Bayesian Hierarchical Modeling for Equitable Marathon Qualifying Standards

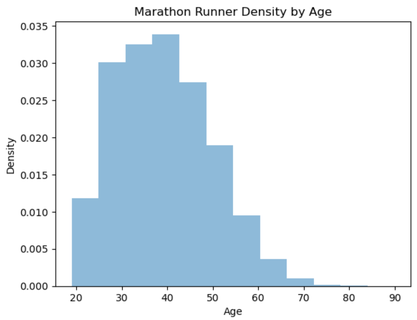
Spencer Siegel
siegels@stanford.edu

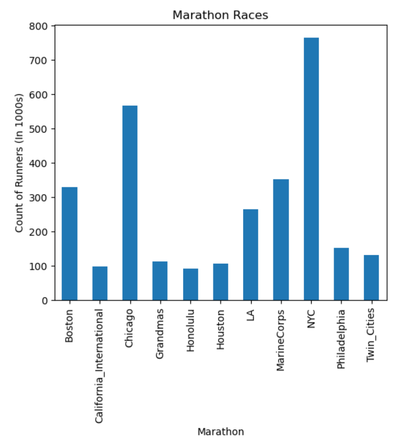
Introduction:

The Boston Marathon qualifying standards have received quite a bit of scrutiny over the years regarding fairness. The current standards are somewhat arbitrary which has frustrated various runners who also may feel that performances at certain courses/conditions should be viewed differently. This study extends my prior research by fitting a Bayesian Hierarchical Model (BHM) to predict marathon performance from 11 of the most popular marathons in the United States. Using features such as the individual runner, marathon location, marathon year, and a cubic spline on age, I have fit a Bayesian Hierarchical Model to predict marathon finish time and generate gender-aging curves. I have found that my model specification for this hierarchical model has outperformed Ridge and CatBoost regression.

Data:

Researchers at Colorado School of Mines, under the supervision of Dr. Dorit Hammerling, have scraped marathon data from 2000-2019 for 11 of the major United States marathons (Boston, Chicago, California International, Grandmas, Honolulu, Houston, Los Angeles, Marine Corps, New York, Philadelphia, Twin Cities). This data includes runner name, runner age (as an integer), marathon location, and marathon year. I had to perform extensive data work with roughly 3M observations due to the absence of unique runner IDs.





Features:

The features I used for this modelling task were age, global runner id (categorical), marathon location (categorical), and marathon location-year (categorical interaction). For age, I utilized a cubic spline with 5 degrees of freedom to allow for a flexible, nonlinear relationship between age and marathon finish time. I derived global runner id from my data processing step. I included marathon location as a feature to address course difficulty and marathon location-year to address race conditions.

References:

[1] G David Garson. Hierarchical linear modeling: Guide and applications. Sage, 2013.
[2] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. Journal of Machine Learning Research, 2013.
[3] Carl Munkby. Evaluation of probabilistic programming frameworks, 2022.

Models:

Ridge Regression:

$$T_{ijk} = \theta_0 + \sum_{l=1}^{5} S(a_{ijk}; 5)_l \, \theta_l + M_{jk} + \lambda \|\theta\|_2^2$$

Where T_{ijk} in the finish time in minutes for runner i in race j in year k
θ_0 is the intercept
$S(a_{ijk}; 5)$ is the cubic spline on age with 5 degrees of freedom.
λ is the L2 regularization hyperparameter
*Note that ridge regression is our baseline and does NOT use the individual runner effect

CatBoost Regression:

CatBoost is a gradient boosting decision tree algorithm that I used (with squared error as the lost function). The algorithm sequentially constructs decision trees as the "weak" learner. For each tree, it calculates the gradient of the loss function with respect to the model parameters and moves in the direction of the negative gradient for the next tree (additive model). CatBoost handles categorical features (like global runner id in this use case) through target encoding.

Bayesian Hierarchical Linear Model (BHM):

$$T\_\{ijk\} \sim TruncatedNormal(\theta_0 + \sum_{\{l=1\}}^{5} S(a_{\{ijk\}}; 5)_l \theta_l + \epsilon_i + \epsilon_j + \epsilon_{k|j}, \sigma^2)$$

Where T_{ijk}, θ_0, $S(a_{ijk}; 5)$ have the same interpretation as in the ridge setting.
$\epsilon_i \sim N(0, \sigma_r)$ is the random effect intercept on runner i
$\epsilon_j \sim N(0, \sigma_m)$ is the random effect intercept on marathon j
$\epsilon_{k|j} \sim N(0, \sigma_y)$ is the random effect on year k nested within marathon j

This is a very flexible linear modeling technique that allows for both fixed and random effects to be used in a Bayesian setting [1]. Random effects are often used on categorical variables in which one is trying to control and where the categories in the training data represent a subset of the total population. Furthermore, using a random effect is a method to account for non-independent data. This model was fitted using numpyro, a probabilistic programming language which uses a JAX backend [2]. It utilizes stochastic variational inference for converging, an optimization technique which iterates between subsampling the data and adjusting the hidden structure based on the subsample [3].

Experiments:

I split the data into a train (1.43M obs.), validation (178,915 obs.), and test set (178,915 obs.). After convergence of BHM, I generate samples of the full posterior of model parameters and use these in prediction on the test set. For each model, I create aging curves by isolating the age variable and predicting on runners between 18-85. I then used these aging curves to construct new qualifying standards (The one's from the BHM are shown on the right). I started by setting the 18-34 age group time equal to the Boston Marathon Qualifying (BQ) standards and then incremented for each age group based on the shape of the age curve for the given model.

Results:

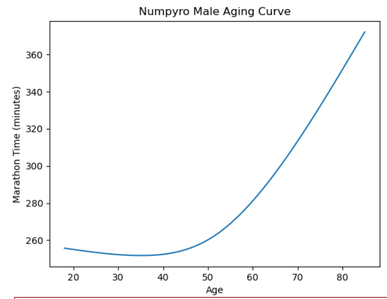
The BHM outperforms both Ridge and CatBoost on the test set by a significant margin. Note that the big train/test discrepancies for CatBoost and BHM are due to inclusion of the individual runner ids.

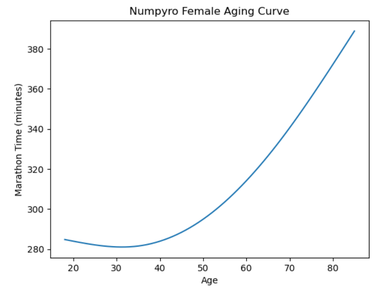
Male Data

	Train RMSE	Train MAE	Test RMSE	Test MAE
Ridge	45.564	36.681	45.648	36.745
CatBoost	29.223	22.439	38.816	30.283
BHM	19.583	14.621	36.774	27.613

Female Data

	Train RMSE	Train MAE	Test RMSE	Test MAE
Ridge	42.865	34.665	42.752	34.584
CatBoost	27.079	21.114	36.532	28.763
BHM	17.219	13.036	34.786	26.636





Discussion:

A Bayesian Hierarchical Modeling approach to predicting marathon performance appears promising. The performance gap between CatBoost and BHM indicates that the linearity assumption between the features and response is appropriate and that the hierarchical structure is effective. Furthermore, the results suggest that race location and year should be considered when evaluating one's performances. The BHM coefficients suggest Honolulu marathon is over 17 minutes slower than an average course (Boston 2004 race conditions had a similar result). Also, the results suggest that the current standards are too strict on young runners relative to middle-aged runners in particular.

Boston Marathon Qualifying (BQ) Standards vs BHM Recommended:

Age Group	18-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80+
Male BQ	3:00	3:05	3:10	3:20	3:25	3:35	3:50	4:05	4:20	4:35	4:50
Male BHM	3:00	2:59	3:00	3:03	3:09	3:20	3:37	3:49	4:07	4:27	4:49
Female BQ	3:30	3:35	3:40	3:50	3:55	4:05	4:20	4:35	4:50	5:05	5:20
Female BHM	3:30	3:30	3:33	3:38	3:45	3:55	4:06	4:19	4:34	4:50	5:08

Future Work :

I would love to gather more data particularly on older runners to have stronger confidence on the right tail of the aging curve. I would also love to model the survivorship of marathon runners because many of the slower runners will likely self-select out once they are too old. Lastly, I would like to test different hierarchical specifications including fitting a separate standard deviation for each runner that is dependent on age.