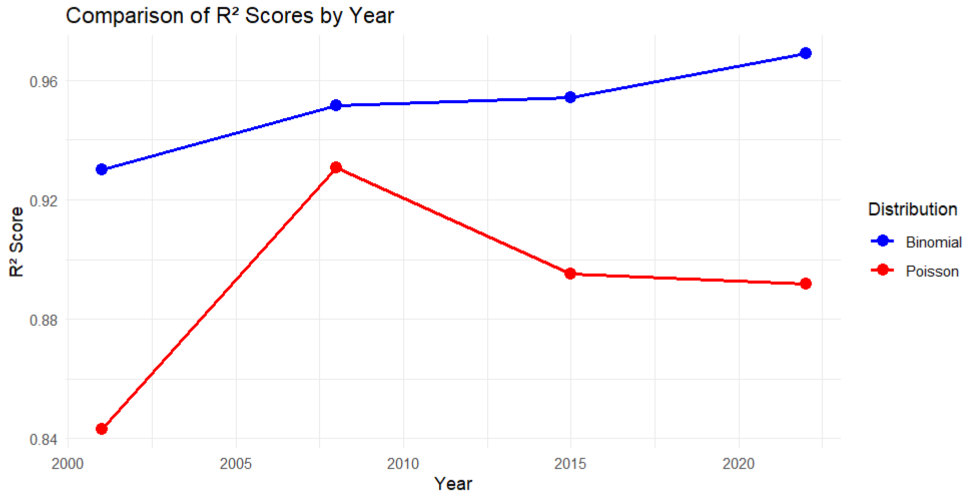
Results

The results of this study were that the Binomial distribution had an average R^2 score of 0.9512 and the Poisson distribution had an average R^2 score of 0.89028. Over the years observed not once did the Poisson distribution have a higher R^2 score than the Binomial distribution. The main correlation between the two was that they would both consistently predict with similar accuracy in terms of their distribution, if the Binomial distribution performed better as did the Poisson distribution, the Poisson distribution was just less accurate in comparison to the Binomial.

For the Binomial and the Poisson distribution, I applied them to the dataset and applied a linear regression comparing the dataset to each distribution. From there I grabbed the R² score for the linear regression, then I took the mean of the R² scores and compared them using confidence intervals and a t-test.

Year-by-Year Comparison

Year	Max-Kp Mean	P-Estimation	R2 Score (Binomial)	R2 Score (Poisson)
2001	3.0164	0.3668478361	0.9302	0.84322
2008	2.46174	0.273527	0.9515	0.9308
2015	3.29863	0.3665144697	0.9543	0.8953
2022	3.183561644	0.3537290	0.9688	0.8918



Confidence Intervals

Confidence intervals for the R² Scores were calculated for both distributions:

Binomial Distribution:

- 95% Confidence Interval: [0.92587,0.97653]
- 99% Confidence Interval: [0.932466553,0.969933]

Poisson Distribution:

- 95% Confidence Interval: [0.832,0.9479]
- 99% Confidence Interval: [0.847146,0.932853]

Let it be noted that at 95% and 99% there is overlap in the confidence intervals for both distributions, therefore simply looking at the confidence intervals does not directly show there is a significant difference between the two. Let it also be noted that the range for the Poisson distribution is larger over both intervals in comparison to the Binomial distribution as well as starting significantly lower.

Hypothesis Testing

The hypothesis test compared the R^2 scores of the Binomial and Poisson distributions

- Ho: $\mu1 - \mu2=0$
- Ha: $\mu1 > \mu2$ (Right Tailed)
- Significance Level: 0.05
- Test Statistic $t = 2.6169$
- Rejection Region: $t > 1.94318$

Since $t = 2.6169$ is greater than the critical value $t = 1.94318$, we reject Ho at the 95% confidence level. This indicates a significant difference between the R^2 scores of the Binomial and Poisson distributions, with the Binomial distribution providing a better fit for the Kp max dataset because it has a higher score.

The Binomial distribution might better fit the data for the Kp-Max because the Binomial distribution models the number of successes in a fixed number of independent trials, where each trial has the same probability of success. This could align with how geomagnetic events might occur as independent occurrences where each period could be treated as a trial with a certain probability of reaching a high Kp value.

Conclusion

There is strong evidence ($t = 2.6169$, $t > 1.94318$) to reject the null hypothesis that there is no difference in the R^2 scores between the Binomial and Poisson distributions when applied to the Kp max dataset. We are 95% confident that the Binomial distribution will have a greater R^2 score when applied to the Kp-Max.

Based on the data analysis and this study's design, it is reasonable to conclude that the Binomial distribution is a better fit for modeling the Kp max data. This finding can be generalized to similar datasets representing maximum geomagnetic activity values due to the consistency of the results across multiple years and varying solar activity levels.

The implication of the Binomial distribution providing a good fit implies that geomagnetic events occur in a predictable, independent manner with a constant probability of occurrence during each period under consideration.

This study, however, had several limitations, the most significant being the small sample of years analyzed. While selecting years from different solar activity phases provides some diversity, a more comprehensive study would include data from all years in the solar cycle. Other limitations include the exclusive focus on the maximum daily Kp values and the assumption of independence among data points, which might not fully account for autocorrelations in geomagnetic activity. Future research could address these limitations by analyzing additional years, incorporating more sophisticated statistical models, and examining other metrics of geomagnetic activity, such as daily averages or cumulative Kp values, to provide a more nuanced understanding of the underlying patterns.

References

[1] Space Weather Live the Kp-Index

https://www.spaceweatherlive.com/en/help/the-kp-index.html

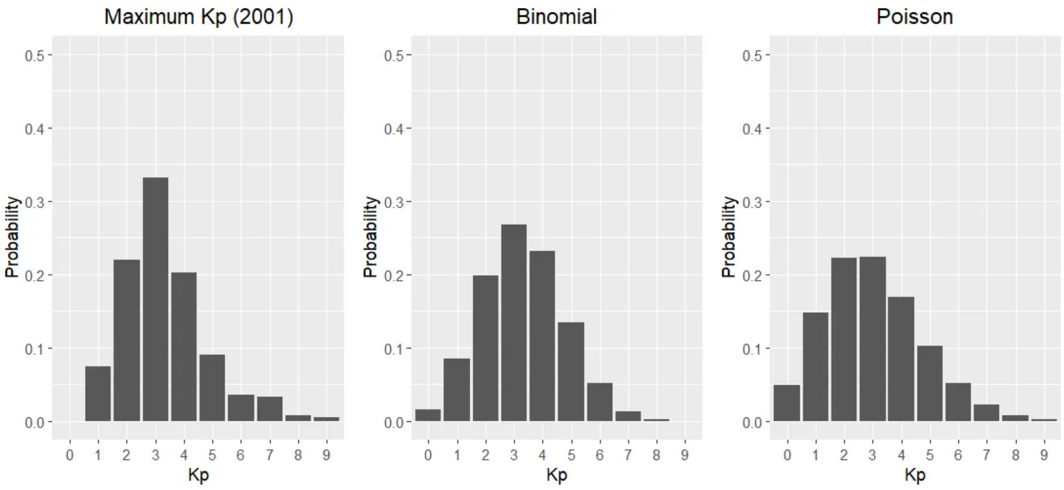
[2] Space Weather Prediction Center (2024). Planetary K-Index

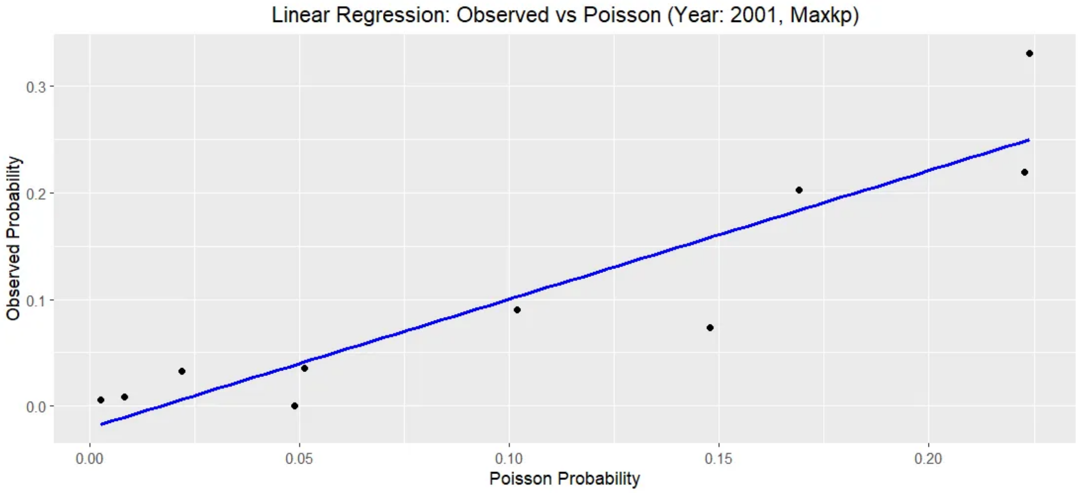
https://www.swpc.noaa.gov/products/planetary-k-index

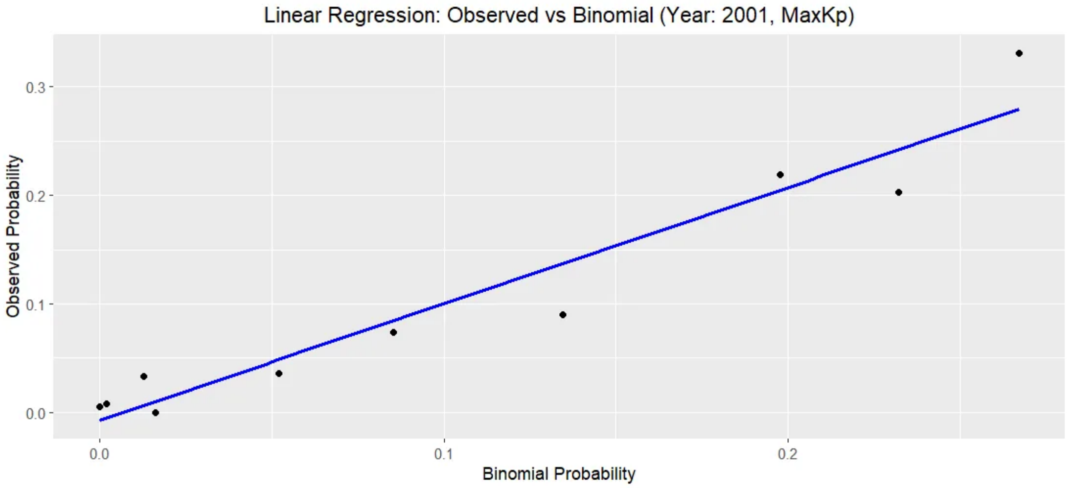
[3] Jemma.mobi (2024). Statistics of the Kp-value.

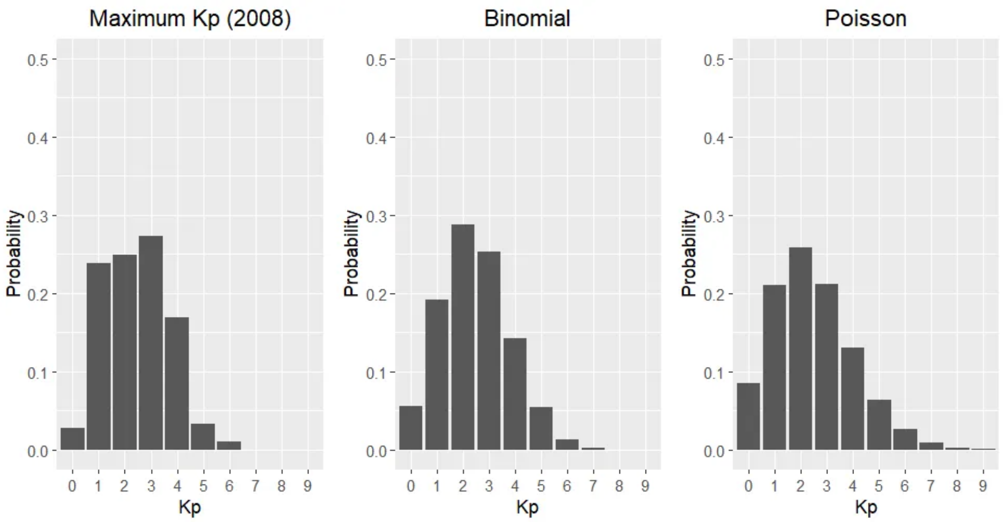
https://jemma.mobi/kp-historia?e

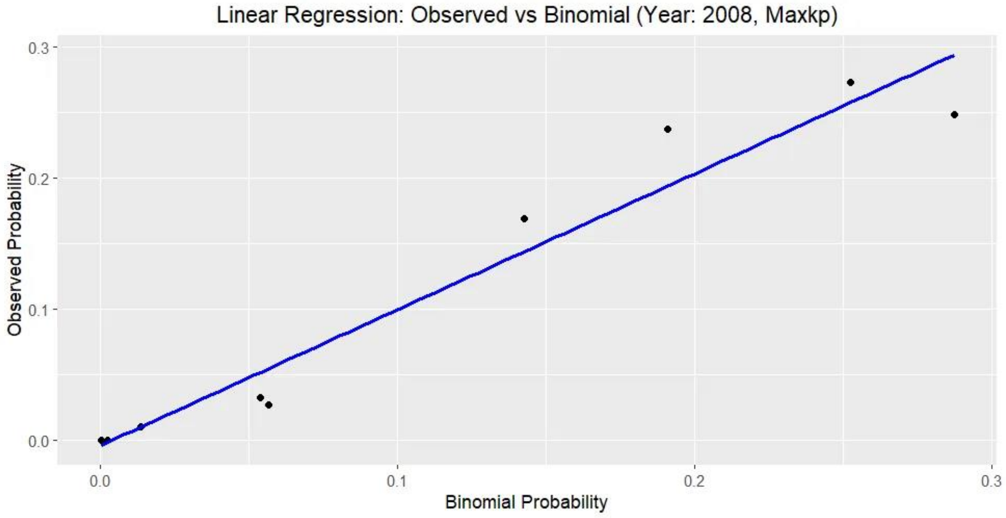
Appendix

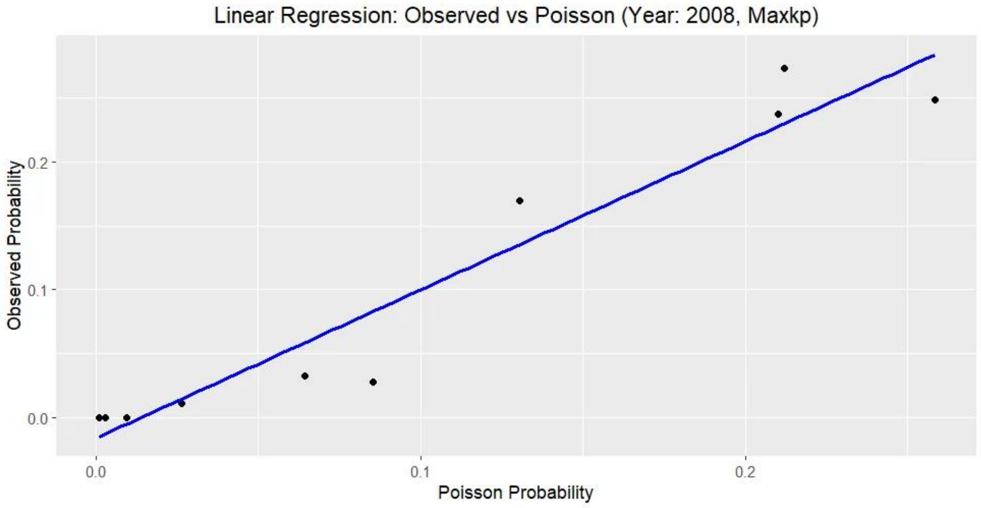

Maximum Kp (2001)
Binomial
Poisson
Probability
Kp
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
9

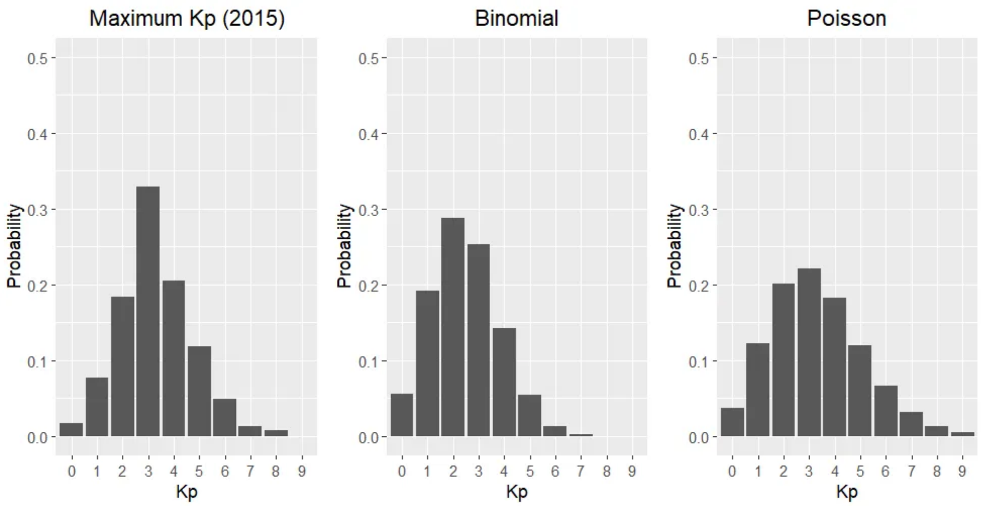

Linear Regression: Observed vs Poisson (Year: 2001, Maxkp)
Observed Probability
Poisson Probability
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.20

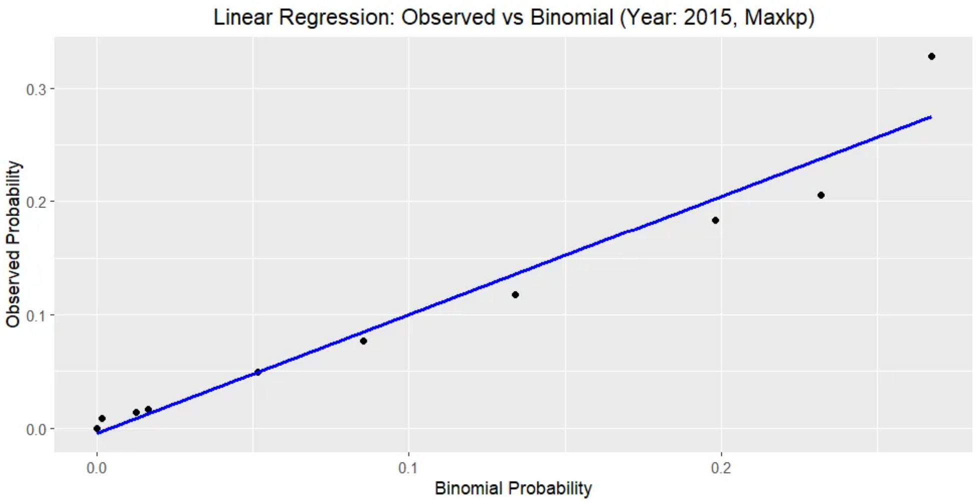

Linear Regression: Observed vs Binomial (Year: 2001, MaxKp)
Observed Probability
Binomial Probability
0.0
0.1
0.2
0.3
0.0
0.1
0.2

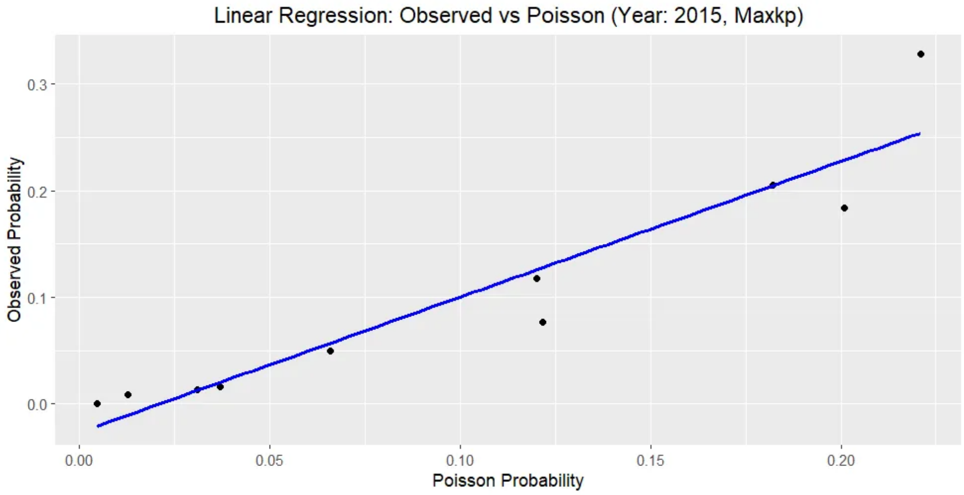

Maximum Kp (2008)
Binomial
Poisson
Probability
0.5
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9
Kp

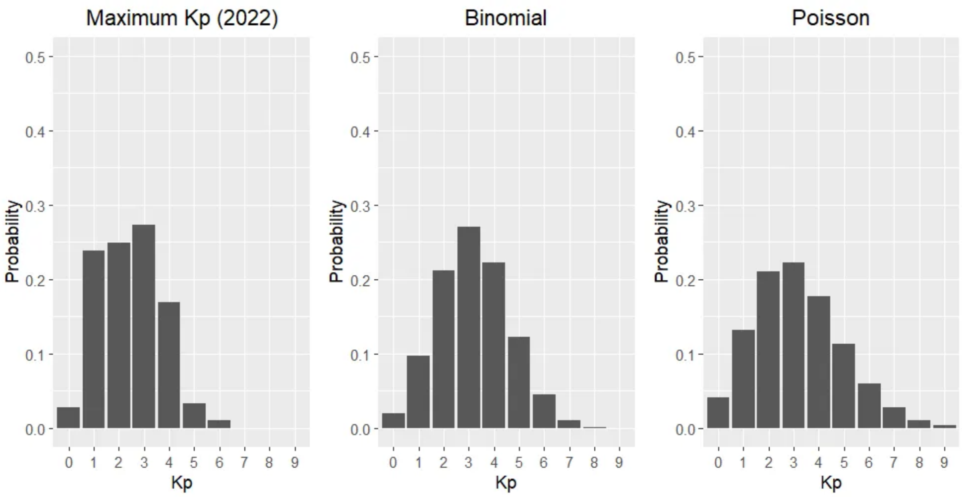

Linear Regression: Observed vs Binomial (Year: 2008, Maxkp)
Observed Probability
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
Binomial Probability

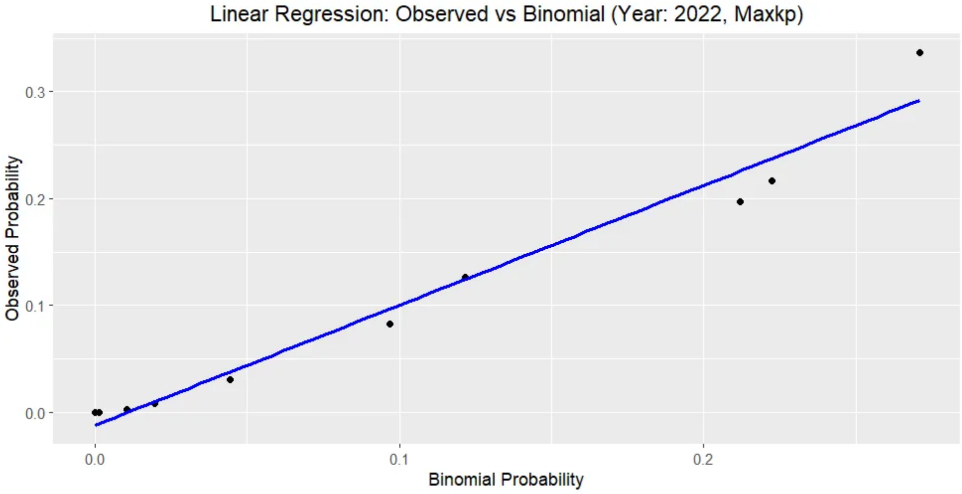

Linear Regression: Observed vs Poisson (Year: 2008, Maxkp)
Observed Probability
0.2
0.1
0.0
0.0
0.1
0.2
Poisson Probability

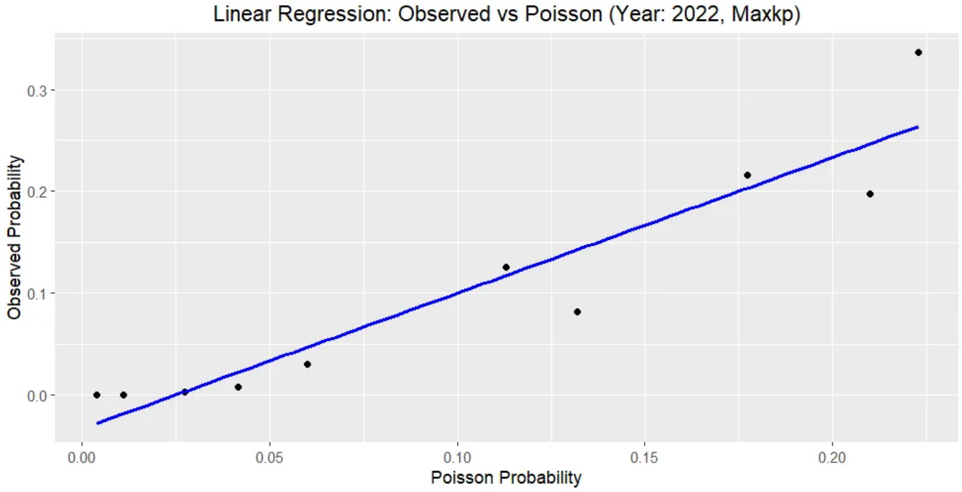

Maximum Kp (2015)
Binomial
Poisson
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0 1 2 3 4 5 6 7 8 9
Kp


Linear Regression: Observed vs Binomial (Year: 2015, Maxkp)
Observed Probability
0.0
0.1
0.2
0.3
0.0
0.1
0.2
Binomial Probability


Linear Regression: Observed vs Poisson (Year: 2015, Maxkp)
Observed Probability
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.20
Poisson Probability


Maximum Kp (2022)
Binomial
Poisson
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0 1 2 3 4 5 6 7 8 9
Kp


Linear Regression: Observed vs Binomial (Year: 2022, Maxkp)
Observed Probability
0.0
0.1
0.2
0.3
0.0
0.1
0.2
Binomial Probability


Linear Regression: Observed vs Poisson (Year: 2022, Maxkp)
Observed Probability
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.20
Poisson Probability