Predicting the Limits of Records in Athletics

A PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Long Chen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Yongcheng Qi

June, 2014

© Long Chen 2014
ALL RIGHTS RESERVED

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor, Dr. Yongcheng Qi, for the help and support he gives on my course work, project and my future study over the past two years. I also want to thank my committee members, Dr. Barry James and Dr. Zhuangyi Liu, from whom I learned and enjoyed math and statistics courses. Furthermore, I would like to thank all the people in the seminar course who helped me with my project and gave me suggestions.

To my dearest parents,
for their love and support.

Abstract

Extreme value theory can be used to predict the occurrence of rare events, such as extreme flood, large insurance losses, stock market crash, or human life expectancy. In this project, we apply the extreme value theory to the athletic events. For athletic events, we mainly focus on the estimation of best athletic performance in near future using extreme value theory. Two types of estimation methods will be used, namely, moment method and maximum likelihood method (MLE). We will give estimation of future athletic record using both methods and compare the results.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Every four years, athletes from all over the world gather at the Olympic Games to compete for medals. It takes athletes years of professional training to have a chance to break the existing world record. So an interesting question is: under the present knowledge of training, material (shoes, suits and equipment), and drug laws, how much more could athletes possibly exceed the current world record in near future?

To answer this question, we can use the extreme value theory, which deals with the issues of extremes. Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables, with a distribution function F. For any positive integer n, let $M_n = \max\{X_1, X_2, X_3, \ldots, X_n\}$ denote the maximum of the n values, then

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, X_2 \leq x, X_3 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) * P(X_2 \leq x) * P(X_3 \leq x) * \cdots * P(X_n \leq x) \\ &= F^n(x). \end{aligned}$$

where $x \in \mathbb{R}$. The minimum can be derived using similar technique, but it's not used in this project thus not discussed. If we know the distribution F, we can easily calculate the exact distribution of M_n, but this is not true in most cases. If we denote x^* as the upper endpoint of distribution F, where $x^* = \sup\{x : F(x) < 1\} \leq \infty$, then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|M_n - x^*| > \epsilon) = 0,$$

thus we know

$$M_n = \max(X_1, X_2, \ldots, X_n) \xrightarrow{p} x^*$$

In fact, if $x < x^*, P(M_n \leq x) = F^n(x) \to 0$, as $n \to \infty$;
If $x \geq x^*, P(M_n \leq x) = F^n(x) = 1$, for all n.
This means no matter what value x takes, $P(M_n < x)$ can be only 0 or 1 as $n \to \infty$, which is degenerate and is not a meaningful study of the distribution of M_n. In order to obtain non-degenerate limit distribution, we need to normalize the distribution.

Theorem 1.1 (**Fisher and Tippett[1], Gnedenko[2]**). *Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with distribution function F. Suppose there exists a sequence of constants $a_n > 0$, and b_n real $(n = 1, 2, \ldots)$, such that*

$$\frac{\max(X_1, X_2, X_3, \ldots, X_n) - b_n}{a_n}$$

has a non-degenerate limit distribution as $n \to \infty$. Then,

$$\lim_{n\to\infty} Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \tag{1.1}$$

where

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma})$$

for some $\gamma \in \mathbb{R}$, with x such that $1 + \gamma x > 0$.

If we take the logarithm of second half of equation (1.1), for any continuity point x satisfying $0 < G_\gamma(x) < 1$,

$$\lim_{n\to\infty} n \log F(a_n x + b_n) = \log G_\gamma(x). \tag{1.2}$$

Since $M_n \xrightarrow{p} x^*$ as $n \to \infty$, then $F(a_n x + b_n) \to 1$ as $n \to \infty$, for each x,

$$\lim_{n\to\infty} \frac{-\log F(a_n x + b_n)}{1 - F(a_n x + b_n)} = 1.$$

Using the result above, equation (1.2) can be rewritten as

$$\lim_{n\to\infty} n\,(1 - F(a_n x + b_n)) = -\log G_\gamma(x), \tag{1.3}$$

or

$$\lim_{n\to\infty} \frac{1}{n\,(1 - F(a_n x + b_n))} = \frac{1}{-\log G_\gamma(x)}. \tag{1.4}$$

Next, we can use inverse functions to reformulate the previous equations and derive more encouraging results. For any nondecreasing function, f, define f^{-1} to be its *left-continuous inverse*. i.e.,

$$f^{-1}(x) = inf(y : f(y) \geq x).$$

Lemma 1.1. *[3] Suppose f_n is a sequence of nondecreasing functions and g is a non-decreasing function. Suppose that for each x in some open interval (a, b) that is a continuity point of g,*

$$\lim_{n\to\infty} f_n(x) = g(x).$$

Let f_n^{-1}, g^{-1} be the left-continuous inverses of f_n and g. Then, for each x in the interval $(g(a), g(b))$ that is continuity point of g^{-1} we have

$$\lim_{n\to\infty} f_n^{-1}(x) = g^{-1}(x). \tag{1.5}$$

Let the function U be the left-continuous inverse of $1/(1-F)$, and we apply Lemma (1.1) to equation (1.4) on the preceding page. It follows that equation (1.4) turns out to be equivalent to

$$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = G^{-1}\left(e^{-1/x}\right) \tag{1.6}$$

for each positive x. Equation (1.6) is equivalent to equation (1.3) or (1.4) but it has simpler form. More encouraging results can be found in the following way:

Theorem 1.2. *[3] Let $a_n > 0$ and b_n be real sequences of constants and G a non-degenerate distribution function. The following statements are equivalent:*

1. $$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x),$$

   for each continuity point x of G.

2. $$\lim_{n\to\infty} n\,(1 - F(a_n x + b_n)) = -\log G_\gamma(x),$$

   for each continuity point x of G for which $0 < G(x) < 1$.

3. $$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = G^{-1}\left(e^{-1/x}\right)$$

   for each $x > 0$ continuity point of $G^{-1}\left(e^{-1/x}\right)$.

Chapter 2

Data Description

In this chapter, the method of data collection will be discussed. Because of the nature of extreme value theory, all the data will be ordered first, and only those data beyond a certain threshold will be selected to conduct further analysis. As our data comes from either the Olympic Games or other top athletic events, athletes competing in this level usually represent the highest level of that country. Under this logic, all the data recorded by the International Association of Athlete Federations (IAAF) will be treated as extreme values.

In this project, we only focus on four athletic events for both male and female athletes: 100 meters run, 200 meters run, 400 meters run, and 800s meter run. We collected data for these four events from IAAF website[4] for the period of year 2002 to year 2006. In some previous papers, people build models using top performance data from each year and try to figure out a model predicting how the world record changes over time. This is intuitively correct, but the result may not be reliable. For example, suppose you have data only from year 2000 to year 2010, and you want to build a model and use it to predict what will happen in the year 2050. If everything remains the same, this might work, but over a long period of time, technology may lead to breakthroughs in equipment or training method, which may have a significant contribution to the world record. Another drawback of this idea is that some top athletes may participate in more than one event each year and thus may contribute many data points to each year. In this case, the data from the same athlete will not be independent and identically distributed, which violates the basic assumption of the extreme value theory. Instead

of predicting records in the far future, Einmahl and Magnus[5] introduced a new idea to overcome the difficulty by using personal best performance, not the top performance in each year.

After data collection, we order and clean the data in each year first, and then we combine them into one column. Let's take the men's 100 meters run as an example. Suppose we have the data from year 2001 to year 2005, we order the data in each year from highest record (shortest time) to lowest record (longest time), and then the lowest record (longest time) of each year will be compared. For the lowest record of each of the five years, we will take the maximum of these five records and use this maximum as lower bound or threshold. All the records lower than this value will be deleted from the dataset. For all the athletes with a better performance than this lower bound, we will compare all his/her records and only keep the best one and delete the rest. In this way, any athlete with a higher-than-threshold record between 2001 to 2005 will be combined into our final dataset, and his/her name will appear only once with his/her best performance. Then we will order the final dataset from highest record (shortest time) to lowest record (longest time) for the future analysis.

Another issue is the order statistics. Since we are conducting analysis using the upper order statistics, which means the bigger the better, our data come from four different running events, in which the less time the athlete uses, the better performance it is. In order to solve this discrepancy, we transform the time into speed, and Einmahl and Magnus[5] convert seconds directly into kilometers per hour.

Because of the imperfection of timing, running times are usually recorded with an accuracy of 0.01 seconds. This will lead data to occur in clusters. For example, more than 10 or 20 athletes could have a record of 10.12 seconds in 100 meters run. If you plot these records, all the 10 or 20 records will concentrate on one point, which will bring problems in analyzing the process. To facilitate the analysis, we will need to "smooth" the data before conducting an analysis. The smoothing method given by Einmahl and Magnus[5] is the following:

Assume we have m athletes with the same personal best of $d = 10.02$ seconds in 100 meters, then the m data points will be smoothed over the interval of $(10.015, 10.025)$

by:

$$d_j = 10.015 + 0.01 * \frac{2j-1}{2m}, j = 1, 2, \cdots, m. \tag{2.1}$$

After cleaning the data, we begin to conduct the analysis, and the details will be discussed in the next chapter.

Chapter 3

Methodology

With data ready to use, we begin to use two approaches, the moment method and the penalized maximum likelihood method, to analyze the data and give an estimate of the right endpoint. Details will be discussed in section 3.1 and section 3.2 on page 10.

3.1 Moment Method

First, let's consider only one athletic event, for example, the 100 meters run for male athletes. Let $X_1, X_2, X_3, \cdots, X_n$ denote the best personal performance data of n athletes. By the assumption of extreme value theory, the n data points are treated as i.i.d. observations from some distribution function F. Denote $X_{n,1} \leq X_{n,2} \leq X_{n,3} \leq \ldots \leq X_{n,n}$ as the corresponding order statistics, so $X_{n,1}$ is the worst performance value, and $X_{n,n}$ is the best performance value, the current world record. For this project, $X_{n,n}$ will be the fastest speed after we transform the data from time to speed.

Among the n order statistics, we usually only use the k upper statistics. We denote the ratio of n and k by t: $t = \frac{n}{k}$. The theoretical criterion for the choice of k is that $k \to \infty$ *and* $t = \frac{n}{k} \to \infty$ as $n \to \infty$. In reality, we choose k that is relatively small compared to n and gives a reliable estimate.

Using t instead of n, equation (1.3) on page 2 will be:

$$\lim_{t\to\infty} t\,(1 - F(a_t x + b_t)) = -\log G_\gamma(x) = -(1+\gamma x)^{-1/\gamma}, \tag{3.1}$$

where $G_\gamma(x) > 0$, $t \in \mathbb{R}^+$ and a_t and b_t are defined by interpolation.

As U is defined to be the left-inverse function of $\frac{1}{1-F}$, Einmahl and Magnus [5] define $b_t = U(t)$ with

$$U(t) = \left(\frac{1}{1-F}\right)^{-1}(t) = F^{-1}\left(1 - \frac{1}{t}\right), \tag{3.2}$$

where -1 still denote the left-continuous inverse. To estimate the right endpoint, we need to estimate γ, a_t and b_t. For $1 < k < n$, Hill[6] defined

$$M_n^{(r)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(\log X_{n,n-i} - \log X_{n,n-k}\right)^r, r = 1, 2. \tag{3.3}$$

The estimator of γ defined by Dekkers, Einmhal, and de Haan[7] is:

$$\hat{\gamma} = M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1}. \tag{3.4}$$

Next, the estimator of a_t is given by:

$$\hat{a} = \hat{a}_{n/k} = \begin{cases} X_{n,n-k}M_n^{(1)}(1-\hat{\gamma}) & \hat{\gamma} \le 0 \\ X_{n,n-k}M_n^{(1)} & otherwise, \end{cases} \tag{3.5}$$

and Einmahl and Magnus[5] estimate b_t using $\hat{b} = \hat{b}_{n/k} = X_{n,n-k}$. From equation (3.2) we can see that $b_{n/k} = U_{n/k}$, and $\hat{b}$ is the empirical analog.

The purpose of this project is to estimate the right endpoint:

$$x^* = \sup\{x : F(x) < 1\}$$

of distribution function F. We assume $\gamma < 0$ when estimating x^*, because once $\gamma > 0$, then x^* will be positive infinity.

Theorem 3.1 (**De Haan, Laurens and Ferreira, Ana[3]**). *For $\gamma \in \mathbb{R}$ the following statements are equivalent:*

1. *There exist real constant $a_n > 0$ and b_n real such that*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right). \tag{3.6}$$

for all x with $1 + \gamma x > 0$.

2. There is a positive function a such that for $x > 0$,

$$\lim_{n\to\infty} \frac{U(nx) - U(t)}{a_n} = \frac{x^\gamma - 1}{\gamma}. \tag{3.7}$$

Based on equation (3.7), for large t, Einmahl and Magnus[5] write heuristically

$$U(tx) \approx U(t) + a(t)\frac{x^\gamma - 1}{\gamma}.$$

Because γ is a negative (otherwise $x^* = \infty$), for large x with $t = n/k$,

$$x^* \approx U\left(\frac{n}{k}\right) + a\left(\frac{n}{k}\right)\frac{1}{\gamma}.$$

Therefore, we can estimate the right endpoint x^* by

$$\hat{x}^* = \hat{b} - \frac{\hat{a}}{\hat{\gamma}}. \tag{3.8}$$

where $\hat{\gamma} < 0$. Since $\hat{a}$ *and* $\hat{\gamma}$ depend on M_n and M_n depends on the choice of k, the x^* will depend on choice of k.

When $\gamma < 0$, Dekkers et al.[7] proved

$$\frac{\sqrt{k}(\hat{x}^* - x^*)}{\hat{a}} \xrightarrow{d} N\left(0, \frac{(1-\gamma)^2(1-3\gamma+4\gamma^2)}{\gamma^4(1-2\gamma)(1-3\gamma)(1-4\gamma)}\right).$$

After we obtain $\hat{x}^*$, we can assess the quality of the world record by measuring the value of $n(1 - F(X_{n,n}))$. In this formula, $1 - F(X_{n,n})$ is the probability of getting a larger-than-world-record data point, and multiplying by the number of data point will give the expected number of exceedances of the current world record $(X_{n,n})$. It may seem natural to measure the quality of the world record using the difference: $x^* - X_{n,n}$, but this quantity could be infinity as γ gets close to 0 or becomes positive. It also ignores the tail behavior of F. Based on equation (3.1) on page 7 with $X_{n,n} = a_t x + b_t$ and $t = n/k$, we have heuristically

$$n(1 - F(X_{n,n})) \approx k\left(1 + \gamma\frac{X_{n,n} - b_{n/k}}{a_{n/k}}\right)^{-1/\gamma}.$$

In this case, we "estimate" the expected number $n(1 - F(X_{n,n}))$ by

$$Q = k\left[max\left(0, 1 + \hat{\gamma}\frac{X_{n,n} - \hat{b}}{a_{n/k}}\right)\right]^{-1/\hat{\gamma}}.$$

The details of Q can be found in Dijk and de Haan[8] or Haan and Ferreira[3]. In general, we estimate γ, the extreme value index, before we estimate the right endpoint x^*. We do not give an estimate of x^* when the estimate of γ is positive or so close to 0 that it is not clear if it is negative.

While the moment method by Einmahl and Magnus[5] gives many good estimates, Hilbe[9] pointed out that some predictions are not consistent with reality: the predicted world record for the marathon was broken in 2008 with an improvement of 27 seconds; while the prediction for the men's 200 meters is 18.63 seconds, a record that is nearly impossible. Next, we will use another approach to tackle the problem of the endpoint estimation.

3.2 Penalized Maximum Likelihood Method

In this section, we will estimate the right endpoint using the method of Penalized Maximum Likelihood. We will start by introducing the likelihood function and Maximum Likelihood Estimation (MLE), and the Penalized Maximum Likelihood Estimation will be introduced afterwards.

Definition 3.1. Likelihood Function:*[10]*

The joint density function of n random variables $X_1, X_2, X_3, \cdots, X_n$ evaluated at $x_1, x_2, x_3, \cdots, x_n$, say $f(x_1, x_2, x_3, \cdots, x_n; \theta)$, is referred to as a **likelihood function**. *For fixed $x_1, x_2, x_3, \cdots, x_n$, the likelihood function is a function of θ and often is denoted by $L(\theta)$.*

If $X_1, X_2, X_3, \cdots, X_n$ represents a random sample from $f(x; \theta)$, then

$$\begin{aligned} L(\theta) &= L(x_1, x_2, x_3, \cdots, x_n \mid \theta) \\ &= f(x_1, x_2, x_3, \cdots, x_n; \theta) \\ &= f(x_1; \theta) * f(x_2; \theta) * f(x_3; \theta) * \cdots * f(x_n; \theta) \\ &= \prod_{i=1}^{n} f(x_i; \theta) \end{aligned}$$

Definition 3.2. Maximum Likelihood Estimator:*[10]*

Let $L(\theta) = f(x_1, x_2, x_3, \cdots, x_n; \theta)$ be the joint pdf of $X_1, X_2, X_3, \cdots, X_n$. For a given

set of observations, $(x_1, x_2, x_3, \cdots, x_n)$*, a value* $\hat{\theta}$ *in* Ω *at which* $L(\theta)$ *is a maximum is called a maximum likelihood estimate (MLE) of* θ*. That is,* $\hat{\theta}$ *is a value of* θ *that satisfies*

$$f\left(x_1, x_2, x_3, \cdots, x_n; \hat{\theta}\right) = \max_{\theta \in \Omega} f\left(x_1, x_2, x_3, \cdots, x_n; \theta\right) \tag{3.9}$$

If Ω is an open interval, and if $L(\theta)$ is differentiable and assumes a maximum on Ω, then the MLE will be a solution of the equation

$$\frac{d}{d\theta} L(\theta) = 0 \ \ (or \ \frac{d}{d\theta} \log L(\theta) = 0)$$

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables with distribution function F having a finite right endpoint. Let's consider a semi-parametric model θ satisfying the following equation:

$$1 - F(x) = c(\theta - x)^{\alpha} + o\left((\theta - x)^{\alpha}\right) \tag{3.10}$$

where $c > 0$ is an unknown constant, $\alpha > 0$ is called exponent of the distribution function F, and θ is the right endpoint of F, i.e., $\theta = \sup\{x : F(x) < 1\}$. In comparison to the previous section, θ here is the same as the x^* of the moment method, and $\alpha = -1/\gamma$.

Similarly to the previous section, let's denote $X_{n,1} \leq X_{n,2} \leq X_{n,3} \leq \ldots \leq X_{n,n}$ as the order statistics of $X_1, X_2, X_3, \cdots, X_n$, so $X_{n,n}$ is the current world record. Let $k = k(n)$ be a sequence of integers with $k/n \to 0$. Hall[11] treated k upper order statistics, $X_{n,n-k+1}, X_{n,n-k+2} \cdots X_{n,n}$ as left censored observations above the threshold $u_n = X_{n,n-k}$. Assume $1 - F(x) = c(\theta - x)^{\alpha}$ temporarily, the likelihood function for $X_{n,n-k+1}, X_{n,n-k+2} \cdots X_{n,n}$, up to a constant scale, is given by

$$L(\theta, c, \alpha) = \prod_{j=0}^{k} \{c\alpha(\theta - X_{n,n-k+j})^{\alpha-1}\}\{1 - c(\theta - X_{n,n-k+j})^{\alpha}\}^{n-k-1} \tag{3.11}$$

Using the method introduced in this section, if we can maximize the equation (3.11), we can obtain Maximum Likelihood (ML) estimators for parameters θ, c, and α (if unknown).

In this project, we focus on the case that $\alpha > 2$ and unknown. For $\alpha > 2$ case, by solving the log-likelihood function, the estimator of θ by Hall is denoted $\tilde{\theta}_H$, which is the smallest solution $(> X_{n,n})$ of the equation

$$(k+1)/\sum_{j=1}^{k} \log \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} - (k+1)/\sum_{j=1}^{k} \frac{X_{n,n-k+j} - X_{n,n-k}}{\theta - X_{n,n-k+j}} - 1 = 0 \tag{3.12}$$

and the estimator for α^{-1} is given by

$$\tilde{\alpha}_H^{-1} = \frac{1}{k+1}\sum_{j=1}^{k} \log \frac{\tilde{\theta}_H - X_{n,n-k}}{\tilde{\theta}_H - X_{n,n-k+j}}$$

Theoretically, this method will give estimates of α and θ; however, there are some drawbacks rooted in this method. This ML method does not work for all possible α, because the likelihood function is unbounded at $\theta = X_{n,n}$. Also, more than one solution will exist for equation (3.12) on the preceding page, and the smallest root yields an inconsistent estimator for α, which will be the case when the estimator of θ is too close to $X_{n,n}$. In addition, there is a significant probability that equation (3.12) on the previous page has no root at all if k is not too large.

To solve this problem of the conventional ML method, we can add a penalty multiplier, say $p(\theta, \alpha, X_{n,n-k}, \cdots, X_{n,n})$, to the likelihood function L. We can define a new likelihood function $L_1(\theta, c, \alpha)$ with $p(\theta, \alpha, X_{n,n-k}, \cdots, X_{n,n}) \to 0$ as $\theta \to X_{n,n}$ such that

$$L_1(\theta, c, \alpha) = L(\theta, c, \alpha) \cdot p(\theta, \alpha, X_{n,n-k}, \cdots, X_{n,n})$$

is bounded globally.

To avoid the unboundedness when θ is getting too close to $X_{n,n}$, we can choose

$$p(\theta, \alpha, X_{n,n-k}, \cdots, X_{n,n}) = \frac{\theta - X_{n,n}}{\alpha(\theta - X_{n,n-k})}$$

as the penalty term, and the penalized likelihood function becomes

$$\begin{aligned} L_1(\theta, c, \alpha) =& c^{k+1}\alpha^k(\theta - X_{n,n})^{\alpha}(\theta - X_{n,n-k})^{\alpha-2} \\ & * \prod_{j=1}^{k-1}\{(\theta - X_{n,n-k+j})^{\alpha-1}\}\{1 - c(\theta - X_{n,n-k})^{\alpha}\}^{n-k-1} \end{aligned} \tag{3.13}$$

for $\theta > X_{n,n}$, and zero otherwise. In this case, $L_1(\theta, c, \alpha)$ is always bounded.

If α is known, say $\alpha = \alpha_0$, we can differentiate the log-likelihood function $L_1(\theta, c, \alpha_0)$ with respect to θ and c. The solution is that

$$\tilde{c} = ((k+1)/n)(\tilde{\theta} - X_{n,n})^{-\alpha_0}$$

and $\tilde{\theta}$ is the solution to the following equation

$$h(\theta) := \frac{\theta - X_{n,n-k}}{\theta - X_{n,n}} + \left(1 - \frac{1}{\alpha_0}\right)\sum_{j=1}^{k-1} \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} - \frac{2}{\alpha_0} - k = 0$$

If both θ and α are unknown, the new ML estimator $(\tilde{\theta}, \tilde{c}, \tilde{\alpha})$ of (θ, c, α), by Qi[12],is defined as the maximizer of $L_1(\theta, c, \alpha)$. By solving the log-likelihood function we have:

$$\tilde{c} = ((k+1)/n)(\tilde{\theta} - X_{n,n-k})^{-\tilde{\alpha}} \tag{3.14}$$

$$\tilde{\alpha}^{-1} = \frac{1}{k}\sum_{j=1}^{k} \log \frac{\tilde{\theta} - X_{n,n-k}}{\tilde{\theta} - X_{n,n-k+j}} \tag{3.15}$$

and $\tilde{\theta}$ is the smallest root to the equation

$$\begin{aligned} g(\theta) := & \sum_{j=1}^{k} \left(\frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} - 1 \right) \\ & - \frac{1}{k} \left(\sum_{j=1}^{k} \log \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} \right) \left(2 + \sum_{j=1}^{k-1} \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} \right) = 0 \end{aligned} \tag{3.16}$$

This new ML method has some advantages over the traditional ML method. This method can be applied to all cases of α, and the new ML estimator exists with probability one and can be uniquely determined, regardless of the sample size; Moreover, it is less biased and gives smaller standard error.

Chapter 4

Analysis

After we clean the data and clarify the methods of analysis, we conduct analysis and demonstrate the results in section 4.2 on page 17.

4.1 Analysis Procedure

Remember we are dealing with the world record of four running events: 100 meters, 200 meters, 400 meters, and 800 meters. The best results of these events are the ones with the smallest values. To facilitate our analysis, we transform the time (in seconds) to speed (in kilometers per hour). After we estimate the ultimate speed of a certain running event, we can transform it back to speed. If we want to give a confidence interval for our estimation result, we need to know the distribution of our estimates in seconds. Einmahl and Magnus[5] have proved that the estimated world record in speed, $\hat{x}^*$, converges to a normal distribution under certain conditions:

$$\frac{\sqrt{k}(\hat{x}^* - x^*)}{\hat{a}} \xrightarrow{d} \mathcal{N}\left(0, \frac{(1-\gamma)^2(1-3\gamma+4\gamma^2)}{\gamma^4(1-2\gamma)(1-3\gamma)(1-4\gamma)}\right) \tag{4.1}$$

Since the transformation from time (in seconds) to speed (in kilometers per hour) is not a linear transformation, we need to figure out the distribution of world records in seconds.

Suppose d (in meters) is the distance of certain running events, and the corresponding time is t (in seconds). Denote the speed (in kilometers per hour) as x, then

$$x = \frac{3.6 * d}{t}.$$

Once we have our estimated world record $\hat{x}^*$, we can calculate the corresponding world record in seconds by

$$t = \frac{3.6 * d}{\hat{x}^*}. \tag{4.2}$$

Let's denote σ_x^2 the variance of equation (4.1), then we can rewrite it as

$$\frac{\sqrt{k}(\hat{x}^* - x^*)}{\hat{a}} \xrightarrow{d} \mathcal{N}\left(0, \sigma_x^2\right) \tag{4.3}$$

where x^* is the limit of endpoint.

Next, we can rewrite equation (4.3) as following:

$$\begin{aligned}
(\hat{x}^* - x^*) &\overset{d}{\approx} \left(\frac{\hat{a}}{\sqrt{k}}\right) * \mathcal{N}\left(0, \sigma_x^2\right) \\
\hat{x}^* &\overset{d}{\approx} x^* + \left(\frac{\hat{a}}{\sqrt{k}}\right) * \mathcal{N}\left(0, \sigma_x^2\right) \\
\hat{x}^* &\overset{d}{\approx} x^* + \left(\frac{\hat{a}\sigma_x}{\sqrt{k}}\right) * \mathcal{N}\left(0, 1\right)
\end{aligned}$$

Together with equation (4.2), we get the following result:

$$\begin{aligned}
t &= \frac{3.6 \cdot d}{\hat{x}^*} \\
&= \frac{3.6 \cdot d}{x^* + \left(\frac{\hat{a}\sigma_x}{\sqrt{k}}\right) * \mathcal{N}(0,1)} \\
&= \frac{3.6 \cdot d}{x^*\left(1 + \left(\frac{\hat{a}\sigma_x}{\sqrt{k}x^*}\right) * \mathcal{N}(0,1)\right)}
\end{aligned}$$

Once we apply the Taylor expansion to the denominator, we derive the following eqaution:

$$\begin{aligned}
t &\approx \frac{3.6 \cdot d}{x^*}\left(1 - \left(\frac{\hat{a}\sigma_x}{\sqrt{k}x^*}\right) * \mathcal{N}(0,1)\right) \\
&\approx \frac{3.6 \cdot d}{x^*} - \left(\frac{3.6 \cdot d\hat{a}\sigma_x}{\sqrt{k}(x^*)^2}\right) * \mathcal{N}(0,1) \\
&\approx \frac{3.6 \cdot d}{x^*} + \left(\frac{3.6 \cdot d\hat{a}\sigma_x}{\sqrt{k}(x^*)^2}\right) * \mathcal{N}(0,1)
\end{aligned}$$

Altogether, we get the distribution of time, t:

$$t \sim \mathcal{N}\left(\frac{3.6 \cdot d}{x^*}, \left(\frac{3.6 \cdot d\hat{a}\sigma_x}{\sqrt{k}(x^*)^2}\right)^2\right) \tag{4.4}$$

Using the above equation, we can give standard error of world record in seconds, thus we can easily calculate the confidence interval for certain estimates.

Using the penalized MLE method, Qi[12] gives the confidence interval as following: If $\alpha_0 > 2$, then

$$(n^{-\gamma_0}k^{1/2+\gamma_0}c^{-\gamma_0}(\tilde{\theta}-\theta_0), k^{1/2}(\tilde{\alpha}^{-1}-\alpha_0^{-1}) \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

where

$$\Sigma = \begin{pmatrix} \gamma_0^{-2}(1+\gamma_0)^2(1+2\gamma_0) & (-\gamma_0)^{-1}(1+\gamma_0)(1+2\gamma_0) \\ (-\gamma_0)^{-1}(1+\gamma_0)(1+2\gamma_0) & (1+\gamma_0)^2 \end{pmatrix}$$

If we only focus on the $\tilde{\theta}$, then we have:

$$\begin{aligned} &n^{-\gamma_0}k^{1/2+\gamma_0}c^{-\gamma_0}(\tilde{\theta}-\theta_0) \sim \mathcal{N}(0,\gamma_0^{-2}(1+\gamma_0)^2(1+2\gamma_0)) \\ &(\tilde{\theta}-\theta_0) \sim \left(\frac{1}{n^{-\gamma_0}k^{1/2+\gamma_0}c^{-\gamma_0}}\right)\cdot\mathcal{N}(0,\gamma_0^{-2}(1+\gamma_0)^2(1+2\gamma_0)) \\ &\tilde{\theta} \sim \theta_0 + \mathcal{N}\left(0, \frac{n^{2\gamma_0}c^{2\gamma_0}}{k^{1+2\gamma_0}\gamma_0^2}(1+\gamma_0)^2(1+2\gamma_0)\right) \end{aligned}$$

If we denote the standard error of $\tilde{\theta}$ as σ_θ:

$$\sigma_\theta = \sqrt{\frac{n^{2\gamma_0}c^{2\gamma_0}}{k^{1+2\gamma_0}\gamma_0^2}(1+\gamma_0)^2(1+2\gamma_0)}$$

and apply the same technique to transform speed to time, we will be able to get the confidence interval of θ using penalized MLE method.

$$\begin{aligned} t &\approx \frac{3.6\cdot d}{\theta_0+\sigma_\theta\cdot\mathcal{N}(0,1)} \\ &\approx \frac{3.6\cdot d}{\theta_0\left(1+\frac{\sigma_\theta}{\theta_0}\mathcal{N}(0,1)\right)} \\ &\approx \frac{3.6\cdot d}{\theta_0}\left(1-\frac{\sigma_\theta}{\theta_0}\mathcal{N}(0,1)\right) \\ &\approx \frac{3.6\cdot d}{\theta_0} - \left(\frac{3.6\cdot d\cdot\sigma_\theta}{\theta_0^2}\mathcal{N}(0,1)\right) \end{aligned}$$

So we have the distribution of time:

$$t \sim \mathcal{N}\left(\frac{3.6\cdot d}{\theta_0}, \left(\frac{3.6\cdot d\cdot\sigma_\theta}{\theta_0^2}\right)^2\right) \tag{4.5}$$

4.2 Analysis Results

4.2.1 Moment Method

As we mentioned in previous chapter, we are predicting the world record in the near fufure, not 50 years from now. We will use the personal bests to conduct analysis, and each athlete only appears once in the dataset of a certain event. The table(4.1) gives a brief summary of our data. The *Depth* is the number of data entries, and this number ranges from 363 up to 970, which means we have enough data to analyze. The *Worst* and *Best* are the technically lowest and highest world record from our dataset. The data here are recorded in seconds:

Table 4.1: Data Summary

	Men			Women		
Event	Depth	Worst	Best	Depth	Worst	Best
100 meters	970	10.30	9.78	578	11.38	10.49
200 meters	512	22.91	19.63	400	23.50	22.00
400 meters	363	46.20	43.62	450	53.20	48.70
800 meters	722	106.60	101.10	503	121.07	113.28

The estimator of γ, the extreme value index, defined by Dekkers, Einmahl, and de Haan[7] is:

$$\hat{\gamma} = M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1}; \tag{4.6}$$

After we run the data in our program, we can obtain the estimation of γ for the 8 different events and get the following table (4.2):

Table 4.2: Estimate of γ

Event	Men	women
100 meters	-0.13	-0.14
200 meters	-0.14	-0.18
400 meters	-0.22	-0.18
800 meters	-0.24	-0.26

We assume that each estimate of γ is strictly less than 0 for all 8 events, so the corresponding estimate of the endpoint is finite: $x^* < \infty$. With the estimator of γ, we

will be able to obtain the estimated extreme speed. Using the transformation defined in section (4.1) and equation(4.4), we give a table of our endpoint estimators, standard error (SE), current best results, and confidence intervals. The "Current Best" are the newest world records by the time this project is done. For each one-sided lower confidence interval, the lower bond is calculated using the standard error and the estimate; while the upper bond is the highest record in the dataset we collected:

Table 4.3: Ultimate world records of male athletes using moment method

	Men			
Event	SE	Estimator	Current Best	Lower C.I. (95%)
100 meters	0.42	9.38	9.58	(8.69, 9.78)
200 meters	0.79	18.83	19.19	(17.53, 19.63)
400 meters	0.69	42.94	43.18	(41.79, 43.62)
800 meters	0.97	100.39	100.91	(98.78, 101.10)

Table 4.4: Ultimate world records of female athletes using moment method

	Women			
Event	SE	Estimator	Current Best	Lower C.I. (95%)
100 meters	0.44	10.02	10.49	(9.30, 10.49)
200 meters	0.90	21.16	21.34	(19.68, 22.00)
400 meters	1.47	47.07	47.60	(44.65, 48.70)
800 meters	2.97	109.00	113.47	(104.12, 113.28)

4.2.2 Penalized Maximum Likelihood Estimation

Using the penalized maximum likelihood method, the estimates of θ and α are guaranteed to exist. By solving equation(3.16) and then using equation(3.14), we can obtain the estimates of θ and α. As mentioned above, the $\tilde{\theta}$ is the smallest solution to the equation(3.16):

$$g(\theta) := \sum_{j=1}^{k} \left(\frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} - 1 \right) - \frac{1}{k} \left(\sum_{j=1}^{k} \log \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} \right) \left(2 + \sum_{j=1}^{k-1} \frac{\theta - X_{n,n-k}}{\theta - X_{n,n-k+j}} \right) = 0$$

The following eight graphs are the $g(\theta)$ of eight different running events, and it is clear that $g(\theta)$ of each event has a unique solution:

From the graphs above, we can see each event will have a unique estimate of $\tilde{\alpha}^{-1}$ and $\tilde{\theta}$. Remember, the γ in moment method is $\gamma = \frac{-1}{\alpha}$, so we can give a table as following:

Table 4.5: Estimate of γ

Event	Men	women
100 meters	-0.20	-0.21
200 meters	-0.14	-0.28
400 meters	-0.22	-0.17
800 meters	-0.21	-0.18

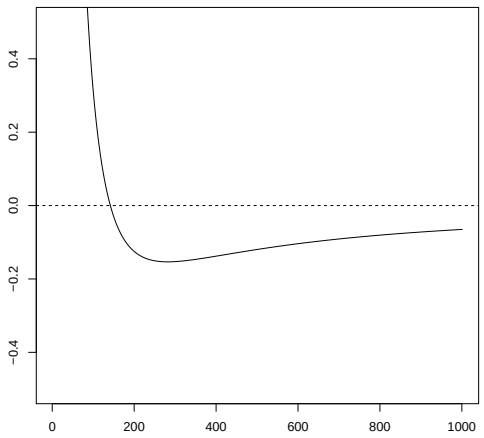
In order to compare with the moment method, we try to use the same k for all events and obtain two table similar to tables(4.3) and table(4.4). The "Current Best" are the newest world records by the time this project is done. For each one-sided lower confidence interval, the lower bond is calculated using the standard error and the estimate; while the upper bond is the highest record in the dataset we collected:

Table 4.6: Ultimate world records of male athletes using penalized MLE

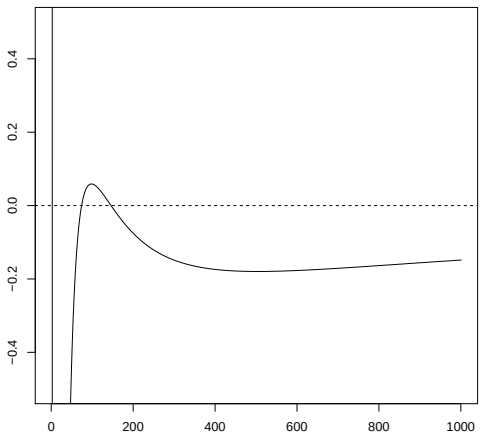
	Men			
Event	SE	Estimator	Current Best	Lower C.I. (95%)
100 meters	0.13	9.60	9.58	(9.38, 9.78)
200 meters	0.68	18.83	19.19	(17.71, 19.63)
400 meters	0.49	42.98	43.18	(42.17, 43.62)
800 meters	0.97	99.82	100.91	(98.22, 101.10)

Table 4.7: Ultimate world records of female athletes using penalized MLE

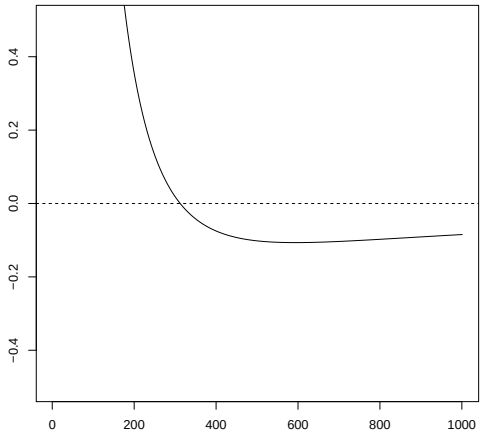
	Women			
Event	SE	Estimator	Current Best	Lower C.I. (95%)
100 meters	0.40	9.81	10.49	(9.15, 10.49)
200 meters	0.23	21.82	21.34	(21.42, 22.00)
400 meters	1.32	46.87	47.60	(44.69, 48.70)
800 meters	2.71	110.07	113.47	(105.61, 113.28)



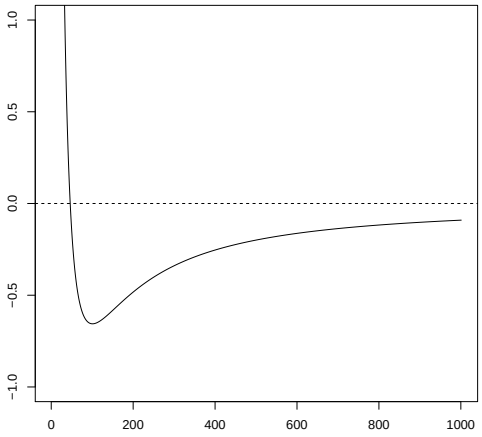
(a) $g(\theta)$ of 100-m male



(b) $g(\theta)$ of 100-m female

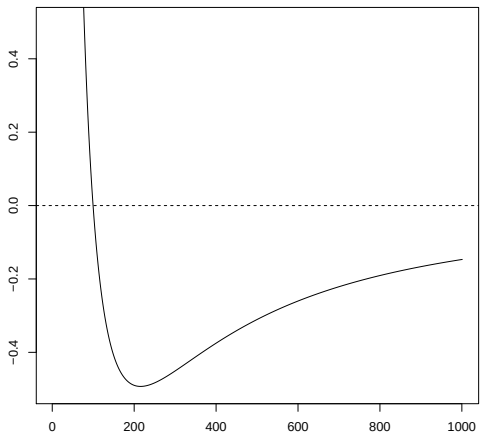


(c) $g(\theta)$ of 200-m male

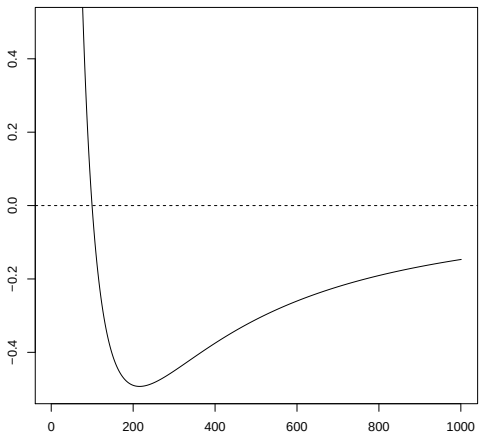


(d) $g(\theta)$ of 200-m female

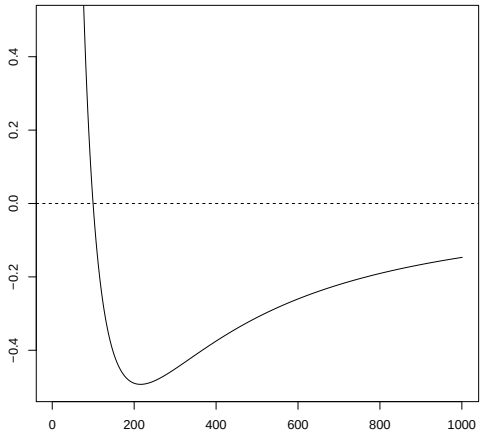
Figure 4.1: $g(\theta)$ of first four running events.



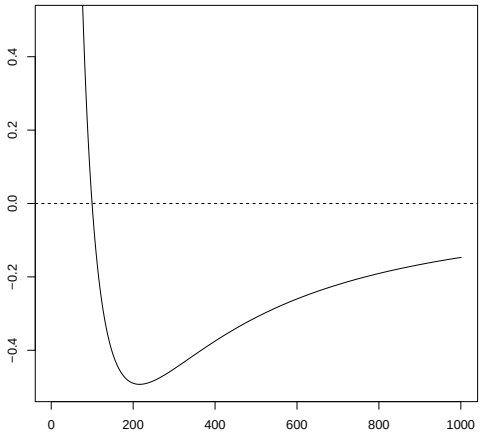
(a) $g(\theta)$ of 400-m male



(b) $g(\theta)$ of 400-m female



(c) $g(\theta)$ of 800-m male



(d) $g(\theta)$ of 800-m female

Figure 4.2: $g(\theta)$ of last four running events.

Chapter 5

Conclusion

In estimating the new world record in the near future, both the moment method and the penalized MLE method can be applied. For the moment method, one important thing is the estimator of γ, which can never be positive if we want a finite estimate of the world record, however, the penalized MLE method works for all cases, and it is guaranteed to have a unique solution. Here is a table of 95% confidence interval of all events using both methods:

Table 5.1: Comparison of Confidence Intervals for Men

		Moment Method	Penalized MLE
Event	Current Best	Lower C.I. (95%)	Lower C.I. (95%)
100 meters	9.58	(8.69, 9.78)	(9.38, 9.78)
200 meters	19.19	(17.53, 19.63)	(17.71, 19.63)
400 meters	43.18	(41.79, 43.62)	(42.17, 43.62)
800 meters	100.91	(98.78, 101.10)	(98.22, 101.10)

Table 5.2: Comparison of Confidence Intervals for Women

		Moment Method	Penalized MLE
Event	Current Best	Lower C.I. (95%)	Lower C.I. (95%)
100 meters	10.49	(9.30, 10.49)	(9.15, 10.49)
200 meters	21.34	(19.68, 22.00)	(21.42, 22.00)
400 meters	47.60	(44.65, 48.70)	(44.69, 48.70)
800 meters	113.47	(104.12, 113.28)	(105.61, 113.28)

In the tables above, the current world record is given in the column "Current Best", and the confidence intervals come from table(4.3), table(4.4), table(4.6), and table(4.7). If the estimation is good, the standard error should be relatively small, and the confidence interval should be relatively narrow.

If we look at the standard error of each event, the estimate given by the penalized MLE method is much smaller than the one calculated using the moment method. Furthermore, although the confidence intervals using the moment method contain all other current world records, they are clearly much wider, thus less useful than the confidence intervals calculated by the penalized MLE method. To summarize, both methods could be used to obtain estimate of future world records, but the penalized MLE method gives a smaller standard error and thus narrower confidence intervals.

References

[1] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press, 1928.

[2] B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):pp. 423–453, 1943.

[3] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer, 2006.

[4] IAAF. *Home of World Athletics – International Association of Athletics Federations*, 2014. [Online; accessed 23-May-2014].

[5] John HJ Einmahl and Jan R Magnus. Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, 103(484), 2008.

[6] Bruce M Hill et al. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5):1163–1174, 1975.

[7] Arnold LM Dekkers, John HJ Einmahl, Laurens De Haan, et al. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

[8] Vincent Dijk and Laurens de Haan. On the estimation of the exceedance probability of a high level. Technical report, Cornell University Operations Research and Industrial Engineering, 1990.

[9] Joseph M Hilbe. Note: Modeling future record performances in athletics. *Journal of the American Statistical Association*, 104(487):1293–1294, 2009.

[10] Lee J Bain and Max Engelhardt. *Introduction to probability and mathematical statistics*, volume 4. Duxbury Press Belmont, CA, 1992.

[11] Peter Hall et al. On estimating the endpoint of a distribution. *The Annals of Statistics*, 10(2):556–568, 1982.

[12] Y Qi. *Penalized maximum likelihood estimation for the endpoint and expoennt of a distribution.* slides 44, University of Minnesota Duluth, June 2013. EVA.

Appendix A

R code

For this project, two methods are used to analyze the dataset, and they are attached in the following two sections.

A.1 Code for the moment method

```
#------------------------------#
#         Project code         #
#------------------------------#

#--------------------------------#
###    For 100 meters male     ###
#--------------------------------#
setwd("C:/Users/Stats Project/Summer Research/data/project data")
rm(list=ls())
data <- read.csv("100m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-105    ### k is the number of upper statistics
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
```

```
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*100*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 100 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("100m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed[1:578],decreasing = F)
n<-length(x)
k<-300    ### k is the number of upper statistics
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
```

```
g1<--.14
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*100*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    200 meters male         ###
#--------------------------------#

rm(list=ls())
data <- read.table("200m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-148
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
```

```
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*200*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 200 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("200m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-50
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
g1<--.18
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
```

```
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*200*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 400 meters male     ###
#--------------------------------#

rm(list=ls())
data <- read.csv("400m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-200
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
g1
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
```

```
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*400*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 400 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("400m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-180
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
g1
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
```

```
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*400*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 800 meters male     ###
#--------------------------------#

rm(list=ls())
data <- read.csv("800m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-250
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
g1
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
```

```
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
st<-(3.6*800*a1*sqrt(v))/(sqrt(k)*x1^2)

#--------------------------------#
###    For 800 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("800m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-50
### moments estimate when r=1
m1<-(1/k)*sum(log(x[(n-k+1):n]/x[n-k]),na.rm=TRUE)
### moments estimate when r=2
m2<-(1/k)*sum((log(x[(n-k+1):n]/x[n-k])^2),na.rm=TRUE)
### 1st estimate of gamma, denoted as "g1"
g1<-m1+1-(1/2)*(1-(m1^2)/m2)^(-1)
g1<--.26
### Estimate a, since g1=-0.0723864, negative, then
a1<-x[n-k]*m1*(1-g1)
### Estimate of b:
b<-x[n-k]
### Estimate of end point, x*:
x1<- b-a1/g1
### Estimate the expected number of exceedances of the current world record,
#denote as Q:
Q1<-k*((max(0,1+g1*(x[n]-b)/a1))^(-1/g1))
### Use v denote the variance:
v<-((1-g1)^2)*(1-3*g1+4*(g1^2))/(g1^4*(1-2*g1)*(1-3*g1)*(1-4*g1))
### SD of time after transform it back
```

```
st<-(3.6*800*a1*sqrt(v))/(sqrt(k)*x1^2)
```

A.2 Code for penalized maximum likelihood method

```
#--------------------------------#
###    For 100 meters male     ###
#--------------------------------#
setwd("C:/Users/Stats Project/Summer Research/data/project data")
rm(list=ls())
data <- read.csv("100m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-105

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 36.81, (36.81+5), tol = 1e-6)
pred
3.6*100/pred
z<-seq(36.81, (36.81+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 100 meters female   ###
#--------------------------------#
```

```
rm(list=ls())
data <- read.csv("100m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed[1:578],decreasing = F)
n<-length(x)
k<-6  ### k is the number of upper statistics
ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 35.945, (35.945+5.0011), tol = 1e-6)
pred
3.6*100/pred

z<-seq(35.945, (35.945+5.0011),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 200 meters male     ###
#--------------------------------#

rm(list=ls())
data <- read.table("200m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-148
```

```
ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 36.679, (36.679+5), tol = 1e-6)
pred
3.6*200/pred

z<-seq(36.679, (36.679+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 200 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("200m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-50

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 32.7273, (32.7273+5), tol = 1e-6)
pred
3.6*200/pred
```

```
z<-seq(32.7273, (32.7273+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 400 meters male     ###
#--------------------------------#

rm(list=ls())
data <- read.csv("400m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-200

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 33.013, (33.013+5), tol = 1e-6)
pred
3.6*400/pred

z<-seq(33.013, (33.013+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
```

```
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 400 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("400m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-180

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 29.569, 33, tol = 1e-6)
pred
3.6*400/pred

z<-seq(29.569,33,by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 800 meters male     ###
```

```
#--------------------------------#

rm(list=ls())
data <- read.csv("800m_male.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
k<-250

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 28.487, (28.487+5), tol = 1e-6)
pred
3.6*800/pred

z<-seq(28.487, (28.487+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")

#--------------------------------#
###    For 800 meters female   ###
#--------------------------------#

rm(list=ls())
data <- read.csv("800m_female.csv", header = TRUE, sep = ",")
x<-sort(data$speed)
n<-length(x)
```

```
k<-35

ftn <- function(theta)
return(sum((theta-x[n-k])/(theta-x[(n-k+1):n])-1)-(1/k)*sum(log((theta-x[n-k])/
(theta-x[(n-k+1):n])))*(2+sum((theta-x[n-k])/(theta-x[(n-k+1):(n-1)]))))
pred<-bisection(ftn, 26.16, (26.16+5.00341), tol = 1e-6)
pred
3.6*800/pred

z<-seq(26.16, (26.16+5),by=.005)

y<-rep(0,length(z))

for (i in 1:length(z)) {
    y[i]<-ftn(z[i])
}
plot(y,ylim=c(-.5,.5),type="l")
```