ISyE 7400 Project Report

Complex Modeling Techniques in Applied Kinesiology

Christopher Shartrand

Professor Roshan Vengazhiyil

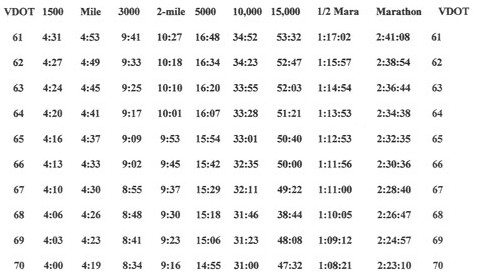
Background / Motivation: During a twenty year period from the mid 1950’s to the mid 1970’s, American distance running achieved new heights in global success at the Olympic level due to the accomplishments of athletes such as Jim Ryun, Frank Shorter, Dave Wottle, and Steve Prefontaine. From their successes, these athletes quickly became household names throughout the nation and are often credited with aiding in the United States running boom of the 1970’s. However large the rise in American distance running, the fall was equally meteoric. By the late 1970’s American distance running had fallen back into mediocrity, which would continue well into the 1990’s.

The rapid demise of quality American distance runners prompted a number of alarming questions, the most important of which was seen as, “How could we have prevented this from happening?” This line of thinking led to the conclusion that the best coaches and experts in the United States truly did not understand the requisite physiology and training for producing high caliber athletes. Beginning in the late 1970’s, kinesiology researcher Dr. Jack Daniels strove to fill in these knowledge gaps through meticulous experimentation. After nearly 20 years of research, he published his findings in his 1998 book, Daniels’ Running Formula.

In his book, he details that the necessary base of properly training endurance running revolves around the understanding and improvement of an athletes’ , or the maximum rate at which an athletes’ body is able to consume oxygen when measured during incremental exercise. As is a measure of the body’s aerobic physical fitness, it is very closely linked with endurance running performance. Determining someone’s true is expensive and time consuming, as it requires specialized equipment during which the incremental exercise is monitored in a laboratory setting (The most common example is measuring using a metabolic cart while increasing the intensity on a treadmill until the individual can no longer sustain the exercise). As such, Daniels developed a simplified method to approximate using a pseudo metric he called a ­. In accordance with the need to be applicable to distance running, the can be calculated by an individual’s run time over a measured distance. The formulaic simplification is seen below in **Equations :**

where is measured in seconds and is measured in meters. Following twenty years of testing and confirmation of these formulas using elite athlete test subjects, he tabulated the results for quick use by the everyday coach. A section of this table can be seen in **Table 1** below:

**Table 1.**

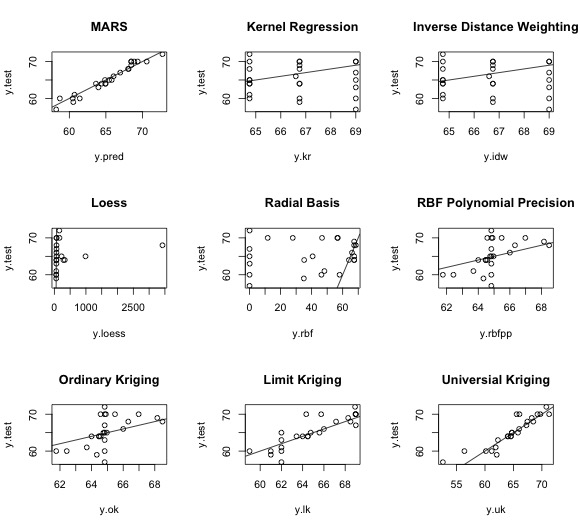


While the value of the is undoubtable as it has helped to aid in the 21st century resurgence of American distance running, it does have several clear drawbacks. First, the values were tested and confirmed using elite runners, whose body capabilities are very much different from the everyday high school runner and even most collegiate athletes. That is, it is unable to capture the individuality and variance of each and every athlete (For example, the personal best times of the author correspond to values of 69, 67, 69, and 63 for the mile, 3000m, 5000m and half-marathon respectively). Furthermore, it is incapable of accounting for the coaching factor. That is, each coaching philosophy values training systems that possibly benefit the running of better times at different distances. Finally, it is extremely “expensive” to conduct multiple tests for an athletes’ , when the goal is training to improve performance. That is, forcing an athlete to run multiple races in a short period of time to find their over multiple distances can greatly hamper important training that will lead to performance improvement. Therefore the question arises, “Is it possible to use advanced numerical and statistical methods to more accurately predict an athletes’ performance based on a single observation when compared to the basic tables?”

The Data: Unlike the examples presented in class, the applied analysis of kinesiology in this setting requires different necessities for the training and testing datasets. As previously stated in the motivation section, the dataset must be collected in such a way that both the individuality factor and the coaching factor can be captured in the modeling estimation in order to improve prediction of the athlete’s performances. Furthermore the training and testing datasets must have some amount of time spacing so that the modeling estimation is not confounded by short-term effects (injuries, seasonal weather conditions, differences between indoor and outdoor track facilities, etc). Therefore the data needs to be collected in a pseudo-longitudinal way such that the same set of athletes and their corresponding times are tracked over a multiple year period, under the same coaching techniques and similar short-term effects. The data collected for the training set was taken from observed race performances of ten athletes from the Fredonia State University Men’s Track and Field team from the 2013-2014 season. The testing dataset was collected from the same ten athletes based on their performances during the 2014-2015 track and field season. The athletes had the same coach, seasonal weather conditions, and nearly identical competition facilities between the two seasons in an attempt to capture the necessary factors described in the motivation section and control for any outside factors that could effect model estimation.

It is important to make note of the issues regarding space-filling designs for use in this specific analysis. Foremost, a space-filling design used for the testing dataset would lack the key components of athlete individuality and coaching techniques therefore making the “testing” of effectiveness of model estimations less accurate than testing dataset used in this study. Additionally, if a space-filling design were to be considered, it would require a number of complex constraints in order for it to be used. That is, consider in the case of this study, a scaled dataset in . In the scope of observed race distances, this study uses the meter runs. These distances would have to be held constant under the scaled range of . Under this condition, some sort of Sliced Latin Hypercube design would be required for the variation of . However this would lead to arbitrary and nonsensical “observations” in the testing dataset. Because would need to be scaled using the shortest observed time and longest observed time, an “observation” of for could indicate running a meter in roughly 16 minutes, a performance better described as “walking” a meter race. Hence, more complex constraints would be required to make the SLHD workable. For this reason and the reasons detailed previously, a space-filling design was not used for the testing of fitted model estimation in this study.

Results: Model estimation using the training dataset was conducted using nine different methods. These included six numerical methods of Multivariate Adaptive Regression Splines, Kernel Regression, Inverse Distance Weighting, LOESS Regression, Radial Basis, and Radial Basis with Polynomial Precision. The three statistical methods used were Ordinary Kriging, Limit Kriging and Universal Kriging. The models were assessed both visually using fit plots of the estimation versus the observed values and numerically by the . The fit plots of the models can be seen in **Figure 1** below:

**Figure 1.**

Numerous observations can be made from the nine different plots. Foremost, both MARS and Universal Kriging appear to have achieved the most accurate estimation based on their limited dispersion along the theoretical perfect fit line. Evidence that the fit between Kernel Regression and Inverse Distance Weighting is poor can be seen immediately by the grouping of its predictions. Upon further investigation, the three groups of predictions that the modeling techniques have created are based solely on the distances run, meters. This leads to high and renders the prediction ability of these models useless for obvious reasons. The LOESS estimation initially looks promising until the observation of the extreme outlying predictions. Additionally the LOESS method was completely incapable of estimating three other performances. Singularities in the dataset matrix caused these issues and as such, eliminate it from consideration for the best model. The remaining techniques used each have their own observable issues in prediction with no specific reason as to why, dissimilar from what has been previously covered. Radial Basis with Polynomial Precision and Ordinary Kriging was used to confirm proper modeling and R coding. The two techniques are identical and as such, yielded the same results. Numerical comparisons of nine methods based on can be observed in **Table 2** below:

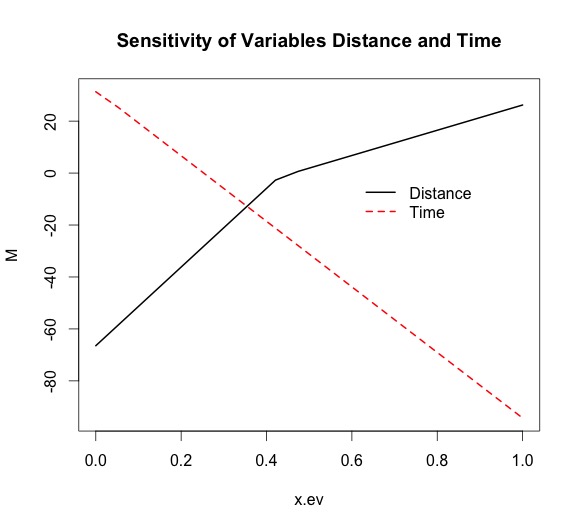
**Table 2.**

|  |  |
| --- | --- |
| Modeling Method |  |
| MARS | 0.8777997 |
| Kernel Regression | 4.734648 |
| Inverse Distance Weighting | 4.734648 |
| LOESS Regression | N/A1 |
| Radial Basis | 36.41985 |
| Radial Basis with Polynomial Precision | 3.492863 |
| Ordinary Kriging | 3.492863 |
| Limit Kriging | 2.148368 |
| Universal Kriging | 1.857107 |

1Failed prediction of three observations due to singularity issues

Based on these two measures of model performance, the MARS model was chosen for the remaining analysis and prediction in this study.

Next, it was important to investigate the sensitivity of the two predictor variables and . By intuition alone, it should be clear that both variables are significant as it is necessary to have both in order to estimate your using the . Regardless, it is prudent to confirm this numerically by the sensitivity of the main effects of the two variables. Computation of these main effects was plotted for visualization and can be seen in **Figure 2** below:

**Figure 2.**

Therefore, our intuition is confirmed, as both and are significant factors in the prediction of .

Returning to the objective of the study, it was desired to discover a prediction method for an athlete’s performance from a single observation that is more accurate than the simple tables detailed by Jack Daniels’ in his book. Using the predicted values from the MARS model, I broke up the observations in the testing dataset among the three distances of meters. From there I created three simple linear regression models by regressing the observed time run at that distance against the predicted values. Next, using the formula, I calculated a value from a new observation (time over one of the race distances), outside of both the training and testing datasets. Then using the inverse.predict() function from the library ChemCal, I used that value to predict the athlete’s time for the other two race distances that were not observed.

For example, during the 2015-2016 season, one of the ten athletes used in the datasets was observed running a m in 8:36. This time over the distance corresponded to a value of 71. Using inverse prediction for both the m and m predicted times of 3:48 with a confidence interval of [3:44,3:52] and 14:28 with a confidence interval of [14:28,14:28] due to a small standard error. According to Daniels’ table values, the predictions based on a of 71 would be 3:57 for the m and 14:44 for the 5000m. The athlete was later observed to have run 4:02 for the 1500m and 14:32 for the 5000m. Similar prediction for three more athletes has been detailed in **Table 3** below:

**Table 3.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Athlete | Observed | Model Prediction 1 | Model Prediction 2 | Daniels Prediction 1 | Daniels Prediction 2 | Observed Time  1 | Observed  Time  2 |
| 1 | 68 | 8:51 | 15:13 | 8:48 | 15:18 | 8:50 | N/A |
| 2 | 65 | 4:18 | 9:10 | 4:16 | 9:09 | 4:17 | 9:15 |
| 3 | 67 | 4:07 | 8:58 | 4:10 | 8:55 | 4:08 | 8:59 |

Conclusion / Future Work: Overall from the individual case studies, it appears that the fitted MARS model has slightly better overall predictions of race performance. However it must be noted that in extreme cases, such as the athlete described in the above paragraph, the predictions for m can be drastically overestimated. I believe this to be due, in part, to the data collected. In general with the three distances, a performance at m is more reflective of an athletes ability to perform at m than at m. While the net difference in distance is less between m and m, it is exactly double the race distance, leading to a much different style of racing when compared to the m. In essence, because a m performance is more correlated to a m performance, we observe higher uncertainty in our prediction at m.

It is also pertinent to address potential criticism of the model’s predictive capabilities. Overall, the improvements are small. In the case of Athlete 1 in **Table 3**, the model prediction for m is a 1 second improvement over that of the Daniels’ system. However small these may seem, their effects can often be drastic in training for long distance running. It is often the case that movement between training tiers can be determined by a matter of 1-2 seconds in performance. In such a cases, an incorrect prediction could lead to assigning an athlete to a tier too strenuous or too lax for their current fitness. While these training intensities also only differ by 1-2 seconds, their effect can be quickly compounded. Most long distance runners run anywhere from 80 to 100 miles per week for a total of anywhere from 3500 to 4500 miles per year. Therefore being off in training intensities of 1-2 seconds can lead to ranges of 3500 to 9000 seconds off what training should be. The result is that an athlete could be training too hard, leading to “over-training” stress and/or injury and decreased performance, or training too easily, preventing said athlete from reaching their optimal performance potential.

In terms of future work for this project, a second iteration of the study would involve more data gathering for improved uncertainty and prediction of the coaching factor. In order to improve the uncertainty of the m predictions, I would expand the model to include race performances at m. Similar to the m relationship, m times are much more correlated with m times. As described in the data section, I only collected a training dataset of ten athletes from one season and a testing dataset of those same ten athletes from the following season. In order to better understand the coaching factor, I would propose expanding the training/testing sets to add athletes independent of the current group from a season independent of the initial season used to the training set and their subsequent following season to the testing set. Doing so would not only help to better understand how the coach trains to achieve performances but also improve uncertainty as more observations are added to both datasets.