CentER

Discussion Paper

No. 2009–57

ULTIMATE 100M WORLD RECORDS THROUGH EXTREME-VALUE THEORY

By John H.J. Einmahl, Sander G.W.R. Smeets

July 2009

ISSN 0924-7815

Tilburg University

Ultimate 100m world records through extreme-value theory

John H.J. Einmahl — *Tilburg University*

Sander G.W.R. Smeets — *AZL, Heerlen*

July 10, 2009

Abstract. We use extreme-value theory to estimate the ultimate world records for the 100m running, for both men and women. For this aim we collected the fastest personal best times set between January 1991 and June 2008. Estimators of the extreme-value index are based on a certain number of upper order statistics. To optimize this number of order statistics we minimize the asymptotic mean squared error of the moment estimator. Using the thus obtained estimate for the extreme-value index, the right endpoint of the speed distribution is estimated. The corresponding time can be interpreted as the estimated ultimate world record: the best possible time that could be run in the near future. We find 9.51 seconds for the 100m men and 10.33 seconds for the women.

Running title. Ultimate 100m world records.
Key words and phrases. 100m running, endpoint estimation, statistics of extremes, world record.
JEL codes: L83, C13, C14.

1 Introduction

Extreme-value theory deals with statistical problems concerning the far tail of the probability distribution. It provides solid mathematical tools for semi-parametric inference, in particular for the estimation of extreme quantiles and the endpoint of the support of the distribution. The obtained statistical procedures have many applications, e.g., in hydrology (flood prevention), finance, (re)insurance, meteorology, and engineering.

In this paper we focus on one of the most popular events in athletics: the 100m, for both men and women. We would like to know: how fast can we run? In other words, we are interested in the ultimate world record. Most research concerning ultimate world records considers the development of the world record over time and extrapolates the trend to the future. That approach necessarily uses only a limited number of past world records and therefore gives rather unstable estimates. Our approach, however, is based on as many as possible personal best times of top athletes. As a consequence the estimated ultimate world record tells us what could be achieved 'tomorrow', not what could happen in 500 years from now. Observe that the usual approaches do not tell much about the world record in the near future. The present approach is similar to that in Einmahl and Magnus (2008). We will use, however, more refined methods to estimate the ultimate world records. Also, we base our estimates on a data set which is more up-to-date and, in addition, excludes performances from before 1991 (when modern doping control was introduced), in order to avoid as much as possible doping related times.

The properties of the tail of a distribution depend crucially on one parameter, the so-called extreme-value index. Many good estimators have been derived for this index. All these estimators depend on the number of upper order statistics, say k, used in the estimation. A too low number of order statistics results in a large variance; a too high number of order statistics may cause a large bias. To determine a good value of k we use a method based on minimizing the asymptotic mean squared error.

This paper is organized as follows. In section 2 we will comment on the data that are used. Section 3 briefly reviews the relevant extreme-value theory and discusses the choice of a good k. Our results will be presented in section 4 and in section 5 the results are compared to those in the recent literature.

2 The data

We investigate two events: the 100m both for men and women. For this purpose we collect the fastest personal best times that are set in a certain period. So each athlete only appears once on our list. Including multiple times of one athlete would not be in line with the assumption of independent data, which we will use later on.

The use of doping is an important issue in athletics. A first international legal standard on the use of drugs in sports in general was set by the Council of Europe at the Anti-Doping

Convention in 1989. In 1990 the International Association of Athletic Federations (IAAF) started with out of competition controls. To exclude, as much as possible, doping related times, our observation period therefore starts on 01-01-1991. The observation period ends at 19-06-2008.

For the period before 2001 the data were taken from the Swedish website, `http://hem.bredband.net/athletics/athletics_all-time_best.htm`. This website provides all best times up to 2001. For the time period 01-01-2001 to 19-06-2008 we use the 8 official seasonal time lists of the IAAF website, `www.iaaf.org`. Note that we only consider officially recognized times; so times with a wind speed of more than 2m/s are not taken into account. Times of athletes that are not recognized by the IAAF because of doping use are deleted too.

When combining the lists we have to make sure that no gaps occur. So, when a certain personal best time is included then *all* personal best times lower than this time have to be included too. Therefore we consider the highest personal best time of each list and take the lowest of these times as an upper bound on our combined list. All times higher than this upper bound are removed. The result is then an all time list, without gaps, containing the personal best times up to the aforementioned upper bound time. To obtain the final list it remains to delete all personal bests that are set before our starting date 01-01-1991.

Times for the 100m are available in hundredths of seconds. Since many athletes have equal personal bests the data show ties. To avoid estimation problems the data are smoothed as in Einmahl and Magnus (2008). When m athletes have a personal best of 9.85 then these m results are smoothed equally over the interval (9.845, 9.855) as follows:

$$t_j = 9.845 + 0.01\frac{2j-1}{2m}, \quad j = 1, \ldots, m.$$

As a last step all smoothed times are converted into speeds. Higher speeds correspond to 'better' times.

Table 1 shows the sample size and the time range for each event. Note that the world record of Florence Griffith-Joyner for the 100m women, 10.49 seconds, has been set in 1988 (before 1991) and is therefore not the fastest time in our sample. The current world record for the 100m men of Usain Bolt, 9.69 seconds, dates from August 2008 and therefore falls also outside our sample.

event	sample size	fastest	slowest
100m men	762	9.72	10.30
100m women	479	10.65	11.38

Table 1: Data summary.

3 Extreme-value theory

Suppose that $X_1, X_2, \ldots, X_n$ (the speeds) are n identically, independently distributed observations with continuous distribution function F. Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ denote the corresponding order statistics.

Suppose that there exist a sequence $a_n > 0$ and a sequence b_n such that the maximum $X_{n:n}$, scaled and centered by a_n and b_n, converges in distribution to a non-degenerate distribution G: for all continuity points x of G,

$$\lim_{n\to\infty} P\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x). \tag{1}$$

Then the class of non-degenerate distributions that can occur as a limit in (1) is the class of extreme-value distributions. This class is $G_\gamma(ax + b)$, with $a > 0$ and $b \in \mathbb{R}$, where

$$G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right) \text{ for } 1 + \gamma x \geq 0;$$

$\gamma \in \mathbb{R}$ is the extreme-value index. So, apart from the scale and location constants this class of extreme-value distributions is characterized by the extreme-value index γ. Condition (1) is the so-called extreme-value condition. If the extreme-value condition holds for $G = G_\gamma$, then we say that F is in the max-domain of attraction of G_γ, denoted by $F \in D(G_\gamma)$. See e.g. de Haan and Ferreira (2006) for a thorough account.

As we study the ultimate world records we are interested in the right endpoint of the distribution, $x^* = \sup\{x \in \mathbb{R} : F(x) < 1\}$. This endpoint is finite (infinite) if $\gamma < 0$ ($\gamma > 0$). If $\gamma < 0$, an approximation of x^* can be obtained as follows. Rewrite condition (1) as:

$$\lim_{t\to\infty} t(1 - F(a_t x + b_t)) = -\log G_\gamma(x) = (1+\gamma x)^{-1/\gamma}, \tag{2}$$

with $G_\gamma(x) > 0$ and $t \geq 1$. When t is not an integer, a_t and b_t are defined by intrapolation. For large values of t we then have,

$$1 - F(a_t x + b_t) \approx \frac{1}{t}(1+\gamma x)^{-1/\gamma}.$$

Define $y = a_t x + b_t$ to obtain

$$1 - F(y) \approx \frac{1}{t}\left(1 + \gamma \frac{y - b_t}{a_t}\right)^{-1/\gamma}. \tag{3}$$

Now suppose that $\gamma < 0$. Take $y = x^*$ and note that $F(x^*) = 1$. So we find, with $t = n/k$ for some positive integer k much smaller than n,

$$x^* \approx b_{n/k} - \frac{a_{n/k}}{\gamma}. \tag{4}$$

3.1 Estimators of γ and x^*

First we present an estimator for γ, the so-called moment estimator, introduced in Dekkers et al. (1989). Define, for $k \in \{2, 3, \ldots, n-1\}$:

$$M_n^{(r)} = M_n^{(r)}(k) = \frac{1}{k}\sum_{i=0}^{k-1}(\log X_{n-i:n} - \log X_{n-k:n})^r, \quad \text{for } r = 1, 2.$$

The moment estimator is given by

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

Under certain conditions, including (1) and $k/n \to 0$ and $k \to \infty$ ($n \to \infty$), we have for all $\gamma \in \mathbb{R}$ asymptotic normality of $\hat{\gamma}$, with rate $1/\sqrt{k}$.

Using the usual estimators for $a_{n/k}$ and $b_{n/k}$,

$$\hat{a}_{n/k} = (1 - \min\{0, \hat{\gamma}\})X_{n-k:n}M_n^{(1)} \quad \text{and} \quad \hat{b}_{n/k} = X_{n-k:n},$$

(see, e.g., de Haan and Ferreira, 2006), we obtain, using (4), as an estimator for x^*:

$$\hat{x}^* = \hat{b}_{n/k} - \frac{\hat{a}_{n/k}}{\hat{\gamma}}. \tag{5}$$

To construct a confidence interval for the endpoint x^* we can use that under certain conditions (see Dekkers et al. 1989, or de Haan and Ferreira, 2006):

$$\frac{\sqrt{k}(\hat{x}^* - x^*)}{\hat{a}_{n/k}} \xrightarrow{d} N\left(0, \frac{(1-\gamma)^2(1-3\gamma+4\gamma^2)}{\gamma^4(1-2\gamma)(1-3\gamma)(1-4\gamma)}\right). \tag{6}$$

Clearly the estimators of γ and x^* depend on the choice of the number of upper order statistics k. The next subsection will deal with this issue.

3.2 Optimal threshold selection

It is useful to depict $\hat{\gamma}$ as a function of k. For low values of k the estimator has a large variance whereas for high values of k the estimator may have a large bias. We seek to find a value of k such that the asymptotic mean squared error (AMSE) of $\hat{\gamma}$ is minimized. The AMSE can be written as the sum of the asymptotic variance ($AVar$) and the square of the asymptotic bias ($ABias$).

Suppose $\gamma < 0$. Then

$$AVar(\hat{\gamma}) = \frac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}\frac{1}{k},$$

see, e.g., Beirlant et al. (2004). The asymptotic bias depends on the so-called second order parameter $\rho \leq 0$; ρ appears in the usual strengthening of the extreme-value condition (1). $ABias$ depends on whether $\rho < \gamma$ or $\rho > \gamma$, but in practice it is almost impossible to 'know' which case holds. The estimates of γ in athletics are negative but close to zero, like -0.1 or -0.2, see Einmahl and Magnus (2008). Therefore, and since $\rho \leq 0$ we use the expression for $ABias$ in case $\rho < \gamma$:

$$ABias(\hat{\gamma}) = \frac{1-2\gamma}{\gamma(1-\gamma)}c\left(\frac{n}{k}\right),$$

where c is a function for which $c(x) \to 0$ as $x \to \infty$; see, e.g., Beirlant et al. (2004). So $ABias$ depends not only on γ but also on the function c. In order to eventually estimate the AMSE, we need to estimate $AVar$ and $ABias$. $AVar$ requires an initial estimator of γ; $ABias$ requires in addition an estimator for $c(n/k)$.

To get an initial estimator of γ we plot $\hat{\gamma}$ against k. For small values of k the graph will be highly volatile; for large values of k the estimator typically shows a bias and the graph will typically move downwards or upwards. We will choose the first stable region in the graph and take the average over this region, $\hat{\gamma}_{ini}$, as an initial estimator of γ. This is not so straightforward in practice. In order to make our method less sensitive a range of values for $\hat{\gamma}_{ini}$ will be specified, rather than one singe value.

In Beirlant et al. (2005), under certain conditions the following approximate model is derived:

$$Z_j := (j+1)\log\frac{UH_j}{UH_{j+1}} = \gamma + c\left(\frac{n}{k}\right)\left(\frac{j}{k}\right)^{-\eta} + \varepsilon_j, \quad \text{for } j = 1, \ldots, k,$$

with $UH_j = X_{n-j:n} M_n^{(1)}(j)$ and where the ε_j are mean-zero error terms. In case $\rho < \gamma$ we have $\eta = \gamma$. Given the initial estimator of γ, $c(n/k)$ now can be estimated by 'ordinary least squares' from

$$\tilde{Z}_j := Z_j - \hat{\gamma}_{ini} = c\left(\frac{n}{k}\right)\left(\frac{j}{k}\right)^{-\hat{\gamma}_{ini}} + \varepsilon_j, \quad \text{for } j = 1, \ldots, k.$$

This yields

$$\hat{c}(n/k) = \frac{\sum_{j=1}^{k} \tilde{Z}_j \left(\frac{j}{k}\right)^{-\hat{\gamma}_{ini}}}{\sum_{j=1}^{k} \left(\frac{j}{k}\right)^{-2\hat{\gamma}_{ini}}} .$$

4 Results

First we will test whether the data satisfy the extreme-value condition. Second, we will use the estimators and methods from the previous section to find a final $\hat{\gamma}$ for both events. Finally the endpoint is estimated.

4.1 Testing the extreme-value condition

We will test the extreme-value condition (1), using the test in Dietrich et al. (2002). Figure 1 presents the test statistic and its estimated, asymptotic 95% quantile under the null hypothesis (1). For values of the test statistic lower than its quantile the null hypothesis is not rejected. For the 100m men and 100m women the extreme-value condition is not rejected for almost all values of k, except for some very low values. This means that we can continue our analysis, which is based on (1). The final choices of k, however, should not be very low.

4.2 Estimation of γ and threshold selection

Figure 2 shows the moment estimator $\hat{\gamma}$ versus k. All estimates are negative. Recall that a negative γ guarantees a finite right endpoint of the distribution. Instead of using only one value of $\hat{\gamma}_{ini}$, we consider a range of s_0 values for $\hat{\gamma}_{ini}$. This will make the procedure less sensitive to this initial choice. The 'steps' in this range will be of size 0.01. When choosing the k-region for $\hat{\gamma}_{ini}$ we also keep in mind that the extreme-value condition must be satisfied. For the 100m men we are not sure whether to search for the initial $\hat{\gamma}$ in the first stable region (k from 50 to 80) or the second stable region (k from 110 to 200). The

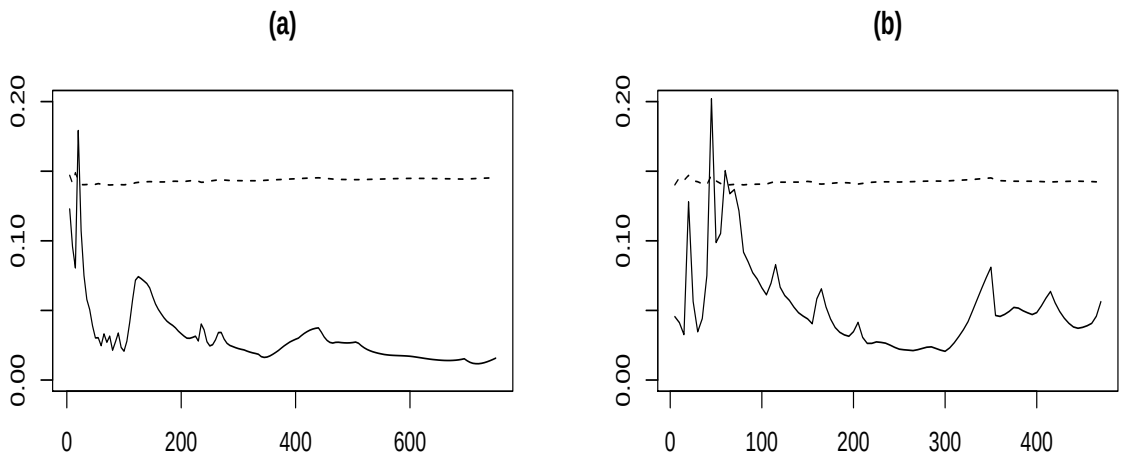

Figure 1: Testing the extreme-value condition (Dietrich et al., 2002) for (a) 100m men and (b) 100m women. The dashed line represents the 95% quantile, the solid line gives the test statistic (both as a function of k).

first region shows volatile behavior whereas the second region may lead to a bias. The first region gives, by averaging, a $\hat{\gamma}_{ini}$ of -0.25; the second one gives a $\hat{\gamma}_{ini}$ of -0.17. Therefore we let $\hat{\gamma}_{ini}$ range from -0.25 to -0.17 ($s_0 = 9$). For the 100m women the first stable region is from about 60 to 100; the second from about 110 to 200. Again the first region shows slightly more volatile behavior. For the first region we get $\hat{\gamma}_{ini} = -0.28$; for the second region $\hat{\gamma}_{ini} = -0.18$. $\hat{\gamma}_{ini}$ will range between those two bounds ($s_0 = 11$).

Consider one of these s_0 values, say $\hat{\gamma}_{ini}^{(s)}$. After estimating $\hat{c}(n/k)$ we are able to estimate the AMSE of the moment estimator, $\widehat{AMSE}(k, \hat{\gamma}_{ini}^{(s)})$. We can plot this estimated AMSE against k for each value of $\hat{\gamma}_{ini}$ in the specified range. This results in s_0 different AMSE-plots. Comparing these plots we notice that the general shape of the plots is the same; it does however happen that the k-region for which the minimum estimated AMSE is attained, differs among the plots. We aim at choosing a k-region for which the AMSE is relatively small, independent of the s_0 initial estimates. Therefore we consider the average AMSE:

$$\overline{AMSE}(k) = \frac{1}{s_0} \sum_{s=1}^{s_0} \widehat{AMSE}(k, \hat{\gamma}_{ini}^{(s)}),$$

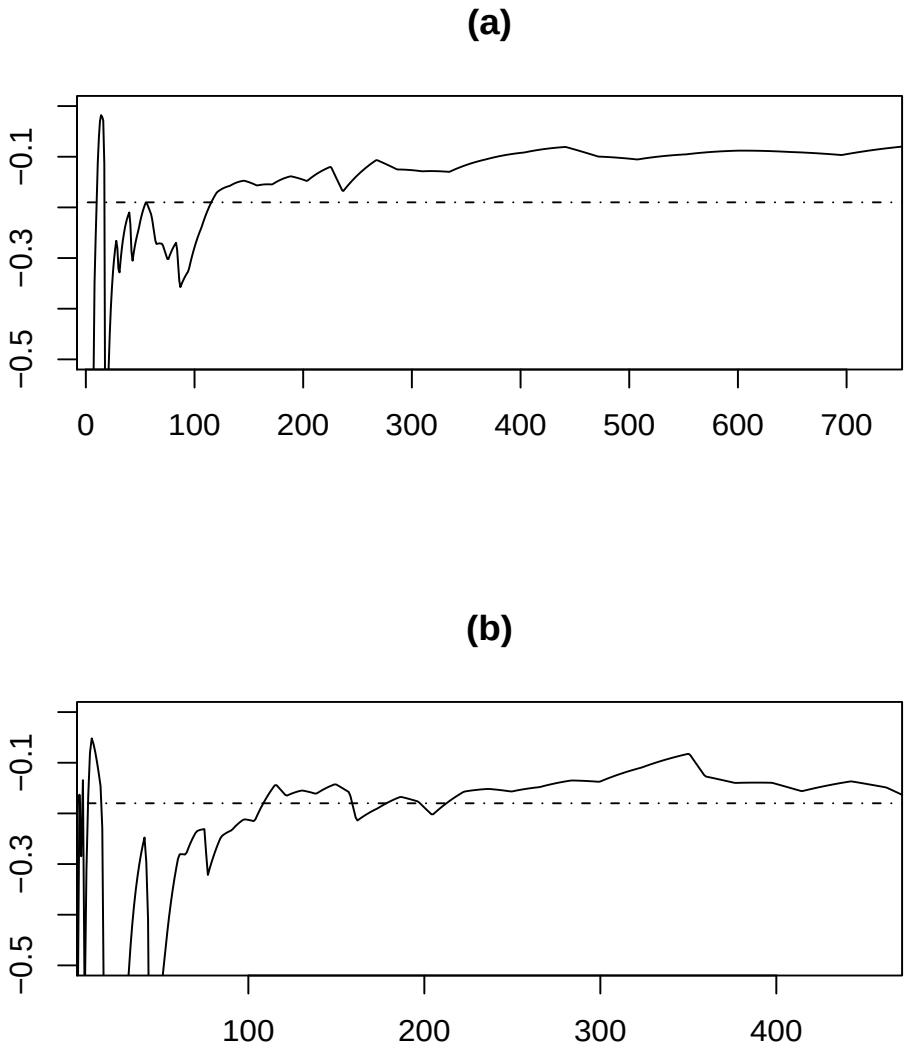

Figure 2: The moment estimator versus k for (a) 100m men and (b) 100m women; the horizontal line shows the final $\hat{\gamma}$.

see Figure 3. By analyzing these average AMSE plots a k-region is chosen for which the estimated average AMSE is small. The vertical dotted lines show the ultimately chosen k-regions. For the 100m men we take k between 45 and 70 and between 101 and 175 and for the women between 80 to 250. The last step is then to average $\hat{\gamma}$ over these k-regions.

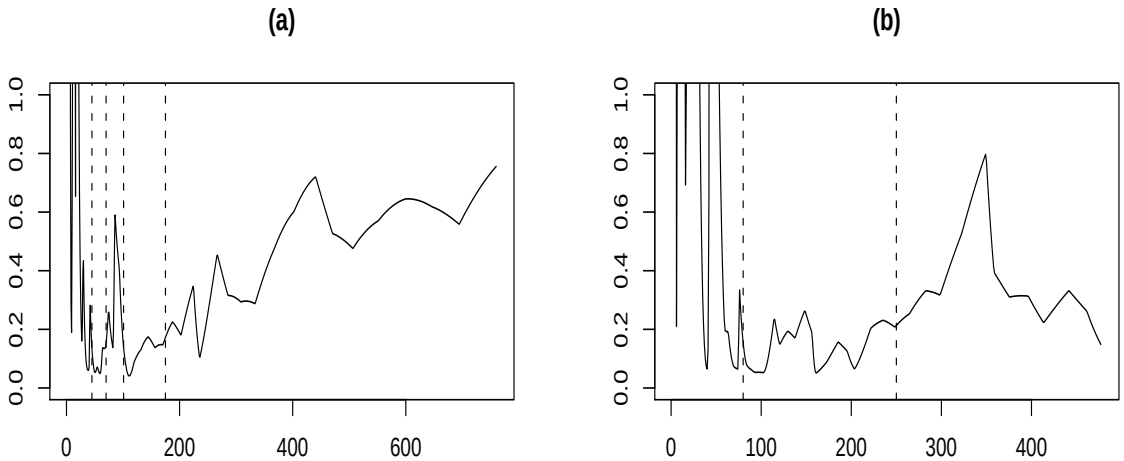

Figure 3: Average estimated AMSE of the moment estimator for (a) 100m men and (b) 100m women.

Table 2 presents the final $\hat{\gamma}$ for both events.

	$\hat{\gamma}_{ini}$	$\hat{\gamma}$
100m men	-0.25 to -0.18	-0.19
100m women	-0.28 to -0.18	-0.18

Table 2: Estimates of γ.

4.3 Endpoint estimation

In this section we will estimate the ultimate world records. To estimate the endpoint we plug the final $\hat{\gamma}$ into the formula for $\hat{x}^*$, given in (5). The thus obtained expression only depends on k through $M_n^{(1)}(k)$ (via $\hat{a}_{n/k}$) and $X_{n-k:n}$. Figure 4 shows the plots of the endpoint estimator. These plots are quite stable. At first glance we get already a good indication of the ultimate world records. For the 100m men the estimated ultimate world record will be between 37.6 km/h (9.57s) and 38.0 km/h (9.47s); the ultimate record for the 100m women will be between 34.7 km/h (10.37s) and 35.0 km/h (10.29s). To arrive

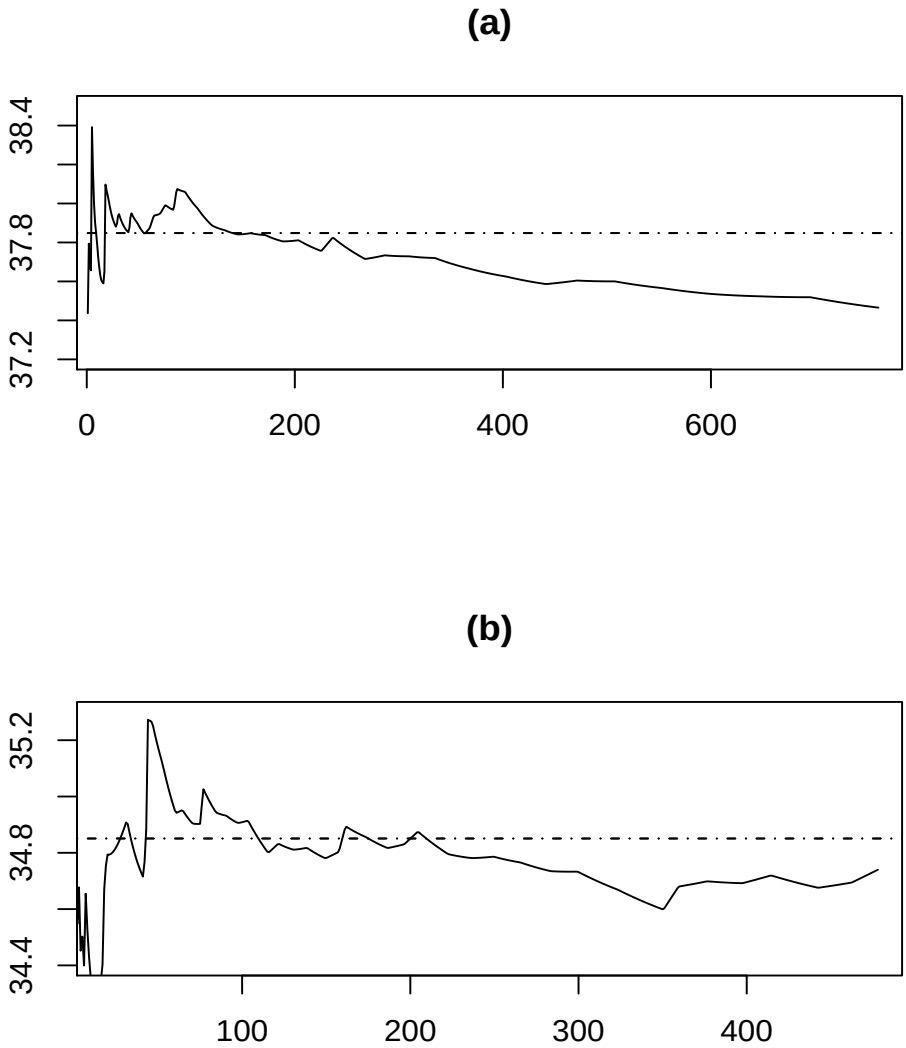

Figure 4: Estimated endpoint $\hat{x}^*$ in km/h with $\hat{\gamma}$ fixed for (a) 100m men and (b) 100m women; the dashed-dotted line represents the final choice of the endpoint.

at a single value, we again use the method of finding a first stable region. Moreover, we keep in mind that within this region the extreme-value condition must be satisfied. The precise k-regions for men and women are 110-200 and 80-210, respectively. Left-sided 95% confidence intervals can be obtained by using (6). Using the final $\hat{\gamma}$, the confidence

intervals depend on k through $\hat{x}^*$, $\hat{a}_{n/k}$ and $\sqrt{k}$. For each k in the k-region for the endpoint estimation, the one-sided confidence interval is computed; the final approximated lower bound for the confidence interval is obtained by averaging over the k-region. Table 3 shows our final results.

	current WR (time)	endpoint (speed)	endpoint (time)	conf. limit (time)
100m men	9.69	37.85	9.51	9.21
100m women	10.49	34.85	10.33	9.88

Table 3: Ultimate world records both given in speed (km/h) and time. The last column presents the lower 95% confidence limit.

The 100m for men and women still show some substantial room for improvement, respectively 0.18 seconds (1.9%) and 0.16 seconds (1.5%). The 95% confidence limits are respectively 0.48 seconds (5.0%) and 0.61 seconds (5.8%) away from the current world record. Recall that the current world records have been set outside our observation period.

5 Comparison to the literature

Our ultimate world records can be compared to those in Einmahl and Magnus (2008). In that paper all data from before 2005 were considered and the obtained ultimate world records are 9.29 seconds (men) and 10.11 seconds (women). These times are about 0.2 seconds sharper than our estimated records. The results might differ because the time horizon of the data differs. In this paper, we have excluded data from before 1991 (recall that modern doping control was introduced in 1991) and have added data from after 2005. An interesting question is whether the removal of times from before 1991 has led to the less sharp ultimate world records on the 100m. To investigate this question we consider data from 1991 to 2005.

Using the same methods as before we find estimated ultimate world records based on these data. Table 4 compares the estimated endpoints for the datasets with different time horizons. The results for the period 1991-2005 are from this section; the period 1991-2008 has been investigated in the previous sections and the results for the period before 2005 are from Einmahl and Magnus (2008). Note that we only present the point estimates. The

results seem to indicate that the doping suspicious data from before 1991 have contributed to the sharper estimated ultimate world records in Einmahl and Magnus (2008). Removing the data from before 1991 and not adding the data from 2005 on gives about the same estimate for the women as earlier in this paper and for the men we find an estimate in between the two earlier estimates; see Table 4.

period/ author	100 m men	100m women
1991-2008	**9.51**	**10.33**
1991-2005	9.37	10.36
-2005	9.29	10.11
Berthelot	9.73	10.43
Denny	9.48	10.39

Table 4: Ultimate world records.

We also briefly comment on two other recent papers, from researchers in biology/physiology. In Berthelot et al. (2008) a regression-type approach is used to find many ultimate world records in sports, including those on the 100m running. The results can be found in Table 4; for the 100m men the estimate is much less sharp than ours and actually already above the current world record. For the 100m women the estimate is somewhat less sharp than ours, but it is indicated in that paper that the suspicious 10.49 world record, set in 1988, has a large influence on the value 10.43. Deleting that world record leads to an estimate of 10.73, much higher than our estimates and above the current world record and above 10.65, the best time since 1991. In Denny (2008) an extreme-value theory related approach (quite different from ours) is followed, but not only humans, but also dogs and horses are studied. The results for the 100m can also be found in Table 4. The 9.48 for the men agrees very much with our 9.51 and so does the 10.39 (calculated excluding the 10.49 world record) for women with our 10.33.

References

[1] J. Beirlant, G. Dierckx and A. Guillou (2005), 'Estimation of the extreme-value index and regression on generalized quantile plots', *Bernoulli*, 11, 949-970.

[2] J. Beirlant, Y. Goegebeur, J. Segers and J. Teugels (2004), '*Statistics of Extremes: Theory and Applications*', Wiley, New York.

[3] G. Berthelot, V. Thibault, M. Tafflet, S. Escolano, N. El Helou, et al. (2008), 'The citius end: world records progression announces the completion of a brief ultra-physiological quest', *PLoS ONE*, 3, e1552; doi:10.1371/journal.pone.0001552.

[4] A. Dekkers, J. Einmahl and L. de Haan (1989), 'A moment estimator for the index of an extreme-value distribution', *The Annals of Statistics*, 17, 1833-1855.

[5] M. Denny (2008), 'Limits to running speed in dogs, horses and humans', *Journal of Experimental Biology*, 211, 3836-3849.

[6] D. Dietrich, L. de Haan and J. Hüsler (2002), 'Testing extreme value conditions', *Extremes*, 5, 71-85.

[7] J. Einmahl and J. Magnus (2008), 'Records in athletics through extreme-value theory', *Journal of the American Statistical Association*, 103, 1382-1391.

[8] L. de Haan and A. Ferreira (2006), '*Extreme Value Theory: An Introduction*', Springer, New York.

Dept. of Econometrics & OR and CentER
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands
j.h.j.einmahl@uvt.nl

AZL
P.O. Box 4471
6401 CZ Heerlen
The Netherlands
sander.smeets@azl.eu