Is a sub 2 hour marathon in the near future? Modeling rare events in sports.

Rodney X. Sturdivant, Ph.D., Baylor University and Nick Clark, Ph.D., West Point

Outline

- Baseball Rare Events (if needed only)
- Background
- Marathon Data
- Simple Model
- Self-Exciting Model
- Further Research
- SCORE

Background

The New York Times

Records Fell at the Track Worlds. A Trend? Not So Fast.

Some have referred to this era as a golden age of better and better times. But a deeper look at the data shows the simple shorthand conclusion is incomplete.

Golden Age?

Are we living in a time of records?

- Idea: seems like an increase in records falling, but is it just the nature of randomness?

How can we address this question?

What would randomness look like?

Rod Aloha 10K Run (San Diego, 2018), 2nd Age Group

Rod Last Marathon (LA, 2018), 1st, Glendora CA Runners

Marathon World Record Data

Men's Marathon world records since 1908

- 50 total
- First 5

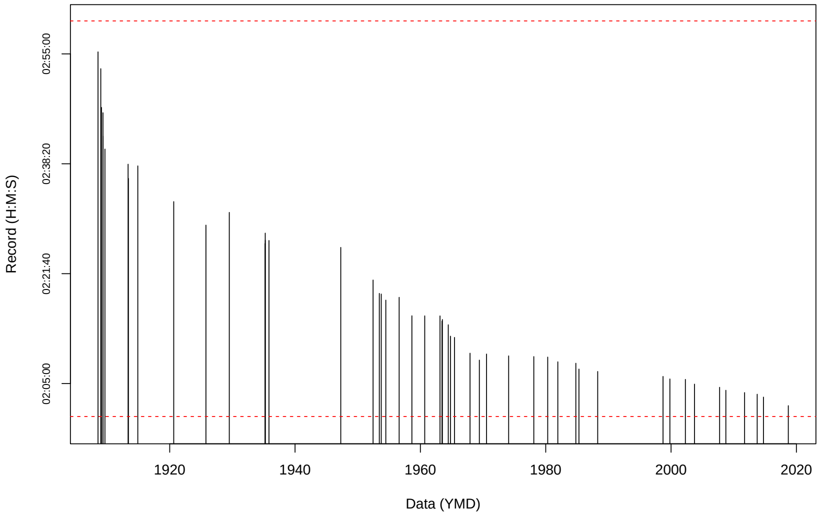
Time	Name	Nationality	Date	Event/Place
2:55:18.4	Johnny Hayes	United States	July 24, 1908	London, United Kingdom
2:52:45.4	Robert Fowler	United States	January 1, 1909	Yonkers,[nb 5]United States
2:46:52.8	James Clark	United States	February 12, 1909	New York City, United States
2:46:04.6	Albert Raines	United States	May 8, 1909	New York City, United States
2:42:31.0	Henry Barrett	United Kingdom	May 8, 1909[nb 6]	Polytechnic Marathon, London, U

- Last 5

Time	Name	Nationality	Date	Event/Place
2:03:59	Haile Gebrselassie	Ethiopia	September 28, 2008	Berlin Marathon
2:03:38	Patrick Makau	Kenya	September 25, 2011	Berlin Marathon
2:03:23	Wilson Kipsang	Kenya	September 29, 2013	Berlin Marathon
2:02:57	Dennis Kimetto	Kenya	September 28, 2014	Berlin Marathon
2:01:39	Eliud Kipchoge	Kenya	September 16, 2018	Berlin Marathon

Visualizing the data

- Two and three hour times shown as horizontal lines
- 2 hour marathon pace: 4:35 per mile
- 3 hour pace: 6:52 per mile



SOME MARATHON RECORD HOLDER STORIES/INFO

MAYBE INCLUDE A COUPLE OF PICTURES OF PEOPLE

ADD SOME SUMMARY STUFF - TRIVIA: COUNTRIES, LOCATIONS OF MARATHON ETC

SIMPLE MODEL

POISSON PROCESS

A model for a series of discrete events where the average time between events is known, but the exact timing of events is "random" meeting the following criteria:

- Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.
- The average rate (events per time period) is constant.
- Two events cannot occur at the same time.

Poisson Process Interarrival Times

The time between events (known as the interarrival times) follow an exponential distribution defined as:

$$P(T > t) = e^{-\lambda t}$$

- T is the random variable of the time until the next event
- t is a specific time for the next event
- λ is the rate: the average number of events per unit of time.

Note the possible values of T are greater than 0 (positive only).

Reasonableness of Exponential Interarrivals

The exponential distribution has certain attributes, for example:

$$E(T) = SD(T) = 1/\lambda$$

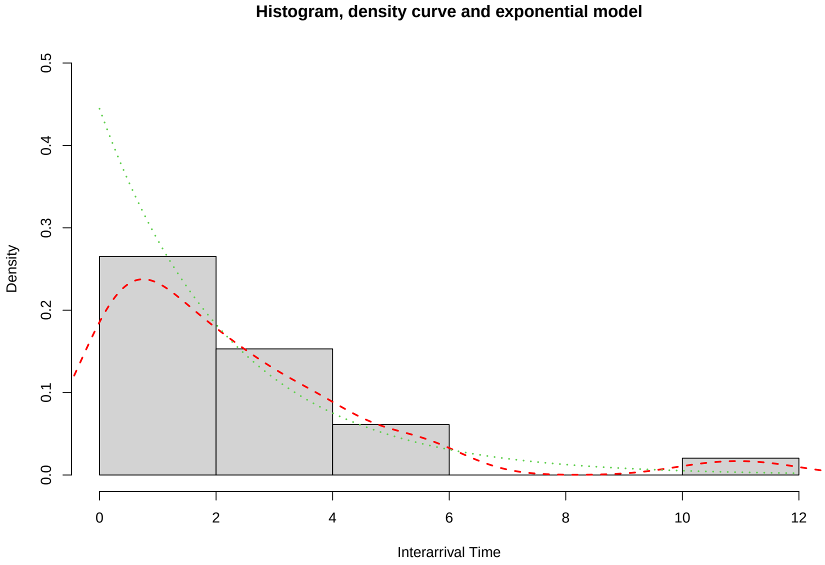
For the time between record data:

- Mean: 2.25 years
- SD: 2.43 years

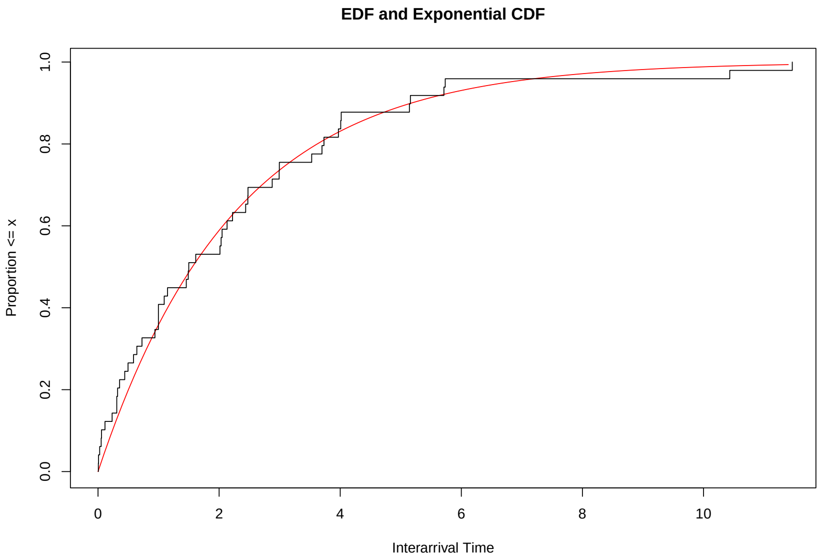
Reasonable... slightly "overdispersed"

Estimating the Model

We estimate (MLE) $\lambda = 1/E(T) = 0.445$



Graphical Assessment of Fit



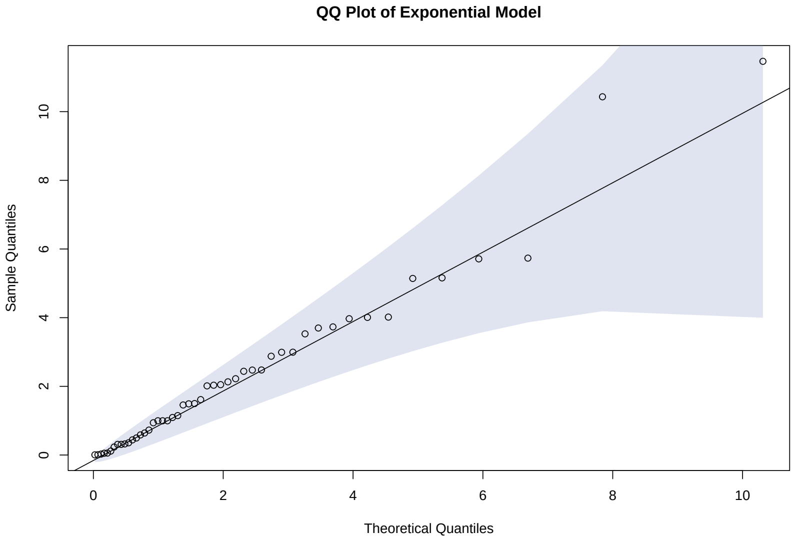
Testing Model Fit

Goodness of fit tests

- Kolmogorov-Smirnov: $p = 0.926$
- Cramer-Von Mises: $p = 0.855$
- Anderson-Darling: $p = 0.563$

All fail to reject the null hypothesis of model fit

Are records then random?



/2022−09−09

What are the poorly fit points?

- 11.5 year gap

Time	Name	Nationality	Date	Event/Place
2:26:42	Sohn Kee-chung	Japanese Korea	November 3, 1935	Tokyo, Japan
2:25:39	Suh Yun-bok	Korea	April 19, 1947	Boston Marathon

- 10.4 year gap

Time	Name	Nationality	Date	Event/Place
2:06:50	Belayneh Dinsamo	Ethiopia	April 17, 1988	Rotterdam Marathon
2:06:05	Ronaldo da Costa	Brazil	September 20, 1998	Berlin Marathon

A “Self-Exciting” Model

Self-Exciting Point Processes

- Events “trigger” more events
- Examples of use include earthquakes, crime waves

Hawkes Processes

- Let H_t be the history of events up to time t. The Hawkes (1971) model of the conditional intensity is:

$$\lambda(t|H_t) = \nu + \sum_{i:t_i<t} g(t - t_i)$$

where ν is the background rate of events and g is the “triggering function”.

Exponential Triggering Function

- The "triggering" function can be further decomposed:

$$g = \mu g^*$$

where g^* is a density function known as the "reproduction kernel" and μ is known as the "reproduction" mean.

- A common choice for the "reproduction kernel" is the exponential density given by:

$$g^*(t) = \beta e^{-\beta t}$$

Fitting the model

Parameter estimates for marathon data (exponential) Hawkes process, using MLE:

- baseline intensity 0.382
- reproduction mean 0.129
- exponential reproduction function rate 3.583

Note the baseline intensity is slightly lower than the constant model rate estimate of 0.445

The estimated reproduction function is then:

$$\begin{aligned} g(t) &= \mu g^*(t) = \mu \beta e^{-\beta t} \\ &= 0.13 * 3.58 e^{-3.58t} \end{aligned}$$

Model implications

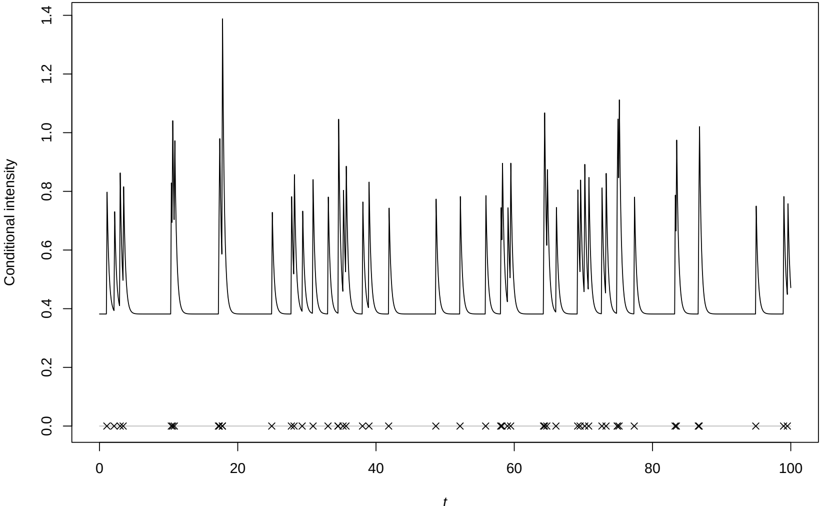
At the instant of the first event (world record), $t = t_1$ so $g(t - t_1 = 0)$ and the reproduction rate is:

$$g(0) = 0.13 * 3.58e^{-3.58 0} = 0.13 * 3.58 = 0.462$$

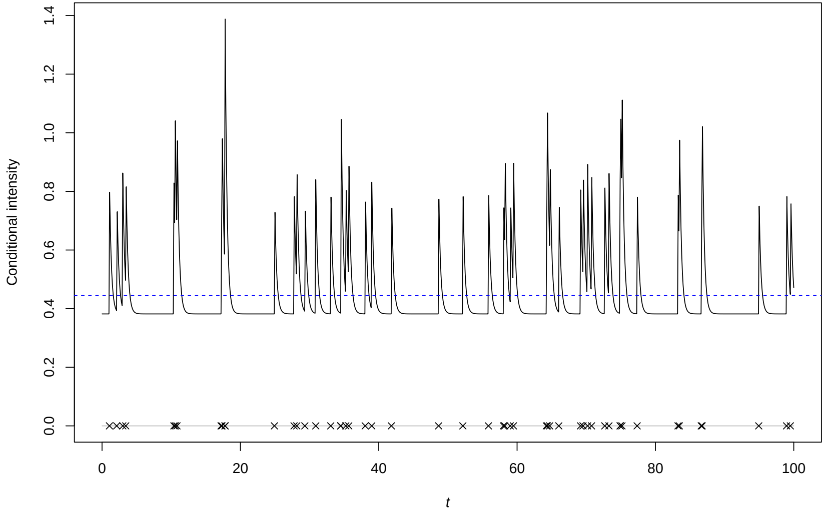
- The rate increases from the baseline rate of 0.382 by this amount at the moment of this occurrence
- The rate then decays back to baseline over time (unless a new event occurs).
- Each new event “excites” the rate to increase and then decay

The Intensity Function over Time

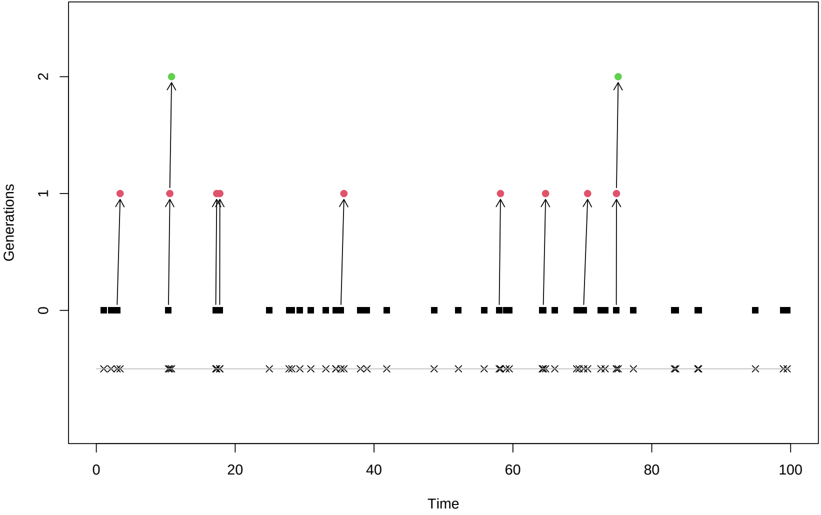
Below is based on a simulation of the intensity function over a 100 year period. NOTE: HERE WOULD BE NICE TO SHOW FOR OUR DATA ALTHOUGH MIGHT NOT GIVE THE FULL PICTURE ANYWAY



Intensity compared to the constant rate model



Process as “Generations”

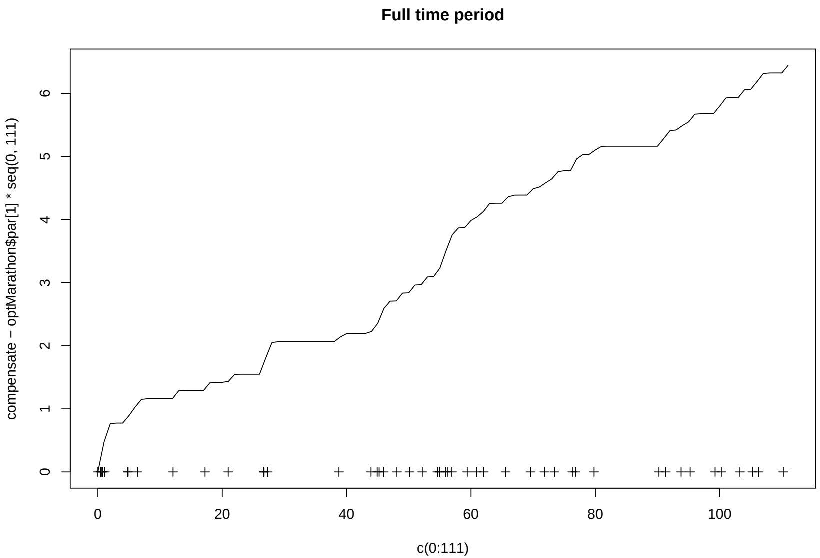


The Compensator Function

NEED TO WORK ON EXPLAINING - BELOW IS THE VERSION TAKING OUT BASELINE. . . MAYBE START WITH CONSTANT (CAN DO Poisson MODEL AND THEN THE BASELINE RATE HERE)

“Residuals”

Compensator Function for Fitted Hawkes Model



Residuals for Fitted Hawkes Model

Possibly Hession/Covariance Uncertainty in parameter estimates

Comparing Models - Residuals

Comparing Models - Overdispersion

Conclusions

Further Work. . .

SCORE

Here several slides - the overall project - the module for this work
(at whatever point we can get it. . .)

References

Data source: Wikipedia (https://en.wikipedia.org/wiki/Marathon_world_record_progression) scraped August 12, 2022

Poisson process: https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

Hawkes, Alan G. 1971. "Spectra of Some Self-Exciting and Mutually Exciting Point Processes." Biometrika 58 (1): 83–90. https://doi.org/10.2307/2334319.

"Hawkesbow" package…