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 assumptions, data exploration, preprocessing and methodology involved.

ASSUMPTIONS

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

- While controlling for the variables that are given in the dataset, any deviation in the prediction of movement could be due to external factors, one of which might indicate dew point effect.
- For this purpose, the residuals of the movement model are treated as coming from normal distribution, and because of this I have used z-score rather than t-score.
- Residuals are also homoscedastic (constant variance) across the levels of independent variables.
- Presence of moisture doesn't affect pitcher grip, and eventually spin or release.
- Correlation check has been done while assuming the absolute value of Pearson Coefficient greater than 0.6 is highly correlated and anything less than or equal to 0.6 is safe.
- Fastballs less than 70mph are treated as either wrong data or pitches from position players, and hence have been removed from the analysis.

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. Pitch characteristics like *HORIZONTAL_BREAK*, *RELEASE_SIDE* & *HORIZONTAL_APPROACH_ANGLE* are affected by the pitcher-handedness between lefty and righty pitchers. For the sake of this analysis, the above variables were negated to allow for all pitch characteristics to be coming from a right-handed pitcher **[In 4-5]**. I then checked for NaN values where only the column *EVENT_KEY* had 7258 rows, which makes sense since it's pitch-by-pitch data for at-bats in multiple games **[In 6]**.

We also check for outliers present in the data using boxplots **[In 10 & 14]** and describe **[In 16]** function in Python. While there are some outliers for the pitch characteristics, I have decided to include most till I can get more information about these pitches (through some video analysis). We do see that there are release speeds lower than 70 mph, and for the purpose of this analysis, those that are fastballs have been removed as per the assumption stated **[In 11-13]**. For the pitch types, 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]**.

I have also checked for correlation among the independent variables **[In 17]**. Of the pair of variables that have an absolute value of pearson's coefficient higher than 0.6 we keep the one variable that has higher correlation with our target variable. This results in *RELEASE_SPEED* getting thrown out of the Induced Vertical model, and *VERTICAL_APPROACH_ANGLE* is removed for the Horizontal Break model. Looking at the different sample of pitches thrown by every *PITCHER_KEY* and the different *PITCH_TYPE_TRACKED_KEY*, we will consider these variables as random effects in the model, to allow for the group-level variability to be captured.

Methodology

When I received the assessment, I was confused on how to approach this problem. Most of the models that I have built in sports have a particular target variable I am predicting on. So it was a nice challenge to estimate 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[1], 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, or other external factors not included in the data).

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.

To estimate this probability, we are assuming that the residuals are normally distributed in the population. After finding the residual of each sample, and calculating the mean and standard deviation of the residual distribution, the question we want to ask is what is the probability of observing a value of the residual as extreme or more extreme as the given value. This is essentially asking, "**How unlikely is a residual of *i* given our current model, and the residual distribution?**". If it's very unlikely, then it provides stronger evidence that some external factor is influencing this observation. If we find this probability to be *p*, then *1-p* basically gives the probability that external factors must have affected this pitch. This *1-p* can then be assumed to be the estimated dew point affect probability. **Again, this is only an estimate and having temperature data can help us pinpoint the probability. A high probability doesn't directly prove the presence of external factors; it could be anomaly, or non linearities not captured by the model or interactions between predictors not considered or other model misspecifications.**

Implementation & Results

To implement this, I built a class **[In 19]**, because I needed to implement the model 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 stands true. From the above Kernel Density and QQ plot **[In 20-21]**, we see that, for both the models, the residuals from the samples show an approximately normal distribution.

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/