When We Get Hooked on Baseball

Seth Stephens-Davidowitz

April 18, 2014

Define $F_{a,s,2014}$ as the number of fans team s has of people age a in year 2014.
Define $P_{s,y}$ as the performance of team s in year y. (Different measures of performance are used, as discussed below.)
The regression is:

$$log(F_{a,s,2014}) = \alpha_a + \gamma_s + \gamma_s * Age + \sum_{j=-4}^{21} \beta_j P_{s,2014-a+j} \quad (1)$$

where α_a is age dummies and γ_s is team dummies. All regressions were done separately, by gender.
In words, every team x age is an observation. And this is regressed on team dummy variables, age dummy variables, and then dummy variables for how good the team was starting 4 years prior to the birth to the age of 21.

What is reported is $\beta_j - \beta_0$ to compare everything to effect on year someone was born.

Starting before birth allows for a sanity check.

More Details About Data: This data is included for everybody age 30-64 and downloaded from Facebook. (Age decisions are for the following reason: Facebook does not include individual-age data above 64. There is just a 65+ category. And, if lower age cutoff than age 30 is used, the final year might come as statistically significant because it correlates with performance 10 years down the road, which could influence 30 year olds.)

Start with men. Figure I uses winning the World Series as a measure of performance. Figure II uses total wins in regular season as measure of performance. Figure III uses both. Note, no matter the measure of performance, the same ages stick out as clearly most influential. However, winning the World Series gives a much sharper picture.

Figure IV shows the estimated effect, not including the team x age trend. Including the trend is the preferred model. But the most important years remain roughly the same.

Figure V shows the effects of winning a World Series at different ages on women's fandom. Note the model differs a decent amount depending on the inclusion of a trend.

In addition, I did this model a whole bunch of different ways, changing ages included, model specification, and other details. I found that the results using women were much less stable and generally would find later years are more important. Results with men always keyed in on the same years. I failed to find much evidence of decreased importance of childhood by age.

Finally, a quick note about data quality. In general, it seemed pretty reliable. Teams that are known to be more popular had more fans. And fans tended to be located in places you would expect. Facebook is known to receive a decent number of fake likes. It is hard to imagine, though, this is generating the patterns seen here. But I did notice a couple anomalies in the data: the Mets, for example, have way more fans in Massachusetts than is reasonable. But, again, I would be pretty surprised if any anomaly was generating the very robust importance of years 8-12 in a boy's baseball life and the outsized importance of Championship seasons.

Please e-mail me with any questions, comments, or requests.

Figure I
Effect of a Championship (Men)

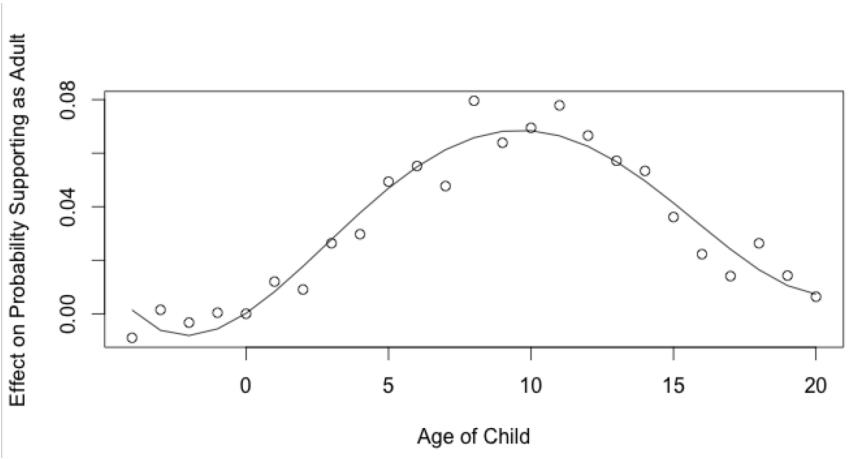


Figure II
Effect of a Win (Men)

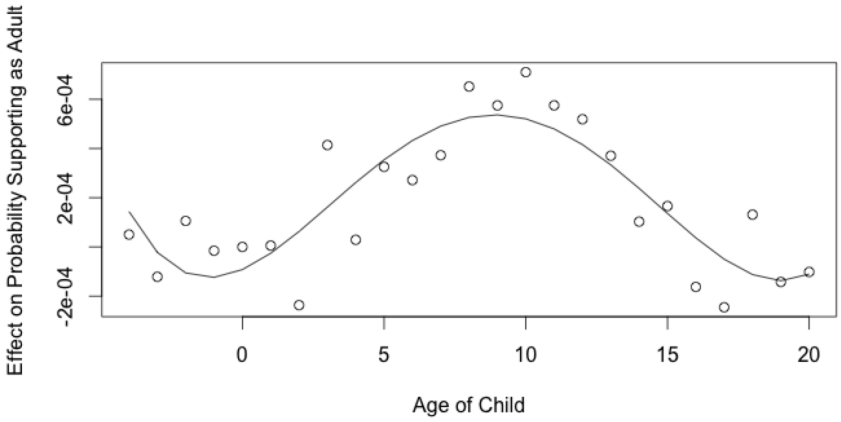
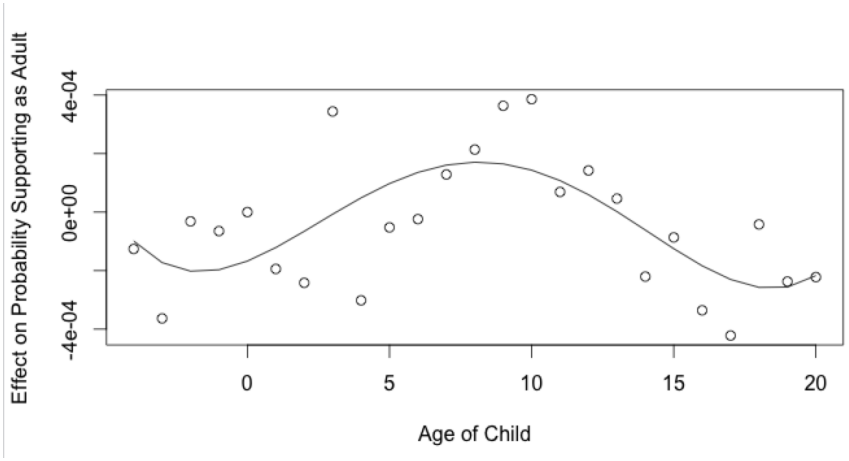
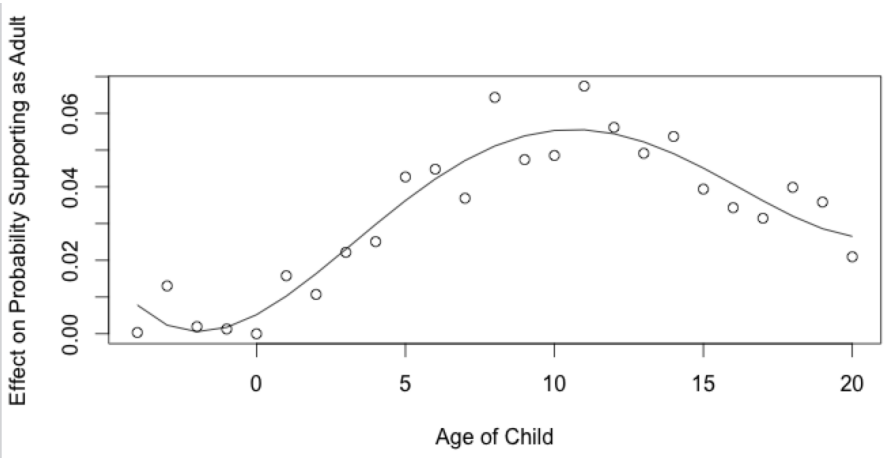


Figure III
Model with Both Wins and Championships (Men)



(a) Win



(b) World Series

Figure IV
Effect of a Championship (Men) – model without a linear trend

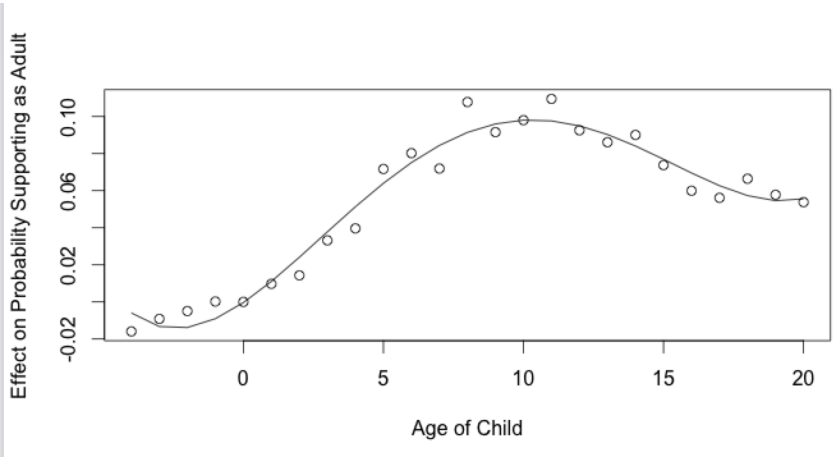
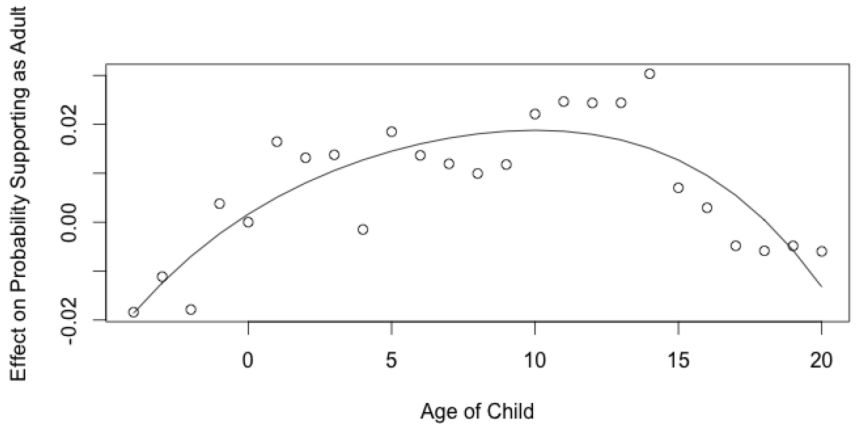
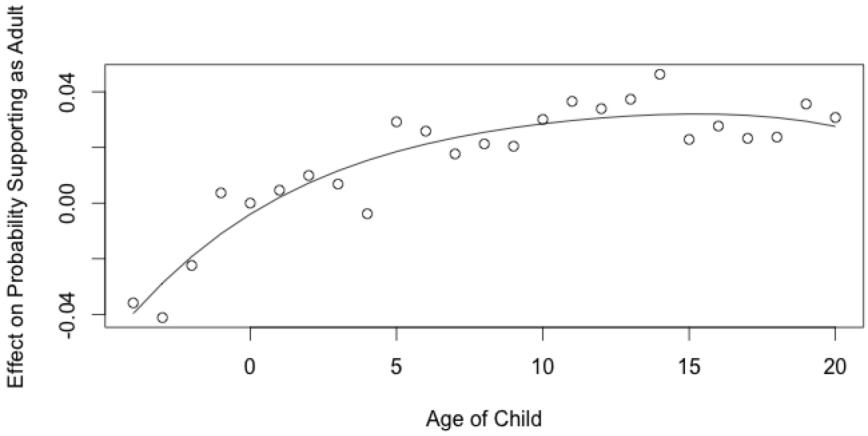


Figure V
Effect of a Championship (Women) – Trend vs. No Trend



(a) Trend



(b) No Trend