830 Semester Project

Andy Beck

December 17, 2018

Introduction

Marathons, races that are 26.2 miles (~42.2km) in length, have seen their popularity in the United States grow in the recent past, peaking in 2014 where an estimated 550,600 runners completed a marathon in the United States [CITE ME]. Goals for these runners can vary from wanting to compete for a victory to just wanting to finish, but in general annecdotal evidence seems to suggest that at least those who run more than one seek to improve upon previous performances. Previous scientific study of marathon runners focused more rigorously on understanding how nutrition (Jeukendrup 2011), running mechanics and energy cost over the course of a marathon (Hausswirth, Bigard, and Guezennec 1997) impact race performance. Little work has been published that evaluates pacing strategy, with most of these articles concluding that steady pacing over the race yields better results than changing pace over the race (Angus 2013). Even less work has been done to evaluate differences in pace strategy across age, gender, different courses, and ability level.

Here in this report, I attempt to quantify how pace profiles differ across these factors. In my first analysis, I incorrectly apply GEE to a dataset containing the half split and full marathon time from 37,250 runners from an anonymized dataset that I found on github. I then use a more appropriate linear regression model and argue that a more robust method like iteratively re-weighted least squares (IRLS) might be more appropriate due to some evidence of deviation from the normality assumption of the model. In subsequent analyses, I use data from the Boston Marathon (2015, 2016, and 2017) to explore how pace profiles: 1) differ acorss age, gender, and ability level; and 2) how pace profiles on the same course vary by year.

Data and Methods

Dataset 1 - Vanderplas Anonymized Marathon Times (VAMT): 37,250 Half and Finish Times

The VAMT dataset, made by Jake Vanderplas, contains 37,250 aggregated and anonymized marathon times scraped from the internet. Along with how long it took the runner to complete the full marathon, we also have the time for the first half of the race, the runner's age, and the runner's time. In addition to these variables, I computed the paces for each half and the full race, whether or not the runner ran a Boston Marathon Qualifier (using the 2019 standard, as I do not know the year of these results), and wheter or not the runner finished within the top 10% of their gender. In this dataset, the average finishing time was approximately 4 hours and 48 minutes (SD = 1 hour 3min 32.14 sec), 4:36:28.6 (1:00:57.38) for men and 5:11:2.52 (1:02:15.14) for women.

Dataset 2: Boston Marathon Results (2015 - 2017)

Results from the 2015, 2016, and 2017 Boston Marathon were acquired from kaggle and include the runner's name, gender, age, city/state/country of residence, country of citizenship, full marathon time, and intermediate splits at 5km, 10km, 20k, 13.1 miles, 25km, 30km, 35km, and 40km.

Year	Gender	N	Average Finishing Time(sd)	Average Half Split(sd)	Average Pace Difference (Full - Half)
2015	Male	14581	3h 36m 35.68s (40m 46.31s)	1h 42m 58.74s (17m 22.22s)	10m 37.22s (12m 22.28s)
	Female	12017	3h 58m 21.55s (36m 43.13s)	1h 54m 2.24s (15m 57.12s)	10m 16.10s (10m 7.26s)
	Total	26598	3h 46m 25.70s (40m 28.12s)	1h 47m 58.33s (17m 37.53s)	10m 27.69s (11m 24.69s)
2016	Male	14463	3h 45m 54.57s (41m 15.71s)	1h 44m 7.79s (17m 21.80s)	17m 38.81s (14m 17.32s)
	Female	12167	4h 5m 54.00s (38m 10.82s)	1h 55m 52.29s (16m 23.64s)	14m 9.60s (11m 23.66s)
	Total	26630	3h 55m 2.58s (41m 6.43s)	1h 49m 49.69s (17m 54.54s)	16m 3.22s (13m .66s)
2017	Male	14438	3h 48m 54.69s (42m 52.34s)	1h 45m 13.11s (17m 48.41s)	18m 27.82s (15m 21.76s)
	Female	11972	4h 9 m 5.48s (38m 28.01s)	1h 56m 52.37s (16m 42.81s)	15m 20.45s (12m 13.61s)
	Total	26410	3h 58m 3.56s (42m 8.88s)	1h 50m 30.11s (18m 15.92s)	17m 2.88s (14m 6.83s)

Methods

Pacing Profiles in VAMT Data

To analyze the differences in pacing profile across age, gender, and ability level in the VAMT dataset, I fit the following linear regression model:

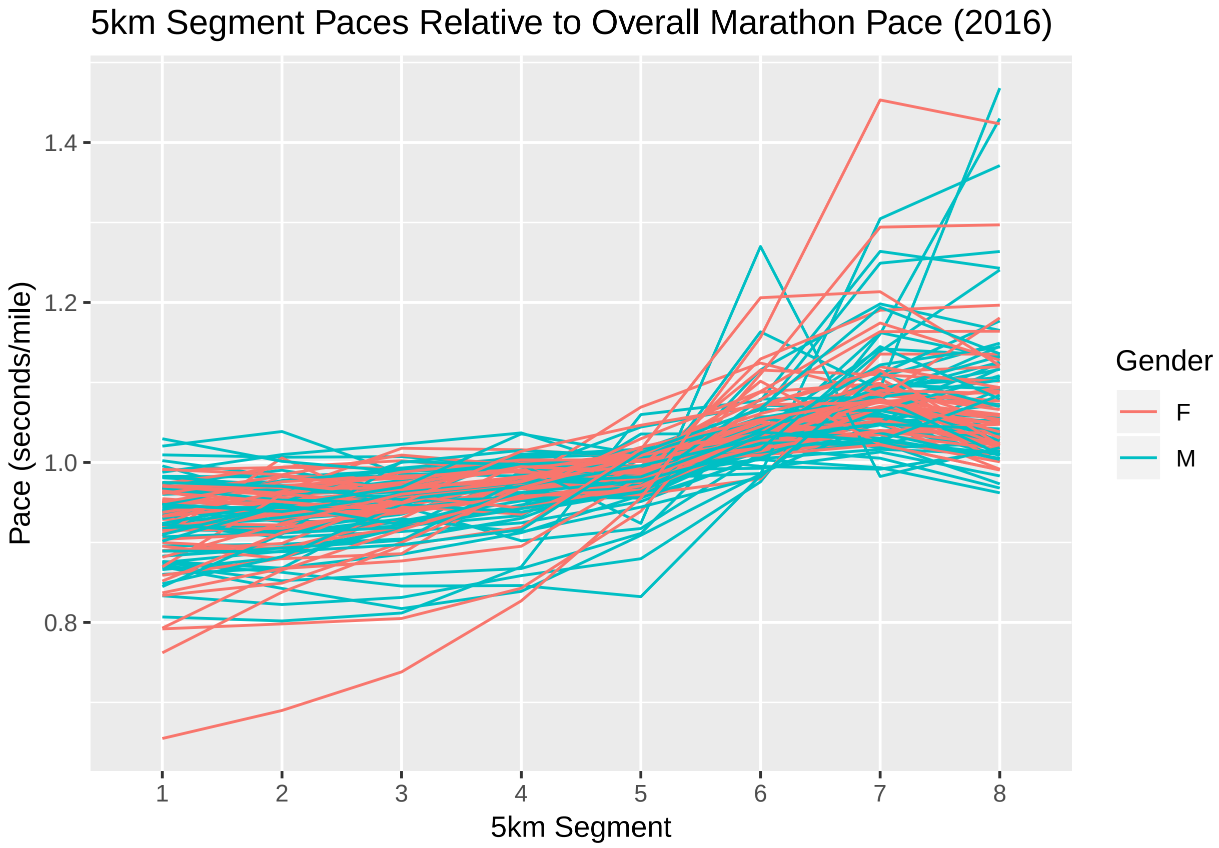
$$y_i = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{age}_i + \beta_3 \text{BQ}_i + \beta_4 \text{gender}_i * \text{age}_i + \beta_5 \text{BQ}_i * \text{age}_i + \beta_6 \text{BQ}_i * \text{gender}_i + \beta_7 \text{age}_i^2 + \epsilon_i$$

where y_i is $\frac{(\text{second half pace}_i - \text{first half pace}_i)}{\text{overall pace}_i}$, $\epsilon_i \sim_{iid} N(0, \sigma^2)$, and age has been normalized. This model was chosen to include the covariates analyized by (Reese and Ward 2015). This dataset only includes the half and full splits, so we don't have a very granular view of how pacing changes on average over the race, but we can nonetheless evaluate how changes in pacing during the second half of the race might be influenced by factors such as age, gender, and ability level. Due to evidence of deviation from normality (see appendix), we also fit a more robust, IRLS model with the same covariates.

Pacing Profiles in Boston Marathon Data

In the Boston Marathon data, we have a much more granular view of how the pacing profile for an individual changes over the course of the race thanks to intermediate splits being available every 5km. We'd like to again look at the relationship between age, gender, ability level, and pacing over the course of the marathon. To do this, we will model the relative pace of each 5km segment (compared to the overall pace of the marathon) as a linear function of age, gender, and whether or not the individual finished in the top 20% of their gender.

Each individual has his or her own unique pacing profile, as demonstrated in the plot below. We'd like our method to allow for variation in each individual's profile, but ultimately our interest is in population level effect estimates for our covariates. For this analysis, we use GEE to estimate the population-averaged effects of our covariates. We fit separate models for each individual year, as there is some evidence that the pacing profiles vary across the years (in particular, the ratio of the pace of the second half over the pace of the first half is on average smaller in 2015 than in 2016 and 2017 (see appendix); this might be due to the weather having been cooler in 2015). For our working correlation model, we use the AR1 correlation structure as our working correlation model, since it seems reasonable to assume that two adjacent segments for an individual are more alike than two segments further apart from each other.



Results

VAMT Dataset Analysis

Linear Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.2715057	0.0008427	322.2011559	0.0000000
factor(gender)W	0.0166321	0.0012502	13.3035549	0.0000000
ageN	0.0022361	0.0007447	3.0027818	0.0026770
factor(BQ)1	-0.0786333	0.0034881	-22.5434162	0.0000000
factor(top10Q)1	-0.1391070	0.0030434	-45.7071546	0.0000000
ageN2	0.0053808	0.0004392	12.2514707	0.0000000
factor(gender)W:ageN	-0.0008208	0.0012371	-0.6634639	0.5070376
ageN:factor(BQ)1	-0.0189990	0.0021943	-8.6585090	0.0000000
factor(gender)W:factor(BQ)1	-0.0024151	0.0043924	-0.5498390	0.5824331

IRLS Model

term	estimate	std.error	statistic
(Intercept)	0.272312172	0.000846227	321.7956322
factor(gender)W	0.016760952	0.001255497	13.3500571
ageN	0.003193084	0.000747824	4.2698322
factor(BQ)1	-0.078846867	0.003502853	-22.5093301
factor(top10Q)1	-0.139962326	0.003056328	-45.7942720
ageN2	0.005003224	0.000441059	11.3436656
factor(gender)W:ageN	-0.000764273	0.001242374	-0.6151714
ageN:factor(BQ)1	-0.019506205	0.002203544	-8.8521981
factor(gender)W:factor(BQ)1	-0.002041735	0.004410996	-0.4628739

From the above table, there seems to be evidence that the difference in second half pace to the first half pace relative to overall pace varies across age, gender, and our two measures of ability (Boston qualifier and top 10 percentile). In particular, this model suggests that female runners, older runners, runners of "lower ability" slow down more in the second half of the race. There does not appear to be a significant interaction between age and gender or gender and the Boston Qualifier indicator. The effect of age appears to be moderated by the Boston Qualifier status, with Boston Qualifiers being less impacted by age.

Boston Marathon Analysis (GEE)

2015

term	estimate	std.error	statistic	p.value
(Intercept)	0.915986685	0.000390824	5.493090e+06	0.000000000
as.numeric(segment)	0.017875302	0.000083746	4.555983e+04	0.000000000
ageN	-0.000009591	0.000171168	3.139887e-03	0.955314170
factor(Gender)M	-0.010508758	0.000591865	3.152513e+02	0.000000000
factor(top20Q)1	0.001294368	0.000307060	1.776921e+01	0.000024939
as.numeric(segment):factor(Gender)M	0.001910463	0.000136445	1.960493e+02	0.000000000

2016

term	estimate	std.error	statistic	p.value
(Intercept)	0.882075634	0.000419700	4.417066e+06	0.000000000
as.numeric(segment)	0.025475254	0.000096066	7.032297e+04	0.000000000
ageN	-0.000074589	0.000182171	1.676469e-01	0.682211621
factor(Gender)M	-0.041400093	0.000647881	4.083313e+03	0.000000000
factor(top20Q)1	0.001095533	0.000384132	8.133718e+00	0.004344956
as.numeric(segment):factor(Gender)M	0.008498512	0.000151494	3.146979e+03	0.000000000

2017

term	estimate	std.error	statistic	p.value
(Intercept)	0.875708992	0.000458533	3.647366e+06	0.000000000
as.numeric(segment)	0.026870979	0.000105971	6.429706e+04	0.000000000
ageN	0.000054232	0.000195433	7.700611e-02	0.781396402
factor(Gender)M	-0.038569999	0.000704828	2.994563e+03	0.000000000
factor(top20Q)1	0.001429656	0.000435590	1.077227e+01	0.001030321
as.numeric(segment):factor(Gender)M	0.007784021	0.000164638	2.235366e+03	0.000000000

term	estimate	std.error	statistic	p.value

Along with each individual covariate, we also include an interaction term between segment and gender. It appears that in all three years, on average runners slowed down as the race progressed (as indicated by the positive coefficient for segment number). Although overall males tended to have lower segment paces relative to their overall times, the positive coefficient for the segment-by-gender interaction suggests that males tended to slow down more over the course of the marathon.

Conclusion and Ideas for Future Work

In this analysis, I evaluated the relationship between pacing profiles for a marathon and covariates gender, age, and proxies for ability/experience. While these results demonstrate that pacing profiles (whether defined by comparing the second half of a race with the first half or by the change in pace over 5km segments of the race) on average vary across age, gender, and ability, these analyses only begin to break the surface and bring forward more questions than they answer. For example, we see variability in pacing profiles across different years of the Boston marathon (appendix), but we have yet to explore how the average pacing profile varies across different courses. I have results from another marathon over the same period of time, but did not have time to figure out the best way to compare these datasets (the main hurdle being different intermediate split times between not only the two races, but differences across years in the same race).

Had this report not been put together for a course on estimating equations, I probably would have explored alternative modeling approaches for the Boston dataset. In particular, I'd be interested in seeing how a mixed-model approach would have fit the data (accounting for individuals having their own unique profiles), or maybe try to reproduce the bayesian approach of (Reese and Ward 2015). Another issue I didn't have time to tackle was missingness. In the appendix, I summarize the missingness in the intermediate splits for the Boston marathon results. It would be interesting to evaluate different impuation methods to try and fill in missing values, which would be similar to work done by (Hammerling et al. 2014), who sought to predict finishing times for individuals who were unable to complete the 2013 Boston Marathon due to the two bombs detonated near the finish line.

Future Work

Idea 1: Longitudinal Study of Pacing Profiles

A question of interest in regards to pace profiles is whether or not these change as a runner becomes more experienced, with the idea being that as a runner runs more races, his or her profile shifts from that having a slower half to one that is more evenly paced (if not running a faster second half). While we could try to use a proxy measure in a cross-sectional study like this (i.e., use placing in the top 10% as a measure of "ability/experience"), it could potentially be more informative to study the pacing profiles of a cohort of runners as they run their first, second, . . . , etc races.

Difficulties in performing such a study include:

1. First-time marathoners differ in their running histories (example: a former collegiate runner might run with a more even pacing profile than someone who picked up running recently with the sole focus of completing the distance).
2. It's highly unlikely that a large group of people will run the same sequence of races, or even the same number of races in a given window of time.
3. Acquiring the data is time consuming, and its not always possible to get the intermediate split times for past races (for example, the author was only able to acquire intermediate split times for 7 of the 10 marathons he's run).

Idea 2: Pacing Profiles of Other Distances (Notably: Half Marathon)

Another question of interest is how do pacing profiles of half marathon runners vary across age, gender, and experience? While only half the distance of a marathon, the half marathon is still long enough to require an informed pacing strategy to avoid hitting the proverbial wall. Along with the difficulties encountered in the above analyses of the full mararhon datasets, half marathons add the additional challenge in that not only do the intermediate splits differ across races, it's unusal to see a split at the half-way point (the author was unable to find any results of his that included a "quarter-marathon" split; most had some combination of 5km/10km/10mile).

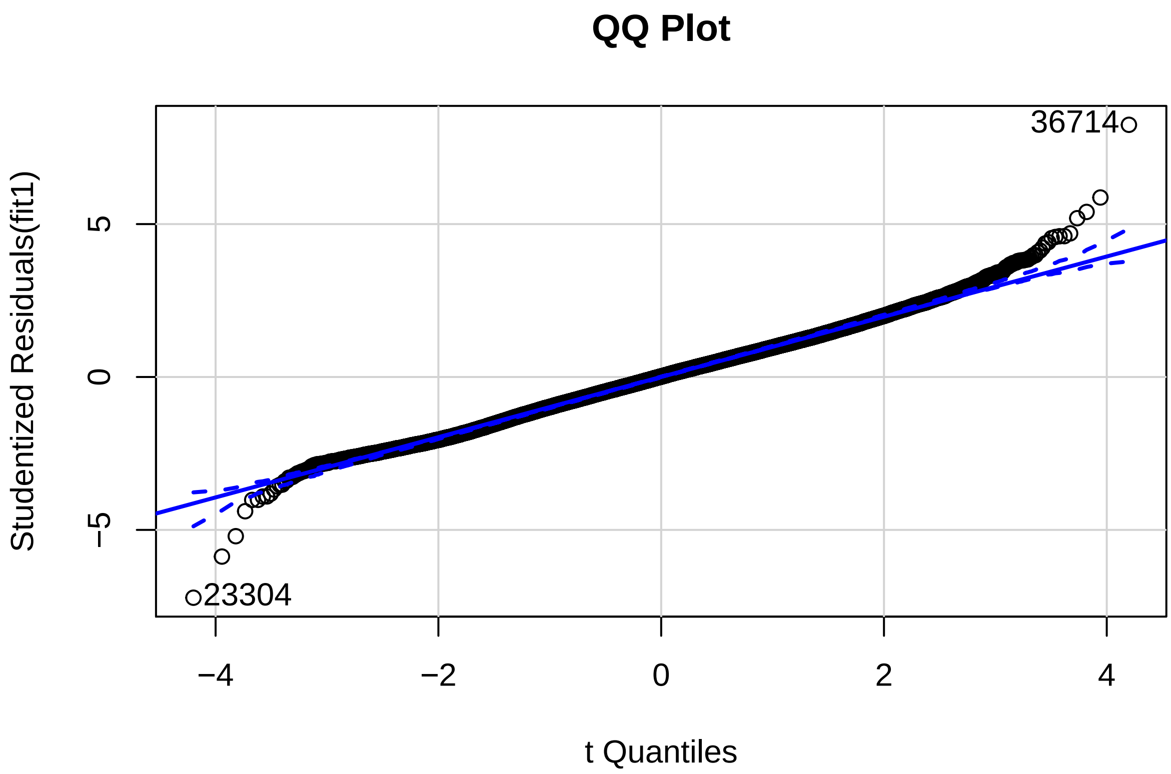
It might also be interesting to compare pacing profiles for individuals who run both half and full marathons. It would be challenging to acquire this dataset, and additional confounders might impact any such analysis (for example, someone training for a full marathon might include a half-marathon in their training, but might race it more conservatively than if the half-marathon had been their goal race).

Idea 3: Using Profiles in Rank Courses by Difficulty

Intuitively, a more difficult course might produce more drastic differences in pacing between the first and second half. Would a tougher course encourage more people to run a more modest pace in the first half, or would runners start at the same pace they'd use on an easier course and then subsequently slow down more over the second half? If the latter was the case, then the average pace profile could be a tool to use in measuring the difficulty of the course, and this would be easier to acquire data to compute (as opposed to trying to gather times for the same set of runners across different races and seeing how their pacing profile on the course of interest differed from the average pacing profile on other courses). It would also be of interest to start with a list of courses ranked under some criterion of difficulty and evaluate the differences of pace profiles across varying difficulties.

Appendix

QQ Plot for VAMT Linear Regression Model



```
## [1] 23304 36714
```

Missingness in the Boston Marathon Data

2015

	nMissing
5K	152
10K	31
15K	18
20K	29
Half	28
25K	31
30K	39
35K	51
40K	56

2016

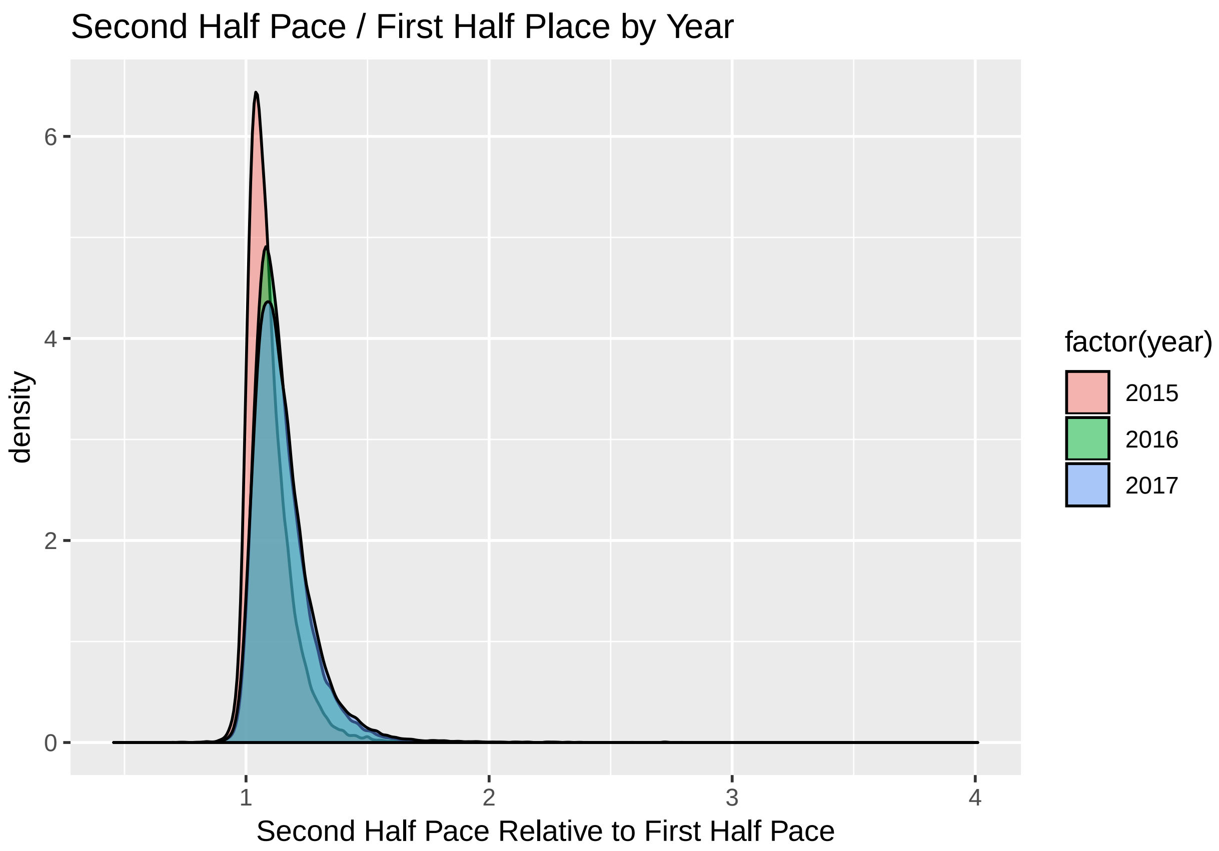
	nMissing
5K	52
10K	29
15K	14
20K	23
Half	17
25K	10
30K	24
35K	12
40K	14

2017

	nMissing
5K	25
10K	54
15K	19
20K	33
Half	17
25K	40
30K	25
35K	23
40K	6

Second Half over First Half Pace By Year in the Boston Marathon Data

As a quick check to see if pacing profiles varied by year in the Boston Marathon data, I computed the pace for the first and second half of the race for each individual, and the ratio of the two (second over the first). While the historgram of the ratios look similar across all three years, note that the average in 2015 is less than both 2016 and 2017. A simple one-way ANOVA test indicates that the average ratio is not the same across all three years, and it would appear that this result is driven by the lower values seen in 2015.



ANOVA: First Half Pace / Second Half Pace ~ Year

term	df	sumsq	meansq	statistic	p.value
factor(year)	2	54.95768	27.4788412	2091.768	0
Residuals	79573	1045.32319	0.0131367	NA	NA

Simple Linear Model: First Half Pace / Second Half Pace ~ Year

Note the similar values of the covariates for 2015 and 2016.

term	estimate	std.error	statistic	p.value
(Intercept)	1.0941802	0.0007031	1556.11690	0
factor(year)2016	0.0516191	0.0009940	51.93076	0
factor(year)2017	0.0591055	0.0009961	59.33890	0

References

Angus, Simon D. 2013. “Did Recent World Record Marathon Runners Employ Optimal Pacing Strategies?” *Journal of Sports Sciences* 32 (1): 31–45. doi:10.1080/02640414.2013.803592.

Hammerling, Dorit, Matthew Cefalu, Jessi Cisewski, Francesca Dominici, Giovanni Parmigiani, Charles Paulson, and Richard L. Smith. 2014. “Completing the Results of the 2013 Boston Marathon.” *PLoS ONE* 9

(4). doi:10.1371/journal.pone.0093800.

Hausswirth, C, A. Bigard, and C. Guezennec. 1997. “Relationships Between Running Mechanics and Energy Cost of Running at the End of a Triathlon and a Marathon.” *International Journal of Sports Medicine* 18 (05): 330–39. doi:10.1055/s-2007-972642.

Jeukendrup, Asker E. 2011. “Nutrition for Endurance Sports: Marathon, Triathlon, and Road Cycling.” *Journal of Sports Sciences* 29 (sup1). doi:10.1080/02640414.2011.610348.

Reese, C. Shane, and Jared Ward. 2015. “Analyzing Split Times for Runners in the 2013 St. George Marathon.” http://www.runblogrun.com/2017/04/17/Jared Ward Thesis.pdf.