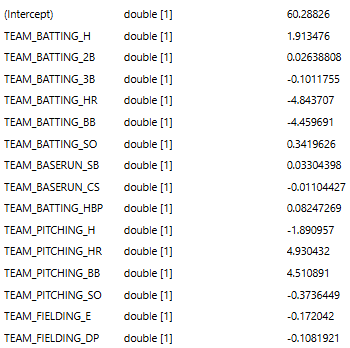
**DATA 621: Moneyball Data Models**

**Lawrence Yu**

**3. Building Data Models**

**Base model:** TEAM\_BATTING\_H + TEAM\_BATTING\_2B + TEAM\_BATTING\_3B + TEAM\_BATTING\_HR + TEAM\_BATTING\_BB + TEAM\_BATTING\_SO + TEAM\_BATTING\_SB + TEAM\_BATTING\_CS + TEAM\_BATTING\_HBP + TEAM\_PITCHING\_H + TEAM\_PITCHING\_HR + TEAM\_PITCHING\_BB + TEAM\_PITCHING\_SO + TEAM\_FIELDING\_E + TEAM\_FIELDING\_DP

Explanation:



This in itself is the original unprepared data. INDEX is omitted because it is not a predictor. Such a model may have a risk of overfitting the original data. In terms of questionable coefficients, triples, doubles, and walks should not be negative for the respective team. Those are on base factors that heavily promote scoring runs. On the other end, pitchers giving up home runs and walks also should not be helping their team win more. These are the most egregious issues within this model and it should not be used beyond as a baseline.

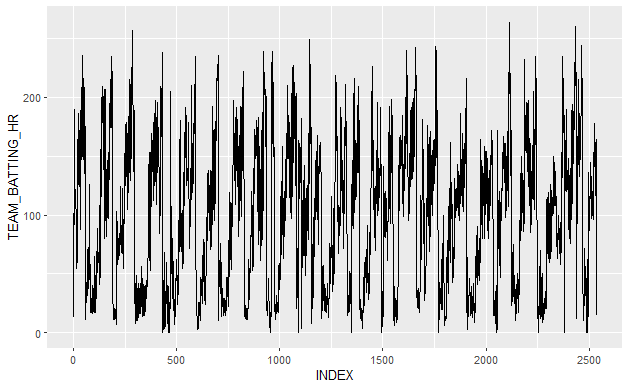
**Model 1:** TOTAL\_BASES + TEAM\_BATTING\_BB + TEAM\_FIELDING\_E + TEAM\_PITCHING\_H + TEAM\_PITCHING\_BB

Reasoning for variables: This model was designed from intuition. The core logic is that offense and defense (including fielding and pitching) should each be represented with their strongest factors. Ultimately, you can't win a game without scoring runs no matter how strong your pitching is. Our belief is that a team's offensive ability will outweigh pitching value unless its pitching is that much better than its offense. This leads to the idea behind the variables chosen.

While we could only include variables from the original data set, we needed to establish context behind our data. For example, home runs are more prevalent in more recent eras, such as the 1980s-2000s, than in the 1800s. They are also the most definitive sources of runs available as they range from 1-4 runs per home run. We performed a quick check to see if time was present in our data and found no sufficient evidence suggesting so.

*# Test to see if data is in time series form using index and home runs*

raw\_training\_df %>% ggplot(aes(x = INDEX, y = TEAM\_BATTING\_HR)) + geom\_line()



There would have been a dramatic increase in home runs around 30% of the way through the data if organized in time order. This roughly would've marked the time when Babe Ruth hit more home runs in a single season than entire teams. Instead, the total home runs per team had no clear pattern when organized in index order. This means that we can proceed with assuming that individual hitting factors will not reflect changes to the game throughout time. Teams can obtain their runs and eventually their wins from speed, power, or contact. These are heavily represented within two factors: total bases and walks. The following is the breakdown of the variables chosen for this model:

TOTAL\_BASES should be relatively time agnostic. This is a statistic we created by assigning values so that singles count as one base, doubles as two, triples as three, and home runs as four. Accruing more bases means you are more likely to be reaching home plate and thus are scoring more runs. The reason this was selected over home runs was because of the time factor where home runs were not the primary source of runs for many years represented within the data.

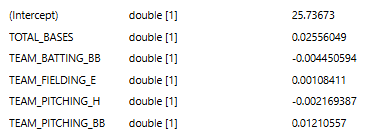
TEAM\_BATTING\_BB is a team's walk total. A team with a lot of walks will have a lot of base runners. This is similar to a single except it is not included in TOTAL\_BASES.

SB\_SUCCESS\_RATE is the amount of stolen bases / total attempts. While stolen bases can increase the value of a hit or walk, being caught also hurts that. We are looking for teams that are more likely to steal successfully than erase their own runners.

TEAM\_FIELDING\_E is the amount of errors made by a team. Errors generally are assigned when a hitter or runner is awarded at least one extra base from a defensive mistake. Teams that accumulate more errors should give up more runs and thus lose from their errors more often.

WHIP\_PROXY takes the concept of walks + hits / innings pitched and adapts it to our available data as walks + hits / 162 games. This rate considers how frequently a team's pitchers give up bases to the other team.

Model concerns and considerations:



Notable omissions are variables that use strikeouts and double plays. Strikeouts are commonly assumed to be the most effective way to keep the other team from scoring. This is true to an extent. However they often require a pitcher to throw more pitches on average and use up that pitcher's arm. A tired pitcher is far more likely to get injured or make a mistake. An early replacement for that pitcher may also be mediocre and blow the game. There is also an idea that players that focus on power may intentionally strike out more often if they can get more bases per hit on average. Banking runs via a home run is worth striking out a little more than the other team.

The results are very peculiar for this model. While these should have been a collection of the most impactful variables, it appears that SB\_SUCCESS\_RATE is being favored too heavily. While it would make sense that more efficient running teams would win more often, this is a factor that has too many possible issues to be considered the main driving force of a win. Teams that steal sparingly, but always succeed do not gain that much value from base stealing. Teams that hit for a lot of power likely do not steal often, yet they should win a lot of games. However, the other coefficients are too muted to properly analyze here. They have appropriate positive/negative correlations based on the initial assumptions, but at very low coefficients. We would likely abstain from using this model unless it proved optimal compared to other options.

**Model 2**: ONE\_BASE:SB\_SUCCESS\_RATE + NUM\_OFFENSIVE\_TIER + TOTAL\_BASES + TEAM\_BATTING\_BB + TEAM\_PITCHING\_SO + LOG\_TEAM\_FIELDING\_E\_CAPPED

Reasoning for variables: This model was based on adding and subtracting variables and comparing results while attempting to control how much specific factors impacted the model. Getting on base was heavily emphasized within the factors which leads to a high risk of multicollinearity. Getting on base is extremely important when scoring runs, so the correlation is to be expected.

ONE\_BASE:SB\_SUCCESS\_RATE compares a team's likelihood of stealing a base successfully against the amount of times players are in position to steal a base. ONE\_BASE references singles + walks since being on first base leads to the highest chance of a steal attempt.

NUM\_OFFENSIVE\_TIER is a numerical representation of a team's total hits performance relative to other teams. -1 means a team is below the 25th percentile, 0 means they are around average, and 1 indicates above the 75th percentile. The goal with this variable is to distinguish teams with extremely good or bad hit tools without giving too much weight to outliers. Teams with an extreme amount of hits are covered by the following factor.

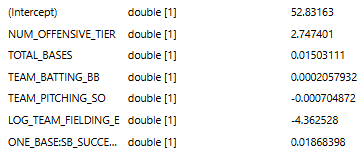
TOTAL\_BASES furthers the focus on getting on base, but incorporates power multipliers for extra base hits. The goal is to provide extra weight for teams that either hit for a lot of power or have an extremely high amount of hits since both possibilities increase this variable and the chances of scoring runs.

TEAM\_BATTING\_BB, or walks, are effectively singles. They are not expected to impact the model as much, but round out the main common ways to get on base. Note that hit by pitch is generally omitted in our models because it happens very sparingly and is generally more of a factor of bad pitching than it is a credit to the hitter.

TEAM\_PITCHING\_SO is included to balance out the focus on hitting a bit. While hitting metrics appear to affect wins more, pitching is not irrelevant to a team's performance. It is directly linked to how other teams hit.

LOG\_TEAM\_FIELDING\_E is used to weed out teams that are very weak fundamentally. In other words, errors are not capped because a truly bad defensive team is more likely to be impacted by its bad defense than a slightly bad one.

Model concerns and considerations:



The coefficients are generally positive or negative as expected. However, that is not true for all and the extreme of each is a bit surprising. NUM\_OFFENSIVE\_TIER has the highest positive coefficient. That is by design because the variable effectively buckets teams into general categories of good or bad if they stray far from average. LOG\_TEAM\_FIELDING\_E is unusually the largest coefficient of all, albeit it is negative as expected. The rationale behind a bad defensive team being more impacted by its defense has merit, but this coefficient seems to carry a bit too much weight. Defensive metrics are difficult to contain within a linear model, so unusual results are not enough to discard a model.

**Model 3:** TEAM\_BASERUN\_SB + TEAM\_PITCHING\_H + TEAM\_PITCHING\_BB + TEAM\_FIELDING\_E + SINGLES + TEAM\_BATTING\_2B + TEAM\_BATTING\_3B + TEAM\_BATTING\_HR

Reasoning for variables: A mix of testing variables and seeing the results of backwards stepwise regression was used to create this set of variables.

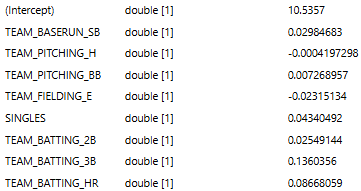
SINGLES + TEAM\_BATTING\_2B + TEAM\_BATTING\_3B + TEAM\_BATTING\_HR: The major unique point of this model was the separation of each hit type from each other. The idea is that not all hit types are equal. Singles can represent contact while home runs are closely correlated with power. Doubles and triples are also important for showing a team's power and speed, albeit to a lesser degree.

TEAM\_PITCHING\_SO, TEAM\_PITCHING\_HR, and TEAM\_BASERUN\_CS were culled by the stepwise function. Caught stealing is connected to the stolen base stat or TEAM\_BASERUN\_SB which is a far better indicator of a team's run scoring ability. Pitching strikeouts and home runs being removed imply that pitching metrics do not have impact wins as much as a team's offense.

TEAM\_PITCHING\_BB and TEAM\_PITCHING\_H are similar to the walks and hits logged for hitters. They are fairly common and not as prone to being skewed by outliers as a result. They would be more reliable than the pitching stats the stepwise function removed.

TEAM\_FIELDING\_E indicates that a team loses a lot more if it makes a lot of errors which covers the impact of fielding. A notable omission is TEAM\_FIELDING\_DP. Double plays turned by the fielders may appear to be a positive action, but it actually implies a team's pitching has given up walks or hits. There is also some multicollinearity there with pitching variables.

Model concerns and considerations:



One of the most surprising results is that TEAM\_BATTING\_3B has the highest coefficient of the offensive variables. This could be a sample size issue or feature as they are typically rarer than other types of hits. In order to get a triple a player is generally going to be very fast and the opposing team typically will make a fielding mistake that is not counted as an error. As a factor that somewhat accounts for both teams' performance, it actually makes some sense as an important variable.

Another curious result is that TEAM\_PITCHING\_BB is positive when giving up walks should allow the opposing team more base runners. It is possible that this doesn't correlate with losing as much because the other team will still need to get that runner home and walks can be handed out strategically in the form of intentional walks to good hitters. We do not believe the coefficient results to be a problem for this model.