
SPORTEREDMÉNYEK MODELLEZÉSE EXTRÉM ÉRTÉK MODELLEKKEL

Szakdolgozat

Szerző: *Csáfordi József András*

Matematika MSc

Sztochasztika szakirány

Témavezető: *Dr. Zempléni András*

Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar



Budapest, 2025

Tartalomjegyzék

1. Bevezetés	3
2. Matematikai bevezető	6
2.1. Extrém érték elmélet	7
2.1.1. Illeszkedésvizsgálat	10
2.2. Simítási módszerek	12
2.2.1. Köbös spline	12
2.2.2. Regularizált log-likelihood	12
2.2.3. Fisher scoring	13
2.2.4. Akaike információs kritérium	15
3. Rekordok	16
4. Nem paraméteres trend	19
4.1. Nem paraméteres trend az átlagban	19
4.1.1. Regularizációs módszerek a simaság érdekében	20
4.1.2. Simítás normális eloszlásra	21
4.2. Kezdőpontok és simítási paraméter kiválasztása	23
4.2.1. Kezdőpontok kiválasztása	23
4.2.2. Fisher Scoring módosítása	24
4.2.3. Simítási paraméter kiválasztása	24
4.3. Alkalmazás	25
4.3.1. Férfi 800 méteres síkfutás	25
4.3.2. Férfi 1500 méteres síkfutás	27
5. Pontfolyamatok	29
5.1. Stacionárius Poisson folyamat	29
5.2. Tolerancia sávok felépítése	31
5.3. Trend Becslése További Adatok Alapján	33
6. Kétváltozós szélsőérték modell	35
6.1. Rendezett szélső értékek	35
6.2. Kétváltozós extrém érték simítás	36
7. Összefoglalás	39
Hivatkozások	40

1. Bevezetés

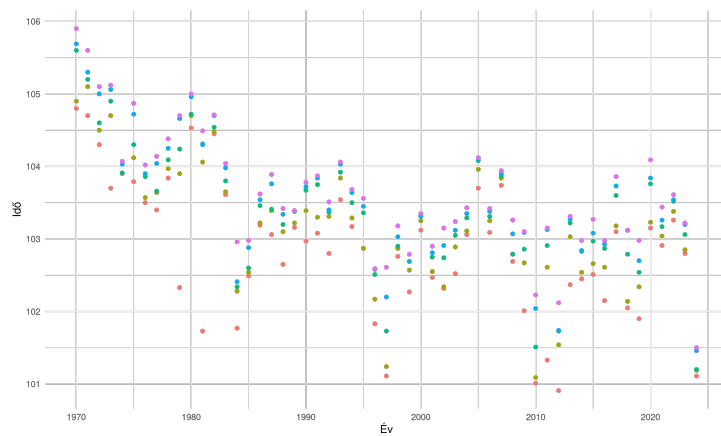
A szélsőértékek mindig is különös figyelmet kaptak az emberi gondolkodásban. Legyen szó a leggyorsabb futásokról, a legnagyobb földrengésekről vagy a legmagasabb hőmérsékletekről. A szélsőségek nemcsak kíváncsiságot keltenek, hanem komoly gyakorlati jelentőséggel is bírnak. A szélsőérték-elmélet egy olyan matematikai eszköztárat kínál, amely lehetővé teszi, hogy megértsük, modellezzük és előrejelezzük az extrém események viselkedését, legyenek azok természetes vagy emberi eredetűek. Ezen elmélet alapjait Emil Julius Gumbel [1] 1958-ban, Maurice Fréchet 1927-ben és Waloddi Weibull 1939-ben írt publicisztikáik fektették le. Stuart Coles [6] műve alapvető elméleti keretet biztosít az extrém értékek elméletének megértéséhez, továbbá részletesen összefoglalja a 20. század legjelentősebb eredményeit és áttöréseit ezen a tudományterületen.

A futószámok elemzése különösen izgalmas területet kínál a szélsőérték-elmélet alkalmazására. A modern technológia révén ma már hatalmas mennyiségű adat áll rendelkezésünkre az egyéni sportteljesítményekről, például a maratoni futások idejéről vagy a sprinterek maximális sebességéről. Ezek az adatok nemcsak az emberi teljesítőképesség határainak feltérképezésében segítenek, hanem arra is választ adhatnak, hogy mennyire valószínű egy új világcsúcs megszületése, vagy milyen körülmények között érhetünk el eddig elképzelhetetlen eredményeket.

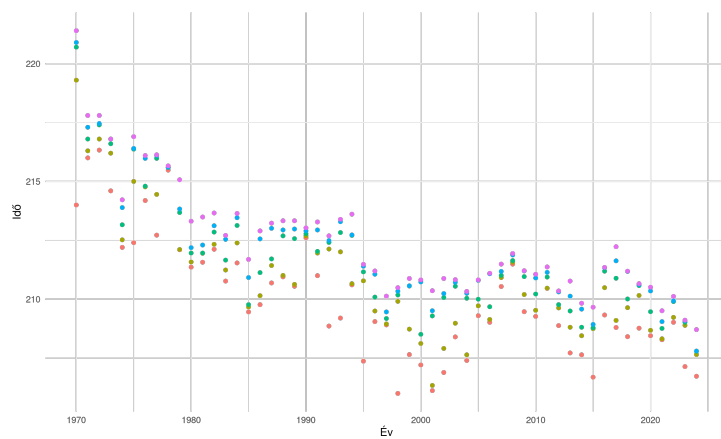
Az extrém értékek vizsgálata lehetőséget nyújt a szélsőséges eseményeket befolyásoló tényezők azonosítására és modellezésére, különös tekintettel az időbeni változásokra. Ennek keretében vizsgálható például, hogyan változnak a szélsőséges teljesítmények az idő múlásával, illetve milyen trendek és szezonális hatások érvényesülnek az adott folyamatokban. Az időbeli dinamikák modellezése kiemelt jelentőséggel bír a sporttudományban, hiszen hozzájárulhat a teljesítményváltozások pontosabb megértéséhez, valamint a sportstratégiák és edzésprogramok hatékonyabb kialakításához.

Ez a tanulmány a szélsőérték-elmélet alapjaira építve vizsgálja a futószámok extrém eredményeit. Célunk nemcsak az eddigi adatok elemzése, hanem annak megértése is, hogy miként használhatjuk fel a szélsőérték-elméletet a jövőbeli teljesítmények modellezésére és előrejelzésére. Az eredmények nemcsak a sporttudományt gazdagíthatják, hanem szélesebb körű betekintést is nyújthatnak abba, hogyan kezelhetjük az extrém eseményeket más területeken is.

Munkám Mohd Bakri Adam 2016-ban megjelent *Modelling record times in sport with extreme value methods* [14] cikkére építkezik az ott bemutatott módszerek részletesebb kifejtése, illetve Adam korábbi meg nem jelent [16] doktori disszertációjának megértése. Célunk, hogy az ott szereplő elemzéseket pár év távlatából megvizsgáljuk az 1500 és 800 méteres férfi síkfutások adatain, amiknek az éves szinten első 5 legjobb eredményeit az {1}. és {2}. ábrákon láthatók.



1. ábra. Férfi 800 méteres síkfutás 5 legjobb eredménye éves szinten.



2. ábra. Férfi 1500 méteres síkfutás 5 legjobb eredménye éves szinten.

A dolgozat kezdeti részében átfogó matematikai bevezetést nyújtunk, amelyben részletesen definiáljuk azokat az alapfogalmakat és elméleti konstrukciókat, amelyek elengedhetetlenek a dolgozat további részeinek megértéséhez. Kiemelten foglalkozunk az extrém érték elméletével minimum adatok tekintetében, továbbá bemutatunk különféle regularizációs technikákat a becslések simításának elősegítése érdekében. A (3). fejezetben a rekordfolyamatokat vizsgáljuk, különös hangsúlyt helyezve a rekordértékek időbeli fejlődésének modellezésére. A (4). fejezetben nemparaméteres statisztikai modelleket ismertetünk, valamint bemutatjuk ezek alkalmazását az előzetesen tárgyalt adathalmazokon. Ezt követően részletesen tárgyaljuk az általános Poisson-folyamatot, amelyhez a becslési pontosság növelése érdekében toleranciasávokat definiálunk, majd a folyamat trendjének megbízható becslésére törekszünk további adatfelhasználás segítségével. A dolgozat utolsó fejezetében a kétváltozós szélsőérték modellezésre fókuszálunk, ahol egymástól független mintahalmazok segítségével végzünk elemzéseket.

2. Matematikai bevezető

Ebben a fejezetben összefoglaljuk a téma megértéséhez szükséges definíciókat, fogalmakat és tételeket. Ezeket az alapokat a dolgozatban nem bizonyítjuk, de számos dolgozat, illetve matematikai bevezető tanulmányban fellelhető, ahol részletesebben végigveszik ezen fogalmakat.

A statisztika területére lépve az egyik első alapfogalom, amellyel találkozunk, a likelihood függvény, valamint az ehhez kapcsolódó maximum likelihood becslés. Ezen fogalmak a dolgozat során többször elő fognak kerülni, így indokolt, hogy definiáljuk.

Definíció 2.1 (*Likelihood-függvény*). Ha az $(Y_1, \dots, Y_n)$ független minta diszkrét (a lehetséges értékeinek a száma véges vagy megszámlálhatóan sok), akkor a likelihood függvénye:

$$L_{n,\vartheta}(k_1, \dots, k_n) = \prod_{j=1}^n \mathbb{P}_{j,\vartheta}(Y_j = k_j) \quad (k_1, \dots, k_n \in \mathbf{H}),$$

ahol $\mathbf{H}$ a diszkrét valószínűségi változók értékkészlete. Ha az $(Y_1, \dots, Y_n)$ független minta abszolút folytonos, és Y_j sűrűségfüggvénye (a P_ϑ valószínűség mellett) $f_{j,\vartheta}$, akkor a minta likelihood-függvénye:

$$L_{n,\vartheta}(t_1, \dots, t_n) = \prod_{j=1}^n f_{j,\vartheta}(t_j) \quad (t_1, \dots, t_n \in \mathbb{R}).$$

A klasszikus likelihood-elemzés feltevése, hogy a minta független és azonos eloszlású. Ugyanakkor a valós adatok megfigyelése között lehetnek időbeli trendek, autokorrelációk vagy változó eloszlású folyamatok, így a klasszikus iid feltételezés nem teljesül. Ilyen esetekben szükségessé válhat nemparaméteres trendkövetés.

Definíció 2.2 (*Maximum likelihood becslés*). A ϑ maximum likelihood-becslése (ML-becslése) az $X_1, \dots, X_n$ mintából $\hat{\vartheta}$, ha $\hat{\vartheta}$ maximalizálja a $\vartheta \mapsto L_{n,\vartheta}(X_1, \dots, X_n)$ függvényt, ahol $L_{n,\vartheta}$ a minta likelihood-függvénye. Azaz, ha

$$L_{n,\hat{\vartheta}}(X_1, \dots, X_n) \geq L_{n,\vartheta}(X_1, \dots, X_n) \quad \text{minden } \vartheta \in \Theta\text{-ra.}$$

Definíció 2.3 (*Sztocasztikus rendezés*). X sztochasztikusan kisebb, mint Y , $X < Y$, ha $F_Y(a) \leq F_X(a)$, minden $a \in \mathbb{R}$.

A sztochasztikus rendezéssel a 6. fejezetben találkozunk újra. Ennek a definíciónak köszönhetően tudjuk vizsgálni a kétváltozós szélső érték modelleket.

A rekordfolyamatok vizsgálata során gyakran előfordul, hogy nem minden extrém érték figyelhető meg teljes mértékben. Ilyen esetekben cenzorált adatokról beszélünk, amelyek

elemzéséhez speciális statisztikai módszerekre van szükség. Az egyik legismertebb ilyen eszköz a Kaplan-Meier becslés.

Definíció 2.4 (*Kaplan–Meier becslés*). Legyen S_t túlélésfüggvény, t_i az időpont, ahol legalább egy esemény bekövetkezett, d_i az események száma a t_i időpontban, és n_i a túlélő egyedek száma a t_i időpontban. A Kaplan-Meier becsléssel S_t -t tudjuk becsülni.

$$\hat{S}_t = \prod_{i:t_i < t} \left(1 - \frac{d_i}{n_i}\right)$$

2.1. Extrém érték elmélet

Az extrém érték elméletéhez kapcsolódó legjelentősebb áttörések a 20. század közepére tehetőek, amikor is Emil Julius Gumbel 1958-ban publikálta *Statistics of Extremes* [1] című könyvét, melynek fő alapja Fischer, Gnedenko és Tippett extrém érték témakörben végzett publikációik. Mivel ebben az időszakban a kutatások célja nem kifejezetten a sporteredmények voltak, hanem általános extrém érték elemzési kérdések, így a vizsgálatok elsősorban maximális értékekre koncentráltak. Ugyanakkor érdemes megjegyezni, hogy számos sportágban épp a maximális eredményt elérő versenyző lesz a győztes (például atlétikai dobószámok). Ugyanakkor lehetséges minimum adatok elemzése, maximum adatokra vonatkozó módszerek alkalmazásával, hiszen egy $X_1, X_2 \dots$ valószínűségi változók sorozatára az alábbi egyenlőség teljesül:

$$\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n),$$

így az előjel váltással megengedhető lesz, hogy maximális adatokat elemezzünk a megadott módszerek alkalmazásával.

Legyenek $X_1, X_2, \dots$ független, azonos eloszlású valószínűségi változók, F_x eloszlásfüggvénnyel. Legyen $\{m_n\}$ az x_i -n vett minimumok sorozata ($m_n = \min(X_1, \dots, X_n)$). Ha létezik $\{a_n\} > 0$, $\{b_n\}$ konstans sorozat úgy, hogy

$$Pr \left(\frac{m_n - b_n}{a_n} \leq x \right) \longrightarrow H(x), \text{ ahogy } n \longrightarrow \infty,$$

ahol $H(x)$ nem elfajuló eloszlás, akkor $H(x)$ az alábbi eloszlások egyikének affin transzformáltjába fog esni:

- 1-es típusú (vagy Gumbel-eloszlás):

$$H(x) = 1 - \exp(-\exp(x))$$

- 2-es típusú (vagy negatív Fréchet-eloszlás):

$$H_\alpha(x) = \begin{cases} 1 - \exp(-(-x)^{-\alpha}), & x < 0 \\ 1, & x \geq 0 \end{cases}$$

- 3-as típusú (vagy Weibull-eloszlás):

$$H_\alpha(x) = \begin{cases} 1 - \exp(-x^\alpha), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Az eloszlásokban szereplő α paraméterre feltételezzük, hogy $\alpha > 0$, amely a modell értelmezhetőségéhez szükséges. Az alábbi három eloszlást extrém érték eloszlásoknak nevezzük minimum értékekre. Az 1-es típusú eloszlást Emil Julius Gumbel-ről nevezték el, aki a már említett könyvében foglalkozott vele, míg a 2-es típusú eloszlást Maurice Fréchet 1927-ben, a 3-as típusú eloszlást pedig Waloddi Weibull publikálta 1939-ben, akikről később az eloszlások az elnevezésüket kapták.

Az általánosított extrém érték eloszlás minimum adatokra GEV_{min} egyesíti a három különböző eloszlást és affin transzformáltjaikat egy 3-paraméteres eloszláscsalád formájában.

Definíció 2.5. *Általános extrém érték eloszlás eloszlásfüggvénye minimum adatokra. (GEV_{min})*

$$H(x) = 1 - \exp \left\{ - \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \quad (1)$$

ahol $\mu \in (-\infty, \infty)$ helyparaméter, $\sigma \in (0, \infty)$ skálaparaméter, és $\xi \in (-\infty, \infty)$ alakparaméter.

Az alakparaméter ξ meghatározza az X változó lehetséges értékhelyét a GEV_{min} eloszlásfüggvény alapján a következők szerint:

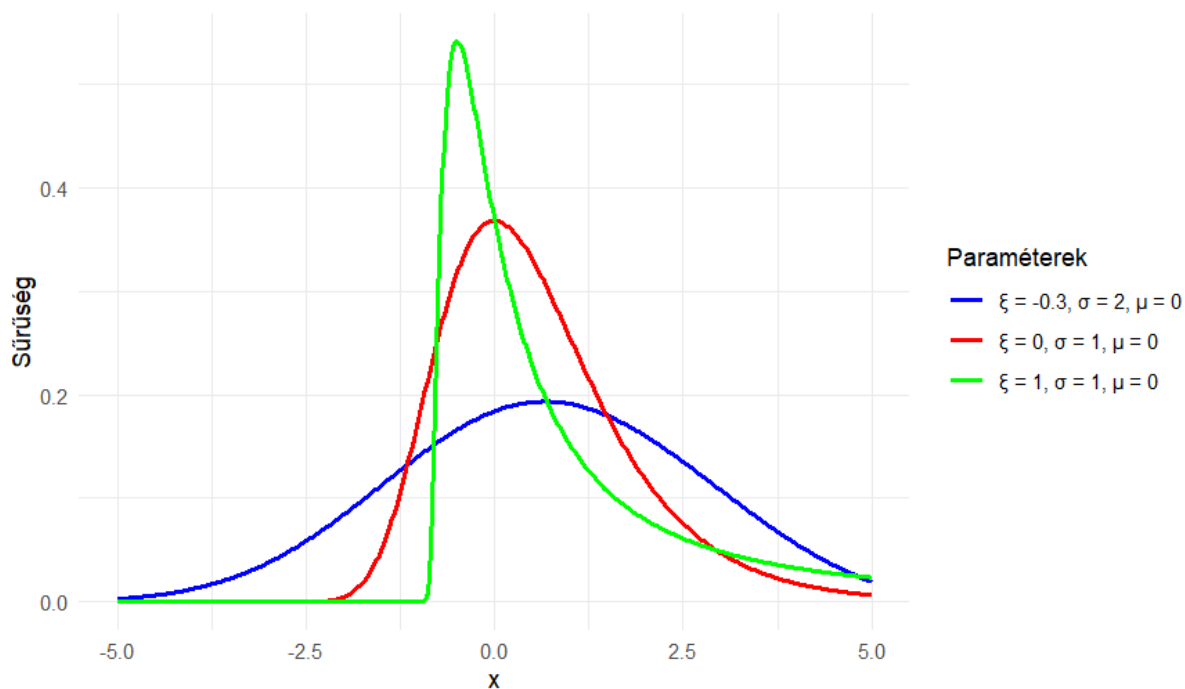
$$\xi > 0\text{-ra,} \quad -\infty < X \leq \mu + \frac{\sigma}{\xi}$$

$$\xi < 0\text{-ra,} \quad \mu + \frac{\sigma}{\xi} \leq X < \infty$$

$$\xi = 0\text{-ra,} \quad -\infty < X < \infty$$

GEV_{min} sűrűségfüggvénye:

$$h(z) = \frac{1}{\sigma} \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1-1/\xi} \exp \left\{ - \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}.$$



3. ábra. Különböző paraméterekkel generált extrém érték eloszlás maximum adatokra a három különböző típusú eloszlásra.

Tegyük fel, hogy az $\{X_i\}_{i=1}^n$ megfigyelési sorozat egy idősort alkot, amelyet egymást követő egyenlő hosszúságú blokkokra osztunk. Legyen m egy pozitív egész szám, amely a blokkok hosszát jelöli, azaz minden blokk m darab X_i értéket tartalmaz. Ekkor a teljes sorozat $n = \lfloor N/m \rfloor$ darab blokkból áll, és ezekhez a blokkokhoz rendeljük hozzá a minimális értékeket:

$$Z_i = \min(X_{(i-1)m+1}, \dots, X_{im}), \quad \text{ahol } i : 1 \leq i \leq n.$$

Ezek a blokkok a gyakorlatban az éves szinten legjobb sporteredményeket is jelenthetik, amikor minden év m darab megfigyelést tartalmaz. A Z_i minimális értékeket tartalmazó sorozat $\{Z_1, \dots, Z_n\}$ feltételezhetően független, azonos eloszlású mintát képez, amelyre alkalmazható a GEV_{min} modell. A paraméteres maximum likelihood becsléshez a log-likelihood-függvény a következő formában írható fel:

$$l(z : \mu, \sigma, \xi) = -n \log \sigma - (1 + 1/\xi) \sum_{i=1}^n \log \left[1 - \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]_+ - \sum_{i=1}^n \left[1 - \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]_+^{-1/\xi}$$

A log-likelihood függvény $\xi = 0$ esetében úgy kapható meg, ha $l(z : \mu, \sigma, \xi)$ határértékét

vesszük $\xi \rightarrow 0$ mellett. Ekkor azt kapjuk, hogy:

$$l(z : \mu, \sigma) = -n \log \sigma - \sum_{i=1}^n \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^n \exp \left(\frac{z_i - \mu}{\sigma} \right)$$

Coles 2001-es *An introduction to statistical modeling of extreme values* [6] című könyvében részletesebben is tárgyalja az extrém érték elemzés elméleti alapjait és gyakorlati alkalmazásait. Az ő munkája széles körben elfogadott referenciaként, ezért hivatkozása indokolt a téma mélyebb megértéséhez.

2.1.1. Illeszkedésvizsgálat

Egyszerű eszközökkel mint a Q-Q, illetve a P-P plotok könnyen felmérhetjük, hogy a $\text{GEV}_{\min}$ eloszlás mennyire illeszkedik az éves minimum adatokra. A P-P plot a következőképpen készíthető. z tapasztalati eloszlásfüggvénye a következő:

$$\widetilde{H}(z_{(i)}) = \frac{i}{n+1}, \quad i = 1, \dots, n,$$

ahol $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$ a sorba rendezett mintánk. H megfelelő modell alapú becslései z -nél:

$$\hat{H}(z_{(i)}) = 1 - \exp \left\{ - \left[1 - \hat{\xi} \left(\frac{z_i - \hat{\mu}}{\hat{\sigma}} \right) \right]_+^{-1/\hat{\xi}} \right\}, \quad i = 1, \dots, n, \quad (2)$$

ahol a $\hat{\xi}$, $\hat{\mu}$, és $\hat{\sigma}$ a becsült paraméterek.

Ekkor a P-P plot a következő lesz:

$$\hat{H}(z_{(i)}) \text{ vs } \widetilde{H}(z_{(i)}) \quad i = 1, \dots, n$$

A modell akkor illeszkedik jól, ha a gráf közel van az $y = x$ egyeneshez. A hátránya az, hogy a faroknál nem ad elég információt az illeszkedésről, mert mindkét koordináta konvergál a széleken, ezért kicsi a szórása ezeknek a pontoknak. Ezzel ellentétben a Q-Q plot jobban szemlélteti a farok illeszkedését.

$$\left[\hat{H}^{-1} \left(\frac{i}{n+1} \right), z_{(i)} \right], \quad i = 1, \dots, n,$$

ahol $\hat{H}^{-1}(\frac{i}{n+1})$ a (2). egyenletből:

$$\hat{H}^{-1} \left(\frac{i}{n+1} \right) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left\{ 1 - \left[-\log \left(\frac{i}{n+1} \right) \right]^{-\hat{\xi}} \right\}.$$

Indokolt, hogy a megfelelő eloszlás kiválasztása több különböző objektív módszerből származó információk alapján történjen. A P-P és Q-Q plotok csak szemléletes képet mu-

tatnak az elemzésünkről, de objektív döntések meghozatala nem alapozható rájuk. Az Anderson-Darling tesztstatisztika alkalmas lehet arra, hogy az empirikus eloszlásfüggvény és a feltételes eloszlásfüggvény közti különbséget vizsgáljuk.

Definíció 2.6. *Anderson-Darling teszt*

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x) \cdot \{1 - F(x)\}} dF(x),$$

ahol $F(\cdot)$ az eloszlásfüggvény és $F_n(\cdot)$ az empirikus eloszlásfüggvény.

Az Anderson-Darling teszt súlyfüggvénye olyan módon van megválasztva, hogy nagyobb érzékenységet biztosítson az empirikus eloszlásfüggvény és a feltételezett eloszlásfüggvény közötti eltérésekre, különösen a faroknál. A gyakorlatban az alábbi egyenletet használjuk a tesztstatisztika kiszámításához:

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\log F(x_i) + \log \{1 - F(x_{n+1-i})\}].$$

Az AD-teszt egy olyan tesztstatisztika, amely egyenlő súlyt rendel az eloszlás mindkét végéhez. Extrém érték elméletben viszont gyakran elég az eloszlás felső egyik végére összpontosítanunk, így egy módosított Anderson-Darling teszthez kell folyamodnunk, ami kiemeli a farok felső vagy alsó részének az eltéréseit.

Definíció 2.7. *Módosított Anderson-Darling teszt*

Alsó farok:

$$AU_n^2 = \frac{n}{2} - 2 \sum_{i=1}^n F(x_i) - \sum_{i=1}^n \left\{ 2 - \frac{2i-1}{n} \right\} \log \{1 - F(x_i)\},$$

felső farok:

$$AL_n^2 = -\frac{3n}{2} + 2 \sum_{i=1}^n F(x_i) - \sum_{i=1}^n \frac{2i-1}{n} \log \{F(x_i)\}.$$

Észrevehetjük, hogy a két tesztstatisztika súlyfüggvénye megegyezik az AD teszt súlyfüggvényével, így a két statisztika összege egyenlő lesz az A tesztstatisztikával:

$$A_n^2 = AU_n^2 + AL_n^2$$

A tesztstatisztikák kiszámítása önmagában nem elegendő a modell illeszkedésének értékeléséhez. A számított értékeket valamilyen módon össze kell vetnünk a nullhipotézis alatti eloszlásukkal. Ezt a problémát a p -érték segítségével kezeljük. Ez a p -érték azt a legkisebb szignifikanciaszintet adja meg, amelynél a nullhipotézis még éppen elutasítható. Minél kisebb a p -érték, annál kevésbé valószínű, hogy a megfigyelt adatok illeszkednek a feltételezett eloszláshoz. Az *Assessment of modified Anderson-Darling test statistics*

for the generalized extreme value and generalized logistic distributions [13] cikk részletes útmutatást közöl a GEV eloszláshoz tartozó szimulált kritikus értékekről különböző mintanagyság és szignifikanciaszint mellett.

2.2. Simítási módszerek

Ebben a részben alapvető simítási módszereket definiálunk, amik különösen a 4. fejezetben kerülnek kihasználásra.

2.2.1. Kőbös spline

A kőbös spline részletesebb leírásáról és a következtetések levonásáról lásd Greem és Silverman *Nonparametric regression and generalized linear models: a roughness penalty approach* [3] című cikkét.

Legyenek $t_1, \dots, t_n$ valós számok egy $[a, b]$ intervallumon, ahol $a = t_0 < t_1 < \dots < t_n < t_{n+1} = b$, ahol t_i pontokat *csomóknak* hívjuk. A $g : [a, b] \rightarrow \mathbb{R}$ függvényt kőbös spline-nak hívjuk, ha a következő feltételek teljesülnek:

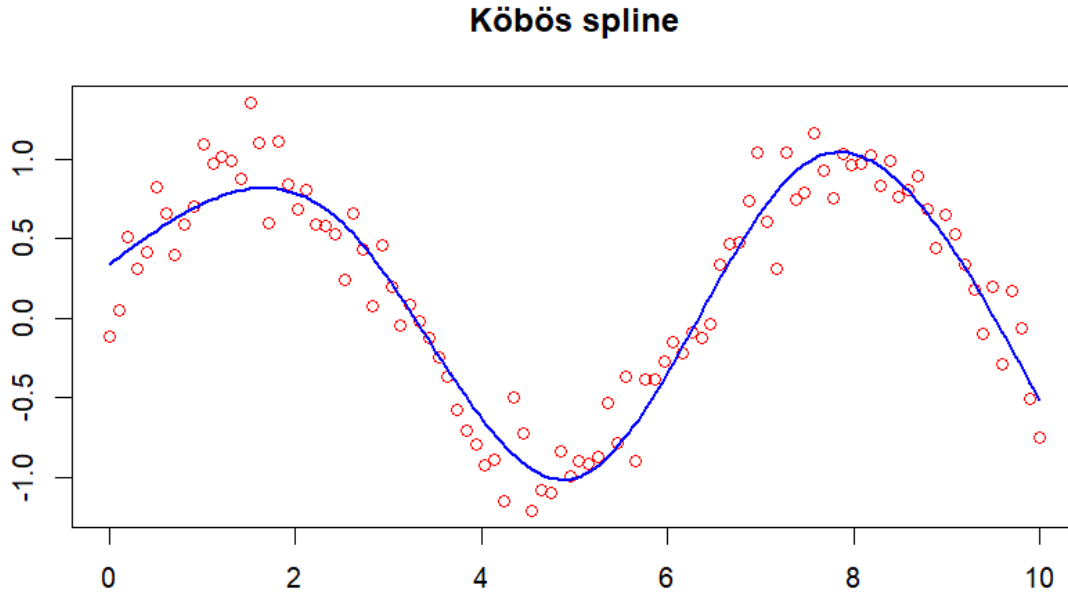
1. g függvény harmadfokú polinom minden $[t_i, t_{i+1}]$, $i = 0, \dots, n$ intervallumon.
2. a $g(t)$ függvény, valamint első és második deriváltja folytonos az $[a, b]$ intervallumon.
Ez azt jelenti, hogy minden belső csomópontban $(t_1, \dots, t_n)$ a g , g' és g'' értékei jobbról és balról megegyeznek.

Továbbá, ha a spline kielégíti az alábbi természetes határfeltételeket: $g''(a) = g''(b) = 0$, akkor spline-t természetes kőbös spline-nak nevezzük. Ezek a feltételek biztosítják, hogy a spline lineáris viselkedést mutasson az $[a, t_1]$ és $[t_n, b]$ szélső intervallumon. Az extrém érték modellezés során a GEV_{min} -eloszlás helyparaméterét időfüggő módon modellezzük, ez lehetővé teszi az évek szerinti simított trendbecslést.

2.2.2. Regularizált log-likelihood

Tegyük fel, hogy $Z_1, \dots, Z_n$ független valószínűségi változók, amelyek időben sorban következnek be, azaz minden Z_i esemény a t_i időpontban következik be, ahol $t_1 < t_2 < \dots < t_n$. Legyen $F(\cdot, \theta)$ egy ismert paraméteres család (például normális, Poisson, exponenciális), amelynek paramétereit két részre bonthatóak: egy időben változó $g(t)$ és egy időben állandó (ψ) . Ekkor $Z_i \sim F(g_i, \psi)$, ahol $g(t_i) = g_i$ egy időben változó paramétert reprezentál. A $g(t)$ függvény szerepe az, hogy időfüggő módon vezérlje az F eloszlás egyes paramétereit. A regularizált log-likelihood függvény a következő módon definiáljuk:

$$l(\mathbf{z}; \mathbf{g}, \psi) - \frac{1}{2} \lambda \int_a^b [g^{(2)}(t)]^2 dt, \quad (3)$$



4. ábra. Az ábrán egy természetes köbös spline illesztés látható, amelyet 100 darab $[0,10]$ intervallumon egyenletesen elosztott pontra szinusz alapú, normális eloszlású zajjal generált adatokra alkalmaztunk. Az illesztéshez 10 csomópontot használtunk, az értékek pedig a becült spline görbe mentén kerültek meghatározásra.

ahol λ egy simítási paraméter $\mathbf{g} = (g_1, \dots, g_n)^T$ simítási függvényekkel és $\mathbf{z} = (z_1, \dots, z_n)$ adattal. Kicsi λ érték esetén a becslés kizárólag a log-likelihood függvénytől függ, míg nagy λ értéknél egy lineáris becslést kapunk g -re, így ez erős büntetéssel jár. Minimalizálása a regularizált log-likelihood függvénynek egy természetes köbös spline függvényt kapunk, amikor minimalizáljuk a g -t, akkor folytonos lesz a második deriváltban. A regularizált log-likelihood függvény maximalizálása megegyezik azzal, mintha maximalizálnánk egy véges-dimenziós rendszert, ami kapcsolódik g_i -hez, ami a köbös spline simítása, ami kiváltja a g görbét. Maximalizálása a (3). egyenletnek g és ψ szerint hasonló, mint a maximalizálása

$$l(\mathbf{z}; \mathbf{g}, \psi) - \frac{1}{2} \lambda \mathbf{g}^T K \mathbf{g},$$

$\mathbf{g}$ és ψ figyelembe vételével, ahol $\mathbf{g} = (g_1, \dots, g_n)^T$. Az utóbbi egyenlet maximalizálása, egy egyszerű Fisher-scoring algoritmussal érhető el, amit a következő alfejezetben tárgyalunk.

2.2.3. Fisher scoring

Mielőtt rátérnénk a Fisher scoring módszer ismertetésére, először egy olyan alapvető definíciót adunk meg, amelyet később az algoritmus kidolgozásánál fogunk alkalmazni.

Definíció 2.8 (*Fisher-információ*). Egy mintaelem Fisher-információja:

$$l_1(\vartheta) = \mathbb{E}_{\vartheta} \left(\left(\frac{\partial}{\partial \vartheta} \log f_{\vartheta}(X_1) \right)^2 \right),$$

ahol f_{ϑ} a likelihood függvény.

A Fisher-scoring módszer (eredetileg Newton-Raphson) egy általános módszer, amelyet Nelder és Wedderburn 1972-ben javasolt a paraméterbecslések numerikus kiértékelésére. (A témával bővebben az alábbi cikk foglalkozik: [3]). Legyen Q egy $n \times (n-2)$ -es felső háromszög mátrix, és R $(n-2) \times (n-2)$ -es szimmetrikus háromszögmátrix. Legyen

$$K = QR^{-1}Q^T. \quad (4)$$

Az általános Fisher-scoring algoritmust regularizált log-likelihood függvényhez az alábbi algoritmus írja le:

1. Kiszámítjuk a K mátrixot a (4). egyenletben.
2. Kiválasztjuk a λ értéket, ahol $\lambda > 0$.
3. Létrehozuk a következőket: $\mathbf{g} = (g_1, \dots, g_n)^T$ és $\psi = (\psi_1, \dots, \psi_b)^T$, ahol $n, b \in \mathbb{N}$
4. Kiszámítjuk $\mathbf{g}_{új}$ -t

$$\mathbf{g}_{új} = \mathbf{g} + [I(\mathbf{g}) + \lambda K]^{-1} \left[\frac{\partial l}{\partial \mathbf{g}}(\mathbf{z}; \mathbf{g}, \psi) - \lambda K \mathbf{g} \right],$$

ahol

$$I(\mathbf{g}) = \begin{pmatrix} I(g_1) & 0 & \dots & 0 \\ 0 & I(g_2) & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & 0 & I(g_n) \end{pmatrix}.$$

$I_{\mathbf{g}}$ a várható információs mátrix a $\mathbf{g}$ simítási függvényhez, amit a $\mathbf{g}$ és ψ értékekből számoljuk ki.

5. Kiszámítjuk $\psi_{új}$ -t a $\mathbf{g}_{új}$ és ψ segítségével:

$$\psi_{új} = \psi + I(\psi)^{-1} \frac{\partial l}{\partial \psi} |_{\psi, \mathbf{g}_{új}},$$

ahol $I(\psi)$ a várható információ az időállandó paraméterre, amit a $\mathbf{g}_{új}$ és ψ segítségével számoljuk ki.

6. Ismételjük meg a 4. és 5. lépést a $g = \mathbf{g}_{\text{új}}$ és $\psi = \psi_{\text{új}}$ értékekkel frissítve, és folytassuk az iterációt mindaddig, amíg a paraméterek változása egy előre meghatározott tűréshatáron (pl. ε) belülre nem esik, azaz amíg teljesül például $\|\mathbf{g}_{\text{új}} - \mathbf{g}\| < \varepsilon$ és $\|\psi_{\text{új}} - \psi\| < \varepsilon$.

2.2.4. Akaike információs kritérium

Munkánk során a korrigált Akaike információs kritériumot (AICc) választottuk a simítási paraméter kiválasztásához. Erről részletesebben a (4). fejezetben beszélünk. Először értsük meg, hogy mire használható az Akaike információs kritérium (AIC), amiből később az AICc-t felhasználjuk. Az adatokra vonatkozó modellek alapján az AIC megbecsüli az egyes modellek minőségét a többi modellhez képest. Az AIC megbecsüli egy adott modell által elvesztett információ relatív mennyiségét, azaz minél kevesebb információt veszít a modell, annál jobb az adott modell, miközben figyelembe veszi a modell illeszkedésének pontosságát és a modell egyszerűségét is, ezzel a túlillesztés és alul illesztés kockázatát kezeli. Hirotugu Akaike-ről nevezték el aki 1973-ban *Information theory and an extension of the maximum likelihood principle* [5] című munkájában kezdett el először foglalkozni ennek megalkotásával.

Definíció 2.9. Akaike információs kritérium

Tegyük fel, hogy van valamilyen statisztikai modellünk az adatainkra. Legyen k a becsült paramétereink száma a modellben. Legyen $\hat{L}$ a likelihood függvény maximális értéke. Ekkor az AIC értéke:

$$AIC = 2k - 2 \ln(\hat{L})$$

Ha kevés mintánk van az AIC olyan modellt választhat, amelyikben túl sok paraméter van, azaz túlilleszt, így ennek megoldására tovább fejlesztették az AIC-t, ami korrelál a kisebb mintaszámmal.

Definíció 2.10. Korrigált Akaike információs kritérium

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1},$$

ahol n a mintánk mérete, míg k a paraméterek száma.

A simítási paraméter λ_g kiválasztása az alapján történik, hogy az AICc értéke melyik paraméter esetében a legkisebb. Ez a kiválasztott paraméter lesz a legjobb választható nem-parametrikus modelleknél.

3. Rekordok

Sporteredményekben a rekordok létfontosságú események a fejlődés jellemzésében, azonban ezek a legszélsőségesebb eredmények is előfordulásuk idején, így természetes összefüggés van a szélsőérték-elmélet és a rekordok elmélete között. A modern atlétikában számos tényező szerepet játszhat egy új rekord beállítása közben. Ebben a fejezetben a rekordteljesítményeket matematikai szemszögből vizsgáljuk meg. Kiindulópontként *Glick Breaking Records and Breaking Boards* [2] című cikkét választottuk, amely alapvető betekintést nyújt a csúcsteljesítmények statisztikai elemzésébe, és megalapozza a további vizsgálódásainkat.

Legyen $Z_1, Z_2, \dots, Z_n$ az éves minimumok sorozata, ahol Z_t az éves minimum a t . évben, majd legyen $Y_t = \min(Z_1, Z_2, \dots, Z_t)$, $t \geq 1$ -re, így $\{Y_t\}$ az éves rekord folyamat. Ezen rekordok nem azonos eloszlásúak, mert egyre jobban tömörülnek az elméleti minimumhoz, ahogy nő t , így Y_t egyre kisebb értékek körül összpontosul, fejlődés látszik az idő előrehaladtával. Tegyük fel, hogy $\{Z_t\}$ függetlenek úgy, hogy $Z_t \sim \text{GEV}_{\min}(\mu(t), \sigma, \xi)$ és a $\mu(t)$, $t = 1, \dots, n$; σ , és ξ paraméterek határozzák meg az éves folyamat eloszlását. Tegyük fel, hogy van adatunk $\{Y_t\}$ -re, de nincs $\{Z_t\}$ -re. Ha $Y_t < Y_{t-1}$, azaz új rekord történt, akkor $Y_t = Z_t$ egy $\text{GEV}_{\min}(\mu(t), \sigma, \xi)$ változóból származó megfigyelés. Ha $Y_t = Y_{t-1}$, azaz nem történt új rekord, akkor $Z_t > Y_t$, így Z_t egy $\text{GEVM}(\mu(t), \sigma, \xi)$ változó cenzorálva Y_t -nél. Az $\mathbf{y} = (y_1, \dots, y_n)$ rekordsorozat log-likelihood függvénye ebből következően:

$$l(\mathbf{y}; \mu, \sigma, \xi) = - \sum_{t=1}^n I_t \log \sigma - \sum_{t=1}^n I_t \left(1 + \frac{1}{\xi}\right) \log \Psi_t - \sum_{t=1}^n I_t \Psi_t^{-1/\xi} - \sum_{t=1}^n (1 - I_t) \Psi_t^{-1/\xi}, \quad (5)$$

ahol $\Psi_t = \left[1 - \xi \left(\frac{y_t - \mu(t)}{\sigma}\right)\right]_+$, $\mu = (\mu(1), \dots, \mu(n))$ és $I_t = 1$, ha t -ik időben rekord történt, ha nem, akkor $I_t = 0$.

Ha $\mu(t)$ paraméteresen megadható, akkor a maximum likelihood becsléseket a (5). egyenlet közvetlen maximalizálásával kaphatjuk. Viszont, ha a $\mu(t)$ nemparaméteres függvényként van megadva, akkor egy regularizált log-likelihood függvénnyel adható meg a $\text{GEV}_{\min}$ eloszlás, ami az alábbi:

$$l(\mathbf{y}; \mu, \sigma, \xi) - \frac{\lambda}{2} \mu^T K \mu, \quad (6)$$

ahol K egy szimmetrikus $n \times n$ -es mátrix, $n-2$ ranggal, amely csak $(1, 2, \dots, n)$ -től és λ -tól függ. λ egy simítási paraméter, amit az Akaike információs kritérium alapján választunk ki. A második tagja a regularizált log-likelihood-nak büntetést szab ki a nagy variabilitást

mutató μ értékekre. (6) egyenlethez maximalizáljuk, hogy megkapjuk a legjobb becült értékeket a $\hat{\mu}, \hat{\sigma}$ és $\hat{\xi}$ a Fisher-féle scoring módszer segítségével. Annak érdekében, hogy meghatározzuk a $\{Y_t\}$ folyamat várható rekordértékét, feltételezzük, hogy a $\{Z_t\}$ sorozat független és azonos eloszlású (IID) valószínűségi változókból áll. Jelöljük $H(y)$ -val a GEVmin eloszlásfüggvényt. Ebben az esetben $\mathbb{P}(Y_t > y)$, azaz annak valószínűsége, hogy Y_t meghalad egy adott y értéket, a következőképpen adható meg:

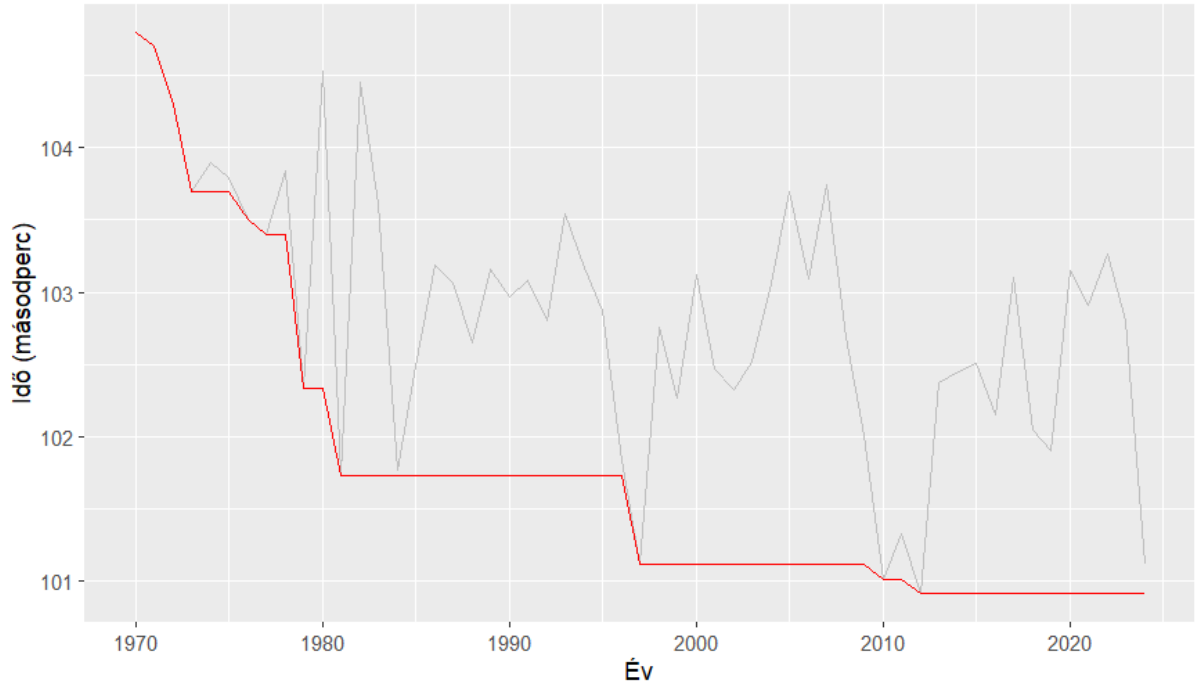
$$\begin{aligned} \Pr(Y_t > y) &= \Pr(Z_1 > y) \Pr(Z_2 > y) \dots \Pr(Z_t > y) \\ &= 1 - H(y) \\ &= \exp \left\{ -t \left[1 - \xi \left(\frac{y - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \\ &= \exp \left\{ - \left[1 - \xi \left(\frac{y - \mu_t^*}{\sigma_t^*} \right) \right]_+^{-1/\xi} \right\}, \end{aligned}$$

ahol $\mu_t^* = \mu - \frac{\sigma}{\xi}(t^\xi - 1)$ és $\sigma_t^* = \sigma t^\xi$, így $Y_t \sim \text{GEVM}(\mu_t^*, \sigma_t^*, \xi)$, és a várható értéke Y_t -nek, $E(Y_t) = \mu_t^* + \frac{\sigma_t^*}{\xi} [1 - \Gamma(1 - \xi)]$. Nem stacionárius esetben, mivel $\Pr(Y_t > y)$ komplex, Az $\{Y_t\}$ eloszlását a $\{Z_t\}$ -ből szimulációval vezetjük le.

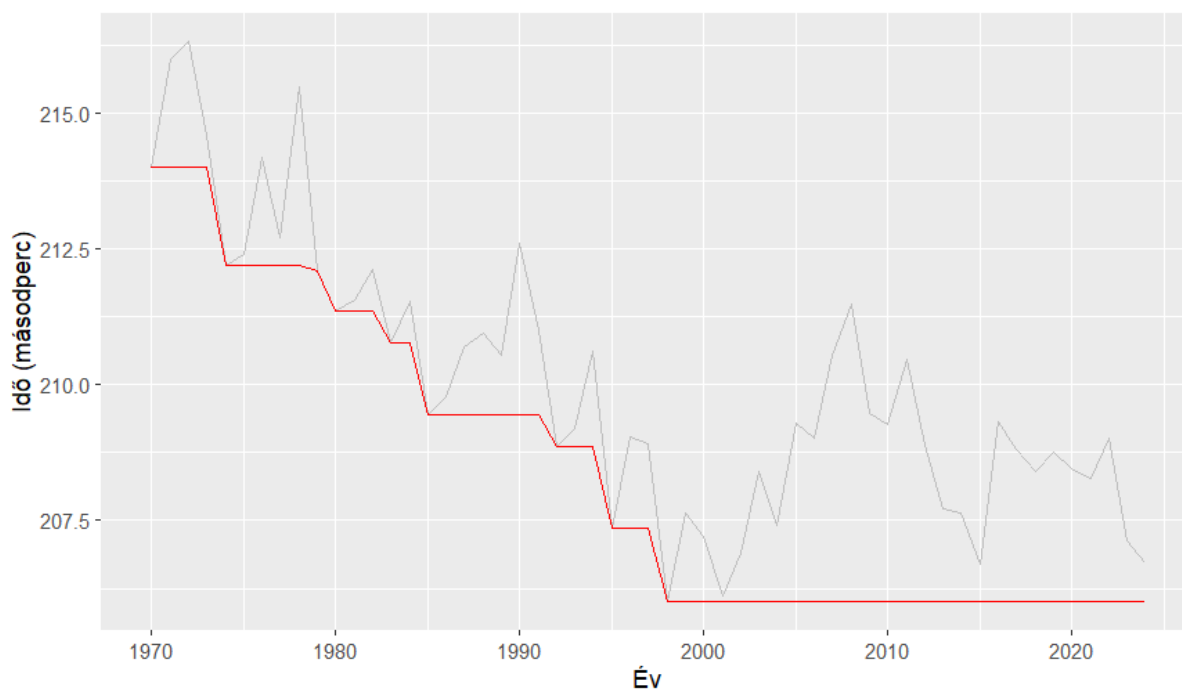
A modell illeszkedését Y_m és becült várható értékének $\hat{E}(Y_m)$ összehasonlításával tudjuk értékelni. Ezt az összehasonlítást bonyolítja $\{Y_t\}$ időfüggése. Ehelyett inkább kiaknázzuk azt a tulajdonságot amelyet, ha $Z_t \sim \text{GEVM}(\mu(t), \sigma, \xi)$, akkor $Z_t = \mu(t) + E_t$, ahol $\mu(t)$ a trend Z_t -ben és $E_t \sim (\text{GEVM})(0, \sigma, \xi)$. Akkor $\{E_t\} = \{Z_t - \hat{\mu}(t)\}$ becült sorozata $t = 1, \dots, n$ -re IID. $\{E_t\}$ várható értékét használjuk, hogy elkészítsük a P-P és Q-Q diagramokat a rekord előrehaladási adatainak illeszkedésének értékelésére, ahogy ezt a következő részben bemutatjuk.

A kezdeti lépés az E_t érték meghatározása. Ha Y_t rekord, akkor $Z_t = Y_t$, azt kapjuk, hogy $E_t = Y_t - \mu(t)$. Különben, ha Y_t nem rekord, akkor $Z_t > Y_t$, így $E_t > Y_t - \mu(t)$. Ezért az $\{E_t\}$ adatai cenzoráltak, ha nem dől meg rekord. Hogy E eloszlását becsljük számolnunk kell a cenzorálással, a standard megközelítéssel, ami a Kaplan-Meier becslés (2.4). Legyen $e_{(1)} < \dots < e_{(m)}$ a megfigyelt rekordok maradékai és d_i az $\{E_t\} > e_{(i)}$ szám. Továbbá legyen $S(y) = \mathbb{P}(Y \geq y) = 1 - H(y)$ az Y túlélésfüggvénye. Ekkor $S(y)$ Kaplan-Meier becslése a $\hat{H}_{KM}(z) = \prod_{i: e_{(i)} < z} \left(1 - \frac{1}{d_i}\right)$. Az eloszlásfüggvény modell alapú becslése E -re $\hat{H}(z) = 1 - \exp \left\{ - \left[1 - \hat{\xi} \left(\frac{z}{\hat{\sigma}} \right) \right]_+^{-1/\hat{\xi}} \right\}$. A modell alapú túlélési valószínűség és az empirikus Kaplan-Meier becslés összehasonlításával kapott P-P és Q-Q diagrammok megszerkeszthetőek. Legyen $e_{(j)}^*$ a j -edik legnagyobb az $e_1, \dots, e_n$ -nek (azaz megfigyelt és cenzorált értékek együtt), akkor a P-P diagram úgy készül, hogy ábrázoljuk $\hat{H}((e_i^*))$ -t szemben $\hat{H}_{KM}(e_i^*)$ -al, $i = 1, \dots, n$ -re. És a Q-Q diagramot, hogy használjuk az exponenciális kvan-

tilis diagramot $\{e_i\}$ -re. A módszer lényege, hogy a transzformált empirikus kvanilisokat, azaz $-\log[1 - \hat{H}(e_i^*)]$ -t hasonlítjuk össze a Kaplan-Meier-féle becslésből származó kvan-tilisek transzformált értékeivel, $-\log[1 - \hat{H}_{KM}(e_{(i)}^*)]$ -t, $i = 1, \dots, n$ -re. A transzformációs exponenciális kvantilis skála az alsó farkot a felső farokká változtatja.



5. ábra. Férfi 800 méteres síkfutás futóidejének fejlődése



6. ábra. Férfi 1500 méteres síkfutás futóidejének fejlődése.

4. Nem paraméteres trend

4.1. Nem paraméteres trend az átlagban

A paraméteres megközelítési modellek gyakran konvergencia problémákhoz vagy gyenge illeszkedéshez vezetnek, ezért vizsgáljuk a nem-parametrikus megközelítést ebben a fejezetben. A célunk simítási technikák kialakítása a regularizált likelihood becslés maximalizálásával. Mivel a különböző módszerek nagyon részletes levezetéseket igényelnének, így csak egy részletesebb áttekintést mutatunk be róluk a fejezetben. Ezekről a módszerekről részletes leírásokat *Smooth Extremal Models in Finance and Insurance* [9], *Penalized likelihood inference in extreme value analyses* [7], és *Generalized Additive Modelling of Sample Extremes* [10] című könyvekben találhatók. A kezdő érték beállítása után kiválasztjuk a simítási paramétert a korrigált Akaike információs kritérium segítségével. A továbbiakban megvizsgáljuk a $GEV_{\min}$ modell egy olyan bővítését, hogy a helyparaméterre úgy tekintünk, mint egy nem-paraméteres módszerrel simítandó függvény.

4.1.1. Regularizációs módszerek a simaság érdekében

Legyen $\lambda > 0$ simítási paraméter és Y valószínűségi változó. Definíció szerint a regularizált négyzetösszeg:

$$S(g) = \sum_{i=1}^n [y_i - g(t_i)]^2 + \frac{1}{2} \lambda \int_a^b [g^{(2)}(x)]^2 dx.$$

A regularizált integrálfüggvény értéke növekszik, ha a g simítása csökken. λ értékét növelve a simább görbét kapunk, de ha $\lambda \rightarrow 0$, akkor érdekesebb körbe keletkezik, ehhez minimalizálni kell g -ben.

Legyen $g_j = g(t_j)$ a kiválasztott simítási függvények minden értéke a csomókban és a második derivált $\gamma_j = g''(t_j)$ a $t_2, \dots, t_{n-1}$ pontokban, és $g''(t_0) = g''(t_{n+1}) = 0$. Ekkor létezik olyan $n \times (n-2)$ -es és $(n-2) \times (n-2)$ -es Q és R mátrix ami csak $t_1, \dots, t_n$ -től függ úgy, hogy

$$Q^T g = R \gamma,$$

ahol R egy pozitív definit, invertálható mátrix. A $Q = [q_{ij}]$ mátrix egy $n \times (n-2)$ -es felső háromszög mátrix lesz:

$$Q = \begin{bmatrix} \frac{1}{h_1} & 0 & \dots & \dots & \dots & 0 \\ -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \dots & \dots & \vdots \\ \frac{1}{h_1} & -\left(\frac{1}{h_2} + \frac{1}{h_3}\right) & \frac{1}{h_3} & 0 & \dots & \vdots \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_3} + \frac{1}{h_4}\right) & \dots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & -\left(\frac{1}{h_{n-3}} + \frac{1}{h_{n-2}}\right) & \frac{1}{h_{n-2}} \\ \vdots & \vdots & \vdots & \vdots & \frac{1}{h_{n-3}} & -\left(\frac{1}{h_{n-2}} + \frac{1}{h_{n-1}}\right) \\ 0 & \dots & \dots & \dots & 0 & \frac{1}{h_{n-2}} \end{bmatrix}$$

ahol $h_i = t_{i+1} - t_i$ és $q_{ij} = 0$, ahol $|j - i| \geq 2$. A Q mátrix oszlopai $j = 2$ -től indulnak. [16]

$R = [r_{ij}]$ egy $(n-2) \times (n-2)$ -es szimmetrikus háromszögmátrix:

$$R = \frac{1}{6} \begin{bmatrix} 2(h_1 + h_2) & h_2 & 0 & \dots & \dots & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & 0 & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & \dots & \dots & h_{n-2} \\ 0 & \dots & \dots & \dots & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{bmatrix}$$

ahol $|r_{ij}| > \sum_{j \neq i} |r_{ij}|$ minden i -re és $r_{ij} = 0$ minden $|j - i| \geq 2$ -re. r_{ij} elemek itt viszont nem csak $j = 2$ -től kezdődnek a mátrixban, de $i = 2$ -től is. [16]

Definiálunk egy K mátrixot a következők szerint:

$$K = QR^{-1}Q^T \quad (7)$$

ahol K egy szimmetrikus $n \times n$ -es mátrix lesz $n - 2$ -es ranggal. K -val képesek vagyunk egy gyors algoritmust adni a köbös spline illesztésre, ami az alábbi tétel felhasználásával lehetséges:

Tétel 4.1. *Ha $g = (g_1, \dots, g_n)^T$ vektor úgy, hogy $g_i = g_{t_i}$ és $\gamma_i = g^{(2)}(t_i)$, akkor és csak akkor határoz meg egy g természetes köbös spline-t, ha a*

$$Q^T g = R\gamma$$

feltétel teljesül. Ekkor a következő regularizációs kifejezést alkalmazzuk a becsült függvény simaságának elősegítésére:

$$\int_a^b [g^{(2)}(t)]^2 dt = g^T K g.$$

A témához kapcsolódóan érdemes kiemelni Fan és Gijbels *Local Polynomial Modelling and Its Applications* [15] című művét, amely átfogó módon tárgyalja a lokális polinomiális modellezés elméleti és gyakorlati aspektusait. Továbbá, Fahrmeir és Tutz *Multivariate Statistical Modelling Based on Generalized Linear Models* [4] című munkája részletesen bemutatja a generalizált lineáris modellekre épülő többváltozós statisztikai modellezési módszereket, amelyek szintén szoros kapcsolatban állnak a dolgozattal.

4.1.2. Simítás normális eloszlásra

A következő alfejezetben bemutatunk egy példát a simítási módszerre normális eloszlás esetén, ezt a megközelítést ugyanakkor a későbbiekben nem alkalmazzuk, csupán szemléltető céllal szerepel. Ha az Y_i valószínűségi változók egymástól függetlenül normális eloszlásúak, $Y_i \sim N(g_i, \sigma^2)$, ahol g_i nem-parametrikus forma és σ időben konstans paraméter, akkor a regularizált log-likelihood függvény a következő:

$$l_\lambda(g, \sigma) = \sum_{j=1}^n \log f[y_j; g(t_j), \sigma^2] - \frac{\lambda}{2} \int_a^b [g^{(2)}]^2 dt, \quad \lambda > 0.$$

Ha $g(t)$ természetes köbös spline, az integrál kifejezhető $g^T K g$ -val.

A várható információ értéke a simítási függvényekre a következők szerint alakul:

$$\begin{aligned}\frac{\partial l}{\partial g_i} &= \frac{1}{\sigma^2} [y_i - g(t_i)] \\ \frac{\partial^2 l}{\partial g_i^2} &= -\frac{1}{\sigma^2} \\ I(g_i) &= E \left(-\frac{\partial^2 l}{\partial g_i^2} \right) = \frac{1}{\sigma^2}\end{aligned}$$

míg a várható információ értéke az időben konstans paraméterre:

$$\begin{aligned}\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \int_{i=1}^n [y_i - g(t_i)]^2 \\ \frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \int_{i=1}^n [y_i - g(t_i)]^2 \\ I(\sigma) &= E \left(-\frac{\partial^2 l}{\partial \sigma^2} \right) = \frac{2n}{\sigma^2}\end{aligned}$$

A Fischer scoring algoritmus normális eloszlásra az alábbi módon néz ki:

1. Kiszámítjuk a K mátrixot a 7 egyenletben.
2. Kiválasztjuk a λ értéket, ahol $\lambda > 0$.
3. Létrehozuk a következőket: $\mathbf{g} = (g_1, \dots, g_n)^T$ és $\sigma = \sigma_1$
4. Kiszámítjuk $\mathbf{g}_{\text{új}}$ -t

$$\mathbf{g}_{\text{új}} = \mathbf{g} + [I(\mathbf{g}) + \lambda K]^{-1} \left[\frac{\partial l}{\partial \mathbf{g}} - \lambda K \mathbf{g} \right],$$

ahol

$$I(\mathbf{g}) = \sigma^{-2} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

és

$$\frac{\partial l}{\partial \mathbf{g}} = \sigma^2 \begin{pmatrix} y_1 - g_1 \\ y_2 - g_2 \\ \vdots \\ y_{n-1} - g_{n-1} \\ y_n - g_n \end{pmatrix}$$

5. Kiszámítjuk $\sigma_{\text{új}}$ -t a $\mathbf{g}_{\text{új}}$ és σ segítségével:

$$\sigma_{\text{új}} = \sigma + I(\sigma)^{-1} \frac{\partial l}{\partial \sigma} = \frac{\sigma^2 + \hat{\sigma}^2}{2\sigma}, \text{ ahol } \hat{\sigma} = \int_{i=1}^n (y_i - g_i)^2.$$

6. Ismételjük meg a 4. és 5. lépést $g = \mathbf{g}_{\text{új}}$ és $\sigma = \sigma_{\text{új}}$ helyettesítéssel, és álljunk meg, ha ezek az értékek már konvergálnak egy korláthoz.

Ennek az algoritmusnak egy általánosított változata a (2.2.3). alfejezetben található.

4.2. Kezdőpontok és simítási paraméter kiválasztása

Legyen $\{Z_t\}$ független éves minimumok sorozata és tegyük fel, hogy extrém érték eloszlást követ minimumokra. Válasszuk μ_t átlagbeli trendet g_t simított függvényre, ahol $t = 1, \dots, n$. Ekkor a 2. fejezetben látott $\text{GEV}_{\min}(\mathbf{g}, \sigma, \xi)$ -t kapjuk. Illetve Z_t -re a következő a várható érték és a variancia:

$$E(Z_t) = g_t + \frac{\sigma}{\xi} [1 - \Gamma(1 - \xi)]$$

$$\text{Var}(Z_t) = \frac{\sigma^2}{\xi^2} \left\{ \Gamma(1 - 2\xi) - [\Gamma(1 - \xi)]^2 \right\}.$$

A kezdőértékek beállításában fog segíteni a normális eloszlás felhasználása, ahol a normális eloszláshoz használunk egy hely- és egy skálaparamétert.

4.2.1. Kezdőpontok kiválasztása

Tegyük fel, hogy az adataink normális eloszlást követnek. $Y_t \sim N(m_t, \mu_N^2)$, ahol a helyparaméter nem parametrikus, $m_t = m(t)$, ahol $m(\cdot)$ egy simítási függvény, és σ_N konstans. Ekkor Y_i sűrűségfüggvénye a következő:

$$f(y_i; m_i, \sigma_N^2) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp \left[-\frac{(y_i - m_i)^2}{2\sigma_N^2} \right], \quad \sigma_N > 0,$$

és a regularizált log-likelihood:

$$l_{\lambda_m}(\mathbf{m}, \sigma_N) = \sum_{j=1}^n \log f(y_j; m_j, \sigma_N^2) - \frac{\lambda_n}{2} \int_a^b [m^{(2)}(t)]^2 dt, \quad \lambda_m > 0.$$

Fisher-scoring módszert alkalmazva a normális eloszláson, meg tudjuk becsülni $\hat{g}_t$ -t és $\hat{\sigma}$ -t m_t és σ_N kezdeti értékéből kiindulva. Ha $Z_t \sim \text{GEV}_{\min}$, akkor momentumok összehasonlításához egy módszer az, ha vesszük $\hat{E}(Z_t) = \hat{m}_t$ -t és $\hat{V}ar(Z_t) = \hat{\sigma}_N^2$ -et, így:

$$\mu_{t,\text{kezdeti}} = \hat{m}_t - \frac{\hat{\sigma}_N}{\hat{\xi}} [1 - \Gamma(1 - \hat{\xi})]$$

és

$$\hat{\sigma}_{\text{kezdeti}} = \sqrt{\frac{\hat{\xi}^2 \hat{\sigma}_N^2}{\Gamma(1 - 2\hat{\xi}) - [\Gamma(1 - \hat{\xi})]^2}}.$$

Ha fixáljuk $\hat{\xi} = \hat{\xi}_{\text{kezdeti}}$ -t, akkor megkapjuk $\hat{\sigma}_{\text{kezdeti}}$ és $\hat{\mu}_{t,\text{kezdeti}}$ értékeket.

4.2.2. Fisher Scoring módosítása

1. Kiszámítjuk a 7 egyenletben leírt K mátrixot.
2. Kiválasztunk egy kezdőértéket a simítási paraméternek λ_g és Ω értéket, ahol $\lambda_g > 0$ és $0 \leq \Omega \leq 1$.
3. Megbecsüljük $\hat{m}_t$ és $\hat{\sigma}_N$ értékeket a $N(m_t, \sigma_N)$ normális eloszlásból
4. Bevezetjük μ_t, σ és ξ értékeket $\hat{m}_t$ -t és $\hat{\sigma}_N$ -t használva, hogy megkapjuk $\xi_{\text{kezdeti}} = \xi$ kezdeti értéket,

$$\mu_{t,\text{kezdeti}} = g_{t,\text{kezdeti}} = m_t - \frac{\sigma_N}{\xi} [1 - \Gamma(1 - \xi)]$$

és

$$\sigma_{\text{kezdeti}} = \sqrt{\frac{\xi^2 \sigma_N^2}{\Gamma(1 - 2\xi) - [\Gamma(1 - \xi)]^2}}$$

5. Kiszámítjuk $g_{\text{új}}$ -t:

$$g_{\text{új}} = g + \left[E \left(-\frac{\partial^2 l}{\partial g \partial g^T} \right) + \lambda_g K \right]^{-1} \left(\frac{\partial l}{\partial g} - \lambda_g K g \right)$$

6. Kiszámítjuk $\theta_{\text{új}}$ -t felhasználva $g = g_{\text{új}}$ -t, ahol $\theta = \begin{pmatrix} \xi \\ \sigma \end{pmatrix}$

$$\theta_{\text{új}} = \theta + \Omega \left[E \left(-\frac{\partial^2 l}{\partial \theta \partial \theta^T} \right) \right]^{-1} \left(\frac{\partial l}{\partial \theta} \right).$$

7. Teszteljük, hogy konvergál-e és legyen $\theta = \theta_{\text{új}}$ és $g = g_{\text{új}}$, majd térjün vissza az 5., 6. és 7. lépésre. Ha a változás kisebb egy előre megadott küszöbnél, akkor megállunk.

Ω az elfogadási skálázása a 6. lépés második felének. A becsült modell illeszkedésének vizsgálatára a P-P és Q-Q plotokat tudjuk használni.

4.2.3. Simítási paraméter kiválasztása

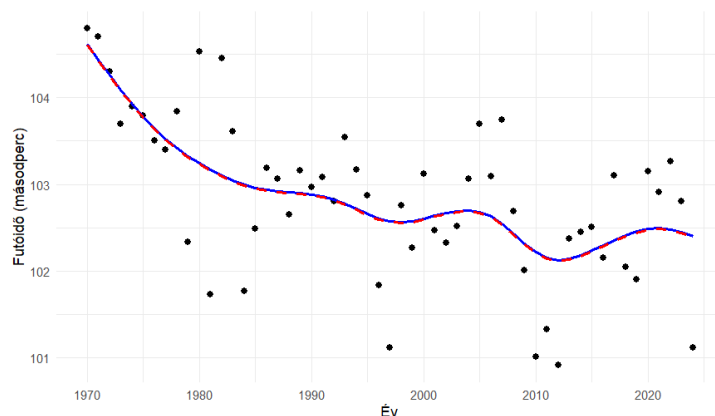
A simítási paraméter λ kiválasztása többféle módon történhet. Ilyen módszerek lehetnek a Cross-validation (CV) vagy az Általánosított Cross-validation (GCV). Mivel a simítási paraméter kiválasztása ezekben az esetekben nagy számítási kapacitást igényelnek extrém értékekre, így alternatív módszerek lehetnek a Mallows' C_p kritérium vagy az Akaike információs kritérium (AIC).

4.3. Alkalmazás

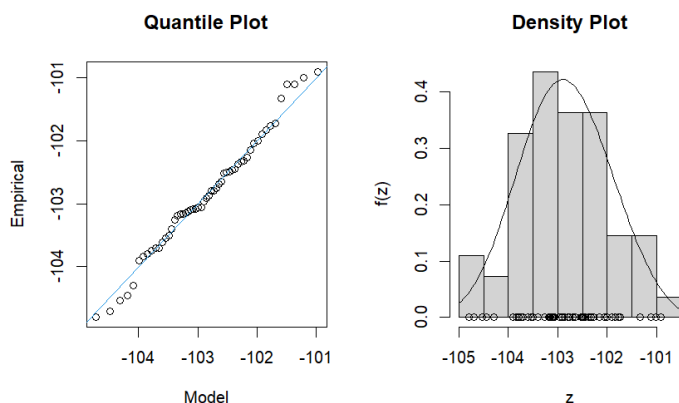
A fejezet további részében a már bemutatott adathalmazon fogjuk tesztelni a nem paraméteres modelleket. Bemutatjuk, hogy a regularizált log-likelihood módszer képes egy alternatívát képezni a paraméteres megközelítésnek. Ebben a dolgozatban csak a helyparaméterre fogjuk alkalmazni ezt a módszert, de a többi extrém érték paraméterre is ki lehet terjeszteni a megvalósítását.

4.3.1. Férfi 800 méteres síkfutás

A {7}. ábrán látható a becsült μ_t a simítási paraméterrel. A becsült μ_t paraméter követi a trendet és teljesen egybeesik a várható értékkel. A {8} ábrán láthatjuk, hogy jól illeszkedik a GEV_{min} eloszlásra. A {9} ábrán a várható értéket hasonlítjuk össze a $\hat{g}_t$ becsléssel. Az ehhez tartozó AICc kritérium értéket a {10} ábrán látható.



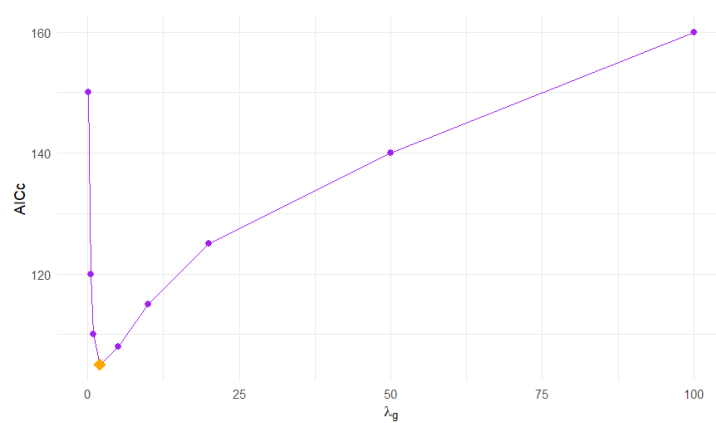
7. ábra. A férfi 800 méteres síkfutás adatpontjai (fekete pont), a μ_t -t becsülő regularizált log-likelihood függvény (kék vonal), és $E(Z_t)$ -t (szaggatott piros vonal) mutatja az ábra az elfogadási tartomány és egy 10 éves visszatérési szint mellett.



8. ábra. A férfi 800 méteres síkfutáshoz tartozó diagnosztikai ábrák



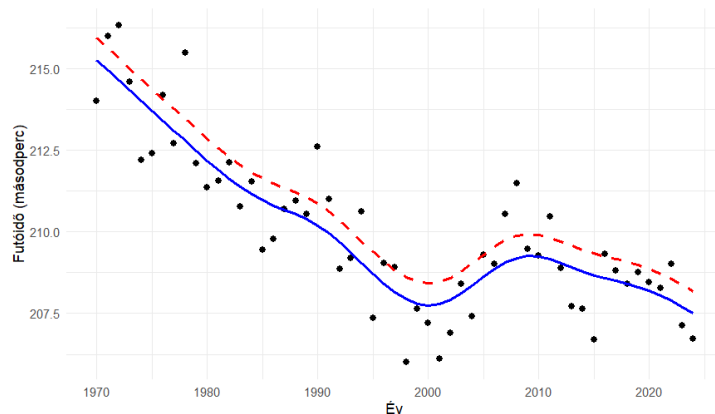
9. ábra. A pontozott vonal az $E(Z_t) = \hat{g}_t + \hat{\sigma} \left[1 - \Gamma(1 - \hat{\xi}) \right] / \hat{\xi}$ várható értéket reprezentálja, míg a piros vonal a $\hat{g}_t$ becslést.



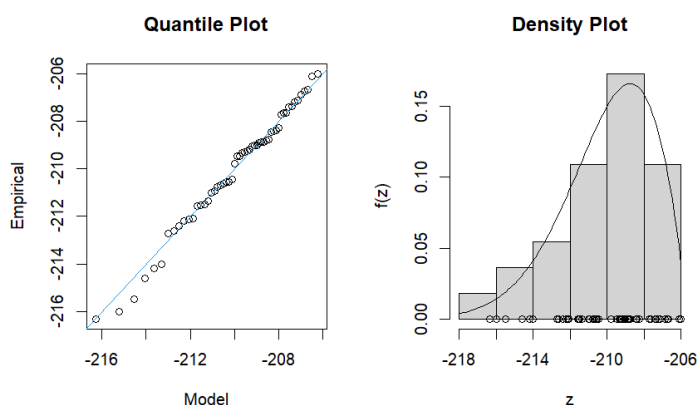
10. ábra. A férfi 800 méteres síkfutáshoz tartozó AICc kritérium a λ_g -vel szemben.

4.3.2. Férfi 1500 méteres síkfutás

A {11}. ábrán látható a becsült μ_t az átmeneti variációval. A becsült μ_t paraméter viszonylag jól követi a trendet, de a várható érték alatt mozog az elfogadási tartományon belül. A {12} ábrán láthatjuk, hogy elfogadhatóan jól illeszkedik a GEV_{min} eloszlásra. A {13} ábrán a várható értéket hasonlítjuk össze a $\hat{g}_t$ becsléssel. Az ehhez tartozó AICc kritérium értéket a {14} ábrán látható.



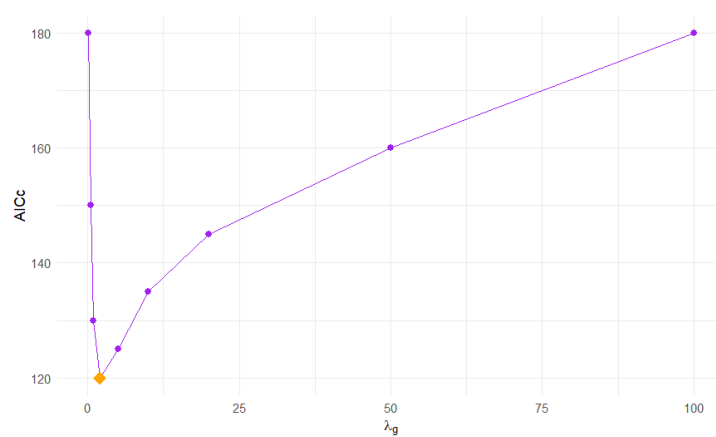
11. ábra. A férfi 1500 méteres síkfutás adatpontjai (fekete pont), a μ_t -t becsülő regularizált log-likelihood függvény (kék vonal), és $E(Z_t)$ -t (szaggatott piros vonal) mutatja az ábra az elfogadási tartomány és egy 10 éves visszatérési szint mellett.



12. ábra. A férfi 1500 méteres síkfutáshoz tartozó diagnosztikai ábrák



13. ábra. A pontozott vonal az $E(Z_t) = \hat{g}_t + \hat{\sigma} [1 - \Gamma(1 - \hat{\xi})] / \hat{\xi}$ várható értéket reprezentálja, míg a piros vonal a $\hat{g}_t$ becslést.



14. ábra. A férfi 1500 méteres síkfutáshoz tartozó AICc kritérium a λ_g -vel szemben.

5. Pontfolyamatok

Az eredeti pontfolyamatok alatt a Poisson-folyamatot és a felújítási folyamatot értjük. Ezek olyan módszerek, amelyek segítségével véletlenszerű pontokat rendelhetünk különböző tartományokhoz, például egy szakaszhoz a valós számegyenesen, vagy téglalapokhoz egy d -dimenziós Euklideszi térben, amelyet $\mathbb{R}^d$ -vel jelölünk.

Egy pontfolyamat akkor van teljesen meghatározva, ha ismert, hogy mekkora a valószínűsége annak, hogy egy adott, véges számú, egymást nem metsző intervallumban hány pont található. Másképp fogalmazva: a folyamat akkor van teljesen definiálva, ha ismerjük a pontok számának közös valószínűségi eloszlását ezekre a diszjunkt tartományokra.

5.1. Stacionárius Poisson folyamat

Vegyünk egy stacionárius folyamatot $\{X_t\}$. Határozzunk meg $A = [0, 1] \times (-\infty, x)$ régióit valamilyen kis r értékre, akkor P_n -t a következőképpen definiálhatjuk: $P_n = \{[i/(n+1), (X_i - b_n)/a_n] : i = 1, \dots, n\}$, ahol $\{a_n > 0\}$ konstansok sorozata és $\{b_n\}$ olyan, hogy:

$$\Pr \left[\frac{\min(X_1, \dots, X_n) - b_n}{a_n} > x \right] \xrightarrow{n \rightarrow \infty} 1 - H(x) = \exp[-\Lambda(x)],$$

ahol $\Lambda(x) = [1 - \xi(x - \mu)/\sigma]_+^{-1/\xi}$ valamilyen $\mu, \sigma > 0$ és ξ -re, illetve H_x nem-elfajuló határeloszlás. Ekkor H_x eloszlása $\text{GEV}_{\min}(\mu, \sigma, \xi)$, aminek bizonyítását Leadbetter, Malcolm R., Georg Lindgren, és Holger Rootzén *Extremes and related properties of random sequences and processes* [12] című könyvében részletesen kifejtették. Legyen P a P_n határértéke, ahogy $n \rightarrow \infty$. Legyen $x_F = \min\{x : F_X(x) > 0\}$, ahol F_X az $\{X_t\}$ eloszlásfüggvénye, és $b_l = \lim_{n \rightarrow \infty} (x_F - b_n)/a_n$. $A \subseteq [0, 1] \times (-\infty, b_l)$ -re legyen $N_n(A)$ azon P_n pontok száma, amelyek A -ba esnek és legyen $N(A)$ jelölje a P pontok számát A -ban. P nem-homogén Poisson-folyamat, a következő intenzitással:

$$\lambda(t, r) = \frac{1}{\sigma} \left[1 - \xi \left(\frac{r - \mu}{\sigma} \right) \right]_+^{-1-1/\xi}.$$

Most, ha $A = [t_0, t_1] \times (-\infty, r)$, $0 < t_0 < t_1 < 1$ -el és $r < b_l$ -el, $N_n(A)$ limitáló határeloszlása $\text{Poisson}(\Lambda(A))$, ahol $\Lambda(A) = (t_1 - t_0) \left[1 - \xi \left(\frac{r - \mu}{\sigma} \right) \right]_+^{-1/\xi}$. Most a nem-stacionárius esetet vesszük figyelembe, $\mu(t)$ helyparaméterrel. Ennek a nem homogén Poisson folyamatnak az intenzitás mértéke ekkor, $\lambda(t, r) = \frac{1}{\sigma} \left[1 - \xi \left(\frac{r - \mu(t)}{\sigma} \right) \right]_+^{-1-1/\xi}$, ahol az időt $[0, 1]$ -re skálázzuk minden t -re $\mu(t)$ -ben.

Tegyük fel, hogy t_0 és t_{m+1} idők között m egymást követő tényleges rekord sorozatának ideje és értéke $(t_1, r_1), \dots, (t_m, r_m)$, ahol r_i a t_i időpontban kapott rekordérték, ahol

$r_1 > \dots > r_{m-1} > r_m$. Feltehető, hogy az aktuális rekord t_0 -ban r_0 . Ennek a tényleges rekordsorozatnak a valószínűségét szakaszokban vezetjük le. A Poisson-folyamatot használva annak a valószínűsége, hogy a (t_0, t_1) időintervallumban nem lesz új rekord:

$$\exp \left[- \int_{t_0}^{t_1} \int_{-\inf}^{r_0} \lambda(t, r) dr dt \right] = \exp \left\{ - \int_{t_0}^{t_1} \left[1 - \xi \left(\frac{r_0 - \mu(t)}{\sigma} \right) \right]_+^{-1/\xi} dt \right\}.$$

Annak a valószínűsége, hogy 1 rekord van az $(r_i - \delta r, r_i)$ intervallumban, a $(t_i, t_i + \delta t, t_i + \delta)$ időintervallumban,

$$\approx \lambda(t_i, r_i) \delta r \delta t \exp \left\{ - \int_{t_i}^{t_i + \delta t} \int_{r_i - \delta r}^{r_i} \frac{1}{\sigma} \left[1 - \xi \left(\frac{r - \mu(t)}{\sigma} \right) \right]_+^{-1-1/\xi} dr dt \right\}$$

kis $\delta r > 0$ és $\delta t > 0$ -ra.

A valószínűsége, hogy nincs rekord a $(t_i + \delta t, t_i)$ idősorban

$$\exp \left\{ - \int_{t_i + \delta t}^{t_{i+1}} \int_{-\infty}^{r_i} \frac{1}{\sigma} \left[1 - \xi \left(\frac{r - \mu(t)}{\sigma} \right) \right]_+^{-1-1/\xi} dr dt \right\}.$$

Így, az együttes valószínűsége r_i rekordnak t_i időben és, hogy nincs rekord a t_{i+1} -ig arányosan

$$\lambda(t_i, r_i) \exp \left\{ - \int_{t_i}^{t_{i+1}} \int_{-\infty}^{r_i} \frac{1}{\sigma} \left[1 - \xi \left(\frac{r - \mu(t)}{\sigma} \right) \right]_+^{-1-1/\xi} dr dt \right\}.$$

A rekord előrehaladásának valószínűsége t_0 -tól t_{m+1} -ig arányos ezzel a kifejezéssel:

$$\left[\prod_{i=1}^m \lambda(t_i, r_i) \right] \exp \left\{ - \sum_{i=0}^m \int_{t_i}^{t_{i+1}} \left[1 - \xi \left(\frac{r_i - \mu(t)}{\sigma} \right) \right]_+^{-1/\xi} dt \right\}. \quad (8)$$

Mivel $\mu(t)$ lassan változik azt várjuk, hogy a 8 egyenlet integrálja t -vel is simán változik, ezért a numerikus egyszerűsítéshez közelítjük az integrálok összegét a log-likelihood függvényben. Az integrál közelítésének pontos értékéhez megtörjük a t_j és t_{j+1} közötti időt $t_{j+1} = t_j + \delta k_j$ -val, ahol k_j a partíciók száma t_j -től t_{j+1} -ig δ egyenlő távolságra, és $s_{j,i} = t_j + (t_{j+1} - t_j)(i-1)/k_j, i = 1, \dots, k_j + 1$ azt adja, hogy:

$$\sum_{i=0}^m \sum_{j=1}^{k_i} \delta \left[1 - \xi \left(\frac{r_i - \mu(s_{j,i})}{\sigma} \right) \right]_+^{-1/\xi}. \quad (9)$$

Mint az 3 szakaszban, a $\mu(t)$ nem paraméteres következtetéseit büntetett likelihood kinyerhető, a 5 egyenletben azonos büntetési kifejezéssel.

5.2. Tolerancia sávok felépítése

Annak érdekében, hogy új módszereket dolgozzunk ki annak diagnosztizálására, hogy a modell elfogadható-e vagy sem a tényleges adatokra, két szemponttal kell foglalkoznunk. Az idő, amikor a rekord bekövetkezett, ez t , és az új rekord értékével, ami r , amely megadja, hogy rekord előfordult a t időpillanatban. Első megközelítésünk az, hogy a tolerancia-intervallumokat generálunk a rekord előrehaladására, azzal a feltétellel, hogy az illesztett modell helyes. Az ilyen tolerancia-intervallumok felépítése során az illesztett modellt használjuk a rekord progressziós adatainak replikációjának szimulálására. Az eredményként létrejött pontonkénti tolerancia-intervallumok egy olyan értéktartományt biztosítanak, amelyet a rekord-progresszió értékeinek területeként lehet használni annak megítélésére, hogy az illeszkedett modell elfogadható-e a megfigyelt rekord-progressziós adatokhoz viszonyítva. A megfigyelt rekord-progressziós adatoknak a területen belül kell lenniük annak jelzésére, hogy a modell jó.

Az előbb említett megközelítés kulcsfontosságú eleme az, hogy képesek legyünk szimulálni egy rekord-progressziót az illeszkedett modellre. Hatékonysági okokból csak rekord időket és értékeket szeretnénk szimulálni. A rekord-progresszió előállításához két szakasza van egy adott rekord időpontnak t_j és egy adott rekord értéknek r_j . Az első szakaszban a következő rekord időpontot $t_{j+1}(t_{j+1} > t_j)$ kell előállítani, a második szakaszban pedig a következő rekord értéket $r_{j+1}(r_{j+1} < r_j)$. Továbbá szükség van egy inicializációs szakaszra (t_0, r_0) megadására és egy módra a szimuláció befejezésére. Az alábbiakban részletesen ismertetjük ezeket a szakaszokat. Megközelítésünk alapja mind a három esetben a határértékkel rendelkező nem-homogén Poisson-folyamat (P) használata.

Először is, állapítsuk meg a következő rekord várakozási idő eloszlását azzal a feltétellel, hogy az r_j értékű rekord a t_j időpontban következett be. Jelöljük ezt a várakozási idő valószínűségi változóját W_{j+1} -ként. Ekkor a $Pr(W_{j+1} > w) = Pr(\text{P pontok nélkül } A_j\text{-ben})$, ahol $A_j = [t_j, t_j + w] \times (-\infty, r_j)$.

$$Pr(\text{P pontok nélkül } A_j\text{-ben}) = \exp\{-\Lambda(A_j)\} = \exp\left[-\int_{t_j}^{t_j+w} \int_{-\infty}^{r_j} \lambda(s, r) dr ds\right]$$

T_{j+1} , a $(j+1)$ rekord érték idejének szimulálásához a W_{j+1} várakozási idő túlélőfüggvényének valószínűségi integrál-transzformációját használjuk. Konkrétan, ha u egy $U \sim U(0, 1)$ realizációja, akkor az $u = Pr(W_{j+1} > w)$ egy w realizációt ad a várakozási idő eloszlásból. A következő rekord időpontja ezután $t_{j+1} = t_j + w$, ez pedig egy T_{j+1} realizációja. Ezért megoldjuk t_{j+1} -et $u = \exp\left[-\int_{t_j}^{t_{j+1}} \int_{-\infty}^{r_j} \lambda(s, r) dr ds\right]$ -ra azt kapjuk, hogy:

$$-\log u = \int_{t_j}^{t_{j+1}} \int_{-\infty}^{r_j} \lambda(s, r) dr ds \approx \sum_{i=1}^{k_j+1} \lambda\left\{1 - \frac{\xi}{\sigma} [r_j - \mu(s_i)]\right\}_+^{-1/\xi}. \quad (10)$$

A második szakaszban szükségünk van a V_j érték csökkenésének eloszlásának meghatározására, miután a $(j + 1)$ -ik rekord értéke a t_{j+1} időpontban következik be. Az előző rekord érték, az r_j most a küszöbérték az újabb adatok számára, hogy azok a következő új rekord értéket alkossák: $r_{j+1} = r_j - v$, ahol $v > 0$. A rekord szint csökkenésének valószínűsége nagyobb, mint v az időpontban t_{j+1} pedig $Pr(V_{j+1} < v | T_{j+1} = t_{j+1}) = Pr(R_{j+1} < r_j - v | R_{j+1} < r_j, T_{j+1} = t_{j+1})$, ahol R_{j+1} az új rekord előrehaladás a t_{j+1} időpontban. Habár,

$$Pr(R_{j+1} < r_{j+1} | R_{j+1} < r_j, T_{j+1} = t_{j+1}) = \left\{ 1 - \xi \left[\frac{v}{\sigma - \xi(r_j - \mu(t_{j+1}))} \right] \right\}_+^{-1/\xi}, \quad (11)$$

az Általánosított Pareto eloszlás, $GPD(\hat{\sigma}, \xi)$, $\hat{\sigma} = \sigma - \xi(r_j - \mu(t_{i+1}))$ paraméterrel. Egy új rekord értéket szimulálunk a csökkenő valószínűségi változóból, mely legyen V_i . Az r_{j+1} érték szimulálásához a V_i eloszlásfüggvényének valószínűségi integrál-transzformációját használjuk. Egy $U \sim U(0, 1)$ valószínűsítés használata után az $u = Pr(V_j < v)$ egy megvalósulást ad, ami új rekord érték eloszlást ad egy új rekord előfordulási idejében. A következő rekord értéket ezután generáljuk mint $r_{j+1} = r_j - v$.

A tolerancia sávok konstrukciójának algoritmusát a következőképpen néz ki:

1. Kezdetben: $\sigma = \hat{\sigma}_{\text{rec}}$, $\xi = \hat{\xi}_{\text{rec}}$ és $\mu = \hat{\mu} = \{\hat{\mu}_1, \dots, \hat{\mu}_n\}$
2. Válasszuk ki az $r_0 < (\leq r_1)$ és $t_0 (\leq t_1)$.
3. Első lépés: Szimuláljuk t_{j+1} -et a (10). egyenlet segítségével, használva t_j és r_j értékeket.
4. Második lépés: Szimuláljuk r_{j+1} -et a (11). egyenlet segítségével, használva r_j -t és t_{j+1} -et.
5. Ha $t_{j+1} < T$, ismételjük meg a 3. és 4. lépéseket, ahol T a maximális érdeklődési időpont. Álljunk meg, ha $t_{j+1} \geq T$.

Az algoritmust k -szor megismételjük annak érdekében, hogy k realizációt kapjunk az illeszkedett modell rekord-progresziós folyamatáról. A pontonkénti kvantilizéseket használtuk a megfigyelt rekord-progreszió lehetséges rekordértékének felső és alsó határáként. Ha a valós rekord-progreszió viszonylagosan jól illeszkedik a pontonkénti elfogadási határokhoz, az azt jelzi, hogy a paraméterek becsült értékei megfelelőek a kiválasztott modellhez, és a modell ésszerű leírása az adatoknak.

Mivel már tudjuk, hogy $Z_i \sim \text{GEV}_{\min}(\mu_i, \sigma, \xi)$, akkor $Z_i = \mu_i + E_i$, ahol μ_i a Z_i trendje, és $E_i \sim \text{GEV}_{\min}(0, \sigma\xi)$. Ezután az $\{E_i\}$ sorozat, $i = 1, \dots, n$ független és azonos eloszlású,

mivel különválasztottuk a trendet az adatokból. Az E_i -t használjuk a P-P és Q-Q ábrák készítéséhez a rekord-progresziós adatok jó illeszkedésének értékeléséhez.

Az első lépés az E_i értékének megszerzése. Ha Y_i egy rekord, akkor $y_i < y_{i-1}$, ahol $Z_i = Y_i$, ekkor $E_i = Y_i - \mu_i$. Ellenkező esetben, ha Y_j nem rekord, akkor $y_j = y_{j-1}$, mivel $Z_j > Y_{j-1}$. Ekkor $E_j > y_{j-1} - \mu_{j-1}$. Az E_j -re vonatkozó adatokat ezért cenzúrázzuk, amikor nincs rekord megszakítva. Az E eloszlás becsléséhez figyelembe kell vennünk a cenzúrázást, a szokásos megközelítés a Kaplan-Meier becslés. Kezdetben ábrázoljuk a Kaplan-Meier becslést, $KM_j = \prod_j \left(1 - \frac{d_j}{t_j}\right)$, hogy becsljük az E empirikus túlélőfüggvényét, ahol d_j a j időpontig bekövetkező rekordok száma, és t_j a j -től hátralévő események száma. Az E_i modellezett túlélőfüggvény becslt értéke a $\hat{S}(z) = \exp \left\{ - \left[1 - \hat{\xi} \left(\frac{z}{\hat{\sigma}} \right) \right]_+^{-1/\hat{\xi}} \right\} = 1 - \hat{H}(z)$. Az összehasonlítás a modell alapú túlélő valószínűség és az empirikus Kaplan Meier becslésből készített ábrával a P-P és Q-Q ábrák készítéséhez szükséges. Amikor KM_j -t ábrázoljuk a $S(e_j^*)$ -hez, ahol e_j^* az $e_1, \dots, e_n$ értékek j -edik legnagyobb értéke, vagy ezek átalakításai adnak P-P és Q-Q ábrákat. A P-P ábrát $\hat{H}_{e_i^*} = 1 - \hat{S}(e_i^*)$ versus $1 - KM_i = KM_j$ néven állítják elő az $i = 1, \dots, n$ esetén. Ha a minimális adatokat vizsgáljuk, az alsó farok jó illeszkedésének helyesbítése szükséges a grafikon jobb oldalán. A Q-Q ábra az $\{e_i\}$ exponenciális kvantilis ábráját használja, és a következőképpen néz ki: $-\log[\hat{S}(e_i^*)]$ versus $-\log[KM_i]$ $i = 1, \dots, n$ esetén.

Összefoglalva, hogy áthidaljuk a nem-független és azonosan eloszló (non-iid) nehézségeket, különválasztjuk a trendet az adatokból, hogy az azonosan és függetlenül eloszló (iid) maradékat kapjuk. Ezután a P-P és Q-Q ábrákat az így kapott maradékatokból készítjük el.

5.3. Trend Becslése További Adatok Alapján

Mivel a rekord adatok nem szolgáltatnak elegendő információt, a rekord-progreszió teljes elkülönítéséhez egy trendtől, szükségünk van egy másik, különböző adathalmazra, amely hasonló típusú folyamatból származik, és képes további információkat nyújtani a jobb elkülönítéshez. Ez a további adatforrás képesnek kell lennie arra, hogy reprezentálja az egész adathalmazt, beleértve a rekord adatokat is.[16] megmutatta az extrém értékű adatok elemzésének nehézségeit a parametrikus módszerekhez viszonyítva a nem-parametrikus módszerekhez képest. A nem-parametrikus megközelítés lehetővé teszi, hogy az adatok magukért beszéljenek kevesebb feltételezés alapján. A bonyolult trendek kezelése a nem-parametrikus megközelítés segítségével hatékonyabb, mint a parametrikus megközelítés. Feltételezzük, hogy $Z_1, Z_2, \dots$ az extrém további adatok, amelyek információt nyújthatnak a modelltrendről a rekord-progresziós adatokban. Ebben az esetben az éves minima típusú adatokat fogjuk használni. Adams [16] kimutatta, hogy a

nem-parametrikus megközelítés alkalmazása előnyös a $\mu(t)$ becslésére. Amikor a nem-parametrikus megközelítést alkalmazzák az elemzésben, a $\mu(t)$ értékeket a 9 egyenletben a g_t sima függvényre cserélik az időben $t = 1, \dots, m$ időpillanatokra. A log-likelihood függvény g_t -vel:

$$l(r; g_t, \sigma, \xi) = + \sum_{i=1}^m \log \lambda(t_i, r_i) - \sum_{i=0}^m \sum_{j=1}^{k_i} \delta \left[1 - \xi \left(\frac{r_i - g_i}{\sigma} \right) \right]_+^{-1/\xi}.$$

Annak érdekében, hogy pontos integrál becslést kapjunk, felosztjuk az időt t_i és t_{i+1} között egy kis egyenlő távolságra, δ -ra az $[s_0, s_1]$ partícióktól $[s_{k_i-1}, s_{k_i}]$ partíciókig. A módosított log-likelihood:

$$l(r; g_i, \sigma, \xi) = + \sum_{i=1}^m \log \lambda(t_i, r_i) - \sum_{i=0}^m \sum_{j=1}^{k_i} \delta \left[1 - \xi \left(\frac{r_i - g_{s_j}}{\sigma} \right) \right]_+^{-1/\xi}$$

,ahol az s_0 -tól s_{k_i} -ig terjedő távolság az $i = 1, \dots, n$ időpontoknál a t_i és t_{i+1} közötti távolságot jelenti, az s_0 és s_{k_i} közötti távolságok pedig felosztva lettek $[s_0, s_1], \dots, [s_{k_i-1}, s_{k_i}]$.

6. Kétváltozós szélsőérték modell

Ebben a fejezetben a kétváltozós extrém érték modellek és a hozzájuk kapcsolódó statisztikai eljárásokat járjuk körbe, ahol az adataink sorrendi összefüggéseknek vannak kitéve. Mivel a dolgozat során különböző versenyszámok eredményeit vizsgáltuk, így feltehetjük, hogy X sztochasztikusan kisebb, mint Y . (2.3)

A kezdeti pontunk meghatározásához a független kétváltozós normális eloszlást használjuk, és a két megfelelő simítási paraméter meghatározásához az AIC_C kritérium lesz segítségünkre. A sztochasztikus rendezéssel kapcsolatos részletesebb tanulmányt *Dispersive ordering among linear combinations of uniform random variables* [8] és *Basics of applied stochastic processes* [11] írásokban találhatóak. Továbbá ebben a fejezetben felte tesszük, hogy Z_x és Z_y egymástól független. Az egymástól függő esetek vizsgálata ebben a dolgozatban nem kerül kidolgozásra.

6.1. Rendezett szélső értékek

Tegyük fel, hogy Z_x és Z_y két független minimum blokk, amelyek GEV_{min} eloszlást követnek.

$$Z_x \sim GEV_{min}(\mu_x, \sigma_x, \xi_x) \quad \text{és} \quad Z_y \sim GEV_{min}(\mu_y, \sigma_y, \xi_y),$$

és x_p, y_p kvantilis függvények.

Tétel 6.1. Legyen $Z_x \sim GEV_{min}(\mu_x, \sigma_x, \xi_x)$ és $Z_y \sim GEV_{min}(\mu_y, \sigma_y, \xi_y)$, $x_p = \mu_x + \frac{\sigma_x}{\xi_x} [1 - (-\log p)^{-\xi_x}]$ és $y_p = \mu_y + \frac{\sigma_y}{\xi_y} [1 - (-\log p)^{-\xi_y}]$ megfelelő kvantilisekkel, ahol $p \in [0, 1]$. Továbbá legyen $\xi_y = \xi_x = \xi < 0$. Ha Z_x és Z_y sztochasztikusan rendezett, ahol $Z_x < Z_y$, akkor a GEV_{min} paramétereire az alábbi feltételek teljesülnek:

$$\sigma_x < \sigma_y \quad \text{és} \quad \mu_x + \frac{\sigma_x - \sigma_y}{\xi} < \mu_y.$$

Bizonyítás. Legyen $k_p = \frac{1}{\xi} [1 - (-\log p)^{-\xi}]$, akkor $x_p = \mu_x + \sigma_x k_p$ és $y_p = \mu_y + \sigma_y k_p$ és megköveteljük, hogy $x_p < y_p$ fentálljon minden $p \in [0, 1]$ -re. Ekkor, ha $p \rightarrow 0 \Rightarrow k_0 \Rightarrow \infty$ és $p = 1 \Rightarrow k_1 = \frac{1}{\xi}$. Ha $S(k_p) = y_p - x_p$, akkor $S(k_p) = \mu_y + \sigma_y k_p - \mu_x - \sigma_x k_p$. $k_p = M$ -re, ahol $M \rightarrow \infty$

$$S_M = \mu_y - \mu_x + \sigma_y M - \sigma_x M \sim M(\sigma_y - \sigma_x), \quad \text{ahogy } M \rightarrow \infty.$$

$S(M) > 0$ -ra van szükségünk, ahogy $M \rightarrow \infty$, ekkor megkapjuk, hogy $\sigma_x < \sigma_y$. $k_1 = \frac{1}{\xi}$ esetén:

$$S\left(\frac{1}{\xi}\right) = \mu_y - \mu_x + \frac{\sigma_y}{\xi} - \frac{\sigma_x}{\xi}$$

$S\left(\frac{1}{\xi}\right) > 0$ -ra van szükség, ami esetén azt kapjuk, hogy $\mu_y - \mu_x + \frac{\sigma_y - \sigma_x}{\xi} > 0$. $S(k_p) = 0$ lineáris k_p -ben, így a szélsőértékek a k_p tartományában a k_p végpontjain fordul elő. ■

Következmény 6.2. Legyen $Z_x \sim GEV_{min}(\mu_x, \sigma_x, \xi_x)$ és $Z_y \sim GEV_{min}(\mu_y, \sigma_y, \xi_y)$ esetén teljesül a $\xi_x = \xi_y = \xi$ egyenlőség $Z_y > Z_x$ -re, akkor ha $\text{sign}(\xi) = \text{sign}(\sigma_x - \sigma_y)$, akkor $\mu_y > \mu_x + \eta(\sigma_x, \sigma_y, \xi)$, ahol $\eta(\sigma_x, \sigma_y, \xi) = \frac{\sigma_x - \sigma_y}{\xi}$ és $\eta(\sigma, \sigma, 0) = 0$.

Bizonyítás. Legyen $\xi < 0$ és $\sigma_x < \sigma_y$. Ekkor $\text{sign}(\xi) = -1$ és $\text{sign}(\sigma_x - \sigma_y) = -1$, ahol $\frac{\sigma_x - \sigma_y}{\xi} = \eta(\sigma_x, \sigma_y, \xi)$. Ekkor $Z_y > Z_x$ teljesül, ha $\text{sign}(\xi) = \text{sign}(\sigma_x - \sigma_y)$ és $\mu_y > \mu_x + \eta(\sigma_x, \sigma_y, \xi)$. ■

Általánosságban az adatok egyesítése megduplázza a becsült paraméterek számát, de a sztohasztikus rendezésnek köszönhetően ezeknek a paramétereknek a számát tudjuk korlátozni vagy csökkenteni. A kezdeti lépése az adatok modellezésének a marginális eloszlások használata külön. A 6.1 tétel alapján bebizonyítottuk, hogy $Z_x < Z_y$, akkor teljesül, ha $\xi_x = \xi_y < 0$, így a kétváltozós extrém érték eloszlás független esetben a következő:

$$H(z_x, z_y; \mu_x, \mu_y, \sigma_x, \sigma_y, \xi_x, \xi_y) = H(z_x; \mu_x, \sigma_x, \xi)H(z_y; \mu_y, \sigma_y, \xi)$$

A (2.3)definícióban tárgyalt sztohasztikus rendezés alapjait szolgáltatja, hogy várhatóan ugyan abban a versenyszámban a férfi eredmények jobbak lesznek a női eredményekkel szemben, továbbá a hosszabb távú versenyek több időt vesznek igénybe, mint a rövidebb távok, így a $Z_x < Z_y$ megvalósul.

Először kirajzoljuk a tapasztalati eloszlásfüggvényt mindegyik változóra, és megnézzük, hogy rendezettek-e. (... ábra) Második lépésként a becsült ξ_x és ξ_y paraméterek sign függvényét vizsgáljuk meg, ahol negatív érték várható mindkét esetben. A 6.1 tételből így $Z_x < Z_y$ teljesül, így a $H(z_x, z_y)$ modellt kétváltozós függvény segítségével illesztettük, ahol az általános likelihood függvény a következő:

$$L(z_x, z_y; \mu_x, \mu_y, \sigma_x, \sigma_y, \xi) = L(z_x; \mu_x, \sigma_x, \xi)L(z_y; \mu_y, \sigma_y, \xi)$$

A modell illeszkedésének vizsgálatát pedig a P-P és Q-Q plotokkal tudjuk ellenőrizni.

6.2. Kétváltozós extrém érték simítás

Tegyük fel, hogy $\{Z_x\}$ és $\{Z_y\}$ független, éves minimumok sorozatok és GEV_{min} eloszlást követnek. Ekkor (Z_x, Z_y) , független kétváltozós extrém érték eloszlást követ minimum

adatokra, g_x és g_y simítási függvénnyel a helyparaméterre, a következő eloszlással:

$$H(z_x, z_y) = \left(1 - \exp \left\{ - \left[1 - \xi_x \left(\frac{z_x - g_x}{\sigma_x} \right) \right]_+^{-1/\xi_x} \right\} \right) \times \left(1 - \exp \left\{ - \left[1 - \xi_y \left(\frac{z_y - g_y}{\sigma_y} \right) \right]_+^{-1/\xi_y} \right\} \right),$$

ahol $1 - \xi_x(z_x - g_x)/\sigma_x > 0$ és $1 - \xi_y(z_y - g_y)/\sigma_y > 0$, továbbá $\sigma_x, \sigma_y > 0$ és $-\infty < \xi_x, \xi_y < \infty$. Legyen ez az eloszláscsalád $\text{BEV}_{\min}(g_x, g_x, \sigma_x, \sigma_y, \xi_x, \xi_y)$.

Z_x és Z_y együttes sűrűségfüggvénye a következő:

$$f(x, y; \sigma_x, \sigma_y, \xi_x, \xi_y, g_x, g_y) = \frac{1}{\sigma_x} \frac{1}{\sigma_y} \left[1 - \xi_x \left(\frac{x - g_x}{\sigma_x} \right) \right]_+^{-1-1/\xi_x} \times \left[1 - \xi_y \left(\frac{y - g_y}{\sigma_y} \right) \right]_+^{-1-1/\xi_y} \times \exp \left\{ - \left[1 - \xi_x \left(\frac{x - g_x}{\sigma_x} \right) \right]_+^{-1/\xi_x} \right\} \times \exp \left\{ - \left[1 - \xi_y \left(\frac{y - g_y}{\sigma_y} \right) \right]_+^{-1/\xi_y} \right\}.$$

A regularizált log-likelihood függvény pedig:

$$\begin{aligned} l(\sigma_x, \xi_x, \sigma_y, \xi_y, g_x, g_y) &= \sum_{i=1}^n l_i(\sigma_x, \xi_x, g_{x:i}) + \sum_{i=1}^n l_i(\sigma_y, \xi_y, g_{y:i}) \\ &\quad - \frac{1}{2} \lambda_x g_x^T K_x g_x - \frac{1}{2} \lambda_y g_y^T K_y g_y, \\ l_i(\sigma_x, \xi_x, g_{x:i}) &= -\log \sigma_x - (1 + 1/\xi_x) \log \left[1 - \xi_x \left(\frac{x_i - g_{x:i}}{\sigma_x} \right) \right]_+ \\ &\quad - \left[1 - \xi_x \left(\frac{x_i - g_{x:i}}{\sigma_x} \right) \right]_+^{-1/\xi_x}, \\ l_i(\sigma_y, \xi_y, g_{y:i}) &= -\log \sigma_y - (1 + 1/\xi_y) \log \left[1 - \xi_y \left(\frac{y_i - g_{y:i}}{\sigma_y} \right) \right]_+ \\ &\quad - \left[1 - \xi_y \left(\frac{y_i - g_{y:i}}{\sigma_y} \right) \right]_+^{-1/\xi_y}, \end{aligned}$$

Végül a Fischer-scoring-ot alakítsuk át $\text{BEV}_{\min}$ eloszlásra, ahol a helyparaméter nem paraméteres a trendben.

1. Számítsuk ki K_x és K_y mátrixokat (7). egyenlet alapján.

2. Kiválasztunk kezdőértékeket a simítási paramétereknek $\lambda_{g_x}, \lambda_{g_y}$ és Ω értéket, ahol $\lambda_{g_x}, \lambda_{g_y} > 0$ és $0 \leq \Omega \leq 1$.
3. Megbecsüljük $\hat{m}_{x;t}, \hat{m}_{y;t}$ és $\hat{\sigma}_{x;N}, \hat{\sigma}_{y;N}$ értékeket a $N(m_{x;t}, m_{y;t}, \sigma_{x;N}, \sigma_{y;N})$ kétváltozós normális eloszlásból
4. Bevezetjük $\mu_{x;t}, \sigma_x, \xi_x, \mu_{y;t}, \sigma_y$ és ξ_y értékeket $\hat{m}_{x;t}$ -et, $\hat{\sigma}_{x;N}$ -t, $\hat{m}_{y;t}$ -et és $\hat{\sigma}_{y;N}$ -t használva, hogy megkapjuk $\xi_{x;\text{kezdeti}} = \xi_x$, és $\xi_{y;\text{kezdeti}} = \xi_y$, kezdeti értékeket,

$$\begin{aligned}\mu_{x;t,\text{kezdeti}} &= g_{x;t,\text{kezdeti}} = m_{x;t} - \frac{\sigma_{x;N}}{\xi_x} [1 - \Gamma(1 - \xi_x)], \\ \mu_{y;t,\text{kezdeti}} &= g_{y;t,\text{kezdeti}} = m_{y;t} - \frac{\sigma_{y;N}}{\xi_y} [1 - \Gamma(1 - \xi_y)], \\ \sigma_{x;\text{kezdeti}} &= \sqrt{\frac{\xi_x^2 \sigma_{x;N}^2}{\Gamma(1 - 2\xi_x) - [\Gamma(1 - \xi_x)]^2}}, \\ \sigma_{y;\text{kezdeti}} &= \sqrt{\frac{\xi_y^2 \sigma_{y;N}^2}{\Gamma(1 - 2\xi_y) - [\Gamma(1 - \xi_y)]^2}}\end{aligned}$$

5. Kiszámítjuk $g_{x;\text{új}}$ -t és $g_{y;\text{új}}$ -t :

$$\begin{aligned}g_{x;\text{új}} &= g_x + \left[E \left(-\frac{\partial^2 l}{\partial g_x \partial g_x^T} \right) + \lambda_{g_x} K_x \right]^{-1} \left(\frac{\partial l}{\partial g_x} - \lambda_{g_x} K_x g_x \right), \\ g_{y;\text{új}} &= g_y + \left[E \left(-\frac{\partial^2 l}{\partial g_y \partial g_y^T} \right) + \lambda_{g_y} K_y \right]^{-1} \left(\frac{\partial l}{\partial g_y} - \lambda_{g_y} K_y g_y \right)\end{aligned}$$

6. Kiszámítjuk $\theta_{\text{új}}$ -t felhasználva $g_x = g_{x;\text{új}}$ -t és $g_y = g_{y;\text{új}}$ -t,

$$\theta_{\text{új}} = \theta + \Omega A^{-1} \left(\frac{\partial l}{\partial \theta} \right),$$

ahol

$$A = \begin{bmatrix} E \left(-\frac{\partial^2 l}{\partial \xi^2} \right) & E \left(-\frac{\partial^2 l}{\partial \xi \partial \sigma_x} \right) & E \left(-\frac{\partial^2 l}{\partial \xi \partial \sigma_y} \right) \\ E \left(-\frac{\partial^2 l}{\partial \xi \partial \sigma_x} \right) & E \left(-\frac{\partial^2 l}{\partial \sigma_x^2} \right) & 0 \\ E \left(-\frac{\partial^2 l}{\partial \xi \partial \sigma_y} \right) & 0 & E \left(-\frac{\partial^2 l}{\partial \sigma_y^2} \right) \end{bmatrix},$$

és θ vagy $[\xi, \sigma_x, \sigma_y]^T$, vagy $[\xi_x, \xi_y, \sigma_x, \sigma_y]$, vagy $[\sigma]$, vagy $[\xi_y, \sigma_x, \sigma_y]^T$.

7. Teszteljük, hogy konvergál-e és legyen $\theta = \theta_{\text{új}}$ és $g = g_{\text{új}}$, majd térjün vissza az 5., 6. és 7. lépésre. Ha a változás kisebb egy előre megadott küszöbnél, akkor megállunk.

7. Összefoglalás

A dolgozat során bemutattuk, hogyan alkalmazható az extrém érték elmélet a sport-eredmények, különösen a futószámok legjobb éves teljesítményeinek modellezésére. Az extrém érték elemzés révén lehetőség nyílt a rekordok előfordulásának matematikai leírására, valamint azok időbeli alakulásának vizsgálatára. A felhasznált GEV_{min} és regularizált log-likelihood modellekkel tudtuk kezelni a szélsőértékek sajátos viselkedését és alapot biztosítottak a trendek becsléséhez. A dolgozat egyik fő pontja a nem-paraméteres simítási technikák voltak, melyek lehetővé tették a helyparaméter időbeli változásainak követését. Az így kapott trendfüggvények segítségével a sportteljesítmények értelmezhetőbbé váltak. Az alkalmazott módszerek a 800 és 1500 méteres férfi síkfutás adatsorain végzett elemzések demonstrálták ezek gyakorlati hasznosságát. A dolgozat eredményei további kutatásokban folytatódhatnak, akár a kétváltozós extrém érték modellek helyett többdimenziós esetek vizsgálata, vagy nem független versenyszámok vizsgálata. Illetve a jövőben érdemes lehet más sportágakra, illetve a női versenyszámokra is kiterjeszteni az itt bemutatott módszereket és elemzéseket.

Hivatkozások

- [1] Emil Julius Gumbel. *Statistics of extremes*. Columbia university press, 1958.
- [2] Ned Glick. „Breaking records and breaking boards”. *The American Mathematical Monthly* 85.1 (1978), 2–26. old.
- [3] Peter J Green és Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.
- [4] Ludwig Fahrmeir és tsai. *Multivariate statistical modelling based on generalized linear models*. 8. köt. Springer, 1994.
- [5] Hirotogu Akaike. „Information theory and an extension of the maximum likelihood principle”. *Selected papers of hirotogu akaike*. Springer, 1998, 199–213. old.
- [6] Stuart Coles. *An introduction to statistical modeling of extreme values*. 208. köt. Springer, 2001.
- [7] Francesco Pauli és Stuart Coles. „Penalized likelihood inference in extreme value analyses”. *Journal of Applied Statistics* 28.5 (2001), 547–560. old.
- [8] Baha-Eldin Khaledi és Subhash Kochar. „Dispersive ordering among linear combinations of uniform random variables”. *Journal of Statistical Planning and Inference* 100.1 (2002), 13–21. old.
- [9] Valérie Chavez-Demoulin és Paul Embrechts. „Smooth extremal models in finance and insurance”. *Journal of Risk and Insurance* 71.2 (2004), 183–199. old.
- [10] Valérie Chavez-Demoulin és Anthony C Davison. „Generalized additive modelling of sample extremes”. *Journal of the Royal Statistical Society Series C: Applied Statistics* 54.1 (2005), 207–222. old.
- [11] Richard Serfozo. *Basics of applied stochastic processes*. Springer Science & Business Media, 2009.
- [12] Malcolm R Leadbetter, Georg Lindgren és Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.
- [13] Hongjoon Shin és tsai. „Assessment of modified Anderson–Darling test statistics for the generalized extreme value and generalized logistic distributions”. *Stochastic Environmental Research and Risk Assessment* 26 (2012), 105–114. old.
- [14] MB Adam és Jonathan Angus Tawn. „Modelling record times in sport with extreme value methods”. *Malaysian Journal of Mathematical Sciences* 10.1 (2016), 1–21. old.

- [15] Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.
- [16] Mohd Bakri Adam. „Extreme Value Modelling of Sports Data”. unpublished.

Alulírott Csáfordi József nyilatkozom, hogy szakdolgozatom elkészítése során az alább felsorolt feladatok elvégzésére a megadott MI alapú eszközöket alkalmaztam:

<i>Feladat</i>	<i>Felhasznált eszköz</i>	<i>Felhasználás helye</i>	<i>Megjegyzés</i>
Ábra készítés	Gemini 2.5 Pro	4. fejezet	Kódban szereplő hibák javítása
Adatfeldolgozás	GPT-4o	1. és 2. fejezet	
Nyelvhelyesség ellenőrzése	Writefull	Teljes dolgozat	

A felsoroltakon túl más MI alapú eszközt nem használtam.