Marcus Semien - Hit Streak

Grant Nielson

2024-11-07

Figure 1: Marcus Semien, Texas Ranger

To find the probability of Marcus Semien obtaining a 20-game + hitting streak, I used a Bayesian model to simulate game-by-game hitting average for many 2025 seasons. I included ballpark factors for each venue from the Ranger's actual '25 schedule, and estimated increases and decreases in BB and BA by age. All hitting averages were randomly sampled from Semien's 2024 season and then adjusted based on draws from the ballpark, walk rate, and batting average effects for each sample.

I chose Marcus Semien becuase he's been our fearless leadoff hitter for the past few years; playing excellent defense, going to multiple all-star games, and rarely missing games. I'm sure the Rangers value him just right, but I think he is a little less heralded by fans and in the media compared to Corey Seager, El Bombi, and the revolving door of sensational rookies (Jung, Carter, Langford, Rocker, etc.)

A few assumptions/clarifications:

- **All games were assumed to be independent of each other (no hot hand advantage)**

- **No playoffs**
- **I sampled just from Marcus' 2024 season. If I did it again, I would use techniques from Marcel the Monkey's predictions and have 2024 weighted 5x, 2023 weighted 4x, and 2022 weighted 3x.**
- **After deliberation, I did not include Plate Appearance estimates. I sampled hitting averages directly from games.**
- **I assumed Marcus would play all 162 games. I doubt that will be the case given that most players won't play all games, he missed a few last year, he is only getting older, but given that he played every game in 2023, it's not out of the question.**
- **I did not account for the change in hit probability over the course of a game - third time thru the order penalty, hitting a closer, etc.**
- **I did not account for specific team's pitchers or pitching performance. I did use each team's ballpark effects. The variance from the priors will help capture some of the pitcher variance, but not concerning pitching staff strength.**
- **I did not explicitly account for variance in pitching strength for a team over a series (Ragans or Lugo one day to Brady Singer/Kyle Wright has an effect). However, there is game by game variance which helps capture this.**

Data Formatting

A word on priors:

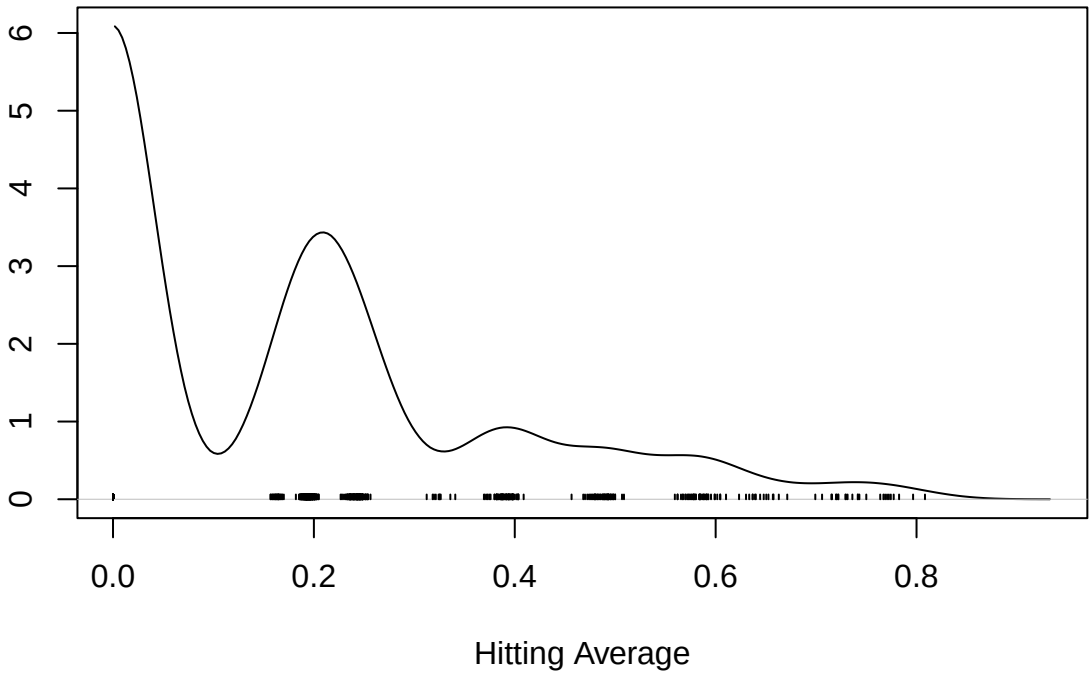
You'll see the BB & BA age effects, and parks effects, are normal with small standard deviations. This is because I took those as calculated values from Statcast and Dynasty Guru. We want them to vary to capture the change in games against different teams, but they are fixed effects in theory.

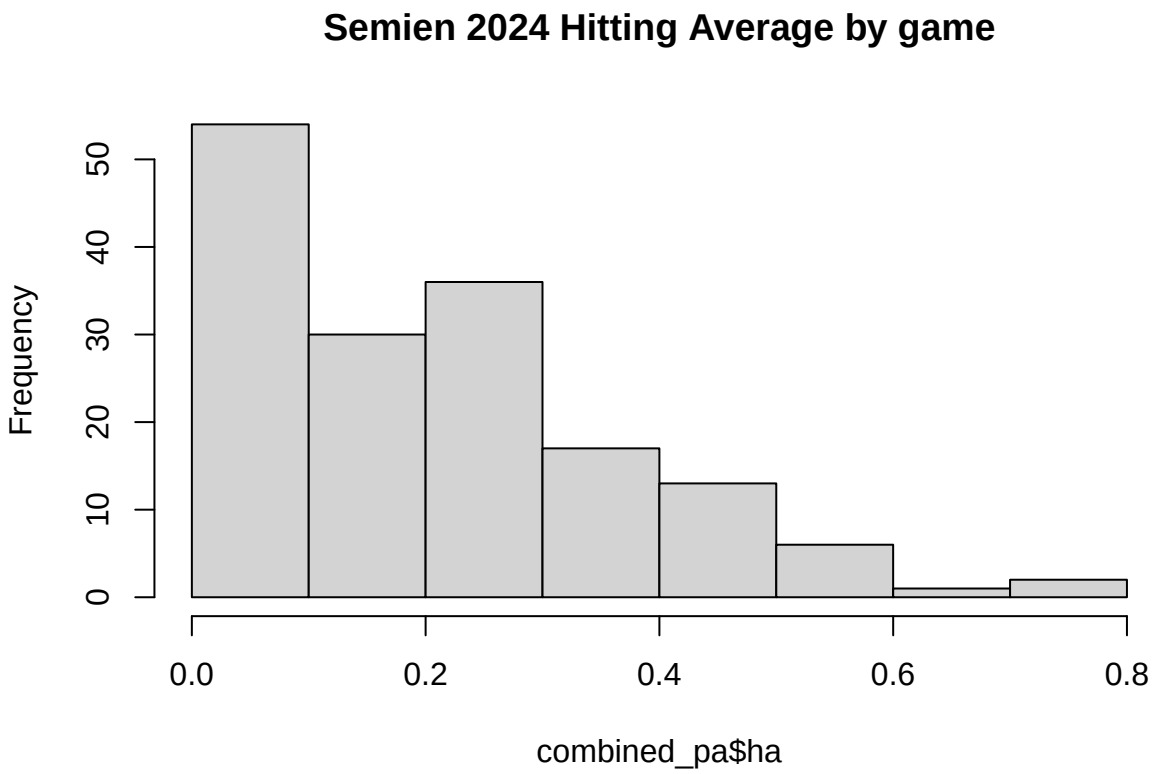
Hitting each game varies quite a bit, which is why I randomly sampled Hitting Average from Semien's 2024 season. The uniform prior is just a sampling tool to get an index. We see in the distribution plot that the posterior distributions match Semien's real HA distribution by game.

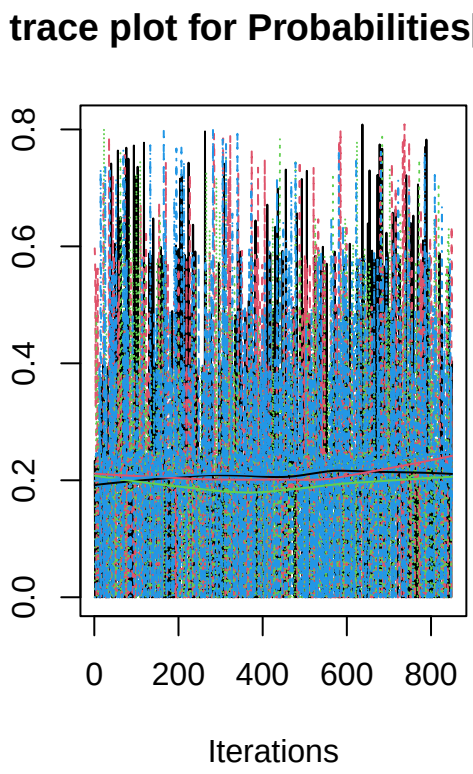
Model

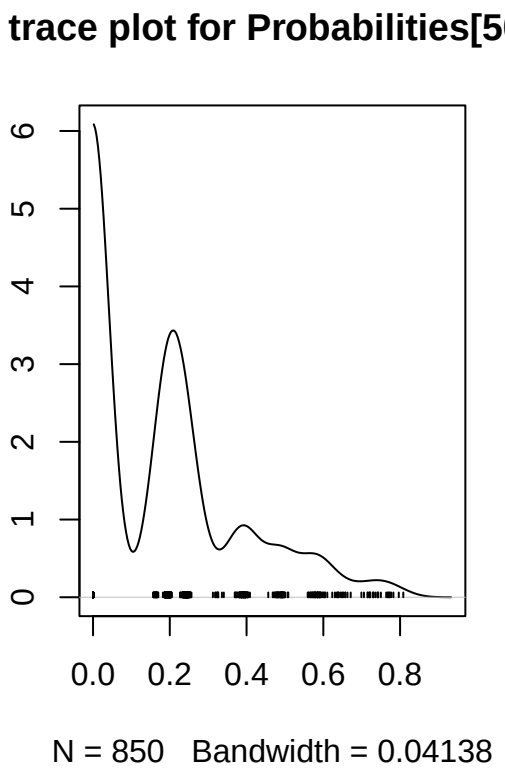
```
## Defining model


## Building model


## Setting data and initial values


##   [Note] 'marcus_pa_home_ha' is provided in 'data' but is not a variable in the model and is being ig


##   [Note] 'marcus_pa_away_ha' is provided in 'data' but is not a variable in the model and is being ig


## Running calculate on model
##   [Note] Any error reports that follow may simply reflect missing values in model variables.


##   [Warning] Dynamic index out of bounds: combined_pa_ha[nimRound(home_index[1])]
##   [Warning] Dynamic index out of bounds: combined_pa_ha[nimRound(away_index[1])]


## Checking model sizes and dimensions
```

```
##   [Note] This model is not fully initialized. This is not an error.
##          To see which variables are not initialized, use model$initializeInfo().
##          For more information on model initialization, see help(modelInitialization).

## ===== Monitors =====
## thin = 1: away_index, ba_age_adjustor_effect, home_index, park_effects, rangers_park_effect, walk_ag
## ===== Samplers =====
## posterior_predictive sampler (34)
##   - away_index
##   - home_index
##   - ba_age_adjustor_effect
##   - walk_age_adjustor_effect
##   - rangers_park_effect
##   - park_effects[]  (29 elements)

## thin = 1: away_index, ba_age_adjustor_effect, ha_away, ha_home, home_index, park_effects, probabilit

## Compiling
##   [Note] This may take a minute.
##   [Note] Use 'showCompilerOutput = TRUE' to see C++ compilation details.

## Compiling
##   [Note] This may take a minute.
##   [Note] Use 'showCompilerOutput = TRUE' to see C++ compilation details.

## [1] 162

## running chain 1...

## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|

## running chain 2...

## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|

## running chain 3...

## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|

## running chain 4...

## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|

##         Length Class     Mode
## samples 4      mcmc.list list
## summary 5      -none-    list
```


Density of Probabilities[50] (They all look about the same)
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
Hitting Average


Semien 2024 Hitting Average by game
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
combined_pa$ha





The densities for the posterior probabilites plots match the actual distribution, which is good.

The convergence plots for each game are similar. Each chain seems to nearly converge, and they are different from each other. This is okay, because this model isn't trying to estimate exact parameters, it's being used as a simulation tool. We wouldn't expect any season to have the same exact hit probability.

Extract Probabilites from model

```
## [1] 0.00415
```

```
## [1] 0.0032
```

```
## 95% credible interval:  0.002509024 0.004084313
```

Analysis

I give Semien a **0.0032** chance of getting a 20-game hit streak next year, with a 95% credible interval of (0.0025, 0.0041). To validate this result, I compared it to the probability of getting the streak using Fan graph's formula to validate. This is a nice formula but assumes that Semien has the same number of plate appearances (4) and hitting average (.215) for every game/at bat:

$$p = \text{probability of success (probability of hitting in a game)}$$

$$k = \text{length of streak}$$

$$G = \text{number of games in a season}$$

$$N = \text{estimated opportunities to begin a streak}$$

$$N = (\text{G - k}) * (1 - \text{p}) + 1$$

Plug In:

$$p = 0.620 = 1 - (1 - 0.215)^4 \quad \text{(Marcus's hitting average (hits/PA) on the season was 0.215)}$$

$$k = 20$$

$$G = 159$$

$$N = 48 = (G - k) \cdot (1 - p) + 1$$

The probability of at least one 20-game streak is:

$$P(\text{at least one 20-game streak}) = 0.003 = N \cdot p^k$$

Having the formula in my credible interval is a validating, although it's worth noting that this wasn't always the case when I changed my seed.

The variance in the simulation makes it easier to get a zero (yet easier to have a big game), which is more representative of a real season. Sometimes, there could be 'clusters' of high variance (good games) for the ha, walk, ba, and park effects estimates that lend themselves to more streaks, but the variance and randomness can also cause lower values than average that keep a streak from happening.

References

Dynasty Guru age curve BB and BA estimates: (https://thedynastyguru.com/2019/02/27/aging-gracefully-approaching-aging-curves-and-advanced-stats-part-ii/)

Fangraphs article on streaks: https://tht.fangraphs.com/the-probability-of-streaks/

Statcast park factors: https://baseballsavant.mlb.com/leaderboard/statcast-park-factors?type=year&year=2024&batSide=&stat=index_wOBA&condition=All&rolling=&sort=8&sortDir=desc