Analysis of Dew Point Affected Pitches **Siddhant Thakur**

This write up explains my thought process while calculating the estimate of dew point affected probability for every pitch in the given dataset. I will be discussing the data exploration, preprocessing, methodology and assumptions involved.

ASSUMPTIONS

- Pitches could come from a position player as well, and ideally we should check for them. (Because some PITCHER_KEY have low number of pitches thrown, I have considered them in this analysis because I don't know who they are, but could be position players)
- High residual may suggest a major effect due to dew point.
- The independent variables are not perfectly correlated.
- The residuals are normally distributed. If they are not, I would use t-score rather than z score.
- Residuals are homoscedastic (constant variance) across the levels of independent variables.

These are the assumptions that I am making while developing the model.

Data Exploration & Preprocessing

The data provided consists of 9889 pitches thrown by 37 pitchers at the Great American Ballpark with 26 columns comprising game events and pitch characteristics. The first thing I noticed was that some of the pitch characteristics varied based on the handedness of the pitcher. Therefore, I wanted to even out the pitcher handedness by considering *THROW_SIDE_KEY* as 'R', and for that I flipped the sign for *HORIZONTAL_BREAK*, *RELEASE_SIDE* & *HORIZONTAL_APPROACH_ANGLE* **[In 4-5]**. After this I checked for NaN values where only the column *EVENT_KEY* had 7258, which makes sense since it's pitch-by-pitch data for at-bats in multiple games **[In 6]**.

I then looked at the number of pitches thrown by every pitcher and of every pitch type. Because every pitcher and pitch type have a different number of pitches thrown, I will consider them as random effects. Also, I have gotten rid of the UN and KN pitch types, as there are only 7 such pitches, and they could possibly be wrong tagging **[In 7-9]**.

We then check for outliers present in the data using boxplots **[In 10 & 14]**, kdeplots **[In 15]** and describe **[In 16]** function in Python. While there are no clear outliers, we see that for release speed there are some outliers with lower than 70 mph, but this could be due to position players throwing pitches. For this initial analysis, I looked at the pitch type for the pitches with release speed less than 70 mph, and got rid of the fastballs **[In 11-13]**. From the KDE distribution of different variables **[In 15]**, we can also see that the distribution for most of the pitch characteristics is not normal but bimodal in some cases.

I also checked for correlation among the independent variables **[In 17]**. There is a 0.5 absolute correlation between parameters like release extension & release height, approach angles with spin rate or release speed, etc. I tried removing these variables from the formula, but it led to an even worse RMSE. Therefore, I decided to stick with the original idea of keeping all the pitch characteristics in the formula.

Methodology

When I received the assessment, I was a little confused on how to approach this problem. Most of the models that I have built on sports have a particular target variable I am predicting on. So, how do we find the pitches affected by dew point without a specific response variable? The statement "pitch movement due to environmental factors" made me realize that along with gravity and air resistance, we may also

have dew point affecting Induced Vertical Break and Horizontal Break. Based on the video mentioned in the reference, I came to the conclusion that as long as we keep release data, spin, release speed and other pitcher dependent variables consistent, any massive change in IVB or HB could be due to external factors, dew point being a major one. Therefore, we can estimate the effect of dew point over pitch movement by holding factors already provided in the data, such as, the pitcher, release characteristics, approach angle, etc consistent. The assumption here is that after doing this, the residuals can inform us about whether those pitches could have been affected by dew point (Note: This will still be an estimate, because some of it could be noise)

We can set up a mixed effect model with the release characteristics, approach angle, spin rate being the fixed effects and pitchers & their pitch types being the random effects. We will model the fixed and random effects against Induced Vertical Break as well as Horizontal Break. We can then interpret the residuals from these models, where a high residual may suggest that the pitch was affected by something not included in the model, possibly dew point. After finding the residual, we can use that to calculate the probability of dew point affecting each pitch, assuming that the residual is present likely due to dew point. For this we are definitely assuming that the residuals are normally distributed in the population. We calculate the mean and standard deviation of the residuals and use that to calculate z-score for each pitch. Since the probability of a pitch not getting affected by dew point would be given by the cumulative distribution function (CDF) of the normal distribution at that z-score, we subtract it by 1 to find the probability of a pitch being affected by dew point > 65 degrees F.

To implement this, I built a class Model **[In 19]**, because I needed to implement it for both Induced Vertical Break and Horizontal Break. The class contains methods for scaling data, building models, getting the summary, getting predictions, computing error, getting residuals and probabilities for each pitch. You can pass the formula and the groups you want as parameters for your mixed effect model, allowing for more than one random effect variable. From this, we can develop two models for both induced vertical and horizontal break as response variables. We then use the probabilities given by the two models and take an average of the two. This will be the final submission for the assessment.

After getting the probabilities, I also wanted to check for normality of the residuals graphically to ensure the assumption made by the mixed effect model. From the above Kernel Density and QQ plot **[In 20-21]**, we see that,

1. For Induced Vertical Break, the residual spread might not exactly follow a normal distribution. This is due to the fact that in the QQ plot, the left tail has a few residuals deviating from the norm, indicating potential outliers.
2. For Horizontal Break, the residuals are more normally distributed compared to the Induced Vertical Break model. The model seems to fit relatively well, but there are still very slight deviations at the left end of the tail in the QQ plot, suggesting again that there may be outliers left.

References,

1. https://www.baseballvmi.com/baseball-humidity#:~:text=Humidity%20actually%20has%20little%20effect,through%20than%20dry%20air%20is.
2. https://www.daytondailynews.com/weather/does-humidity-lead-more-home-runs-baseball/oP4c9mGNHtBAo1gYDpFODJ/
3. https://www.pythonfordatascience.org/mixed-effects-regression-python/