TED WILLIAMS - 1941

The Greatest Season in Baseball History

ABSTRACT

In 1941, Ted Williams put together arguably the greatest season in baseball history by hitting for an unbelievable .406 batting average. This project seeks to assess the Bayesian probabilities of this actually happening from views during the season itself.

Matt Dunman

ISYE 6420

I. INTRODUCTION

In 1941, Boston Red Sox star leftfielder Ted Williams put together one of the greatest seasons in baseball history by eclipsing the monumental .400 batting average plateau, a feat which hasn't been accomplished by anyone since. Not only did "Teddy Ballgame" accomplish this, on the last day of the season with a batting average of .399 and a double-header to play, he sealed the deal by going a combined 6/8 across the two games to end the season with a .406 batting average. For perspective, Ted would hit .356 the following season, and while 50 points lower, he still ended the season as the best hitter in baseball. Not only has no one ended a full season with a .400 batting average in the last 78 years, some believe it may never happen again. This study seeks to explore that magical season, not from an after-the-fact view, but from during the season itself.

Using data pulled from *Baseball Reference*[1], I was able to bring in Ted's previous season stats as a baseline and prior and also the stats from 143-game 1941 season. For the season, I took cumulative data from the end of each month, the last day, and last game of the season to estimate Ted's likely actual batting average and the probability of him actually achieving a batting average of .400 for the season. For example, looking at April I used all of the hitting data from April, whereas for June I combined all data up to that point (April, May, and June), and so on. The purpose of all of this is to relive the experience of the season and to remark at how unbelievable the feat was.

II. DATA INVESTIGATION

Looking at the data available from *Baseball Reference*, there are hundreds of different statistics to view, but for the sake of this project I kept it at simply: Game#, #Games Left, Date, Month, #At-Bats in that game, and #Hits in that game. A snapshot of this data is provided in Table 1.

Game	GamesLeft	Date	Month	AB	H
36	107	1-Jun	June	4	2
37	106	1-Jun	June	5	2
38	105	2-Jun	June	4	1
39	104	5-Jun	June	4	3
40	103	6-Jun	June	4	2

Table 1 - Hitting Data from Five Games in June

The data for Ted's previous season (1940) was simply one line that he batted .344 in 456 at-bats with 193 hits and 368 outs. One season of data is a good amount as it allows steadiness and little wavering with a little extra data, but can be swayed with a substantial amount of data added. Further, in 1939 Ted was a rookie and thus the results were biased too low as he wasn't fully skilled and developed yet. Lastly, the 1940 season is a perfect microcosm of Ted's full career as when he retired in 1960, his full career batting average across 20 years was also exactly .344, making 1940 the perfect candidate for building a prior.

III. MODEL DEVELOPMENT

One of the biggest struggles for this project was to determine an appropriate prior and likelihood, which were eventually decided as Bayesian Conjugates of a prior BattingAverage ~ Beta(193, 368) prior and a likelihood Hits ~ Binomial(BattingAverage, AtBats). Various articles and reports, notably the article *Bayesball*[2] by Ricky Kim cite that Beta priors are ideal for batting averages as they balance the probability of successes and failures and have the ability to absorb likelihoods and data using Bayesian Conjugacy to keep its Beta form as the posterior. While Kim's article looks at comparing different batters, the goal here with this project was to isolate Ted and predict his personal batting average and probabilities.

For the prior, Beta(193, 368) was chosen as it reflects Ted's previous 1940 season with 193 hits and 368 outs in 561 total at-bats. This allows steadiness when adding a few data points, but can be swayed with a lot of new data points. A Binomial likelihood was chosen as I was using cumulative discrete trial data (by summing successes and failures from lumps of time). It is also worth noting that Ted only had 456 at-bats in 1941 as pitchers avoided him due to his success and walked more than any other player in baseball (which doesn't count as an official "at-bat").

With Bayesian Conjugacy, the posterior now evolves for each of the time periods evaluated as Beta(193 + #Hits, 368 + #AtBats - #Hits). For example, adding in April data with 7 Hits in 18 AtBats gave a posterior of Beta(200, 377).

In developing the WinBUGS model, several calculations and adjustments were made given the constraints of a baseball season. As a baseball season is a finite time range and the goal was to

predict a .400 batting average at the end of the season, considerations needed to be made concerning this fact. First, a sampling of Future At-Bats (future_ab) predicts the number of hits to for the rest of the season using a Binomial distribution on the posterior BattingAverage (ba) and the number of at-bats remaining (ab_left). Next, as the posterior doesn't reflect solely the 1941 season, a weighted average Expected Batting Average (ba_exp) was made 1) the future_ab sampling and 2) the actual data up to that point (average). Further, 'average' was stored as current_avg so that it could be seen in the WinBUGS monitor.

For the final calculation, hit400 is a step function comparing ba_exp to .400 to find the probability across iterations of Ted having a .400 average at the end of the season. Lastly, given the actual scenario that Ted went into the final day of the season with a .399 batting average and went for an astonishing 6/8, I put in last_day which assesses the probability that he would go 6/8 given his posterior batting average. The full WinBUGS model with the data for each month, the final day, and the final game (the last day was a double-header) are included in Figures 1 and 2 below.

MODEL

```
model{
   #Beta Prior on Batting Average
   ba ~ dbeta(193, 368)    #Previous season: 193 hits and 368 outs

   #Binomial Likelihood on 'hits' and 'ab' data
   hits ~ dbin(ba, ab)

   #Future At-Bats: Sample using Posterior Batting Average for remaining at-bats in the season
   future_ab ~ dbin(ba, ab_left)

   average <- hits/ab                          # current average
   ab_left <- ab_total - ab                    # total number of at-bats in the season
   ab_todate <- ab/ab_total                    # % of at-bats already occured in the season
   ab_togo <- ab_left/ab_total                 # % of at-bats remaining in the season
   current_avg <- average + ba*0.00000001      # included so that the average can be monitored
   future_atbats <- future_ab/ab_left          # adjusted sample # to % for calculating average

   # Expected batting average: Weighted average of at-bats data and sample of at-bats yet to occur
   ba_exp <- average*ab_todate + future_atbats*ab_togo

   hit400 <- step(ba_exp - 0.4)                # Likelihood of .400 batting average for the season
   last_day <- step(future_atbats - 0.75)      # Likelihood on the final day of hitting 6/8
}
```

Figure 1 - WinBUGS Model

DATA1

list(ab_total = 456)

DATA2

End of April:	list(hits = 7, ab = 18)
End of May:	list(hits = 51, ab = 119)
End of June:	list(hits = 86, ab = 213)
End of July:	list(hits = 113, ab = 276)
End of August:	list(hits = 156, ab = 383)
Going into Final Day:	list(hits = 179, ab = 448)
Going into Last Game:	list(hits = 183, ab = 453)

Figure 2 - WinBUGS Data

IV. RESULTS

Three snapshots of results are provided for this report, taking into account 1 million iterations (after 1000 burn-in). After the first 18 at-bats in April, the probability of Ted hitting .400 for the season was only 3.5% (Figure 3). By the end of July, the probability had increased to nearly 30% (Figure 4). Going into the last day of the season, the probability was 34.3% with the probability of him hitting 6/8 that day to end up with a .406 season average was only 3.4% (Figure 5).

	mean	median	sd	MC_error	val2.5pc	val97.5pc	start	sample	ESS
ba	0.3454	0.3453	0.01976	1.907E-5	0.3072	0.3846	1001	1000000	1073156
ba_exp	0.3472	0.3465	0.02894	2.788E-5	0.2917	0.4057	1001	1000000	1077572
current_avg	0.3889	0.3889	0.0	1.0E-13	0.3889	0.3889	1001	1000000	0
hit400	0.03505	0.0	0.1839	1.795E-4	0.0	1.0	1001	1000000	1049649

Figure 3 - End of April

	mean	median	sd	MC_error	val2.5pc	val97.5pc	start	sample	ESS
ba	0.3656	0.3655	0.01664	1.588E-5	0.3333	0.3985	1001	1000000	1097844
ba_exp	0.3921	0.3925	0.0156	1.525E-5	0.3618	0.4232	1001	1000000	1047067
current_avg	0.4094	0.4094	0.0	1.0E-13	0.4094	0.4094	1001	1000000	0
hit400	0.2996	0.0	0.4581	4.512E-4	0.0	1.0	1001	1000000	1030768

Figure 4 - End of July

	mean	median	sd	MC_error	val2.5pc	val97.5pc	start	sample	ESS
ba	0.3687	0.3686	0.01519	1.509E-5	0.3391	0.3987	1001	1000000	1014204
ba_exp	0.3992	0.3991	0.002842	2.813E-6	0.3947	0.4057	1001	1000000	1020433
current_avg	0.3996	0.3996	0.0	1.0E-13	0.3996	0.3996	1001	1000000	0
hit400	0.3432	0.0	0.4748	4.774E-4	0.0	1.0	1001	1000000	989191
last_day	0.03406	0.0	0.1814	1.87E-4	0.0	1.0	1001	1000000	941198

Figure 5 - Going into the Last Day of the Season

V. CONCLUSION

In conclusion, Ted Williams was an amazing hitter who defied the game and defied probabilities. Given the probabilities and statistics derived in this project, knowledge of the history of baseball before 1941, and knowledge of baseball since 1941, it's truly astonishing that ".406" ended up a reality. Further, these statistics only represent pure probabilities and don't take into account the fact that most pitchers avoided him and wouldn't throw him good pitches to hit, the media and personal pressure to break .400, weather conditions in Boston in late September as the season was ending, and physical / mental fatigue across a season.

As for the project, based on previous baseball analytics I believe the prior chosen was ideal, but there may be an opportunity in the future to optimize the prior for this or other future studies. Finding that perfect middle ground to balance historical data with new incoming data was a struggle for this project, but I believe was addressed properly. In the end, applying Bayesian statistics to see the probabilities evolve through the 1941 season, and knowing that Ted Williams eventually conquered them, has been a truly enjoyable experience for me, and I hope it is for others as well.

REFERENCES

1. "Ted Williams". *Baseball Reference.* https://www.baseball-reference.com/players/w/willite01.shtml

2. Kim, Ricky. "Bayesball: Bayesian Analysis of Batting Average". *Towards Data Science.* 2018, March 4: https://towardsdatascience.com/bayesball-bayesian-analysis-of-batting-average-102e0390c0e4