Statistical Analysis on Physiological Predictors for Running Performance

Juliann Zhou | kyz224@nyu.edu

Abstract

This project will examine the following physiological variables and their correlations with endurance running performance:

- VO2max: the maximum rate of consumption attainable during physical exertion,
- vVO2max: the minimum speed for reaching maximal oxygen uptake in a constant rate exercise)
- %VO2max at lactate threshold (percent VO2max at the exercise intensity at which the blood concentration of lactate increases rapidly)
- running economy (the volume of oxygen consumed per kg of body weight per km).

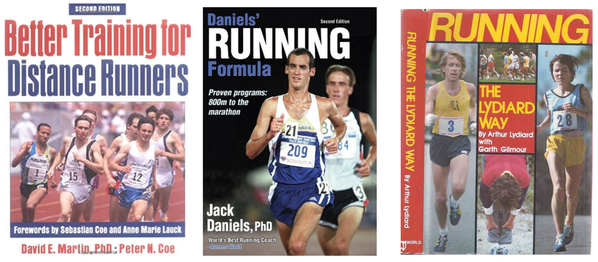

Image: Most influential training books for endurance running with different takes on scientific basis for good training strategies

Dataset Description

For this project, I am using a dataset from the study "Similarities and differences among half marathon runners according to their performance level" by Ana Ogueta-Alday et. al.. It is the appropriate dataset for this project because

1. It has comprehensive records of physiological variables and running performance for 48 subjects (all subjects are 20-50 years-old caucasian males who have a better time than 105 min in the half-marathon).
2. In the original research, the physiological variables are professionally measured through an incremental treadmill test and a submaximal test in the original publication.
3. A Kolmogorov-Smirnov test was also applied to ensure a Gaussian distribution of all results.

Problem Description and Context

The most prominent training strategies for endurance running are based on these physiological variables. Notably, Jack Daniels proposes training at varying speeds/distances based on VO2max. While other strategies, such as Lydiard's, emphasize running high mileage and consistent speed to optimize the improvement in running economy. By analyzing the correlation of physiological variables to running performance, we are able to evaluate different training strategies

Methodology

Overview

For statistical analysis of the dataset, we provide a data summary that includes the mean (±sd), variance, and Pearson correlation coefficient (r) for each physiological variable with running performance. Then, we apply a simple linear and LASSO regression for the variables

Explanation of LASSO regression

I will provide an explanation for the technique and compare it with simple linear regression. In simple linear regression, we minimize the square of the residuals to find a line that generates the minimum sum of squared errors. LASSO (least absolute shrinkage and selection operator) regression adds a penalty equal to the absolute value of the magnitude of the coefficient, thus penalizing less important features of a dataset. As a result, it prevents overfitting models like simple linear regression for a dataset with high variance or small training data and achieves a better prediction. The cost function for LASSO is:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

Cost function of LASSO regression

Here, βj represents the coefficient estimates for different independent variables and describes the weights or magnitude attached to each feature respectively. According to the above equation, the penalty term (the product of λ and the sum of $|\beta j|$) is added to residual sum of squares to regularize the coefficients of the model.

Rationale for model choosing

We use a linear regression model because we hope to find the physiological variables as predictors for the target variable running performance. The dataset satisfies the three prerequisites for applying linear regression:

1. The variance for variables in the dataset is constant,
2. The observations are independent
3. The Kolmogorov-Smirnov test from the original publication confirms normal distributions for variables.

We use LASSO as a regularization technique for linear regression because the sample data exhibits high variance and low bias. By applying LASSO, we minimize the problem of overfitting data. A study by Takayama et al. shows that the physiological variables explain 95% of the variation in running performance for 16 km. Because we have a small set of significant variables that predicts the target variable, we choose LASSO regression over similar regularization techniques such as Ridge regression.

Related Work and Literature

"Test of the classic model for predicting endurance running performance" by James E McLaughlin et al., looks at physiological variables as predictors for running performance with SPSS stepwise analysis, "Relationship between Classic Physiological Variables and Running Performance in Recreational Runners" by Fuminori Takayama et al, conducts single regression analysis on these variables, Among these variables, vVO2max was shown to be the most highly correlated with running performance.

Results

Data Summary: Physiological Variables and Correlation with Running Performance

	Mean (± SD)	Variance	Correlation with Running Performance
VO2MAX	61.5 ± 7.4	55.3	0.75
RE	206.1 ± 17.4	301.2	0.41
%VO2MAX_VT	60.6 ± 6.6	44.0	0.11
%VO2MAX_RCT	87.6 ± 5.1	26.4	0.33
VVO2MAX	19.7 ± 1.9	3.58	0.93

Coefficient estimates: linear vs. LASSO

	Coefficient Estimate
VVO2MAX	27.1
RE	6.47
%VO2MAX_VT	1.36
%VO2MAX_RCT	1.36
VVO2MAX	218.0

	Coefficient Estimate
VVO2MAX	25.8
RE	6.11
%VO2MAX_VT	2.20
%VO2MAX_RCT	2.20
VVO2MAX	230.0

RE measures running economy, %VO2MAX_VT and %VO2MAX_RCT measure %VO2max at ventilatory threshold and respiratory consumption threshold respectively (which are both estimates for lactate threshold), VVO2MAX measures peak speed (speed at VO2max)

For the data summary, we calculate the bivariate Pearson correlation for each variable with running performance. Among these, vVO2max is shown to have the highest correlation. Using python scikit learn, we partition the dataset into training and test data to conduct linear and LASSO regression with λ=0.3 (further increase in λ see no significant increase in r-squared). Then, using test data, we find the models' mean square errors and r-squared. For simple linear regression, the MSE is 79465 and r-squared is 0.78, For LASSO regression, the MSE is 52863 and r-squared is 0.89. As expected, LASSO regression reduces the MSE and increases the r-squared. Both models confirm that vVO2max is most closely correlated with running performance followed by VO2max.

In conclusion, because vVO2max has the highest correlation with running performance. Training methods that optimize the improvement in vVO2max will be the most effective in improving running performance.