MODELS FOR PROGRESSION OF RECORDS

Petr VOLF,ÚTIA AV ČR
E-mail volf@utia.cas.cz

O U T L I N E :

1. Records in case of i.i.d. random variables
2. Records as random point process with increments
3. Regression model for development of best results
4. Probability of record occurrence and increment
5. Application to light athletic data
6. Limitations of model, ideas of improvement

1 Introduction, records in i.i.d. case

Records – maximal values in a series of random variables, $X_1, X_2, \ldots, X_t, \ldots$

Record values $R_1 < R_2 < \ldots$,
their indices $t_1 < t_2 < \ldots$, $(t_1 = 1)$

Case of i.i.d. sequence X_t analyzed by many authors, e.g.
Anděl J. (2001): Mathematics of Chance. Wiley, New York:

- Probability that X_t will be the new record is $\sim 1/t$
- Sequence $\{R_j,\ j = 1, 2, \ldots\}$ behaves as a random point process with intensity $h_x(r)$,
  where $h_x(r)$ is the intensity of distribution of r.v. X_t.

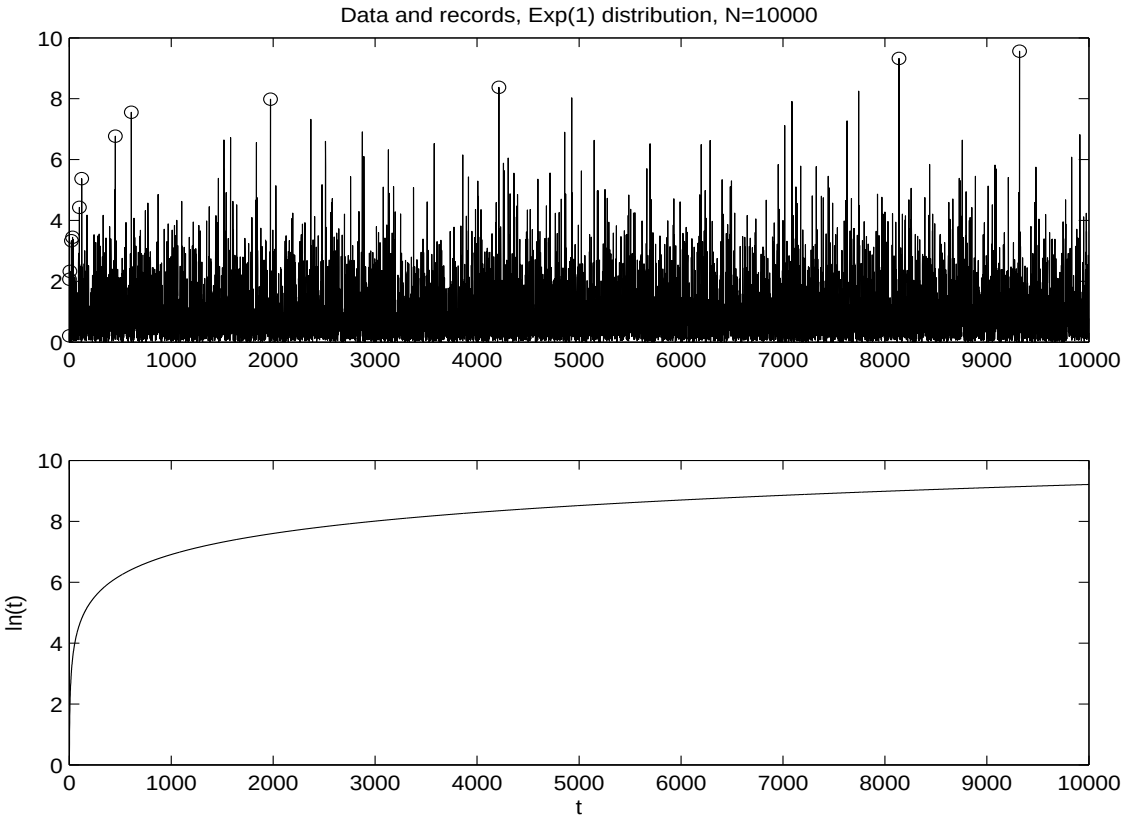

Figure 1: Example of records in sequence of i.i.d. Exp(1) random variables

However, for sports assumption of i.i.d. variables is not adequate.

First, rate of records occurrence is higher then $\sim 1/t$

Improvement (rather 'artificial') – assumption
that number of (high-quality) attempts increases, see

Noubary, R.D. (2005): A Procedure for Prediction of Sports Records, Journal of Quantitative Analysis in Sports

– geometric increase each year:
periods $t = 1, 2, ..T.$ (years) $\rightarrow$ $1, i, i^2, ..., i^{T-1}$

for long-jump men (1962-2004) $i = 1.03$, 43 years $\rightarrow$ 83 "attempts"

Noubary,F. and Noubary,R. (2004). On survival times of sports records. J. of Comp. and Applied Mathematics 169, 227-234.

– model for intensity (number) of attempts, still i.i.d. case

Second, model should reflect increasing level of sports results
(which is also due 'technological' development)

$\sim$ increase of X_t (its mean, quantiles, shift of distribution, ...)

==> more records, without assumption of large increase of number of high-quality attempts and meetings

Hence, other types of models were proposed

Next models describe directly behavior of sequence of records
(i. e. values, increments, times)

REMARK: Athletic record = maximal value (field events),
= minimal value (track events)

2. RANDOM POINT PROCESS MODEL

– describes **intensity** of new record occurrence,

methodology of analysis is borrowed from survival analysis:

Gutierrez,E., Lozano,S. and Salmeron,J.L. (2009).A study of the duration of Olympic records using survival analysis of recurrent events. In: Proceedings of 2-nd IMA Conference on Mathematics in Sports, Groningen 2009, 57-62.

Model allows to incorporate dependence of intensity on influencing factors (e.g. actual record level (relative), last increment, duration of record, seasonal components, ...)

for instance Gutierrez et al (2009) use Cox's regression model.

2.1 Compound point process model

– process of random increments at random times, formally

$$C(t) = \int_0^t Z(s)\,\mathrm{d}N(s) = \sum_{s\le t} Z(s)1[\mathrm{d}N(s) = 1].$$

$Z(s)$ are (nonnegative) random increments,

$N(s)$ is a counting process, mostly non-homogeneous Poisson

If $N(s)$ has intensity $\lambda(s)$, mean and var of $Z(s)$ are $\mu(s)$, $\sigma^2(s)$,
then mean development of $C(t)$ is given as

$$EC(t) = \int_0^t \lambda(s)\,\mu(s)\,\mathrm{d}s, \qquad \mathrm{var}\,C(t) = \int_0^t \lambda(s)\,\left(\mu^2(s) + \sigma^2(s)\right)\mathrm{d}s.$$

Frequent question: existence of finite limit value
(an ultimate record)? – at least in the mean sense.

. . . here, when both $\mathrm{E}C(t)$ and $\mathrm{var}\,C(t)$ tend to finite limits

Discrete-time version of process of increments:

– compound process changes to a Markov, **random walk model** given by:
probabilities $p(t)$ of new record occurrence (in period t)
and random variables $Z(t)$ of record improvement

Terpstra, J.T. and Schauer, N.D. (2007): A Simple Random Walk Model for Predicting Track and Field World Records, Journal of Quantitative Analysis in Sports

use logistic

$$p(t) = \frac{exp(\alpha_1 + \alpha_2 \cdot t)}{1 + exp(\alpha_1 + \alpha_2 \cdot t)}$$

and exponentially distributed $Z(t)$ with $\mathrm{E}Z(t) = \exp(\beta_1 + \beta_2 \cdot t)$,

==> negative β_2 corresponds to bounded $\mathrm{E}C(t)$, $\mathrm{var}(C(t))$.

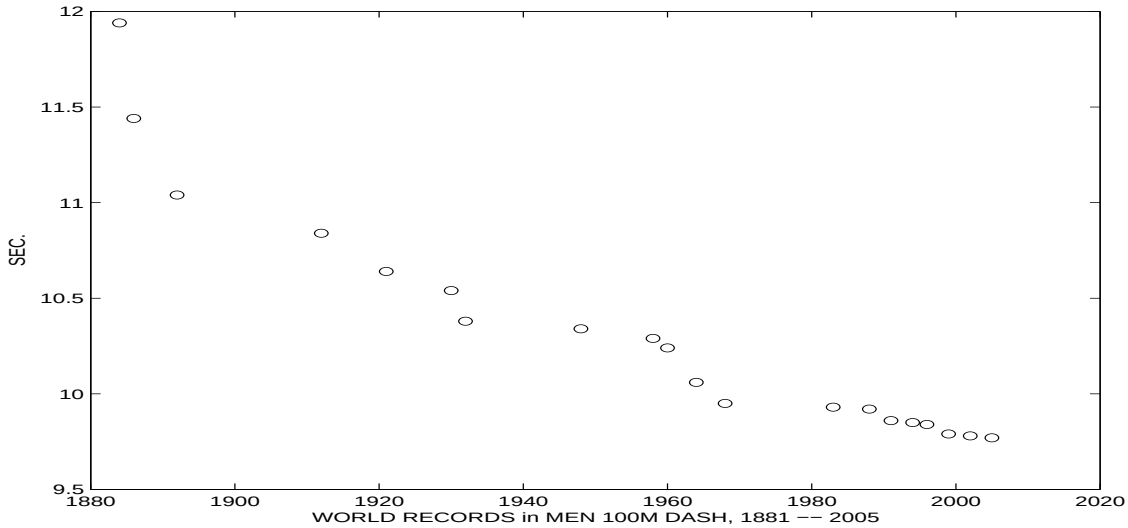

Figure 2: 100m records to 2005

Terpstra and Schauer (2007) use (rather 'nice') data of records in 100m dash men.
Results (years counted as 1884=0.01, 0.02,..., 2005=1.22):
$\alpha_1 = -2.8121, \alpha_2 = 1.7525, \beta_1 = -0.7797, \beta_2 = -2.3983.$

Example of 'not so nice' data – long jump of men,

Results (length measured in cm, years 1901=0.01,...,2008=1.08):
$\alpha_1 = -1.7571, \alpha_2 = -0.1057, \beta_1 = 2.0056, \beta_2 = 0.5032$

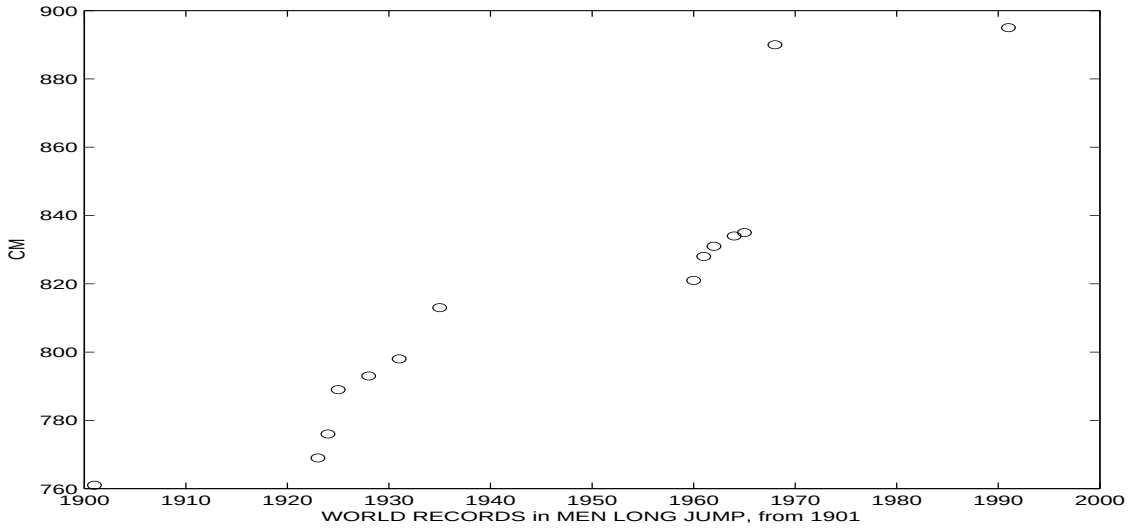

Figure 3: Long-jump records

3. MODELS FOR INCREASE OF PERFORMANCE

Use of more data than just records
– the best (or K best) results of each year

- nonlinear regression (on time)
- time series, dynamic models (& Bayes?)

- - - - - - - - - -

Regression: choice of trend and of error distribution

TREND functions:

- Linear function for local data fitting,
- Exponential-decay function $A + B \cdot \exp(at)$, $a < 0, A > 0$, and $B < 0$ for track events

  (– and similar curves)

- S-shaped curves, for instance Gompertz curve:

$$m(t) = a + b \exp\{-\exp(c(t-d))\},$$

  with $c > 0$, then limit $m(\infty) = a + b$,
  $b < 0$ yields decreasing curve, inflexion is at $t = d$,
  (limit $m(-\infty) = a$)

Distribution of errors:

- Normal
- Gumbel
- Generalized Extreme Value:

$$F(x) = 1 - \exp\{-[1 + k(x - \mu)/\delta]^{1/k}\},$$

for x: $[.] > 0$, $\delta > 0, k \neq 0$.

Selected references:

Smith, R.L. (1988): Forecasting records by maximum likelihood. J.A.S.A. 83, 331388.

Kuper, G.H. and Sterken,E. (2006): Modelling the development of world records in running. CCSO Working paper 2006/04, Univ. of Groningen.

3.1 My suggestion of REGRESSION MODEL

for 1 best result of each year, with exponential-decay trend, log-normal errors, time-dependent variance:

$X(t)$ – the best year result at year t, $t = 1, \ldots, T$,

$Y(t) = lnX(t)$ for field events, $Y(t) = -lnX(t)$ for track events,

$$Y(t) = m(t) + \sigma(t) \cdot \varepsilon(t),$$

where $\varepsilon(t)$ are i.i.d. $\mathcal{N}(0, 1)$,

$$m(t) = A + B \cdot e^{at}, \quad \sigma(t) = C + D \cdot e^{bt}$$

so that $a, b < 0$ ensure $\mathrm{E}Y(t) \to A$, $\sigma(t) \to C$ for $t \to \infty$

For fixed a, b the rest of model is linear,
– standard (weighted LSE and MLE) methods are used
for estimation of parameters A,B and C,D, resp.

4. PROCESS OF RECORDS

Let variables $Y(t)$ have distributions with cdf, density $F_t(y), f_t(y)$.

Let R be actual record (after year t). Then the probability that new record occurs in year $t+k, k = 1, 2, ..)$ is

$$p(k, t, R) = \prod_{j=1}^{k-1} F_{t+j}(R) \cdot (1 - F_{t+k}(R)),$$

new record level is then given by probability density

$$g_k(r, t, R) = \frac{f_{t+k}(r)}{(1 - F_{t+k}(R))}, \quad \text{for } r > R,$$

4.1 Records as Markov chain:

Again, let actual record be R_t at time t.
Then probability $P(R_{t+1} = R_t) = F_{t+1}(R_t)$,
transition to new record $r > R_t$ is given by density $f_{t+1}(r)$.

PREDICTION based on this Markov scheme:

Assume that data are given and model evaluated up to T

Trend of $Y(t)$ (=model) can be extrapolated to $t > T$

We generate, year by year, random trajectories of the Markov process of records described above, starting from value R_T at T

From a set of such trajectories, sample characteristics of future process of records can be computed, e.g. means, variances, quantiles (both of number of new records and of record improvement)

5. ANALYSIS OF DECATHLON DATA

The series of world records from 1920 can be found for instance in materials of IAAF on its Web

We used data from 1950, however, best year marks before 1974 is hard to find, therefore a part of data has been prepared artificially:

Missing best results were created by one step of the EM algorithm:

$\hat{Y(t)} = E(Y(t)|Y(t) < R_t)$, where R_t is actual record at t,

– for $Y(t) = ln(X(t))$.

year	mark	year	mark	year	mark	year	mark
1950	7287	1974	8229	1986	8811	1998	8755
1952	7582	1975	8429	1987	8680	1999	8994
1955	7608	1976	8634	1988	8512	2000	8900
1958	7989	1977	8400	1989	8549	2001	9026
1959	7839	1978	8493	1990	8574	2002	8800
1960	7981	1979	8476	1991	8812	2003	8807
1963	8010	1980	8667	1992	8891	2004	8893
1966	8120	1981	8334	1993	8817	2005	8732
1967	8235	1982	8774	1994	8735	2006	8677
1969	8310	1983	8825	1995	8695	2007	8697
1972	8466	1984	8847	1996	8824	2008	8832
		1985	8559	1997	8837	2009	8790

Table 1: World records and best year marks, decathlon men, from 1950.

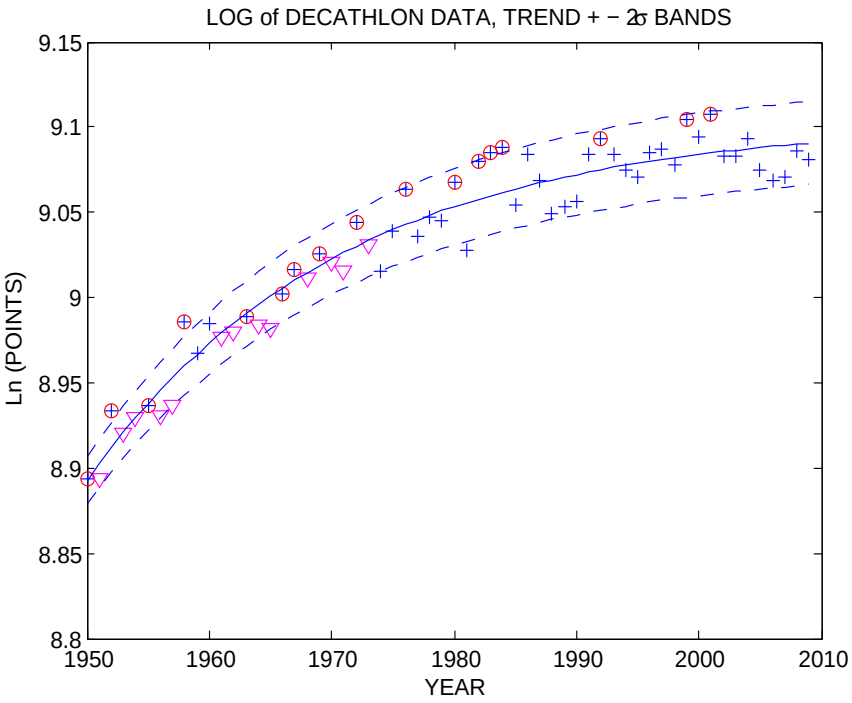

Figure 4: Log of decathlon best results with trend $\pm 2\sigma(t)$

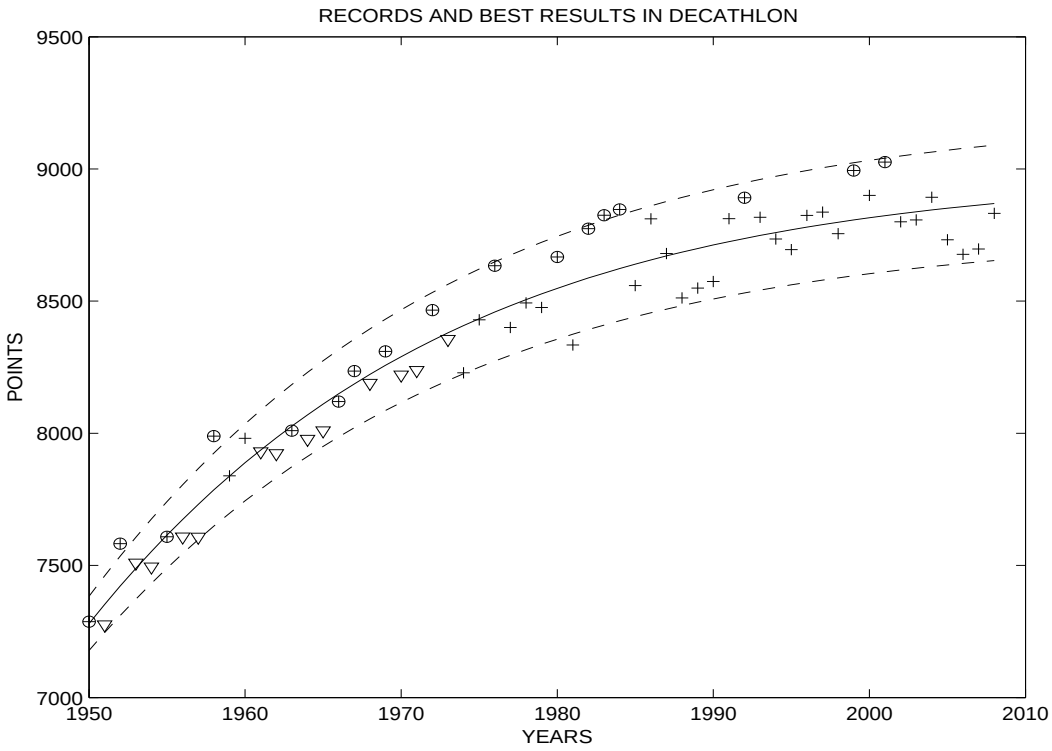

Figure 5: Decathlon best results with trend $\exp(m(t) \pm 2\sigma(t))$

Optimal values of parameters of model were

$$A = 9.1045\,(0.0048),\ \ B = -0.2203\,(0.0094),$$

$$C = 0.0127\,(0.0023),\ \ E = D/C = -0.4861\,(0.0968),$$

$$a = -0.047\,(0.0020),\ \ b = -0.050\,(0.0073),$$

half-widths of 95% asymptotic confidence intervals are in parentheses

Limit distribution of $X(t)$ is lognormal with $\mu = A, \sigma = C$

– Such distribution is almost symmetric,

$\mathrm{E}X \sim 8996$, $\mathrm{median}(X) = exp(A) \sim 8996$, $\mathrm{std}(X) \sim 114$

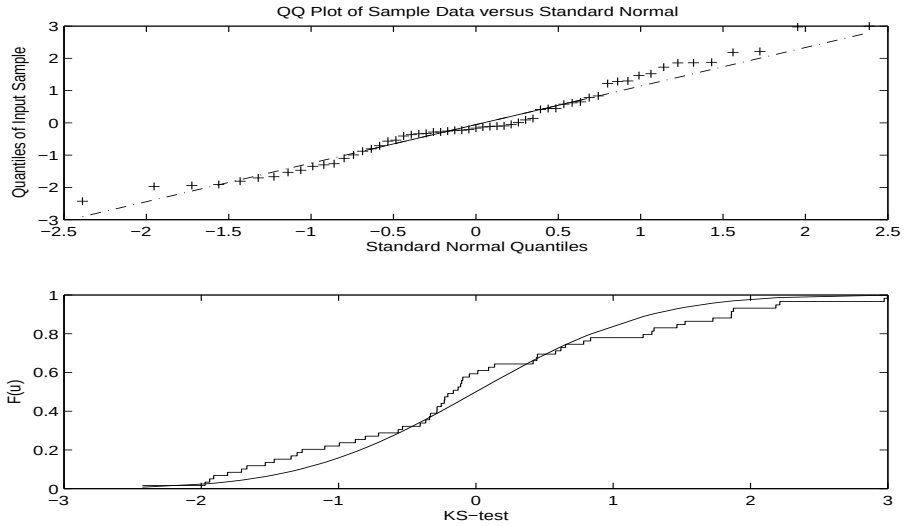

Figure 6: QQ-plot for model of decathlon data (upper tail seems to be wider than Gauss)

KS-test:

max abs difference: 0.0987, approx. crit. value (n=60): 0.1753

Tests of independence of errors, P-values: 0.44, 0.83

(series above and bellow median, series up and down)

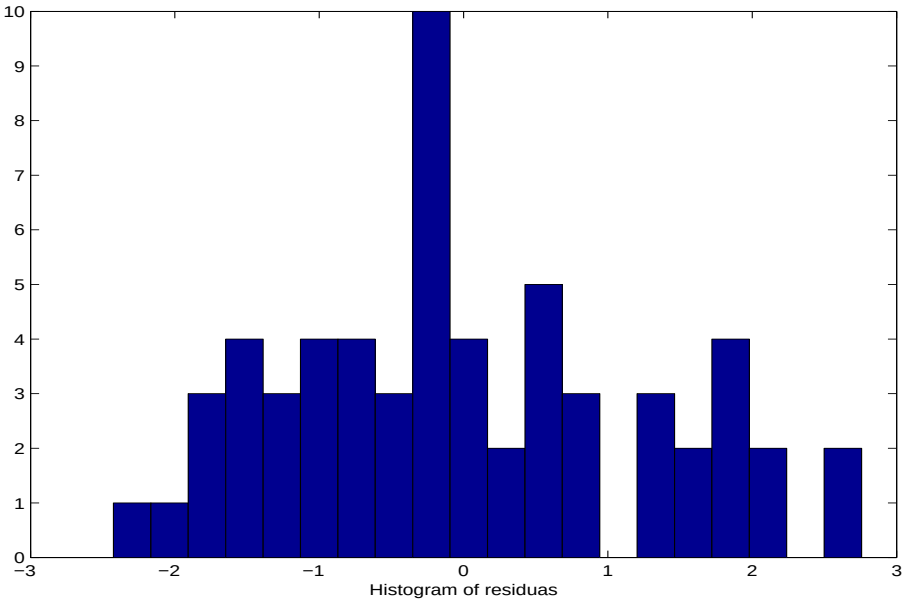

Figure 7: Histogram of residuas (in model for $Y(t)$)

Prediction:

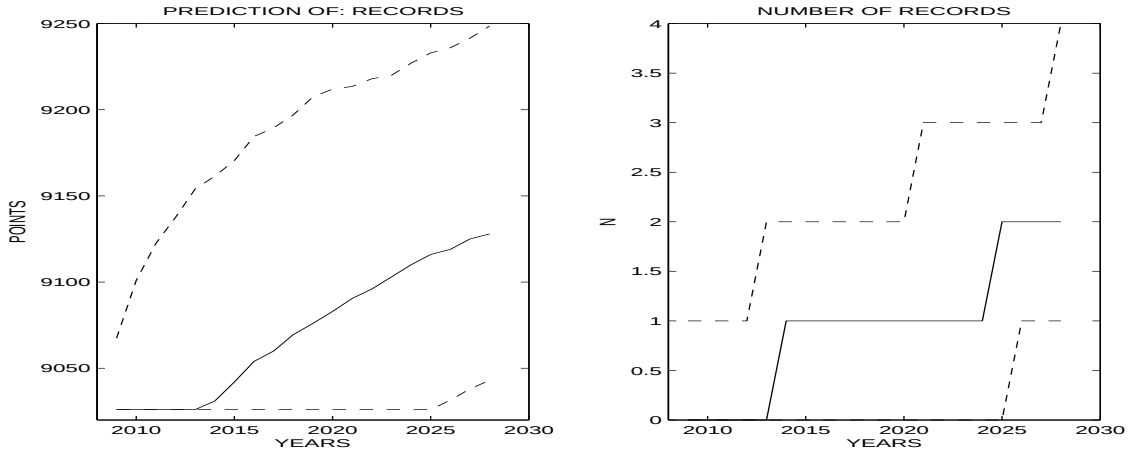

Figure 8: Prediction of record development (left) and number of records (right): medians, 5% and 95% quantiles, results from 1000 Markov chain randomly generated paths, starting from 2008 with actual record $R = 9026$ points (of R. Šebrle, from 2001). It suggests that actual record has chance about 0.5 to be improved before 2015, with value about 9050 points

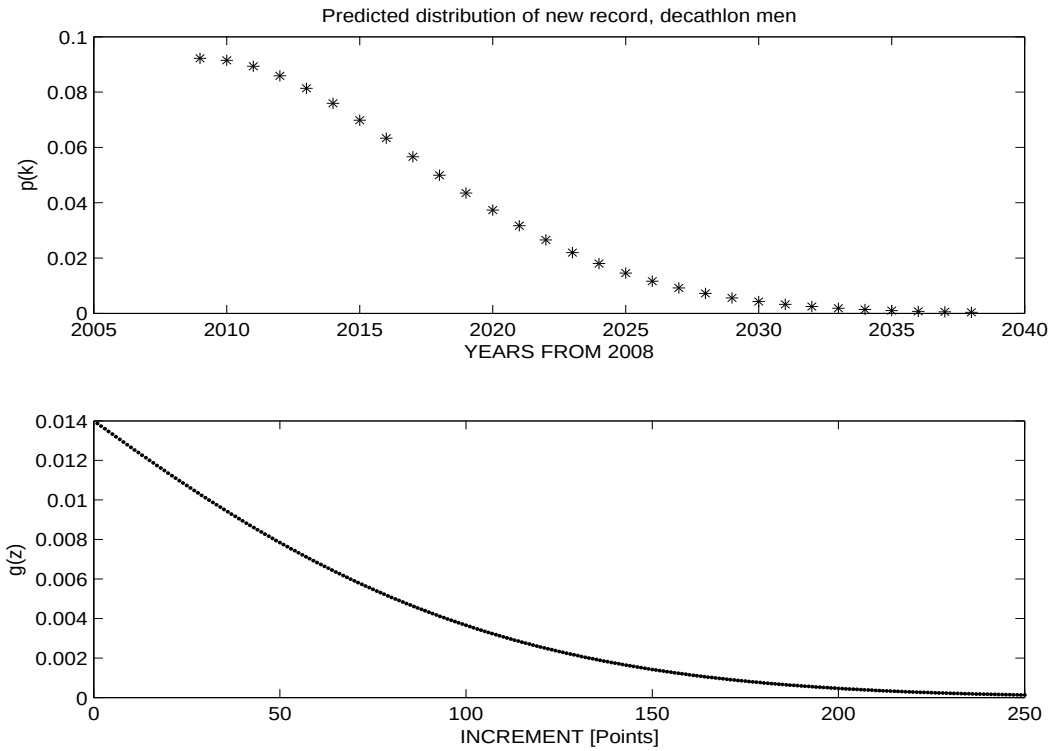

Figure 9: Probability distributions of new record year $p(k, R)$ (above) and record improvement density $g(z, R)$ (below) – it looks like exponential distribution with mean ~ 57, median ~ 40

2 Model limitations – demonstrated on data:

A) 100m dash men

year	mark	year	mark	year	mark	year	mark
1884	11.94	1972	10.07	1986	10.02	1998	9.86
1886	11.44	1975	10.05	1987	9.93	1999	9.79
1892	11.04	1976	10.06	1988	9.92	2000	9.86
1912	10.84	1977	9.98	1989	9.94	2001	9.82
1921	10.64	1978	10.07	1990	9.96	2002	9.78
1930	10.54	1979	10.07	1991	9.86	2003	9.93
1932	10.38	1980	10.02	1992	9.96	2004	9.85
1948	10.34	1981	10.00	1993	9.87	2005	9.77
1958	10.29	1982	10.00	1994	9.85	2006	9.77
1960	10.24	1983	9.93	1995	9.91	2007	9.74
1964	10.06	1984	9.96	1996	9.84	2008	9.69
1968	9.95	1985	9.98	1997	9.86	2009	9.58

Table 2: World records and best year marks, 100m dash men.

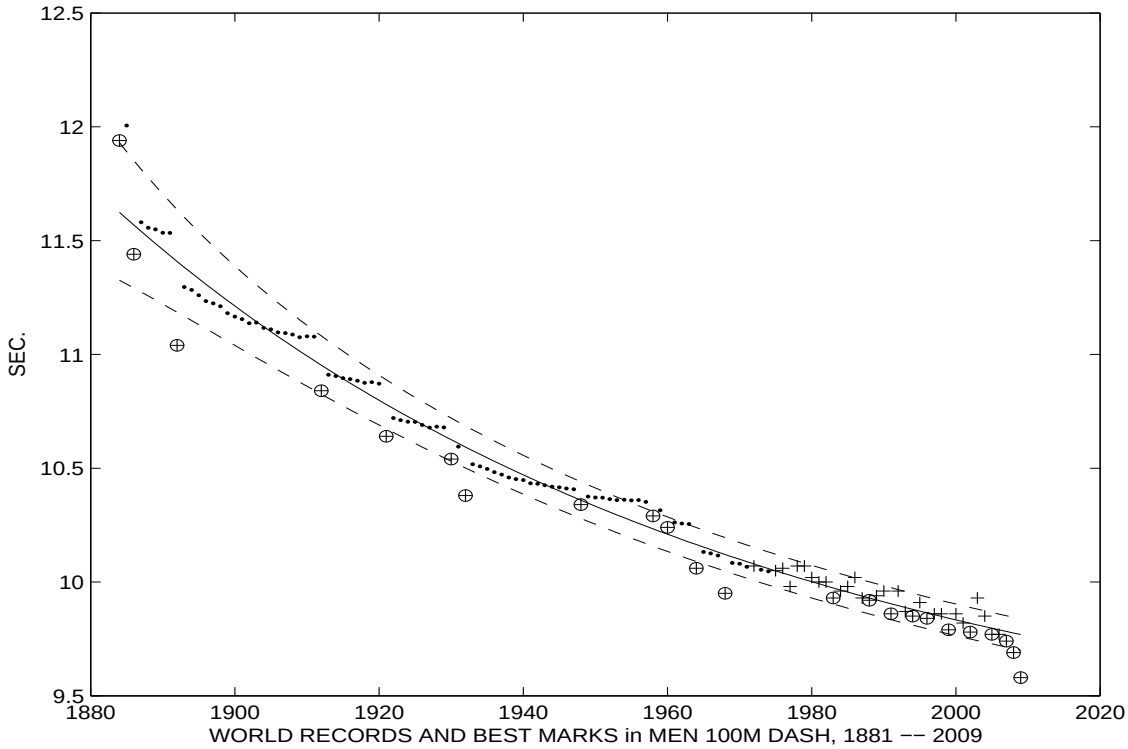

Figure 10: 100m dash men data with trend $\exp(m(t) \pm 2\sigma(t))$
(compare electronically and manually measured times)

B) Long jump men, the same analysis:

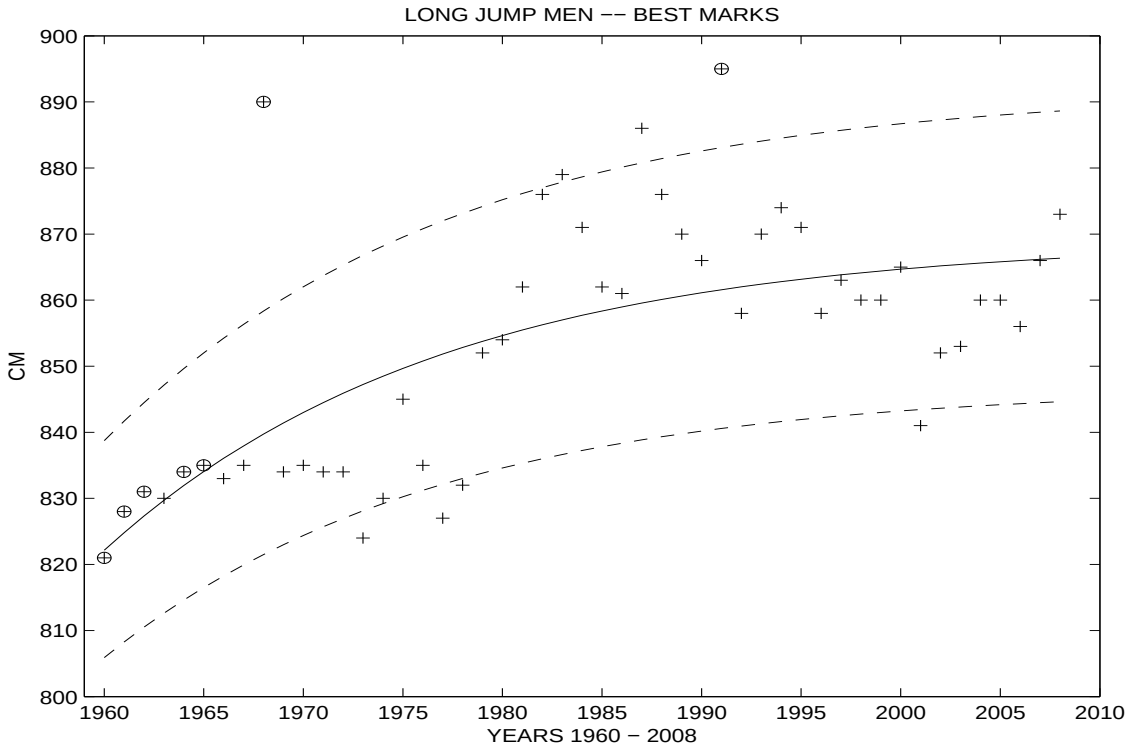

Figure 11: Long-jump data with trend $\exp(m(t) \pm 2\sigma(t))$

year	mark	year	mark	year	mark	year	mark
1960	821	1973	824	1986	861	1998	860
1961	828	1974	830	1987	886	1999	860
1962	831	1975	845	1988	876	2000	865
1963	830	1976	835	1989	870	2001	841
1964	834	1977	827	1990	866	2002	852
1965	835	1978	832	1991	895	2003	853
1966	833	1979	852	1992	858	2004	860
1967	835	1980	854	1993	870	2005	860
1968	890	1981	862	1994	874	2006	856
1969	834	1982	876	1995	871	2007	866
1970	835	1983	879	1996	858	2008	873
1971	834	1984	871	1997	863	2009	874
1972	834	1985	862				

Table 3: World records and best year marks, long jump men, from 1960.

Results for 100m:

$A = -2.2094\,(0.0024)$, $B = -0.2461\,(0.0049)$, $C = 0.0035\,(0.0004)$,
$E = D/C = 2.8565\,(0.0824)$, $a = -0.011\,(0.0009)$, $b = -0.050\,(0.0018)$.
$\mathrm{E}X_\infty = 9.1104$, median $= exp(-A) = 9.1103$, $\mathrm{std}(X_\infty) = 0.0319$.

Results for long-jump:

$A = 6.7674\,(0.0054)$, $B = -0.0589\,(0.0125)$, $C = 0.0130\,(0.0026)$, $E =$
$D/C = -0.2422\,(0.1047)$, $a = -0.060\,(0.0117)$, $b = -0.050\,(0.00144)$.
$\mathrm{E}X_\infty = 862.12$, median $= exp(A) = 869.05$, $\mathrm{std}(X_\infty) = 11.30$.

Used data sources:

http://www.alltime-athletics.com/

http://en.wikipedia.org/wiki/World_record_progression_long_jump_men

http://www.iaaf.org