Some Statistical Analyses of Track and Field Events

Dale L. Zimmerman

Department of Statistics and Actuarial Science
University of Iowa

Octoer 11, 2023

Overview

I will report on some statistical analyses of interest to runners (or those who pretend to be — like Dr. Z), including:

- A multivariate analysis of national track records
- Analyses of runner behavior in two 100-km "ultramarathon" races
- An analysis of record-breaking applied to track and field events

A multivariate analysis of national track records

References:

- “Multivariate analysis of national track records”, B. Dawkins, *The American Statistician*, **43**, 110–115 (1989).
- “Revisiting Olympic track records: Some practical considerations in the principal component analysis”, D. Naik and R. Khattree, *The American Statistician*, **50**, 140–144 (1996).

Data

- National records for men and women at various track races from 100 meters to the marathon, as of 1984; I will also perform the same analyses for data current through 2017
- Data from the 55 countries for which a complete set of records for the "flat" races is available (i.e. no hurdles or steeplechase)
- Events included were 100m, 200m, 400m, 800m, 1500m, 3000m (women only), 5000m (men only), 10000m (men only), and marathon

E.g., men's data (as of 1984)

Country	100m	200m	400m	800m	1500m	5000m	10000m	Marathon
	(secs)	(secs)	(secs)	(mins)	(mins)	(mins)	(mins)	(mins)
Argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
⋮								
USA	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
USSR	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55
W. Samoa	10.82	21.86	49.00	2.02	4.24	16.28	34.71	161.83

Relationships among the variables: correlation matrix

Data as of 1984:

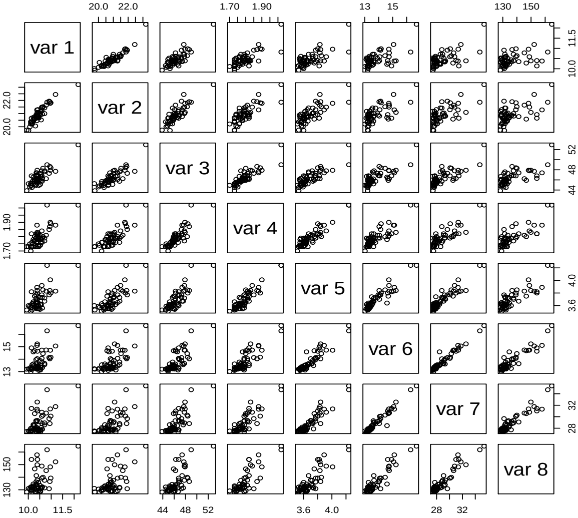
$$\begin{pmatrix} 1.00 & .92 & .84 & .76 & .70 & .61 & .63 & .52 \\ & 1.00 & .85 & .81 & .77 & .70 & .70 & .60 \\ & & 1.00 & .87 & .84 & .79 & .79 & .70 \\ & & & 1.00 & .92 & .86 & .87 & .81 \\ & & & & 1.00 & .93 & .93 & .87 \\ & & & & & 1.00 & .97 & .93 \\ & & & & & & 1.00 & .94 \\ & & & & & & & 1.00 \end{pmatrix}$$

Correlation matrix comparision from 1984 to now

$$\text{1984 data:} \quad \begin{pmatrix} 1.00 & .92 & .84 & .76 & .70 & .61 & .63 & .52 \\ & 1.00 & .85 & .81 & .77 & .70 & .70 & .60 \\ & & 1.00 & .87 & .84 & .79 & .79 & .70 \\ & & & 1.00 & .92 & .86 & .87 & .81 \\ & & & & 1.00 & .93 & .93 & .87 \\ & & & & & 1.00 & .97 & .93 \\ & & & & & & 1.00 & .94 \\ & & & & & & & 1.00 \end{pmatrix}$$

$$\text{Current data:} \quad \begin{pmatrix} 1.00 & .83 & .64 & .62 & .68 & .63 & .57 & .53 \\ & 1.00 & .79 & .69 & .76 & .69 & .64 & .64 \\ & & 1.00 & .62 & .72 & .76 & .74 & .71 \\ & & & 1.00 & .87 & .77 & .73 & .72 \\ & & & & 1.00 & .92 & .89 & .95 \\ & & & & & 1.00 & .98 & .96 \\ & & & & & & 1.00 & .96 \\ & & & & & & & 1.00 \end{pmatrix}$$

Relationships among the variables: scatterplot matrix



Objectives

Objectives:

- Determine a few (at most 2 or 3) linear combinations of the variables that explain most of the variation in the data (dimension reduction)
- Give intuitively reasonable interpretations to those linear combinations

The method of principal component analysis (PCA) achieves these objectives.

A brief diversion: PCA methodology

- PCA transforms the original set of p correlated variables to a new set of p *uncorrelated* variables that are linear combinations of the original variables, called principal components, listed in decreasing order of importance according to how much of the variation in the original variables they explain.
- Basic set-up:
    - Vector of p variables $\mathbf{X}$
    - $\mathbf{X}$ has variance-covariance matrix $\boldsymbol{\Sigma}$
    - $\boldsymbol{\Sigma}$ has eigenvalue/eigenvector pairs $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the $\mathbf{e}_i$'s are orthonormal

PCA methodology, continued

- Total variation among the variables: $\mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \lambda_i$
- 1st PC is the linear combination $\boldsymbol{\ell}_1^T \mathbf{X}$ that maximizes $\mathrm{var}(\boldsymbol{\ell}^T \mathbf{X})$ $(=\boldsymbol{\ell}^T \boldsymbol{\Sigma} \boldsymbol{\ell})$ over $\{\boldsymbol{\ell} : \boldsymbol{\ell}^T \boldsymbol{\ell} = 1\} \Rightarrow \boldsymbol{\ell}_1 = \mathbf{e}_1$
- 2nd PC is the linear combination $\boldsymbol{\ell}_2^T \mathbf{X}$ that maximizes $\mathrm{var}(\boldsymbol{\ell}^T \mathbf{X})$ $(=\boldsymbol{\ell}^T \boldsymbol{\Sigma} \boldsymbol{\ell})$ over $\{\boldsymbol{\ell} : \boldsymbol{\ell}^T \boldsymbol{\ell} = 1,\ \boldsymbol{\ell}^T \boldsymbol{\Sigma} \boldsymbol{\ell}_1 = 0\} \Rightarrow \boldsymbol{\ell}_2 = \mathbf{e}_2$
- etc., *p*th PC is the linear combination $\boldsymbol{\ell}_p^T \mathbf{X}$ that maximizes $\mathrm{var}(\boldsymbol{\ell}^T \mathbf{X})$ $(= \boldsymbol{\ell}^T \boldsymbol{\Sigma} \boldsymbol{\ell})$ over $\{\boldsymbol{\ell} : \boldsymbol{\ell}^T \boldsymbol{\ell} = 1,\ \boldsymbol{\ell}^T \boldsymbol{\Sigma} \boldsymbol{\ell}_i = 0 \text{ for } i = 1, \ldots, p-1\}$ $\Rightarrow \boldsymbol{\ell}_p = \mathbf{e}_p$
- Each maximized value $\boldsymbol{\ell}_i^T \boldsymbol{\Sigma} \boldsymbol{\ell}_i = \mathbf{e}_i^T \boldsymbol{\Sigma} \mathbf{e}_i = \lambda_i$

Results of PCA (for data as of 1984)

Unfortunately a PCA is not invariant to the units in which the variables are measured. Naik and Khattree argue for the PCA to be based on the data transformed to speed, i.e. distance per second, so that the variables are all "on an equal footing."

PC	100m	200m	400m	800m	1500m	5000m	10000m	Marathon	Variation explained
1st	.32	.32	.31	.31	.34	.41	.41	.38	82%
2nd	.60	.47	.23	.06	-.08	-.30	-.30	-.42	12%

Interpretations of PC's:

- 1st PC: overall athletic excellence
- 2nd PC: differential achievement in long-distance vs. short-distance events

Results of PCA for current data

PC	100m	200m	400m	800m	1500m	5000m	10000m	Marathon	Variation explained
1st	.19	.29	.31	.22	.37	.44	.45	.44	82%
2nd	.45	.61	.36	.08	.01	-.22	-.33	-.36	9%

Interpretations of PC's:

- 1st PC: overall athletic excellence
- 2nd PC: differential achievement in long-distance vs. short-distance events

So the interpretations as the same now as they were before, but the 2nd PC explains a little less variation.

France vs. Kenya

France and Kenya had the 8th and 9th largest values of the 1st PC as of 1984, but were at opposite ends of the spectrum with respect to the 2nd PC:

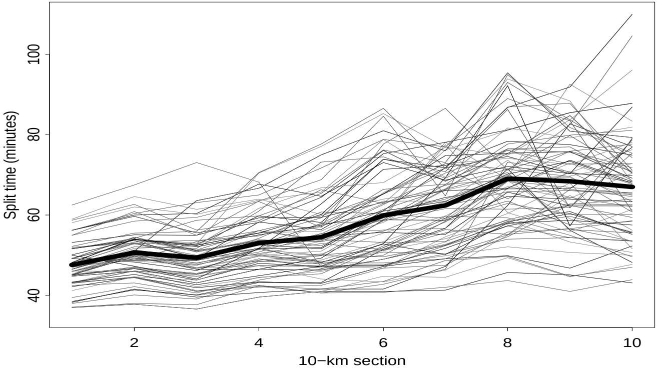
Country	100m	200m	400m	800m	1500m	5000m	10000m	Marathon
	(secs)	(secs)	(secs)	(mins)	(mins)	(mins)	(mins)	(mins)
France	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.38	129.75

Currently, France and Kenya have the 4th and 2nd largest values of the 1st PC.

An analysis of split times for a 100-km race

- Data are from a 100-km race held in 1984 in the U.K.
- Datum is the “split time” for each of 80 competitors in each 10-km section of the race
- Age of 76 of the competitors is an observed covariate
- Objectives are to understand the effect of age on performance, and to understand how performance varies with section and with performance on previous sections

Profile plot of the race



How does performance vary across sections?

- the overall mean profile increases over the first 80 km (more so from 50-80 km than from 0-50 km), then levels off or decreases slightly ("the kick")
- the variance of split times appears to increase as the race progresses
- the behavior of many runners is more erratic, in the sense that successive same-runner split times fluctuate more, in the later sections of the race

Variances and correlations

27									
.95	**35**								
.84	.89	**49**							
.78	.82	.92	**59**						
.60	.63	.75	.89	**91**					
.60	.62	.72	.84	.94	**150**				
.52	.54	.60	.69	.75	.84	**108**			
.45	.48	.61	.69	.78	.84	.78	**152**		
.51	.51	.56	.65	.73	.77	.69	.75	**145**	
.38	.40	.44	.49	.52	.64	.72	.65	.77	**167**

- Variances tend to increase as race progresses

100-km race data: Variances and correlations

27									
.95	35								
.84	.89	49							
.78	.82	.92	59						
.60	.63	.75	.89	91					
.60	.62	.72	.84	.94	150				
.52	.54	.60	.69	.75	.84	108			
.45	.48	.61	.69	.78	.84	.78	152		
.51	.51	.56	.65	.73	.77	.69	.75	145	
.38	.40	.44	.49	.52	.64	.72	.65	.77	167

- Correlations are positive and quite large

100-km race data: Variances and correlations

27									
.95	35								
.84	.89	49							
.78	.82	.92	59						
.60	.63	.75	.89	91					
.60	.62	.72	.84	.94	150				
.52	.54	.60	.69	.75	.84	108			
.45	.48	.61	.69	.78	.84	.78	152		
.51	.51	.56	.65	.73	.77	.69	.75	145	
.38	.40	.44	.49	.52	.64	.72	.65	.77	167

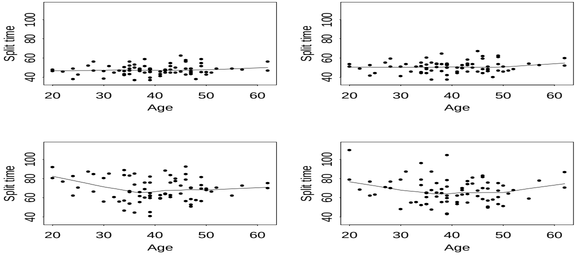
- Correlations between split time for any fixed 10-km section and successive split times tend to decrease

100-km race data: Variances and correlations

27									
.95	35								
.84	**.89**	49							
.78	.82	**.92**	59						
.60	.63	.75	**.89**	91					
.60	.62	.72	.84	**.94**	150				
.52	.54	.60	.69	.75	**.84**	108			
.45	.48	.61	.69	.78	.84	**.78**	152		
.51	.51	.56	.65	.73	.77	.69	**.75**	145	
.38	.40	.44	.49	.52	.64	.72	.65	**.77**	167

- Correlations between split times of consecutive sections are not as large late in the race as they were earlier

Split times versus age (top two plots, first two sections; bottom two plots, last two sections)



There appears to be a quadratic relationship in the later sections of the race (Youth is wasted on the young!)

Additional findings

- A PCA of the covariance matrix indicates that the 1st PC is essentially their average speed for the race, and the 2nd PC is the difference in performance in the last two sections relative to the first 5 or 6 sections (i.e., how much faster the kick is relative to speed earlier in the race). Together, the first 2 PC's explain 85% of the overall variation.
- If we regress each split time on all previous split times, we find that usually the only previous split time that is significantly associated (partially correlated) with a given split time is the immediately previous one (e.g., only the 3rd split time has a significant effect on the 4th split time), and this association is positive.

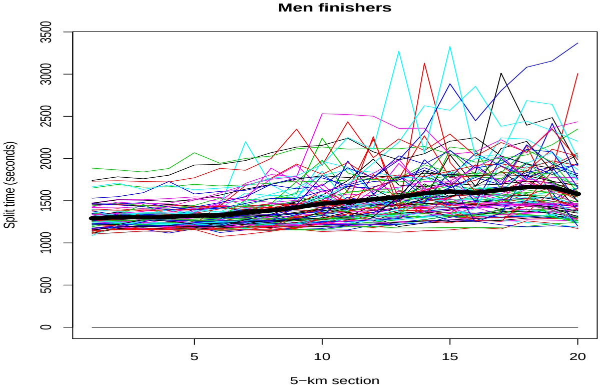
- An interesting exception to the above rule: The last split time has a significant partial correlation with the 5th split time, and this partial correlation is negative. That is, competitors who run slowly on the 5th section, relative to other competitors and also to their own performance on sections 6 through 9, run relatively faster on the last section. A possible physical explanation is that saving energy near the midpoint of the race enables a competitor to run relatively faster at the end.

Analysis of another 100-km race

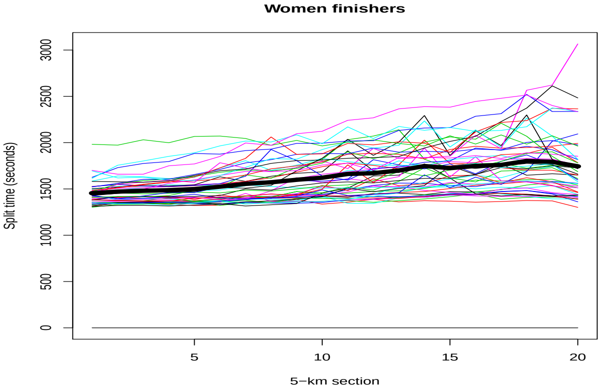
- The 2014 IAU (International Association of Ultrarunners) World Master's Championship 100-km race was held in Doha, Qatar in November 2014.
- The race consists of 20 laps around a 5-km track, so we have 20 5-km split times for each participant.
- 95 male and 48 female participants; ages not recorded.
- The data were downloaded from http://www.stuweb.co.uk/race/VW/splits.html

My objectives were to see which, if any, of the results that emerged from the 1984 data also occur with these data, and to see if there were any interesting differences between men and women.

Profile plot of men finishers



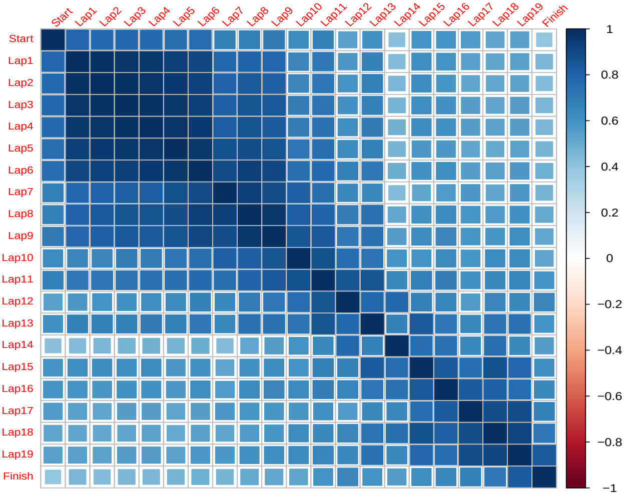
Profile plot of women finishers



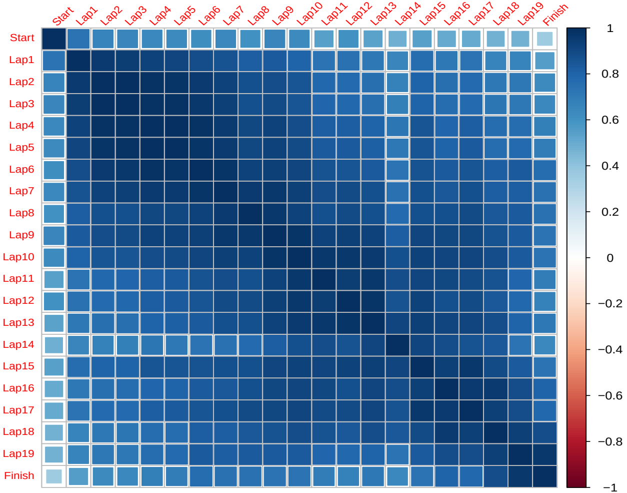
Findings

- The mean profile for men is lower than that for women, but their overall shapes are similar (and similar to that of 1984).
- The variances increase over time (5-fold for men, 8-fold for women), and men's variances are larger than females except for the 6th split.
- The correlations for women are stronger overall than those for men.

Men’s correlations



Women's correlations



Additional findings

- The principal components are similar for men and women: 1st PC ≈ overall average, 2nd PC ≈ contrast between the first and second halves of the race, 3rd PC ≈ contrast between last 10-15 km and the rest (the kick)
- Regressing each split time on the previous split times indicates that in the latter half of the race, a given split time is is significantly associated (partially correlated) with the previous two split times, and this association is always positive.

Record-breaking in athletic events

There is much interest in predicting when a world or Olympic record will be broken in a given event.

Example: Best long jumps, by year, 1962-1999 (* denotes record):

Year	Distance (m)	Year	Distance (m)	Year	Distance (m)
1962	8.31*	1975	8.45	1988	8.76
1963	8.30	1976	8.35	1989	8.70
1964	8.34*	1978	8.32	1990	8.66
1965	8.35	1979	8.52	1991	8.95*
1966	8.33	1980	8.54	1992	8.58
1967	8.35	1981	8.62	1993	8.70
1968	8.90*	1982	8.76	1994	8.74
1969	8.34	1983	8.79	1995	8.71
1970	8.35	1984	8.71	1996	8.58
1971	8.34	1985	8.62	1997	8.63
1972	8.34	1986	8.61	1998	8.60
1974	8.30	1987	8.86	1999	8.60

Data source: www.alltime-athletics.com

When will the world record of 8.95 m, set in 1991 and still holding, be broken?

To address such a question, statisticians can use extreme value theory, in which observations from the tail of a distribution are used for inference.

Reference: “Tail modeling, track and field records, and Bolt’s effects,’ by R.D. Noubary, *Journal of Quantitative Analysis in Sports*, (2010), **6**, Issue 3, Article 9.

A smidgen of extreme value theory

- Suppose $X_1, \ldots, X_n$ are a random sample from a distribution F
- If x is not extreme, we can estimate $F(x)$ consistently by the proportion of observations in the sample that equal or fall below x.
- But what if x is so extreme that all the observations equal or fall below it?
- In that case, under rather mild conditions, an almost surely consistent estimate of $P(X > x) = 1 - F(x)$ is

$$(m/n)(x/X_{(m+1)})^{-1/\hat{a}}$$

where $X_{(1)} > X_{(2)} > \cdots > X_{(n)}$ are the decreasing order statistics, $\hat{a} = (1/m) \sum_{i=1}^{m} \log X_{(i)} - \log X_{(m+1)}$, and m is a sequence of integers chosen such that $m \to \infty$ and $(m/n) \to 0$.

Application to long jump data

Taking $m = 10$ (reasons for this choice are given in the article), the model for the upper tail ($x > 8.70$) is

$$P(X > x) = (10/38)(x/8.70)^{-100}.$$

Thus, the probability of a new world record in 2000 was

$$P(X > 8.95) = (10/38)(8.95/8.70)^{-100} = .0155.$$

In 2000, the probability of a new world record occurring within the next 20 years was

$$1 - (1 - .0155)^{20} = 0.2683.$$

Thus, it should come as no surprise that the world record of 8.95 m (set by Mike Powell of the USA in 1991) is still standing.