The History of Home Runs

Alyssa June Pacleb
12/12/19
SDS 358

TEXAS
The University of Texas at Austin

Statistics
+ Data
Sciences

Introduction

Objectives: Observe the linear trend predicting number of **home runs** hit every **year** change through time controlling for **salary** in thousands. Observing how **league** moderates the effect of year on home runs and how which **hand** the player bats with moderates the effect of year on home runs. Baseball is a sport rich in statistics. I want to see how a "flashy" statistic such as home runs has changed through time.

Hypotheses: Home runs will increase as years increase; the linear segment spanning over the more recent model space should have a larger slope and have a significant difference between the end of the first linear segment and start of the second linear segment. I predict that the league the batter is in and the hand the batter hits with will moderate the effect of year on home runs. Additionally, I predict that batters in the American League and batters that switch between both hands will perform better.

Methods

Sample: Briefly describe your sample data. This sample consists of 2,708 different batters from 1985 to 2016. Each distinct year and batter resulted in 12,726 distinct observations. Using a Cook's plot, there were two observations that were visually different from the other observations and were removed from the model resulting in 12,724 observations. Year and salary are measured in years and dollars, respectively. The two possible values for league are American and National League. The three possible values for batting hand are right, left, and both.

Analysis Method: R and RStudio were used to analyze the data using segmented regression with interactions. Functions like the linear model function were used.

Descriptives

Response Variable:

	Mean	Standard Deviation
Home runs	0.122	0.953

Explanatory Variables:

	Mean	Standard Deviation
Year	0.122	0.953
Salary	33000	3172.81

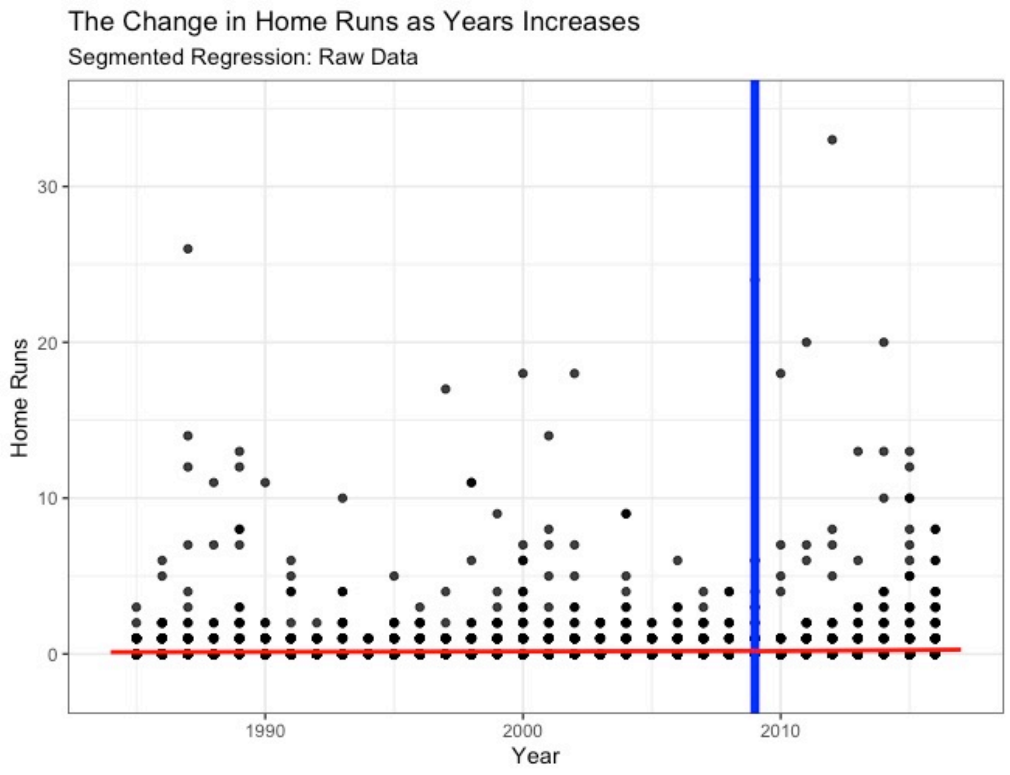
	AL	NL
League	6192	6534

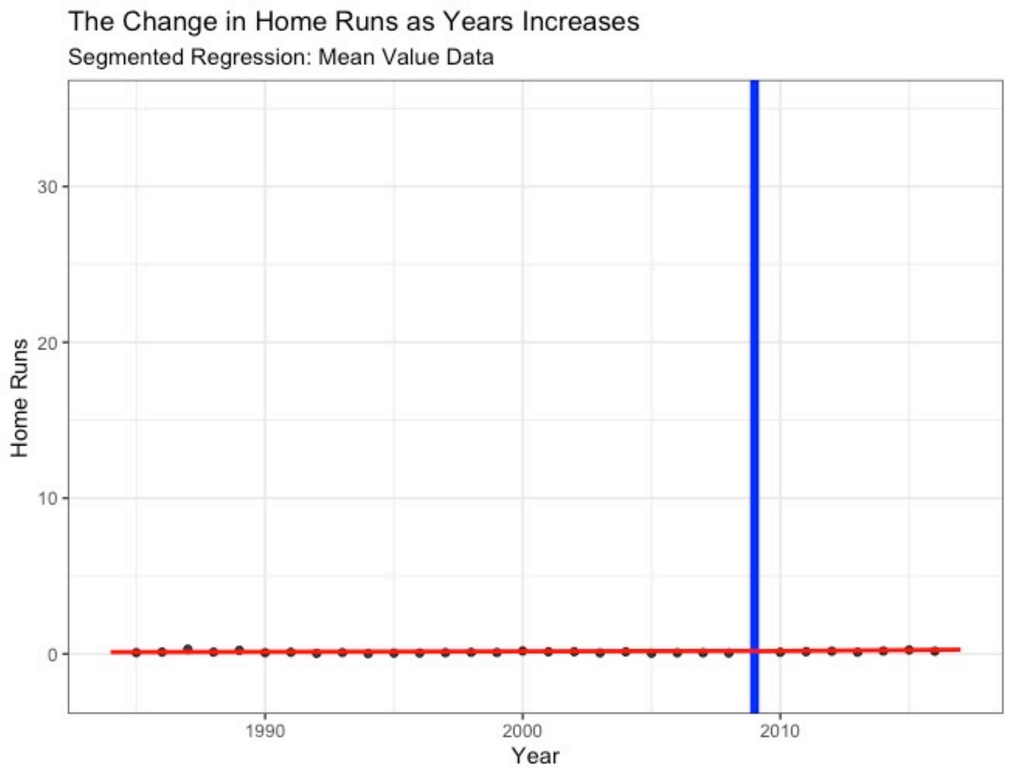
	Right	Left	Both
Batting Hand	8874	3489	363

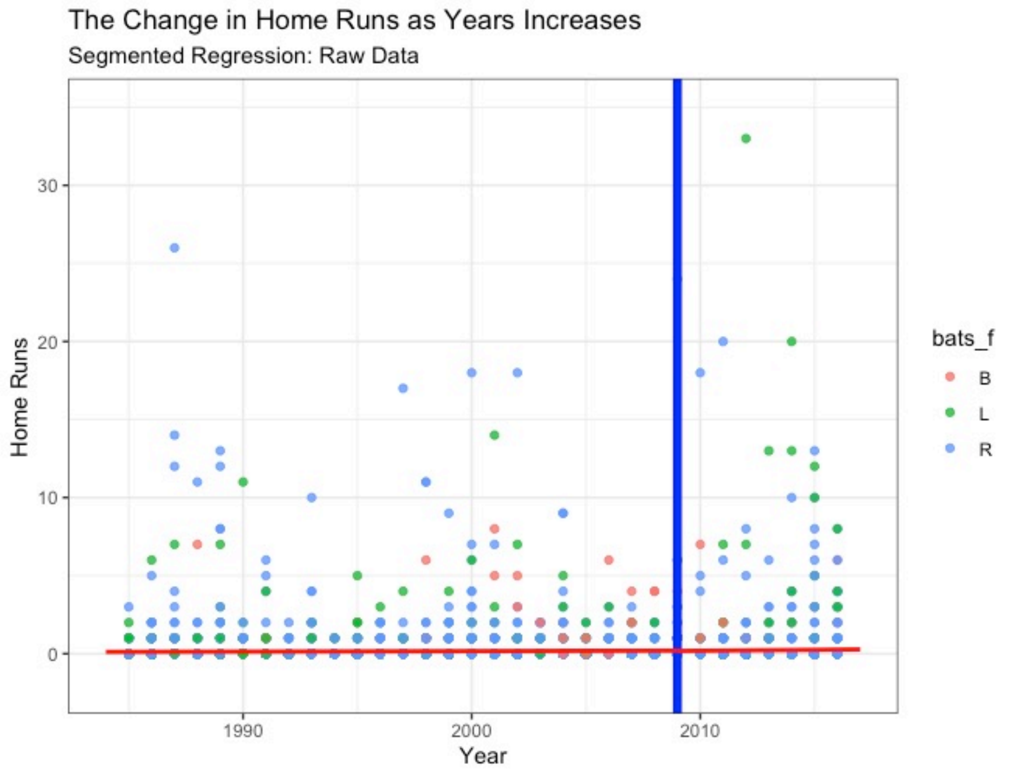
Results

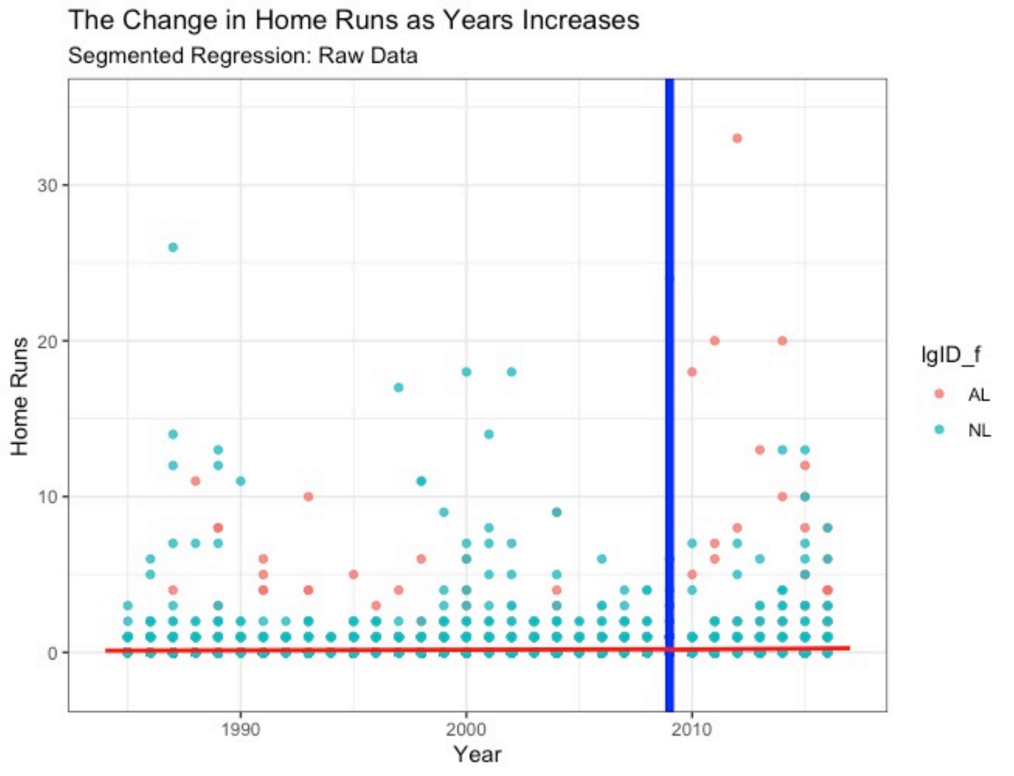
Results table:

	Estimate	T-value	P-value
Intercept	-42.23	-2.604	< 0.05
Segment 1 Slope	0.0213	2.616	< 0.05
Jump	-0.1866	-0.830	0.406
Difference between Segment 1 and 2	-0.0102	-0.225	0.822
Segment 2 Slope	0.0111	0.247	0.805
Salary (in thousands)	0.00001147	4.345	< 0.05
Segment 1 * Left hand	-0.0228	-2.707	< 0.05
Segment 1 * Right hand	-0.0248	-3.042	< 0.05
Segment 1 * National League	-0.0037	-1.315	0.189
Jump * Left hand	0.2629	1.145	0.252
Jump * Right hand	0.2812	1.255	0.209
Jump * National League	-0.0483	-0.714	0.475
Segment 2 * Left hand	0.0210	0.458	0.647
Segment 2 * Right hand	-0.0043	-0.096	0.923
Segment 2 * National League	-0.0079	-0.593	0.553


The Change in Home Runs as Years Increases
Segmented Regression: Raw Data
Home Runs
0
10
20
30
Year
1990
2000
2010

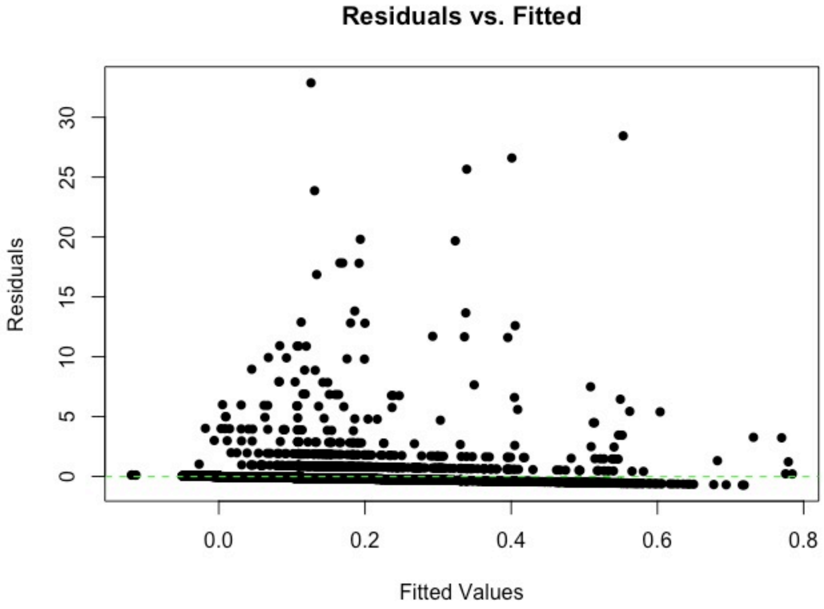

The Change in Home Runs as Years Increases
Segmented Regression: Mean Value Data
Home Runs
0
10
20
30
Year
1990
2000
2010

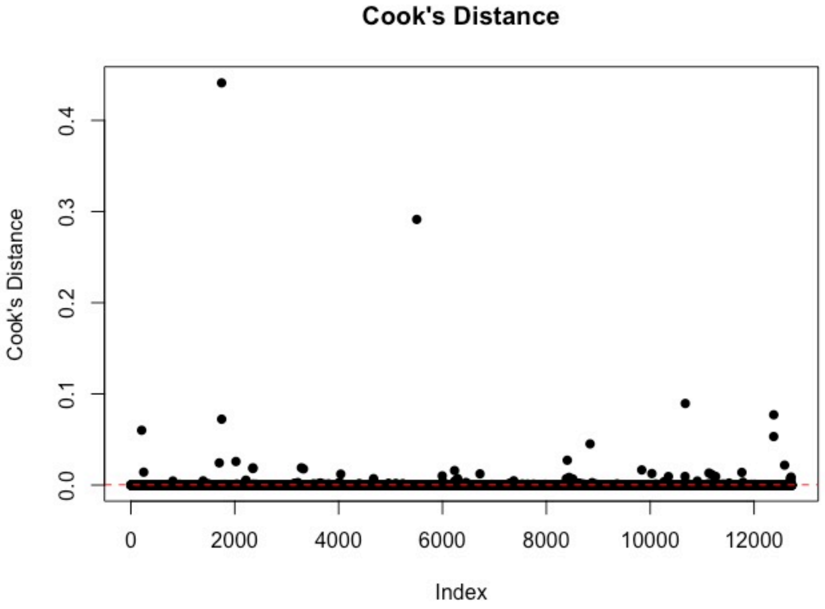

The Change in Home Runs as Years Increases
Segmented Regression: Raw Data
Home Runs
0
10
20
30
Year
1990
2000
2010
bats_f
B
L
R


The Change in Home Runs as Years Increases
Segmented Regression: Raw Data
Home Runs
0
10
20
30
Year
1990
2000
2010
lgID_f
AL
NL

Assumptions

Assumptions: A **Residuals vs Fitted plot** was made to check for non-linearity and unequal residual variances. A **Cook’s Plot** was made to check for outliers and two outliers were removed.





Discussion

Interpretation: The smallest RMSE from different models with different breakpoints was 0.947. The model with the smallest RMSE had a breakpoint of 2009, so the year 2009 was used to separate the dataset. The overall model significantly fits the data and 1.014% of the variation in home runs is explained by the model, F(16, 12707) = 8.139, $p < 0.05$. The intercept of the model was -42.23, meaning that a batter hypothetically playing in year 0 would bat -42.23 home runs, t(12707) = -2.604, $p < 0.05$. For every one unit increase in year before 2009, home runs significantly increased by 0.0213, t(12707) = 2.616, $p < 0.05$. There is a non-significant main effect of the year 2009 increasing the number of home runs per year; after 2009, the number of home runs hit decreases by 0.1866, t(12707) = -0.830, p = 0.406. The increase in home runs each year between segments is non-significant, t(12707) = -0.225, p = 0.822. For every one unit increase in year after 2009, home runs non-significantly increase by 0.0111, t(12707) = 0.247, p = 0.805. The control variable, salary (in thousands), had a significant effect on home runs, b = 1.147×10^{-5}, t(12707) = 4.345, $p < 0.05$.

Of the interactions in this model, the slope of home runs was only moderated by the player's batting hand before 2009. When going from a batter who bats with both hands to a batter that only uses their left hand, the slope for home runs before 2009 decreases by 0.0228, t(12707) = -2.707, $p < 0.05$. Similarly, when going from a batter who bats with both hands to a batter that only uses their right hand, the slope for home runs before 2009 decreases by 0.0248, t(12707) = -3.042, $p < 0.05$.

Upon further investigation of the simple slopes, there was a significant impact of year on home runs for those who switched batting hands (b = 0.019, t(12707) = 2.429, $p < 0.05$). There was a non-significant impact of year on home runs for those who used their left hand (b = -0.003, t(12707) = -1.265, p = 0.2058). There was a significant impact of year on home runs for those who used their right hand (b = -0.005, t(12707) = -3.185, $p < 0.05$).

Limitations: The number of the batters who could switch batting hand was disproportionately less than those who only used either their left or right hand. This difference could have impacted the results. Inflation is not accounted for when describing salary. The Residual vs Fitted plot was not as random as I'd like, but this was expected due to the nature of the response variable.

Implications: Batting average was not readily available which would have been a good measure of batting performance.

This dataset is designed for to make database querying simple (primary and foreign keys) instead of statistical analysis. Other software was needed to easily join the columns of interest from different files into one file.

Segmented regression was the most correct way to assess how home runs change throughout the model space (time).

References:

The data was taken from a baseball database compiled by author and journalist, Sean Lahman.

http://www.seanlahman.com/baseball-archive/statistics/