© Statistical Research and Training Center

مجله‌ی پژوهش‌های آماری ایران

J. Statist. Res. Iran 3 (2006): 23-46

دوره‌ی ۳، شماره‌ی ۱، بهار و تابستان ۱۳۸۵، صص ۲۳-۴۶

پژوهشکده‌ی آمار

Some Statistical Methods for Prediction of Athletic Records

G. R. Dargahi-Noubary

Bloomsburg University

Invited Paper

Abstract. Prediction of the sports records has received a great deal of attention from researchers in different disciplines. This article reviews some of the methods developed by statisticians and offers few improvements. Specific methods discussed include trend analysis, tail modeling, and methods based on certain results of the theory of records for independent and identically distributed attempts. To make the latter theory applicable, and to account for factors affecting the records, adjustments are made to the data in the form of increase in participation or attempts. Models utilized for this purpose include geometric increase, logistic increase, and increase as a non-homogenous Poisson process. A method for prediction of ultimate record is also included together with demonstrating examples using data for men's long jump and 400 meter run.

Keywords. sport records; prediction; trend analysis; tail modeling; theory of records; Poisson processes.

1 Introduction

The athletic ability of human beings is an issue of great interest to physiologists, physical educators, health professionals, sport fans, and general public. Records set in different sports shed light on human strengths and limitations and provide data for scientific investigations and training or treatment programs. Research in this area can be divided into two categories; short-term prediction and long-term (ultimate record) prediction (see Terpstra and Schauer, 2007; Solow and Smith, 2005; Noubary, 2005; Gulati and Padgett, 2003; Bennett, 1998; Blest, 1996; Noubary, 1994; Tryfos and Blackmore, 1985, and references therein). In what follows we present some of the statistical methods developed for prediction of records and offer some new insights.

2 Methods Based on Trend Analysis

Sports records have improved during the years and often faster than our expectation. To analyze this a large number of investigators have utilized models that are made up of a deterministic term $Z(t, \theta)$, to account for the trend and a stochastic component $x(t)$ to account for the variation, that is

$$y(t) = Z(t, \theta) + x(t). \tag{1}$$

Here t represents time and θ represents the unknown parameters. In most cases $x(t)$'s are assumed to be independent and identically distributed *iid* random variables. For deterministic component $z(t, \theta)$ many different forms are suggested (see Blest, 1996, for the list).

For example, Smith (1988) has considered model (1) assuming that $x(t)$'s are *iid* random variables. For $x(t)$ particular distributions considered were normal, Gumbel, and the generalized extreme-value. For $Z(t, \theta)$ the following linear, quadratic and exponential-decay models were examined.

$$\begin{aligned} Z(t,\theta) &= \theta_0 - \theta_1 t, \qquad \theta_1 > 0 \\ &= \frac{\theta_0 - \theta_1 t + \theta_2 t^2}{2}, \qquad \theta_1 > 0 \\ &= (\theta_0 - \theta_1)\frac{1-(1-\theta_2)^t}{\theta_2}, \qquad \theta_1 > 0, \quad 0 < \theta_2 < 1 \end{aligned} \tag{2}$$

Using the maximum likelihood method and numerical procedures these models were applied to the data for mile and marathon races (Smith, 1988). The normal distribution was found to be the most appropriate among the three distributions although it was noted that the choice of distribution was not crucial for forecasting purposes. Smith also noted that the quadratic or exponential model do not provide a significant improvement over the linear model.

When estimating the limiting time (ultimate records) from exponential-decay model the standard errors were so large that made the predictions meaningless. It is not clear whether this problem was due to the choice of model or the estimation procedure. Smith has implied that the choice of model may be insignificant and acknowledged the wide variability of estimates corresponding to different error distributions and different portions of the series. In fact, he is doubtful that the use of such methods, in general, and model (2) in particular, can produce meaningful performance estimates for the distant future.

In an attempt to overcome these difficulties Noubary (1994) considered an innovative model comprised of an envelope function and a stationary stochastic process in multiplicative form. This model and its statistical inference are discussed in the next section.

2.1 Multiplicative Models

Although useful, additive models of the form (1) may not appropriate for sports data as they imply no dependency or association between variation in $x(t)$ and change in $Z(t, \theta)$. In fact, it is reasonable to expect decrease in variability in the latter portion of the data as performances get closer to the ultimate record and significant improvements become less likely. Also, since most world-class runners remain competitive for a number of years (usually between three and six) some dependency may exist between adjacent performance measures.

Noubary (1994) has suggested use of the models of the form

$$\log y(t) = \theta_0 - \theta_1 t + x(t), \qquad \theta_1 > 0$$

$$\log y(t) = \theta_0 - \theta_1 t + \theta_2 \log t + x(t), \qquad \theta_1 > 0 \tag{3}$$

where $\{x(t), t = 1, 2, \ldots\}$ is a zero-mean stationary process. Note that these models can alternatively be written in multiplicative forms as

$$y(t) = \theta_0^* e^{-t\theta_1} x^*(t), \qquad \theta_1 > 0$$

$$y(t) = \theta_0^* t^{\theta_2} e^{-t\theta_1} x^*(t), \qquad \theta_1 > 0$$

where $\theta_0 = \log \theta_0^*$ and $x(t) = \log x^*(t)$. Here both means and variances vary with time. That is, unlike additive model where variance of $y(t)$ is independent of t, here it decreases as t increases. As a result compared to the additive models, the standard errors of the future records are smaller and therefore the likelihood of obtaining a meaningful prediction is higher.

2.2 Statistical Inference

Suppose that an observed series $\{y(t); t = 1, 2, \ldots, N\}$ is generated by the regression model

$$y(t) = \sum_{k=1}^{p} \theta_k Z_k(t) + x(t),$$

where θ_k's are unknown parameters and $x(t)$ is a zero-mean stationary process possessing a continuous spectrum. Let $y = (y(1), y(2), \ldots, y(N))^{\mathrm{T}}$, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$ and Z^{T} be a matrix with $Z_k(t)$ as the entry in the kth row and the tth column. Also, let $\Sigma = \{\gamma(s-t); s, t = 1, 2, \ldots, N\}$ denote the autocovariance matrix of $x(t)$. Then the best linear unbiased estimator of θ is given by

$$\hat{\theta} = (Z^{\mathrm{T}}\Sigma^{-1}Z)Z^{\mathrm{T}}\Sigma^{-1}y.$$

In practice Σ is often unknown and $\hat{\theta}$ is unavailable. Even if Σ is known its inversion may introduce computational problems, especially for long series. To avert these difficulties a common approach is to replace $\hat{\theta}$ by $\tilde{\theta}$ the simple least squares estimator

$$\tilde{\theta} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y,$$

which does not involve Σ. Since $\tilde{\theta}$ is easy to calculate, one may ask if any precision is lost by using it. It has been shown that loss of precision depends on the function $Z_k(t)$. In fact, it has been known for some time (Grenander, 1954) that in certain cases $\tilde{\theta}$ is efficient in the sense that

$$\{\mathrm{var}(\tilde{\theta})\}\{\mathrm{var}(\hat{\theta})\}^{-1} \to I_p \quad \text{as} \quad N \to \infty$$

where $\mathrm{var}(\tilde{\theta})$ and $\mathrm{var}(\hat{\theta})$ are the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$, and I_p is the unit matrix of order p. An important specific case where the required conditions are satisfied occurs when $\Sigma_k\theta_k Z_k(t)$ is a polynomial in t (Hannan, 1960, p. 122). It is easy to show that these conditions are still satisfied if $\log t$ is added to a polynomial. For both cases the limiting form of $\mathrm{var}(\tilde{\theta})(\mathrm{var}(\hat{\theta}))$ is given by

$$V = 2\pi f(0)(Z^{\mathrm{T}}Z)^{-1},$$

where $f(w)$ denotes the spectral density function of the $x(t)$ process. If additionally $x(t)$'s satisfy a Lindeberg type condition, then asymptotically (Anderson, 1971, Theorem 10.2.11)

$$\tilde{\theta} \sim \mathrm{N}(\theta, V).$$

Noubary has demonstrated the application of these results using the data from 400 and 800 meter races (Table 6). He has shown that of the models considered, (3) provides the best fit for both events. The 400 meter data were well fitted using only the non-random part of model (3) with residuals being random and normal. The prediction intervals obtained embraced the recent fastest times. The 800 meter data also fitted well by model (3) with a moving average $MA(2)$ process representing the random part. The residuals were random and normal. The prediction intervals embraced the recent fastest times too. Also compared to other models, (3) provided a smaller mean square error and narrower prediction intervals both for 400 and 800 meter runs.

We end this section by mentioning that Noubary and Shi (1998) have shown the additive models can be converted to difference equations, by considering $Z(t, \theta)$'s as their complementary solutions. They have also shown that all suggested forms are special cases of the so-called exponential model.

3 Methods Based on Tail Modeling

In this approach the probabilities of future performances are calculated using models for the upper (lower) tail of the distribution for performance measures. Since performance measures above a threshold carry more information regarding the future performance this method is more appealing. Many of the proposed methods assume that the tail belongs to a given parametric family and carry out the inference using excesses, that is the performance measures greater than some predetermined value y_0. It is shown that the natural parametric family of distributions to consider for excesses is the generalized Pareto distribution (GPD) taking the form

$$H(y; \sigma, k) = 1 - \left(1 - \frac{ky}{\sigma}\right)^{\frac{1}{k}}$$

where $\sigma > 0$, $\infty < k < \infty$ and the range of y is $0 < y < \infty$ ($k \leqslant 0$), $0 < y < \sigma/k$ ($k > 0$). This is motivated by the following considerations.

- The GPD arises as a class of limit distributions for the excess over a threshold, as the threshold is increased toward the right-hand end of the distribution, i.e., the tail.

- If Y has the distribution $H(y; \sigma k)$ and $y' > 0$, $\sigma - ky' > 0$, then the conditional distribution of $Y - y'$ given $Y > y'$ is $H(y; \sigma - ky', k)$. This is a "threshold stability" property; if the threshold is increased by an arbitrary amount y', then the GPD form of the distribution remains unchanged.

- If N is a Poisson random variable with mean λ and $Y_1, Y_2, \ldots, Y_N$ are independent excesses with distribution function $H(y; \sigma, k)$, then

  $$P(\max(Y_1, Y_2, \ldots, Y_N) \leqslant y) = \exp\left\{-\lambda\left(\frac{1 - ky}{\sigma}\right)^{\frac{1}{k}}\right\}$$

  has the generalized extreme value distribution. Thus, if N denotes the number of excesses in, say, a year and $Y_1, Y_2, \ldots, Y_N$ denote the excesses, then the annual maximum has one of the classical extreme value distributions.

- The limit $k \to 0$ of the GPD is the exponential distribution.

In most applications the excesses are treated as independent random variables. When fitting GPD the parameters are estimated by maximizing the likelihood function using the observations that exceed a chosen threshold y_0. Note that choice of threshold is, to a large extent, a matter of judgment depending on what is considered large or small or an exceptional performance. Like generalized extreme value distributions, GPD includes three specific forms

1. Long tail Pareto,

2. Medium tail exponential,

3. Short tail distribution with an endpoint,

and most classical distributions fall in domain of attraction of one of these models. Note that, like most asymptotic results application of this approach is not free of problems. Here the obvious problems are the choice of a parametric family, determination of the threshold value, and the problem related to the intractable likelihood equations. To avoid the latter, Pickands has introduced a non-parametric method for inference regarding the parameters of the generalized Pareto distribution. Noubary (1984) applied his method to 100, 200, 400, and 800 meter runes and obtained predictions using data from Olympic games. Unfortunately, depending on the spacing between the most recent records, application of this method may lead to some unacceptable predictions.

An estimate of tail without appealing to the likelihood principle has also been proposed by Davis and Resnick (1984). Their estimate is easier to use and is applicable to a wide class of distribution functions. This estimator of the tail is essentially the same as the one proposed in Hill (1979) They both assume a tail model of the form $F(y) = cy^{-\alpha}$ for $y > y_0$ when y_0 is known. From a random sample of size n the estimates of the parameters are obtained using the upper $m = m(n)$ order statistics. Here m is a sequence of integers chosen such that $m \to \infty$ and $m/n \to 0$.

Noubary (2007) has applied this approach to men's long jump (Table 1 and Figure 1) and 400 meter run. To choose $m(n)$, let n denote the sample size and $m = m(n)$ the number of order statistics such that $m \to \infty$ and $m/n \to 0$.

Clearly a clever choice of m can improve the prediction. One obvious choice is $m(n) = \sqrt{n}$, but there are "better" choices. Assume that the data contains r records. Let T_r be the time between the last and penultimate records and t_r, the time the last record has held to date. Then it can be shown that the following choice proposed by Tata (1986) satisfies the above two conditions.

Table 1. Long jump best annual distances 1962-1999 (* values are records)

8.31*	8.30	8.34*	8.35*	8.33	8.35	8.90*	8.34	8.35	8.34	8.34	8.30
8.45	8.35	8.32	8.52	8.54	8.62	8.76	8.79	8.71	8.62	8.61	8.86
8.76	8.70	8.66	8.95*	8.58	8.70	8.74	8.71	8.58	8.63	8.60	8.60

$$m = \sqrt{eT_r} + \sqrt{t_r} = \sqrt{2.718282T_r} + \sqrt{t_r}$$

For long jump,

$$t_r = 1999 - 1991 = 8, \quad T_r = 1991 - 1968 = 23, \quad m = \sqrt{23e} + \sqrt{8} = 10.74 \approx 10$$

This led to the following tail model

$$P(Y > y) = \frac{10}{38\,(y/8.70)^{-100}} \tag{4}$$

Using this, the values of $P(Y > 8.95)$ and $P(Y > 9.00)$ are 0.0155 and 0.00887 for one year and $1 - (1 - 0.01555)^{10} = 0.1446$ and 0.0852 for ten years respectively. Also, the return period of Y to exceed 8.95 is $1/0.0155 = 64.5$ years, which may seem too long. To see whether probabilities obtained from this model are reasonable, consider the second best performance (distance) 8.90 and its probability,

$$P(Y > 8.90) = 0.0271$$

This corresponds to a return period of about 39 years. Data in Table 1 indicates that this record was set in year 1968 exactly 39 years before 2007 and during

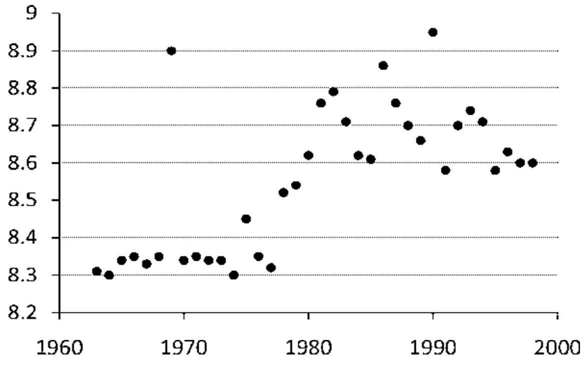

Figure 1. Long jump

this period it was exceeded only once. This agrees with the observation and indicates that (4) is an acceptable model. Also, the fact that the last two records (8.95 and 8.90) are significantly greater than the third best record 8.35 indicates that medium or long tail models provide a better fit than a short tail model.

For men's 400 meter run the fastest times were recorded every year since 1860. The last three records, 43.80, 43.29 and 43.18 were set in years 1968, 1998, and 2000 respectively. Using this information we get $m = 8$ and the following tail (lower tail) model.

$$P(Y > y) = \frac{8}{141\,(44.40/y)^{-90.91}}$$

From this model, the values of $P(Y < 43.10)$ and $P(Y < 43)$ are 0.0038 and 0.0031 for one year, and 0.0374 and 0.03057 for 10 years respectively.

4. Methods Based on Theory of Records

This section presents methods for short-term prediction of records based on results of theory of records. This theory has a large number of exact and asymptotic results regarding the number of records, record times, time interval between records (inter-record times), and record values. Some of the results are non-parametric, and as such are easier to apply. Here we consider few relevant results of this theory and refer the readers to Ahsanullah (1995), Arnold et al. (1998), Glick (1978), and Gulati and Padgett (2003) for other results and their details.

4.1 Short-Term Prediction

To address certain questions regarding the prediction of records, Noubary (2005) has developed a method utilizing the following three results of the theory of records for independent and identically distributed sequence of observations.

(a) If there is an initial sequence of n_1 observations and a batch of n_2 future observations, then the probability for this additional batch to contain a new record is $n_2/(n_1 + n_2)$.

(b) As sample size $n \to \infty$, the frequency of the records among observations indexed by $an < i < bn$ tends to a Poisson count with mean $\ln(b/a)$.

(c) If $F(y) = 1 - \exp\{-y\}$, $y > 0$, and Y_N denote the record values, and if $D_1 = Y_{N_1}$, $D_r = Y_{N_r} - Y_{N_{r-1}}$, $r \geqslant 2$, then the improvements

$D_1, D_2, \ldots$ are independent and identically distributed random variables with common distribution function $F(y)$.

Clearly, the results of theory of records for independent and identically distributed sequences are not directly applicable to sports since, in most cases, sport records are more frequent than what the theory predicts. To account for this Noubary (2005) has treated the problem as if either participation has increased with time, or more attempts have taken place so that the probability of setting a new record was increased. Berry (2002) used the male population of the world as a predictor or an adjusting factor. Using the coefficient of determination as a measure of fit, he found $R^2 = 81.3\%$ for the Olympic winning times in the 100 meter dash. For other events he found R^2 values as high as 95.4%.

Now, although population as a predictor produces surprisingly good results, one should expect even better results if it is replaced by predictors such as the population of participants or the number of attempts as they are more precise indicators of how many times a record is challenged. Berry used the following exponential model for the growth of the world's male population since 1900.

$$\text{Population in Year } t = 1.6 \exp\{0.0088(t - 1900)\}$$

Note that this model can be approximated by a geometric increase with annual rate of $\exp\{0.0088\} = 1.0088388$ since year 1900.

Now, consider a situation where data representing the number of participants is available. For example, Table 5 displays the number of participants of the Boston Marathon for the period 1970-2003. The year 1970 is selected as the starting point because during this year a qualifying time was introduced. As can be seen participation has steadily increased during the years. Using regression, one finds the following quadratic model for the number of participants with $R^2 = 0.938$.

$$\text{Number of Participants in Year } t = -1294 + 1088\, t - 57.5\, t^2 + 1.25\, t^3$$

When fitting this model, we replaced the data for the year 1996 with by the average of the two neighboring values since 1996 was the 100th anniversary of the Boston marathon and more than 38,000 runners were allowed to participate. For this situation, one simple approach would be to model the increase and use that together with result (a) above for prediction. We think that this is reasonable as it will make up for factors that increase the probability of setting new records. To clarify, suppose that in a certain year the best record for men's 100 meter run was s seconds and the probability of setting a new record was p. Suppose further that few decades later the population has tripled. If we divide this population to three subpopulations, then each subpopulation could set a new record with probability p. Thus the probability that at least one subpopulation sets a new record is $1-(1-p)^3$. As an example, for $p=0.05$, this probability is 0.143. We can also look at this as if in a major competition the runners who have potential to break the record get three tries instead of one.

To demonstrate the application of the results (a), (b), and (c) consider, once more, the data for long jump in Table 1. For this event 5 records have been set in the 38 year period (1962-1999). For Independent and identically distributed sequence of length n, the expected number of records is equal to $\sum_{j=1}^{n} 1/j$. This is because the first observation is always a record. The second observation is a record with probability $1/2$ and the jth observation is a record with probability $1/j$. Using this, we see that approximately 83 attempts are needed to produce 5 records as

$$1+\frac{1}{2}+\cdots+\frac{1}{83}=5$$

Since we have 38 years of data, the extra $45=83-38$ attempts need to be distributed over the 38 years of observations and in an increasing format. The problem that remains is to decide about the nature of such distribution. One possibility is to assume that the number of participants or attempts is proportional to the population size at time t. But, as pointed out earlier this approach does not use information from the sports itself and the way records were set. In other words, it is the same for all sports regardless.

As noted, the exponential model for the growth of the world's male population mentioned above can be approximated by a geometric increase. Thus, this seems a reasonable choice. Suppose, for example that i is the geometric rate of increase in participation, or in number of attempts. This means that the number of attempts in any year is i times the number of attempts in the year before. For long jump data the value of i can be found by solving the equation

$$1+i+i^2+\cdots+i^{37}=83$$

This gives $i=1.04$, which means a 4% rate of improvement or equivalently 4% more attempts per year.

To find the probability of a new record during the future 1 and 10 years (in this case year 2000 and the period 2000-2009) we use result (a) and replace $n_1=83$ and $n_2=(1.04)^{38}=4.44$ for 1 year, and $n_1=83$ and $n_2=(1.04)^{38}+\cdots+(1.04)^{47}=53.24$ for 10 years respectively. The resulting probability estimates are 0.051 and 0.391 respectively.

We can also apply result (b) assuming a geometric increase. According to this result the frequency of the records among observations 84 to $137(84+53)$ has approximately a Poisson distribution with mean of $\lambda=\ln(137/84)=0.489$.

Using this, the probabilities of no record and one record during the period $2000 - 2009$ are 0.613 and 0.300 respectively. Also, $1 - 0.613 = 0.387$ is an estimate for the probability of at least one record in 10 years period 2000-2009.

Finally let us demonstrate application of the result (c). First, note that for the long jump data, assuming an exponential distribution for distances beyond for example, 8.25 is reasonable in view of the threshold theory described in Section 3 (Pickands, 1975; Smith, 1987). Recall that according to the threshold theory, the tail of most classical distributions (values beyond a large threshold) takes only one of three possible forms known as generalized Pareto distribution. These forms include the long-tail Pareto, medium-tail exponential and short-tail distribution with an end-point. For performance measures above a high threshold the exponential distribution is either the best model, or because it represents the medium tail is a good approximation for the other two tail behaviors. Here subtracting 8.30 from all distances and dividing the resulting values by 0.195 (standard deviation) provides a sample from the density $f(y) = \exp\{-y\}$, $y \geqslant 0$. The probability of occurrence of a record larger than m_0 in the next n_2 years can then be calculated using the following relation obtained by combining results (a) and (b).

$$P(m > m_0) = \frac{n_2}{n_1 + n_2} \exp\left\{ \frac{-(m_0 - 8.95)}{0.195} \right\}$$

Note that here 8.95 is the value of the last (5th) record. As an example for $m_0 = 9$, the probability estimates are respectively 0.0329 and 0.3024 for the future 1 and 10 years, assuming a geometric increase.

We end this section by noting that, rather than geometric increase, we can following a general approach for modeling population increase, consider models such as Logistic or Gompertz or more generally a model of the form

$$y_{n+1} - y_n = H(y_n) = r^* f(y_n)(1 - g(y_n)).$$

Here y_n denote the number of participants or number of attempts at year n (generation n). One of the simplest and frequently used models that contain a formulation that avoids indefinite growth and represent effects of overcrowding is when r is a linear function of the last year's participation. This choice of r leads to a model of type

$$y_{n+1} - y_n = r^* y_n \left(\frac{1 - y_n}{h} \right) = H(y_n)$$

known as Logistic equation. Here, r^* represents the rate of growth and h represents the carrying capacity. For long jump h may be the maximum number of individuals who qualify to participate in an event such as the Olympics. Models of this type are reasonable for sports where usually rapid initial improvements are followed by much slower advances.

Noubary (2005) has considered the following simpler model instead that exhibits the same behavior as the Logistic equation

$$y_{n+1} = y_n \exp\left\{r^* \left(\frac{1-y_n}{h}\right)\right\}. \tag{5}$$

Applying (5) to the long jump, the number of attempts in future 1 and 10 years period are respectively 4.12 and 48.76 for $y_0 = 1$, $r^* = 0.04$, and $h = 50$. The corresponding numbers using the logistic equation are 4.02 and 47.42. We note that the probability estimates obtained from these models are smaller than that for the geometric increase.

4.2 Prediction Based on Maximum Likelihood Estimate of Number of Attempts

When applying results (a), (b), and (c) we estimated n_1 based on the expected number of records. Instead we can base our estimate on probability of occurrence of r records in a series of length n. This allows us to apply maximum likelihood method and obtain a statistically better estimate for n_1 , the number of attempts.

Let $P_{r,n}$ denote the probability that a series of length n contains exactly r records. It is easy to see that $P_{1,n} = 1/n$ and $P_{n,n} = 1/n!$. Moreover noting that the nth observation is either a record or not, the remaining probabilities can be calculated recursively as

$$\begin{aligned} P_{r,n} &= \frac{n-1}{n} P_{r,n-1} + \frac{1}{n} P_{r-1,n-1} \qquad (6) \\ P_{1,1} &= 1, \\ P_{r,0} &= 0, \qquad r \leqslant n \end{aligned}$$

Note that (6) can also be written as

$$P_{r,n} = \frac{1}{n} \sum_{j=r-1}^{n-1} P_{r-1,j}$$

For example

$$P_{2,n} = \frac{1}{n} \sum_{j=1}^{n-1} P_{1,j} = \frac{1}{n} \sum_{j=1}^{n-1} \frac{1}{j} \approx \frac{1}{n} \{\ln(n-1) + \gamma\}$$

Table 2. Maximum likelihood values of n

r	n	**Max. Prob.**
1	1	1.0
2	2	0.5
3	8	0.325694
4	25	0.253788
5	73	0.214182
6	204	0.188597
7	565	0.170410
8	1557	0.156648
9	4275	0.145767
10	11710	0.136886
11	32022	0.129456
12	87464	0.123122

Also using the properties of the Stirling numbers it is shown (Andel, 2001) that as $n \to \infty$

$$P_{r,n} \frac{1}{(r-1)!\; n}\{\ln(n) + \gamma\}^{r-1} \tag{7}$$

To demonstrate the application, consider the long jump data for the period 1962-2006. The observed number of records is still $r = 5$. For this data application of maximum likelihood yields 73 attempts with $P_{r,n} = 0.214182$ (see Table 2). This leads to

$$1 + i + i^2 + \cdots + i^{44} = 73$$

and $i = 1.0206$. The values of n_2 for the future 1 and 10 years (in this case year 2007 and the period 2007-2016) are respectively 2.503 and 27.485. The probabilities of a new record during the year 2007 and before 2017 are then 0.033 and 0.274 respectively.

Table 2 provides maximum likelihood estimate of n for $r = 1, 2 \ldots, 12$ together with the maximum value of the $P_{r,n}$. As can be seen the value of n increases rapidly. For r-values greater than 12 one could apply the following observation. Noting that $2/1 = 2$, $8/2 = 4$, $25/8 = 3.125$, $73/25 = 2.92$, $204/73 = 2.795$, $565/204 = 2.77$, $1557/565 = 2.756$, $4275/1557 = 2.746$, $11710/4275 = 2.739$, $32022/11710 = 2.735$, $87464/32022 = 2.731$, we conjecture that the ratio is tending to $e = 2.718$. This is also evident from (7) as for larger n the maximizing value of n is $\exp\{r - 1 - \gamma\}$. Thus, for example, an approximation for n when $r = 13$ is $(87464)(2.718) = 237727$.

Next, we apply result (b) assuming a geometric increase. Recall that the frequency of the records among observations 74 and 101 (sum of 74 and 27) has approximately a Poisson distribution with mean $\lambda = \ln(101/74) = 0.311$.

Table 3. Medians of waiting times between successive records and their ratios

Record Number	2	3	4	5	6	7	8
Median(W_r)	4	10	26	69	183	490	1316
Med(W_r)/ Med(W_{r-1})		2.50	2.60	2.65	2.65	2.68	2.69

Using this, the probabilities of zero and one record in the 10 years period (2007-2016) are respectively 0.733 and 0.229. Again $1 - 0.733 = 0.267$ is an estimate for probability of at least one record in a 10 years period.

If rather than geometric increase we consider a slower arithmetic increase the resulting probabilities will be smaller. We think that the geometric increase is suitable for sports with a large number of records whereas the arithmetic increase is suitable for sports with only a few records.

4.3 Waiting Time Analysis

Let W_r denote the waiting time between the $(r-1)$th and rth records. Although the expected waiting time to even the second record is infinite, both the median and the mode of the waiting times are finite. In fact, the following approximate relationship exists between successive waiting time medians.

$$\frac{\text{Median}(W_{r+1})}{\text{Median}(W_r)} \approx e = 2.718$$

Table 3 shows the exact values of the medians and their approximate values. As can be seen the approximations are good even for $r = 4, 5, 6, 7, 8$.
For example, after seeing the second record, the median wait time to the third record is 10 observations (attempts). Other results regarding W_r include a law of large numbers, $\log(W_r/r) \to 1$, and a result indicating that $\log W_r$ is approximately equivalent to the arrival time sequence of a Poisson process. Since sports records are more frequent than records generated by independent and identically distributed sequences, it is possible to model $\log W_r$ as a non-homogeneous Poisson process (see Section 4.4).

Now recall that for long jump the 5th record was set in 1991. Using the maximum likelihood 73 attempts is needed to produce 5 records and these should have occurred during the period 1962-1991 (30 years). This leads to geometric increase with rate $i = 1.055$. Noting that the waiting time to the 6th record is 183 attempts, it takes (in median sense) 49 years for a new record to be set. This means waiting till the year 2040. Recall that the return period of the present record (8.95) was found to be 64.5 years based on the tail model obtained in Section 3.

Table 4. Data for pole vault

Improvement (feet)	Number of years
14 to 15	13
15 to 16	22
16 to 17	1.5
17 to 18	7
18 to 19	10
19 to 20	10

We end this part by noting that, rather than records and waiting times between them, one could consider improvements of equal size and analyze the corresponding waiting times. This seems a reasonable approach since as records improve, increase in number of attempts could offset the decrease in number of record breaking performances. For example, consider the rise in Pole Vault records and their waiting times shown in Table 4.
Here one can consider smaller improvements and apply some of the classical statistical methods. In the case of Pole Vault, for example, the goal of such analysis should be to predict the number of years it would take to go from 20 to 21.

4.4 Attempts as Non-homogeneous Poisson Process

Rather than geometric increase, it is also possible to assume that the number of attempts to break a record is governed by a non-homogeneous Poisson process. Survival of sport records under this assumption is investigated in Noubary and Noubary (2004) where explicit formula is derived for a practical case. The following is a brief description of this approach.

Let $R > 0$ and $S > 0$ be two random variables with respective distribution functions $F_R(\cdot)$ and $F_S(\cdot)$. Suppose R, the record in a given sport, is subject to set of attempts S occurring according to a point process $\boldsymbol{P}$. Then the record breaks if the value of S exceeds R. The value of S is a function of the type of sport, number of participants, prize, training, environmental factors such as temperature, altitude, etc., and factors important to the athletes and the public. The value of R depends on similar factors such as the type and popularity of the sport, amount of rewards or prizes, number of formal competitions, etc. The probability of breaking a record in a single attempt, denoted by p is then

$$P(S > R) = p = 1 - \int_0^{\infty} F_S(x)\ dF_R(x)$$

When applying this model, one is frequently interested in the probability of breaking a record in a specified interval, say $(0,t]$, where 0 represents the beginning of the period. Assume that T is the length of time a record is (or will be) held, then the probability of record being broken in the time interval $(0,t]$, denoted by $F_T(t)$, is

$$F_T(t) = P(T \leqslant t) = 1 - P(T > t) = 1 - L_T(t)$$

where $L_T(t) = P(T > t)$, $L_T(0) = 1$ is the survival function. If R is a record subject to a sequence of attempts $S_1, S_2, \ldots, S_n$, then

$$L_T(t) = \sum_{r=0}^{\infty} P(N(t) = r)\bar{P}(r) \tag{8}$$

where $\{N(t), t \geqslant 0\}$ is a general counting process of attempts and $\bar{P}(r) = P(\max(S_1, S_2, \ldots, S_r) < R)$, $r = 1, 2, \ldots, n$, with $\bar{P}(0) = 1$. Note that $\bar{P}(r)$ presents the probability of surviving the first r attempts. For attempts governed by a homogenous Poisson process with rate λ, we have from (8)

$$L_T(t) = \sum_{r=0}^{\infty} \frac{\exp(-\lambda t)(\lambda t)^r}{r!}\bar{P}(r)$$

If we further assume that the attempts are independent and identically distributed random variables, we get

$$L_T(t) = \sum \frac{\exp(-\lambda t)(\lambda t)^r}{r!}(1-p)^r = \exp(-\lambda t p)$$

Thus, given the mean rate of attempts and a time period of interest then $L_T(t)$ can be calculated for any p. Hence for this situation the main problem is that of estimating the p, i.e. the probability of breaking a record in a single attempt.

4.4.1 Record Survival

Suppose $\boldsymbol{P}$ is Poisson with time-dependent rate $\lambda(t) > 0$ and

$$\Lambda(t) = \int_0^t \lambda(u)\, du$$

Since $(T > t)$ if and only if $\max(S_1, S_2, \ldots, S_n) < R$, the survival probability $P(T > t)$ is given by

$$P(T > t) = \int_0^{\infty} \sum_{n=0}^{\infty} e^{-\Lambda(t)} \frac{(\Lambda(t))^n}{n!} (F_S(x))^n dF_R(x) \tag{9}$$

Note that $(F_S(\cdot))^n$ is the distribution function of $\max(S_1, S_2, \ldots, S_n)$. The expression (9) can also be written as

$$P(T > t) = \int_0^\infty \exp\{-\Lambda(t)(1 - F_S(x)\} \, dF_R(x) \tag{10}$$

If $R = R_0$ is given (e.g. R_0 is the present record), then

$$P(T > t|\ R = R_0) = \exp\{-\Lambda(t)(1 - F_S(R_0))\}$$

Now, it is clear that for the general case (10) requires knowledge of both $F_R(\cdot)$ and $F_S(\cdot)$. Fortunately, there is an important case discussed below where calculations can be carried out with less information and more ease. This case is based on the viewpoint that the strength or importance of a record in a given sport is measured by the number of attempts required to break it.

Suppose that $\boldsymbol{P}$ has been observed throughout the time interval $(-\tau, 0]$, where 0 represent the present time. Suppose also that the largest performance value (records) in this interval is used as a reference for determining further records. Then

$$\begin{aligned} F_R(x) = P(R < x) &= \sum_{n=0}^{\infty} P(\max(S_1, S_2, \ldots, S_n) < x | N(\tau) = n)) P(N(\tau) = n) \\ &= \sum_{n=0}^{\infty} e^{-\Lambda(\tau)} \frac{(\Lambda(\tau))^n}{n!} (F_S(x))^n = \exp\{-\Lambda(\tau)(1 - F_S(x))\} \end{aligned}$$

and application of (10) yields

$$P(T > t) = \frac{\Lambda(\tau)[1 - \exp\{-(\Lambda(\tau) + \Lambda(t))\}]}{\Lambda(\tau) + \Lambda(t)} \tag{11}$$

With confidence given by the right-hand side of (11), there will be no value in $(0, t]$ greater than the maximum value in $(-\tau, 0]$. This implies that, for this situation the survival probability depends only on the rate of attempts.

4.4.2 Examples

Recall the following exponential model for the growth of the world's male population suggested in Berry (2002).

$$\text{Population in Year } t = 1.6 \exp\{0.0088(t - 1900)\}$$

Let us assume that the number of attempts in year t is proportional to the population size at that year. Then using (11) we have the following results.

(*a*) The best record of a 100 years period has 80% chance of surviving an additional 10 years.

(*b*) The best record of a 50 years period has 65% chance of surviving an additional 10 years.

(*c*) The best record of a 10 years period has 24% chance of surviving an additional 10 years.

Suppose now that attempts are made randomly throughout $(-\tau, t]$. If n_1 attempts are made in the interval $(-\tau, 0]$ and n_2 attempts are made in the interval $(0, t]$ then the probability of no new record in $(0, t]$ is

$$P(\max(S_1, S_2, \ldots, S_{n_1}) = \max(S_1, S_2, ...S_{n_1+n_2})) = \frac{n_1}{n_1 + n_2}$$

Thus, corresponding to (a), (b), and (c) above, the 10 years survival probabilities are respectively $100/110 = 91\%$, $50/60 = 83\%$, and $10/20 = 50\%$. However if we assume attempts with a geometric increase of rate 1.0088388, then, for example, corresponding to (a) we have 86% which is closer to 80%. Note that the record of the last 10 years may or may not be the same as the record of the last 20 years. This is one reason for the reduction in survival probability. As mentioned earlier, rather than the general population it is more realistic to consider a model for the population of participants or even the population of participants who have the potential to break records. Recall the regression model for the participation in Boston Marathon (Table 5)

$$\text{Number of Participants in Year } t = -1294 + 1088t - 57.5t^2 + 1.25t^3$$

Using this model the survival probabilities for 5 years period (2003-2008) and 10 years period (2003-2013) are respectively

$$P(T > 5) = 0.632 \quad \text{and} \quad P(T > 10) = 0.422$$

Moreover

$$P(T > 10|\ T > 5) = 0.667$$

We end this section by making a remark regarding the limit of human abilities as it relates to the idea of a possible ultimate record. The problem of estimating the ultimate record is discussed in the next section. In terms of what is discussed here, the ultimate record is the one that will survive forever, i.e. its survival probability is 1. Since it is generally believed that every record will eventually be broken, it is probably more practical to think of a survival probabilities larger than, say 90%, or survival times greater than 50 or 100 years.

Table 5. Data for Boston Marathon 1970-2003

Year	Winner	Time	Time (Minutes)	Number of Participants
1970	Ron Hill	2:10:30	130.50	1174
1971	Alvaro Mejia	2:18:45	138.75	1067
1972	Olavi Suomalainen	2:15:39	135.65	1219
1973	Jon Anderson	2:16:03	136.05	1574
1974	Neil Cusack	2:13:39	133.65	1951
1975	Bill Rodgers	2:09:55	129.92	2395
1976	Jack Fultz	2:20:19	140.32	2188
1977	Jerome Drayton	2:14:46	134.77	3040
1978	Bill Rodgers	2:10:13	130.22	4764
1979	Bill Rodgers	2:09:27	129.45	7927
1980	Bill Rodgers	2:12:11	132.18	5471
1981	Toshihiko Seko	2:09:26	129.43	6881
1982	Alberto Salazar	2:08:52	128.87	7647
1983	Greg Meyer	2:09:00	129.00	6674
1984	Geoff Smith	2:10:34	130.57	6924
1985	Geoff Smith	2:14:05	134.08	5595
1986	Rob de Castella	2:07:51	127.85	4904
1987	Toshihiko Seko	2:11:50	131.83	6399
1988	Ibrahim Hussein	2:08:43	128.72	6758
1989	Abebe Mekonnen	2:09:06	129.10	6458
1990	Gelinda Bordin	2:08:09	128.15	9412
1991	Ibrahim Hussein	2:11:06	131.10	8686
1992	Ibrahim Hussein	2:08:14	128.23	9629
1993	Cosmas Ndeti	2:09:33	129.55	8930
1994	Cosmas Ndeti	2:07:15	127.25	9059
1995	Cosmas Ndeti	2:09:22	129.37	9416
1996	Moses Tanui	2:09:15	129.25	38708
1997	Lameck Aguta	2:10:34	130.57	10471
1998	Moses Tanui	2:07:34	127.57	11499
1999	Joseph Chebet	2:09:52	129.87	12797
2000	Elijah Lagat	2:09:47	129.78	17813
2001	Lee Bong-Ju	2:09:43	129.72	15606
2002	Rodgers Rop	2:09:02	129.03	16936
2003	R. Cheruiyot	2:10:11	130.18	17567

5 Long-Term Prediction

Prediction of ultimate record can be carried out using models such as exponential-decay model discussed in Section 2. Some authors have tried models such as

$$y = \frac{b + ct}{1 + t} \quad \text{or} \quad y = \frac{a + bt + ct^2}{1 + t + t^2}$$

and have used the fact that as $t \to \infty$, $y \to c$. However as pointed out earlier, application of such models usually results in predictions with large standard errors which are not useful or even acceptable (Smith, 1988). Section 5.1 describes a method based on tail modeling. This is a better approach, as it avoids the above mentioned problem and provides a confidence interval for the ultimate record and uses information related to more recent records.

5.1 Estimation of Ultimate Record

Let $Y_1, Y_2, \ldots, Y_n$ be the order statistics corresponding to the data, that is,

$$Y_1 \leqslant Y_2 \leqslant \cdots \leqslant Y_n$$

Let u denote the minimum value of the Y (ultimate record). Assuming that the distribution function $F(y)$ has a lower endpoint and the following condition is satisfied

$$\lim_{t \to 0+} \frac{1 - F(ty + u)}{1 - F(y + u)} = y^{-k}$$

for all $y > 0$ and some $k < 0$, it is shown (De Haan, 1981) that the statistics

$$\frac{\ln m(n)}{\ln\left\{(y_{m(n)} - y_3)/(y_3 - y_2)\right\}}$$

converge to k as $n \to \infty$. Here $m(n)$ is an integer depending on n such that $m(n) \to \infty$ and $m(n)/n \to 0$ as $n \to \infty$. Under these conditions a level $(1-p)$ confidence interval for u is (see De Haan (1981) for details)

$$\left(\frac{Y_1 - (Y_2 - Y_1)}{(1-p)^{-k} - 1}, Y_1\right) \tag{12}$$

When estimation of the maximum value of Y is of interest, the confidence interval takes the form

$$\left(Y_1, \frac{Y_1 + (Y_1 - Y_2)}{(1-p)^{-k} - 1}\right)$$

Since k is unknown, for large n, we may estimate the confidence interval by using in place of k the value of the converging statistics given by (12).

To demonstrate, suppose we wish to estimate the ultimate record for an event such as men's 400 meter run. For this event the best times for each year is available since 1860. Here we have a relatively large sample and thus can apply the above mentioned asymptotic result. The only problem we need to address relates to selection of $m(n)$. When n is not very large, we may use the following discussed earlier

$$m(n) = \sqrt{eT_r} + \sqrt{t_r} = \sqrt{2.718282T_r} + \sqrt{t_r}$$

where T_r denotes the time between the last and penultimate records and t_r, the time the last record has held to date. For the long jump

$$m(n) = \sqrt{23e} + \sqrt{8} = 10.74 \approx 10$$

and $Y_{10} = 8.70$. Using this and the last three records, $Y_1 = 8.95$, $Y_2 = 8.90$, and $Y_3 = 8.86$ we get $k = \ln(10)/\ln(4) = 1.66$ and

$$\frac{Y_1 + (Y_1 - Y_2)}{(1-p)^{-k} - 1} = 8.95 + \frac{0.05}{(0.95)^{-1.66} - 1} = 8.95 + 0.56$$

Thus, based on data for 1962-1999, a 95% confidence interval for the ultimate distance is

$$(8.95, 9.51).$$

For 400 meter run (Table 6) the last three records are 43.18, 43.29, and 43.80 and were set in years 2000, 1998, and 1968, respectively. Using this information we get $m(n) = 8$ and $k = 1.8187$ resulting in a 95% confidence interval

$$(42.77, 43.18).$$

References

Ahsanullah, M. (1995). *Record Statistics*. Nova Science, Commack, NY.

Andel, J. (2001). *Mathematics of Chance*. Wiley, New York.

Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

Arnold, B.C.; Balakrishnan, N.; Nagaraja, H.N. (1998). *Records*. Wiley, New York.

Ballerini, R.; Resnick, S. (1987). Records in the presence of a linear trend. *Adv. in Appl. Probab.* **19**, 801-828.

Bennett, J. (1998). *Statistics in Sports.* Arnold, New York.

Berry, S.M. (2002). A statistician reads the sports pages. *Chance* **15**, 49-53.

Blest, D.C. (1996). Focus on sport lower bounds for athletic performance. *The Statistician* **45**, 243-253.

Box, G.E.O.; Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

Chatterjee, S.; Chatterjee, S. (1982). New lamps for old: An exploratory analysis of running times in Olympic Games. *J. Roy. Statist. Soc. Ser C* **31**, 14-22.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer, London.

Davis, R.; Resnick, S. (1984). Tail estimates motivated by extreme-value theory. *Ann. Statist.* **12**, 1467-1487.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* Wiley, New York.

Glick, N. (1978). Breaking records and breaking boards. *Amer. Math. Monthly* **85**, 2-26.

Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Stat.* **25**, 252-272.

Gulati, S.; Padgett, W.J. (2003). *Parametric and Nonparametric Inference for Record-breaking Data.* Springer, London.

de Haan, L. (1981). Estimation of the minimum of a function using order statistics. *J. Amer. Statist. Assoc.* **76**, 467-469.

Handelman, G.H.; Smith, D.C. (1980). Comparison of running and swimming records. *Sprotswissenschaft* **10**, 161-168.

Hannan, E.J. (1960). *Time Series Analysis.* Methuen, London.

Mengoni, L. (1973). *World and National Leaders in Track and Field Athletics, 1860-1972.* Ascoli Piceno, Italy.

Morton, R.H. (1983). The supreme runner: what evidence now? *Aust. J. Sport Sci.* **3**, 7-10.

Noubary, R. (1984). On estimation of the best attainable time for men's track events in Olympic Games. *Bull. Iranian Math. Soc.* **11**, 53-55.

Noubary, R. (1994). An envelope function model for forecasting athletic records. *J. Forecast.* **13**, 11-20.

Noubary, R. (2005). A procedure for prediction of sports records. *J. Quant. Anal. in Sports* **1**, 1-12.

Noubary, R. (2007). Tail modeling and athletic performances. Submitted for publication.

Noubary, F.; Noubary, R. (2004). On survival times of sports records. *J. Comput. Appl. Math.* **169**, 227-234.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 131-199.

Shapiro, S.S.; Wilk, M.B. (1965) An analysis of variance test for normality (complete sample). *Biometrika* **52**, 591-611.

Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**, 1174-1207.

Smith, R.L. (1988). Forecasting records by maximum likelihood. *J. Amer. Statist. Assoc.* **83**, 331-338.

Solow, A.R.; Smith, W. (2005). How surprising is a new record. *Amer. Statist.* **59**, 153-155.

Tata, M.N. (1986). Estimating the maximum of the support of a beta distribution. Int. Congr. Math., University of California, Berkeley (Personal communication).

Terpstra, J.T.; Schauer, N.D. (2007). A simple random walk model for predicting track and field World records. *J. Quant. Anal. in Sports* **3**, 1-16.

Tryfos, P.; Blackmore, R. (1985). Forecasting records. *J. Amer. Statist. Assoc.* **80**, 46-50.

Appendix

Table 6. Data for 400m run (time is in seconds): 1860-1988

year	time	year	time	year	time	year	time	year	time
1860	53.7	1861	50.2	1862	53.2	1863	51.7	1864	51.7
1865	50.2	1866	52.5	1867	51.4	1868	50.0	1869	51.9
1870	50.7	1871	50.2	1872	49.5	1873	50.3	1874	50.2
1875	50.5	1876	50.5	1877	50.1	1878	51.3	1879	48.9
1880	49.3	1881	48.3	1882	49.9	1883	49.0	1884	48.9
1885	48.5	1886	49.5	1887	49.9	1888	49.7	1889	48.2
1890	48.7	1891	49.1	1892	49.2	1893	48.9	1894	48.7
1895	48.2	1896	48.5	1897	48.7	1898	48.5	1899	49.1
1900	47.5	1901	49.3	1902	49.3	1903	48.7	1904	48.9
1905	48.2	1906	48.5	1907	48.5	1908	47.9	1909	48.3
1910	48.5	1911	48.5	1912	47.7	1913	46.9	1914	48.1
1915	47.7	1916	47.1	1917	48.7	1918	47.3	1919	48.9
1920	48.1	1921	47.7	1922	47.7	1923	47.9	1924	47.4
1925	47.6	1926	48.3	1927	47.5	1928	47.0	1929	47.4
1930	47.6	1931	47.1	1932	46.1	1933	46.6	1934	46.5
1935	46.8	1936	46.1	1937	46.6	1938	46.3	1939	46.0
1940	46.4	1941	46.0	1942	46.6	1943	47.5	1944	47.5
1945	46.7	1946	45.9	1947	45.9	1948	45.7	1949	46.2
1950	45.8	1951	46.0	1952	45.9	1953	45.9	1954	46.1
1955	45.4	1956	45.2	1957	46.0	1958	45.4	1959	45.8
1960	44.9	1961	45.7	1962	45.5	1963	44.6	1964	44.9
1965	45.5	1966	44.7	1967	44.5	1968	43.8	1969	44.4
1970	44.9	1971	44.2	1972	45.0	1973	45.2	1974	45.2
1975	44.93	1976	44.26	1977	45.36	1978	45.47	1979	44.00
1980	44.60	1981	45.12	1982	45.00	1983	45.44	1984	44.27
1985	44.96	1986	44.45	1987	44.32	1988	43.29		

Gholam-Reza Dargahi-Noubary
Department of Mathematics.
Computer Science and Statistics.
Bloomsburg University.
Bloomsburg, PA 17815.
e-mail: *rnoubary@bloomu.edu*