Using Predictive Models to Assess Future Value of MLB Players

Connor Morehouse

February 2021

Outline

- Goal
- The data
- The model
- Example analysis

Goal

- Use features of the data to create a model for an offensive metric that we can then use to infer future performance of a player.

Goal

- Use features of the data to create a model for an offensive metric that we can then use to infer future performance of a player.
- Specifically we will be looking at batting average, the proportion of hits to total at bats.

The Data

- Very large data set
  - 89 columns
  - 4.9 million rows
  - 2 GB file

The Data

- Very large data set
  - 89 columns
  - 4.9 million rows
  - 2 GB file
- Needed to reduce size of data set.
  - Reduced number of columns to 50 immediately.
  - Reduced total rows to only the last pitch of the at bat.
  - Only needed 3 columns for the model, and need a couple others for calculations.

The Data

- Very large data set
  - 89 columns
  - 4.9 million rows
  - 2 GB file
- Needed to reduce size of data set.
  - Reduced number of columns to 50 immediately.
  - Reduced total rows to only the last pitch of the at bat.
  - Only needed 3 columns for the model, and need a couple others for calculations.
    - ★ Launch angle, Exit velocity, Hit direction
    - ★ x and y cords
    - ★ events, batter ids, and year

The Model

- We used an General additive model(GAM).

The Model

- We used an General additive model(GAM).
- General means that we can model an outcome with a distribution of errors that is non-normal.

The Model

- We used an General additive model(GAM).
- General means that we can model an outcome with a distribution of errors that is non-normal.
  - Put simply, this model is a less strict about the assumptions made about the data and can therefore be used on a wider array of data, which we do need for out data.

The Model

- We used an General additive model(GAM).
- General means that we can model an outcome with a distribution of errors that is non-normal.
  - Put simply, this model is a less strict about the assumptions made about the data and can therefore be used on a wider array of data, which we do need for out data.
- Additive means that for each input variable, called a predictor, we fit a smooth function to the data, then sum the impact of those smooth functions into one model.

The Model

- We used an General additive model(GAM).
- General means that we can model an outcome with a distribution of errors that is non-normal.
  - Put simply, this model is a less strict about the assumptions made about the data and can therefore be used on a wider array of data, which we do need for out data.
- Additive means that for each input variable, called a predictor, we fit a smooth function to the data, then sum the impact of those smooth functions into one model.
  - A benefit of using general additive models is the trend line we use doesn't have to be linear, but can be a smooth function.

The Model

- We used an General additive model(GAM).
- General means that we can model an outcome with a distribution of errors that is non-normal.
  - Put simply, this model is a less strict about the assumptions made about the data and can therefore be used on a wider array of data, which we do need for out data.
- Additive means that for each input variable, called a predictor, we fit a smooth function to the data, then sum the impact of those smooth functions into one model.
  - A benefit of using general additive models is the trend line we use doesn't have to be linear, but can be a smooth function.
  - A smooth function is a function that is continuous on a restricted interval.

Why we used a GAM

- Our data doesn't have normally distributed errors, therefore we couldn't use regular linear regression.

Why we used a GAM

- Our data doesn't have normally distributed errors, therefore we couldn't use regular linear regression.
- Our data has predictors that have non-linear relationships to our outcome variable, a GAM allows us to have non-linear predictors.

Why we used a GAM

- Our data doesn't have normally distributed errors, therefore we couldn't use regular linear regression.
- Our data has predictors that have non-linear relationships to our outcome variable, a GAM allows us to have non-linear predictors.
- Lastly, GAMs have the ability to have regularization, which is a method to prevent over fitting model and maintain model accuracy.

Creating the model

The purpose of our model is to calculate the probability of a hit based on the predictors, but before we use the model we can simplify some of the data.

Creating the model

The purpose of our model is to calculate the probability of a hit based on the predictors, but before we use the model we can simplify some of the data.

- Firstly, strikeouts always have a hit probability of 0 (25% of ABs).

Creating the model

The purpose of our model is to calculate the probability of a hit based on the predictors, but before we use the model we can simplify some of the data.

- Firstly, strikeouts always have a hit probability of 0 (25% of ABs).
- Due to the high launch angle of popups, they have a tendency to have their launch angle misread by the ball tracking system. This combined with their extremely low hit probability led us to predict all popups to a hit probability of 0 (5.5% of ABs).

Creating the model

The purpose of our model is to calculate the probability of a hit based on the predictors, but before we use the model we can simplify some of the data.

- Firstly, strikeouts always have a hit probability of 0 (25% of ABs).
- Due to the high launch angle of popups, they have a tendency to have their launch angle misread by the ball tracking system. This combined with their extremely low hit probability led us to predict all popups to a hit probability of 0 (5.5% of ABs).
- Ground balls are also known to cause problems for the ball tracking system due to their very low launch angles, so we used a model that didn't include launch angle to predict the hit probability of ground balls (32% of ABs).

Creating the model

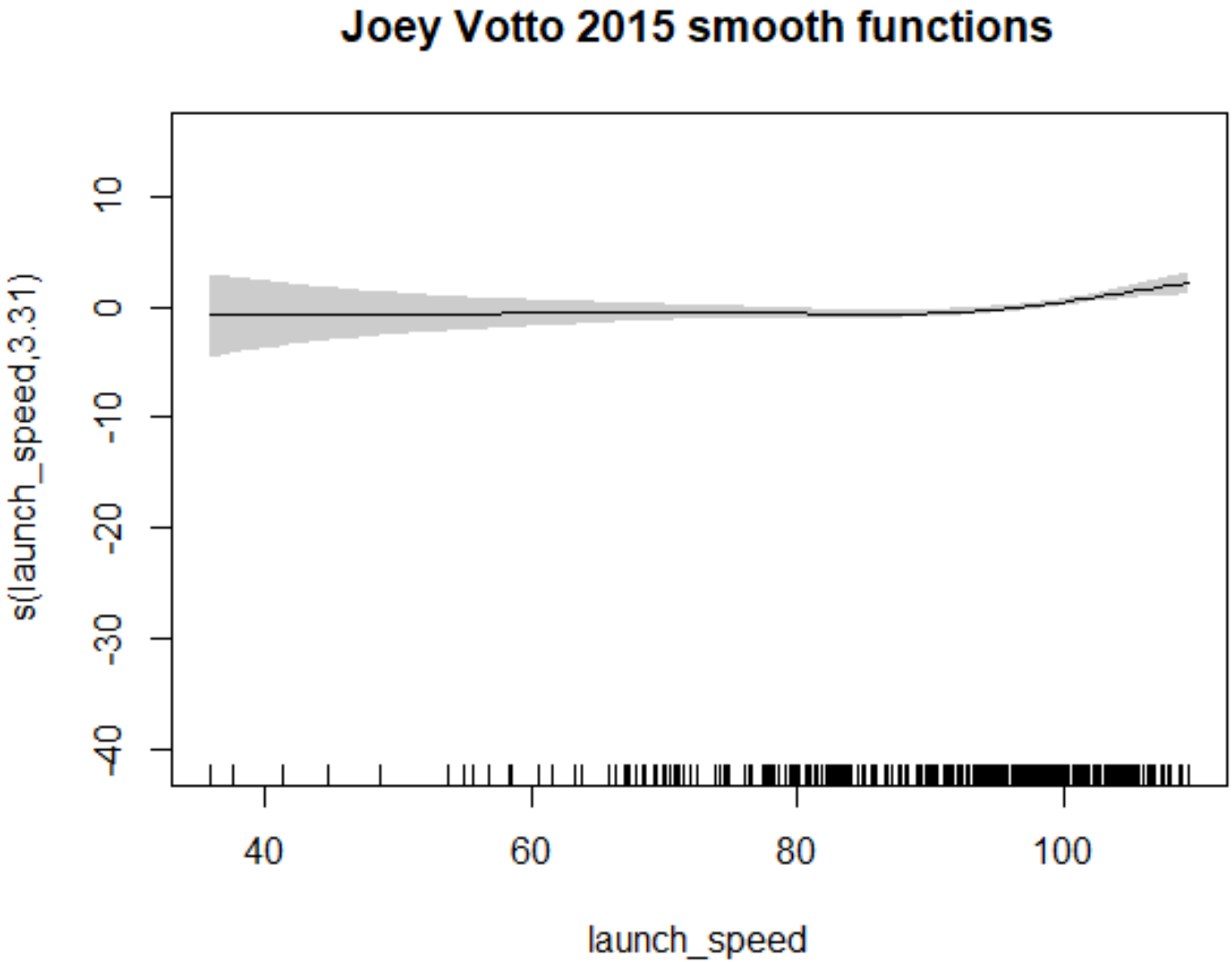
The purpose of our model is to calculate the probability of a hit based on the predictors, but before we use the model we can simplify some of the data.

- Firstly, strikeouts always have a hit probability of 0 (25% of ABs).
- Due to the high launch angle of popups, they have a tendency to have their launch angle misread by the ball tracking system. This combined with their extremely low hit probability led us to predict all popups to a hit probability of 0 (5.5% of ABs).
- Ground balls are also known to cause problems for the ball tracking system due to their very low launch angles, so we used a model that didn't include launch angle to predict the hit probability of ground balls (32% of ABs).
- All other batted balls were predicted with a model that used launch angle, hit direction, and exit velocity.

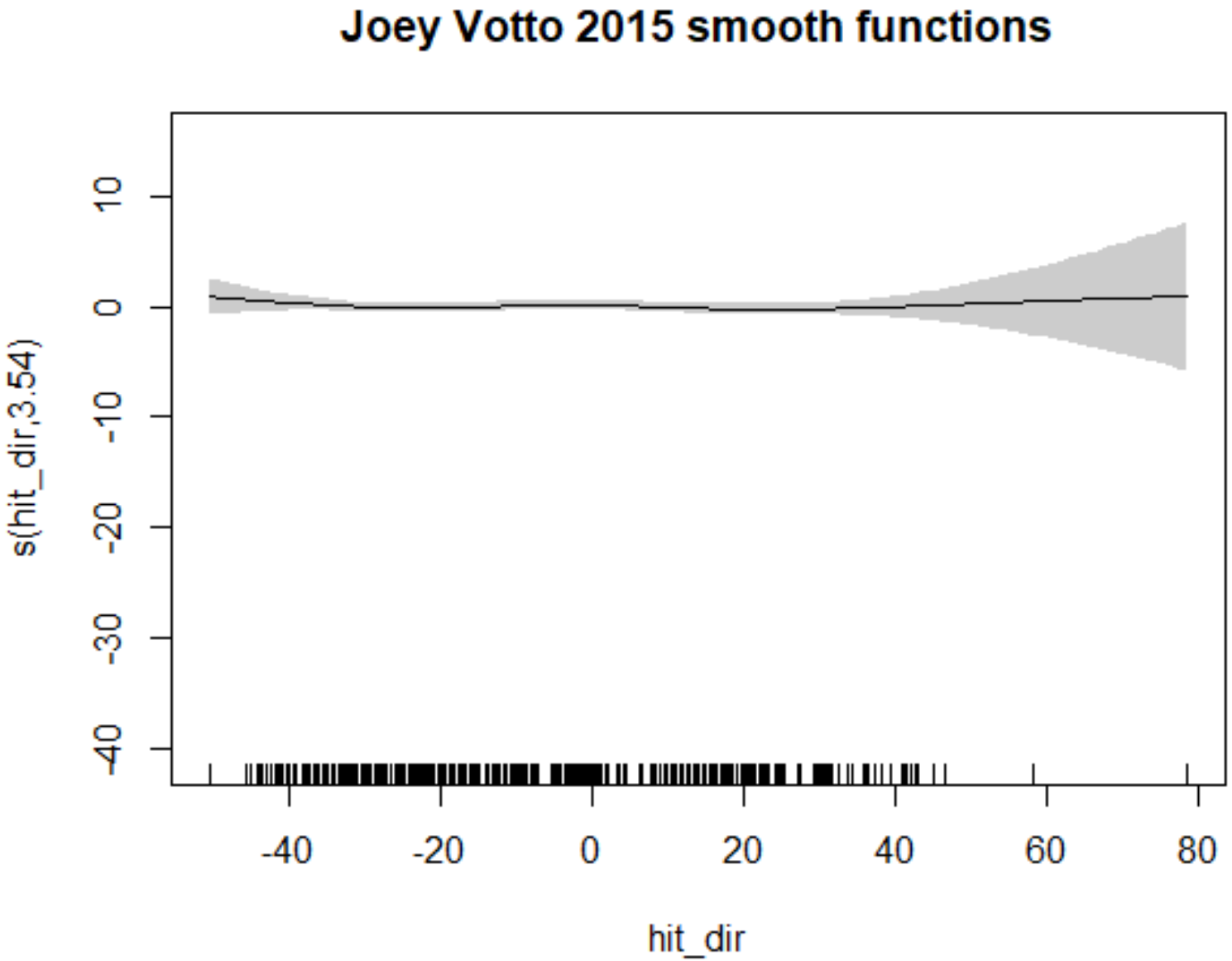
Creating the model

- We created two models for each player each year, one for ground balls and one for all other ABs.
- We used the models from the previous year to predict the next years values.

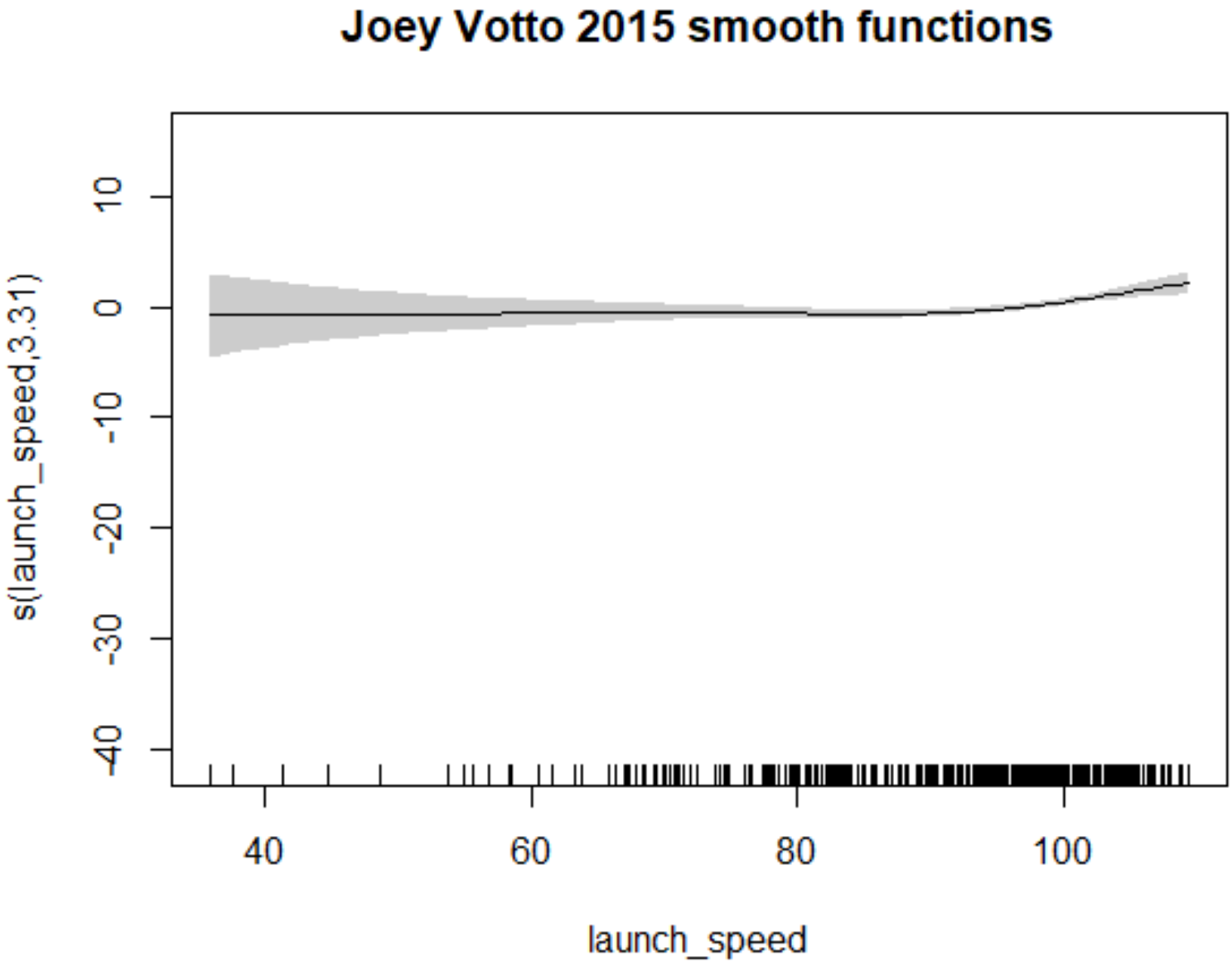
Checking the model



Checking the model



Checking the model



Difference

- We assume the difference between the observed batting average and the predicted batting average to be due to random variance, or 'luck'.

Difference

- We assume the difference between the observed batting average and the predicted batting average to be due to random variance, or 'luck'.
- Because of this, we always expect the observed batting averages to move back towards the predicted batting average over time.

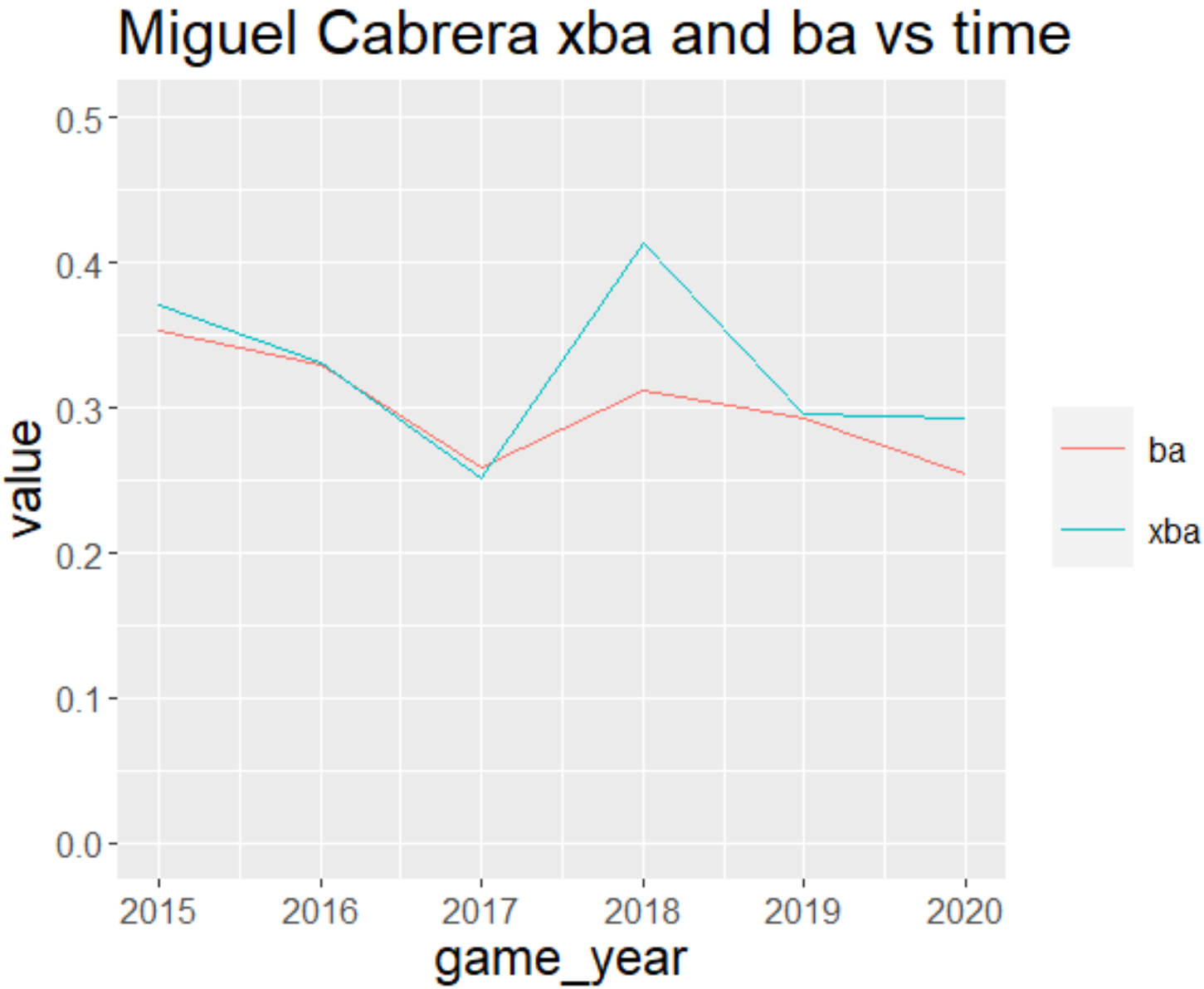
Difference

- We assume the difference between the observed batting average and the predicted batting average to be due to random variance, or 'luck'.
- Because of this, we always expect the observed batting averages to move back towards the predicted batting average over time.
- There are many other reasons for the difference between batting average and expected batting average, like a new training method or new approach, but the model is retrained every year in an attempt to minimize this effect.

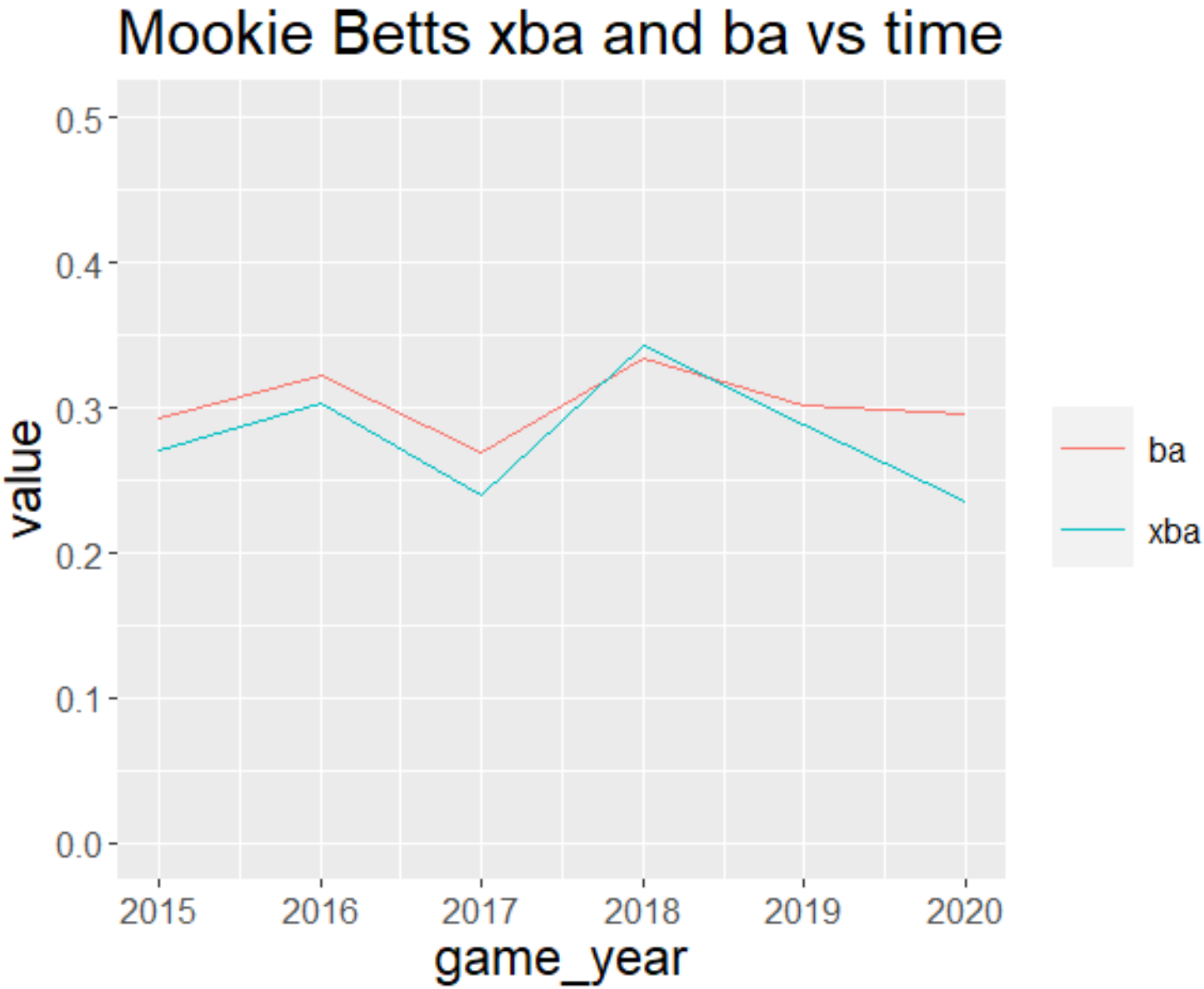
Difference

- We assume the difference between the observed batting average and the predicted batting average to be due to random variance, or 'luck'.
- Because of this, we always expect the observed batting averages to move back towards the predicted batting average over time.
- There are many other reasons for the difference between batting average and expected batting average, like a new training method or new approach, but the model is retrained every year in an attempt to minimize this effect.
- There are also other factors that we didn't include in this model, like speed or age, that could be the cause of difference between observed and predicted batting average.

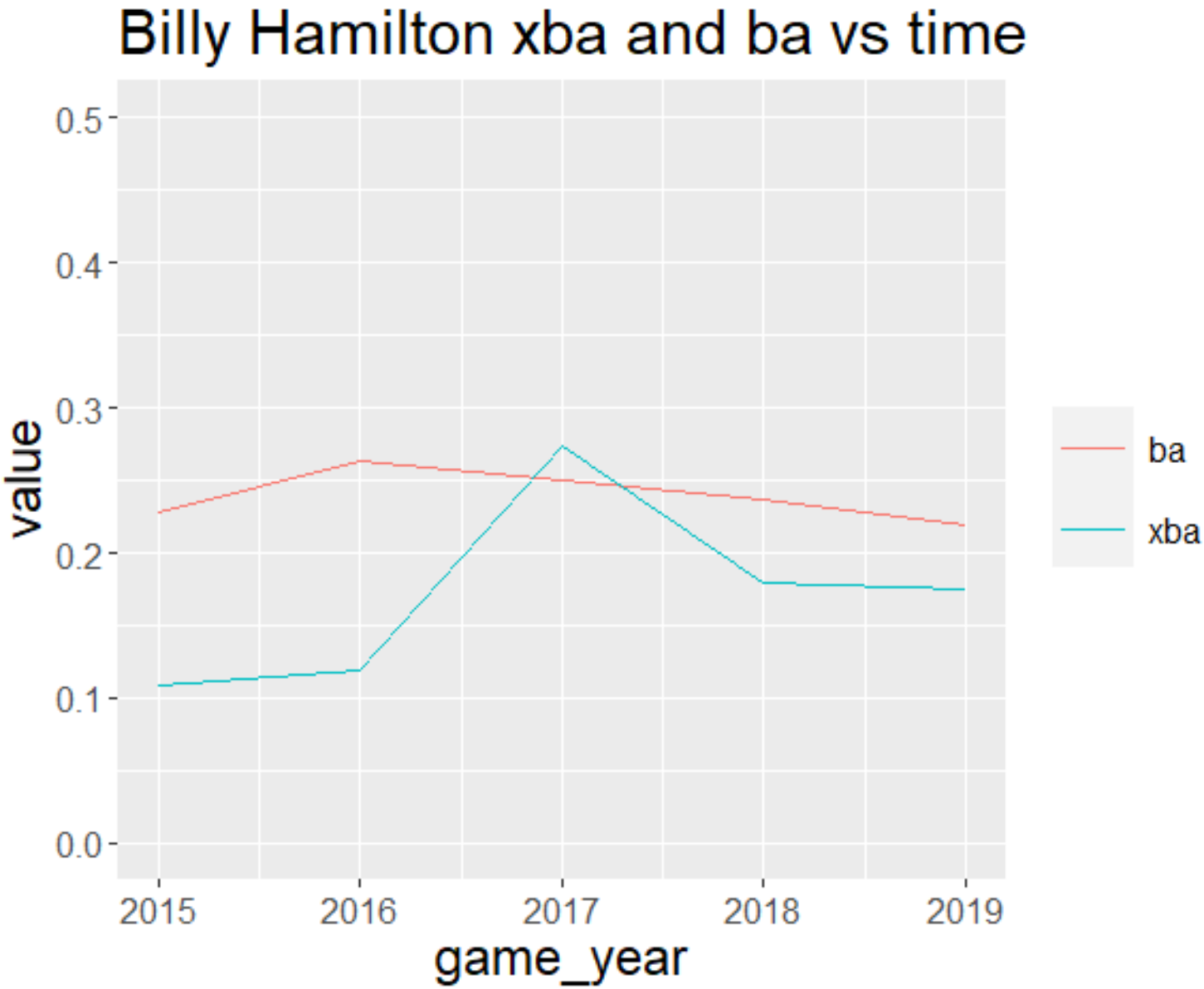
Example analysis - Positive



Example analysis - Negative



Model missteps



Good case for prediction