

Nemparaméteres próbák, lineáris modell egyváltozós esetben

## 1. Előjelpróba

Azt, hogy két mennyiség közül tipikusan melyik a nagyobb (erről volt szó az egyoldali Kolmogorov–Szmirnov-próbánál), úgy is vizsgálhatjuk, hogy nem azt kérdezzük, hogy egy adott  $t$  értéknél milyen valószínűséggel kisebbek, hanem hogy egymáshoz hasonlítjuk őket, és azt kérdezzük, hogy annak valószínűsége, hogy az első nagyobb lesz, mint a második, ugyanaz-e, mint annak valószínűsége, hogy a második nagyobb, mint az első.

Ha például kétszer megismételjük ugyanazt a kísérletet, függetlenül, ugyanolyan körülmények között, akkor ez a két valószínűség a szimmetria miatt ugyanaz, a két kísérlet sorrendje valójában felcserélhető. Ha viszont például  $X$  egy véletlenszerűen választott fővárosi munkavállaló jövedelme, míg  $Y$  egy véletlenszerűen választott vidéki munkavállaló jövedelme, akkor  $\mathbb{P}(X > Y) > \mathbb{P}(X < Y)$ , hiszen bármelyik eset előfordulhat, de (feltételezve, hogy a fővárosban tipikusan nagyobbak a jövedelmek), az első eset valószínűbb. Vagy például egy tóparti szálloda forgalma legyen májusban  $X$ , júniusban  $Y$ . Ekkor  $\mathbb{P}(X > Y) < \mathbb{P}(X < Y)$ , hiszen feltehetően a júniusi forgalom nagyobb (járványmentes évben). Ráadásul ez a kérdés akkor is értelmes, ha  $X$  és  $Y$  nem függetlenek egymástól.

Az alábbi eljárásokat a  $t$ -próbával összehasonlítva mondhatjuk, hogy egy részletesebb képet kapunk, az  $X, Y$  összefüggésére is kérdezhetünk, míg az, hogy melyik eloszlásnak nagyobb a várható értéke, erről nem ad információt (bár ha a különbség várható értékét hasonlítjuk össze 0-val, az már mondhat erről valamit). További szempont, hogy ezeknél az eljárásoknál, a tapasztalati eloszlásfüggvényen alapuló Kolmogorov–Szmirnov-próbához hasonlóan az egyes megfigyelések eloszlásáról semmit nem kell feltételeznünk, sem esetleges normális eloszlást, sem véges szórást.

Legyen  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem feltétlenül független), és folytonos eloszlásúak.

Kétoldali nullhipotézis:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

Legyen  $W$  az olyan párok száma, amikre  $Y_i > X_i$ . A  $H_0$  nullhipotézis teljesülése esetén ez binomiális eloszlású,  $n$  renddel és  $p = 0,5$  paraméterrel, hiszen minden pár esetében egymástól függetlenül  $0,5$  valószínűséggel teljesül az egyenlőtlenség. (Mivel az eloszlások folytonosak, annak valószínűsége, hogy  $X = Y$ , nulla lesz, ezzel nem kell számolnunk.) Mivel a binomiális eloszlás független azonos eloszlású indikátorok összege, érvényes rá a centrális határeloszlástétel, így a nullhipotézis esetén  $W$ -ből a várható értékét ( $n/2$ -t) levonva, majd a szórásával ( $\sqrt{np(1-p)} = \sqrt{n/4}$ ) osztva a kapott valószínűségi változó a standard normális eloszláshoz tart. Ezt közelítésként használva feltételezzük, hogy az alábbi  $z$  mennyiség eloszlása közelítőleg standard normális eloszlás, ha  $n$  elég nagy, és a standard normális eloszlás kvantiliseit, vagyis a  $z$ -próba kritikus értékeit használjuk a próbához.

A próbastatisztika legyen tehát

$$z = \frac{W - n/2}{\sqrt{n/4}}. \tag{1}$$

Elutasítjuk a nullhipotézist, ha  $|z| > \Phi^{-1}(1 - \alpha/2)$ , különben elfogadjuk. A  $p$ -érték, ugyanúgy, ahogy a  $z$ -próbánál,  $2(1 - \Phi(|z|))$  lesz.

Az egyoldali esetben:  $H_0 : \mathbb{P}(X > Y) \geq \mathbb{P}(X < Y)$ .

$$H_1 : \mathbb{P}(X > Y) < \mathbb{P}(X < Y).$$

Elutasítjuk a nullhipotézist, ha  $z > \Phi^{-1}(1 - \alpha)$ , különben elfogadjuk. A  $p$ -érték ilyenkor  $1 - \Phi(z)$ .

## 2. Wilcoxon-próba

Az előző hipotézisvizsgálati feladatban egy másik eljárást is gyakran használnak.

Legyen  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem feltétlenül független), folytonos eloszlásúak.

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

- Hagyjuk el azokat a párokat, ahol  $X_j = Y_j$ . Marad  $k$  pár.
- A megmaradt  $k$  párt állítsuk az  $|Y_j - X_j|$  szerint növekvő sorrendbe. Minél nagyobb az eltérés, annál nagyobb súllyal fog számítani.
- Minden párra számítsuk ki, hogy hányadik ebben a sorrendben, legyen ez  $R_j$ . Az 1 a legkisebb,  $k$  a legnagyobb különbség. Ha egyenlők vannak, mindegyik azonos sorszámot kapjon, a megfelelő sorszámok átlagát.
- Ezt az  $R_j$  rangot szorozzuk meg  $Y_j - X_j$  előjével, majd ezeket adjuk össze:

$$W = \sum_{j=1}^k \operatorname{sgn}(Y_j - X_j) \cdot R_j.$$

- A  $W$ -t a Wilcoxon-próba kritikus értékeihez hasonlíthatjuk.
- Ha a mintaelemszám elég nagy, a

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}}$$

mennyiségre kétoldali  $z$ -próbát alkalmazhatunk, a kritikus érték ebben az esetben  $1 - \Phi^{-1}(1 - \alpha/2)$ , ahol  $\alpha$  a szignifikanciaszint.

- Itt is lehet egyoldali ellenhipotézist is vizsgálni, akkor az egyoldali Wilcoxon-próba kritikus értékére van szükség, illetve a közelítő esetben az egyoldali  $z$ -próbának megfelelően járhatunk el.

**Wilcoxon-próba, példa.** Hat tóparti szálloda májusi és júniusi bevételét mutatja az alábbi táblázat (millió forintban). Az egyes szállodák bevételét egymástól függetlennek tekintjük, de a májusi és a júniusi érték egy szálloda esetében összefügghet. Nincsenek egyenlő értékek a párokon belül, így nem kell elhagyni mintaelemeket.

Kétoldali ellenhipotézist vizsgálunk. Legyen  $X$  a májusi,  $Y$  a júniusi bevétel:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

szálloda	A	B	C	D	E	F
májusi bevétel ( $X_j$ )	20,3	19,3	16,5	22,4	23,8	18,5
júniusi bevétel ( $Y_j$ )	25,2	22,9	14,3	26,3	21,7	22,1
a különbség abszolút értéke ( $ X_j - Y_j $ )	4,9	3,6	2,2	3,9	2,1	3,6
rang ( $R_j$ )	6	3,5	2	5	1	3,5
a különbség előjele ( $\operatorname{sgn}(Y_j - X_j)$ )	+1	+1	-1	+1	-1	+1

Ezután:

$$W = \sum_{j=1}^k \operatorname{sgn}(Y_j - X_j) \cdot R_j = 6 + 3,5 - 2 + 5 - 1 + 3,5 = 15.$$

Bár most a mintaelemszám nem elég nagy, a példa kedvéért a közelítő összeget használva  $k = 6$ -tal (hiszen hat pár van):

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}} = \frac{15}{\sqrt{6 \cdot 7 \cdot 136}} = 1,57.$$

A kétoldali  $z$ -próba esetén az  $\alpha = 0,05$ -höz tartozó kritikus érték:  $1 - \Phi^{-1}(0,025) = 1,96$ . Mivel tehát  $|z|$  kisebb a kritikus értéknél, a nullhipotézist elfogadjuk, annak valószínűsége, hogy a májusi bevétel nagyobb a júniusinál, nem tér el szignifikánsan annak valószínűségétől, hogy a júniusi szignifikánsan nagyobb a májusinál.

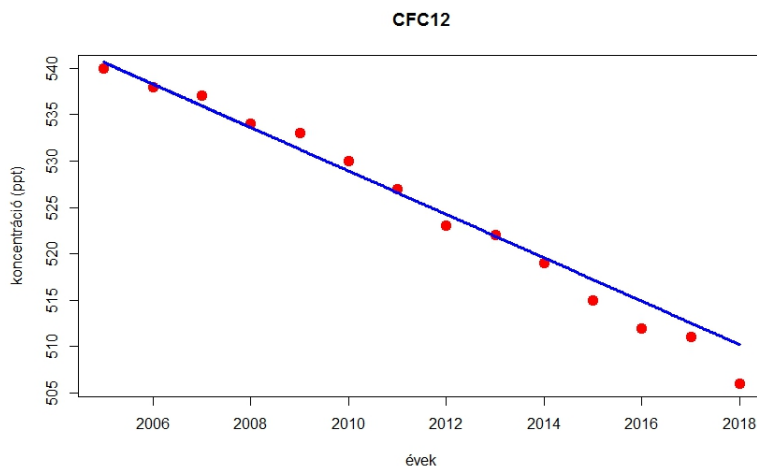
Az előjelpróbával az (1) egyenlet alapján ( $n = 6$  a párok száma):

$$z = \frac{W - n/2}{\sqrt{n/4}} = \frac{4 - 3}{\sqrt{6/4}} = 0,817$$

adódik, hiszen ott  $W$  az olyan párok száma, ahol  $Y_j > X_j$ . Erre is  $z$ -próbát végezhetünk, a kritikus érték most is  $1,96$ , ezzel az eljárással is elfogadjuk a nullhipotézist.

### 3. Lineáris regresszió

A lineáris regresszió során az a célunk, hogy egy  $f(y) = x$  függvényt, melynek értékét néhány  $x_1, x_2, \dots, x_n$  pontban ismerjük, a „lehető legjobban” közelítsünk egy egyenessel (1. ábra). Ehhez szorosan fog kapcsolódni az együtthatók becslése az úgynevezett lineáris modellben.



1. ábra. A CFC-12 (freon) gáz koncentrációja az Antarktison és az adatokra illesztett egyenes (forrás: ESRL, USA)

Egyenes illesztése a **legkisebb négyzetek módszerével**. Adottak tehát  $(x_1, y_1), \dots, (x_n, y_n)$  pontok. A leggyakrabban használt módszer esetén az illesztés hibája az egyenes által megadott  $ax_i + b$  értékeknek és a mért  $y_i$  értékeknek a különbségének négyzetösszege. Az, hogy ez milyen  $a, b$  számokra a legkisebb, egy többváltozós szélsőérték-keresési feladat, melynek megoldását az alábbi állítás adja meg.

**3.1. Állítás (Lineáris regresszió).** Legyenek  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  adott számpárok. Azokat az  $a$  és  $b$  együtthatókat keressük, melyre  $a$

$$h^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

mennyiség minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

A példában:  $\hat{a} = -2,63$ ;  $\hat{b} = 5807,7$  (a  $b$  együttható neve: intercept)

### 3.1. Lineáris modell: példa R-ben

A reziduálisok az  $y_i - (ax_i + b)$  hibák, ezekre vonatkozik egy összefoglaló statisztika.

```
> cfc12<-c(540, 538, 537, 534, 533, 530, 527, 523, 522, 519, 515, 512, 511, 506)
> ev<-c(seq(from=2005, to=2018, by=1))
> summary(lm(cfc12~ev))
```

```
Call: lm(formula = cfc12 ~ ev)
```

```
Residuals:      Min       1Q   Median       3Q      Max
 -1.8571  -0.8736   0.2088   0.8709   1.6483
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	<b>5807.73626</b>	159.19290	36.48	1.15e-13 ***
ev	<b>-2.62637</b>	0.07914	-33.19	3.55e-13 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.194 on 12 degrees of freedom
```

```
Multiple R-squared:  0.9892, Adjusted R-squared:  0.9883
```

```
F-statistic: 1101 on 1 and 12 DF, p-value: 3.554e-13
```

## 4. Lineáris modell egyváltozós esetben

A lineáris modellben az az elképzelésünk, hogy az  $Y$  valószínűségi változó az  $X$ -ből úgy kapható, hogy vesszük  $X$ -nek egy lineáris függvényét  $(aX + b)$ , majd egy normális eloszlású hibát hozzáadunk. Az  $(X, Y)$  pár eloszlásából vehetünk  $n$  elemű mintát, itt a párok egymástól már függetlenek. Ugyanakkor azért lesz egy statisztikai feladat, mert sem az  $a$ , sem a  $b$  együtthatókat nem ismerjük, sem pedig a hozzáadott hiba szórásának nagyságát. Az első feladat tehát ezen ismeretlen paraméterek becslése lesz a megfigyelt minta alapján.

**4.1. Definíció (Lineáris modell).** Legyenek  $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$  valószínűségi változók, és tegyük fel, hogy valamely  $a, b$  valós számokra

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol  $\varepsilon_1, \dots, \varepsilon_n$  független  $N(0, \sigma^2)$  normális eloszlású valószínűségi változók. Az így kapott  $(X_i, Y_i)$  párok együttes eloszlását lineáris modellnek nevezzük.

Az  $X_i$  valószínűségi változókat magyarázó változóknak, az  $\varepsilon_i$  valószínűségi változókat hibának szokták nevezni.

#### 4.1. Becslések a lineáris modellben

A maximumlikelihood-módszer alkalmazható a lineáris modell együtthatóinak becslésére. A kapott eredmények megegyeznek a lineáris regresszióban a legkisebb négyzetek módszerével kapott egyenessel az együtthatók esetében. A szórás is ismeretlen paraméter, erre is adhatunk becslést.

A becslések szórása is kiszámítható: ez azt jelenti, hogy a kiszámított becslésnek (mely a minta függvénye, ezért véletlen), mint valószínűségi változónak mennyi a szórása. Ez természetesen a  $\sigma$  paramétertől függ. Ha a becslések szórásának nagyságrendjére vagyunk kíváncsiak, ebben a képletben  $\sigma$  helyére a  $\hat{\sigma}$  maximumlikelihood-becslést írhatjuk.

**4.1. Állítás.** *A lineáris modellben az  $a, b$  együtthatók maximumlikelihood-becslése a következőképpen írható:*

$$\hat{a} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{k=1}^n (X_k - \bar{X})^2}; \quad \hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az  $a$  és  $b$  paramétereknek:

$$\mathbb{E}(\hat{a}) = a; \quad \mathbb{E}(\hat{b}) = b.$$

A hiba szórásának becslése:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}; \quad D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

A lineáris modell becslései érzékenyek a kiugró értékekre, így azokat a becslés előtt érdemes lehet eltávolítani.

#### 4.2. Előrejelzés a lineáris modellben

**4.2. Állítás.** *Legyen  $x^*$  adott szám. A lineáris modellből kapott előrejelzés az  $Y$  véletlen folyamat  $x^*$  pontban felvett értékére:*

$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

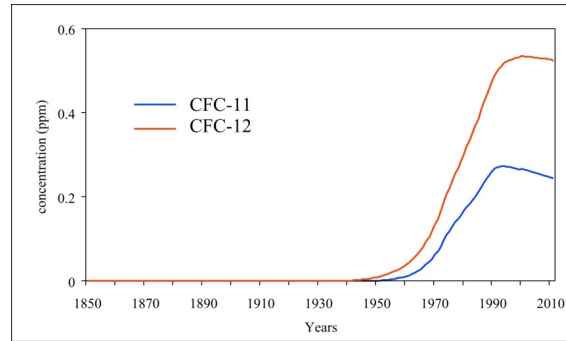
$$D(\hat{a}x^* + \hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

Az előrejelzés szórásának becslésekor a  $\sigma$  értéket gyakran  $\hat{\sigma}$ -val helyettesítik, ez információt ad a becslés pontosságáról. Minél távolabbi pontra készítjük az előrejelzést, annál nagyobb lesz a szórás.

A példában: előrejelzés  $x^* = 2019$ -re:

$$\hat{a} \cdot x^* + \hat{b} = -2,63 \cdot 2019 + 5807,7 = 497,7.$$

Ugyanakkor, ahogy a 2. ábra is mutatja, az, hogy egy folyamat egy rövidebb szakaszon jól közelíthető a lineáris modellel, nem jelenti, hogy nagyobb skálán is érvényes a közelítés (az emelkedés az ipari tevékenység következménye, a csökkenés az adott gázok gyártásának tiltásának következménye – bár a tiltás ellenére a gyártás nem állt le teljesen).



2. ábra. A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

### 4.3. Reziduálisok és $R^2$

A különböző adatsorokra a lineáris modell természetesen különböző mértékben illeszkedik. Az illeszkedés pontosságát elsősorban a reziduálisokon keresztül, vagyis a közelítő egyenes és a megfigyelt érték különbségéből érthetjük meg.

Reziduálisok:  $Y_i - \hat{a}X_i - \hat{b}$  (ezeknek a négyzetösszege minimális)

A teljes ingadozás (total sum of squares):  $\sum_{j=1}^n (Y_j - \bar{Y})^2$ .

Ezt összehasonlíthatjuk a reziduális négyzetösszeggel (residual sum of squares):

$$\sum_{j=1}^n (Y_j - \hat{a}X_j - \hat{b})^2.$$

A kettő hányadosát 1-ből levonva kapjuk az úgynevezett megmagyarázott ingadozás részarányát:

$$R^2 = 1 - \frac{\sum_{j=1}^n (Y_j - \hat{a}X_j - \hat{b})^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2}.$$

Ugyanis a reziduális négyzetösszeg és a teljes ingadozás hányadosa azt mutatja, hogy az  $Y$  tapasztalati szórásnégyzetéből milyen rész adódik az illesztett érték és a megfigyelt érték különbségéből. Az egyből levont érték tehát azt mutatja, hogy a modell jó illeszkedése esetén milyen szórásnégyzet adódna. A reziduális négyzetösszeget egy másik alakba írva látható, hogy az így kapott  $R^2$  valójában a két minta tapasztalati korrelációs együtthatójának négyzete, ezt használjuk definíciónak.

**4.2. Definíció.** A megmagyarázott ingadozás részaránya (coefficient of determination):

$$R^2 = \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\sum_{k=1}^n (X_k - \bar{X})^2][\sum_{k=1}^n (Y_k - \bar{Y})^2]}.$$

Az  $R^2$  értéke 0 és 1 közé esik.

Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. De ez nem minden szempontnak megfelelő mérőszám, és fordítva nem is feltétlenül igaz a következtetés. Például az  $R^2$  érzékeny a kiugró értékekre, néhány kiugró esetén  $R^2$  lecsökken. Vagyis az  $R^2$ -ből nem tudjuk jól eldönteni, hogy a „tipikus” értékek sem illeszkednek jól, vagy esetleg néhány pont kivételével lényegében jó az illeszkedés.

A példában:  $R^2 = 0,98$ , vagyis jól illeszkedik a lineáris modell.

Az R kódban megadott adjusted  $R^2$ : nem csak a reziduálisokat veszi figyelembe, hanem azt is, hogy hány paramétert használtunk (ennek többváltozós esetben van nagyobb jelentősége).

#### 4.4. Konfidenciaintervallumok

Amikor az együtthatókat megbecsüljük, nem csak a becslést, hanem konfidenciaintervallumot is adhatunk. Ugyanis a becslés eloszlása  $t$ -eloszlás, és ez alapján a  $t$ -próba kritikus értékeinek segítségével adhatunk konfidenciaintervallumot.

$1 - \alpha$  megbízhatósági szintű konfidenciaintervallum  $a$ -ra:

$$\left( \hat{a} - t_{n-2,\alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \hat{a} + t_{n-2,\alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right),$$

ahol  $t_{n-2,\alpha}$  az  $f = n - 2$  szabadsági fokú  $\alpha$  szignifikanciaszintű kétoldali  $t$ -próba kritikus értéke.

Az  $x^*$  pontban az előrejelzett érték becslése  $\hat{a} \cdot x^* + \hat{b}$ .

$1 - \alpha$  megbízhatósági szintű konfidenciaintervallum  $ax^* + b$ -re, azaz az  $x^*$ -ban felvett érték várható értékére:

$$\left( \hat{a}x^* + \hat{b} \pm t_{n-2,\alpha} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right).$$

Minél távolabbi pontban készítjük az előrejelzést, annál hosszabb lesz a konfidenciaintervallum.

#### 4.5. Az egyenes meredekségére vonatkozó próbák

A lineáris modell fő egyenlete:  $Y_i = aX_i + b + \varepsilon_i$ . Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól? A lineáris modellen belül ez a kérdés felel meg annak, hogy van-e egyáltalán összefüggés a két vizsgált mennyiség között.

$$H_0: a = 0 \quad H_1: a \neq 0$$

A nullhipotézis teljesülése esetén csak  $Y_i = b + \varepsilon_i$ , vagyis ez normális eloszlású,  $X_j$  eloszlásáról azonban nem volt feltételünk. Ezzel együtt, ha a nullhipotézis igaz, akkor az alábbi mennyiség  $t$ -eloszlású  $f = n - 2$  szabadsági fokkal, ezért erre  $t$ -próbát végezhetünk.

Kétoldali  $t$ -próbát végezhetünk az alábbi próbastatisztikával és  $f = n - 2$  szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha  $|t| > t_{n-2,\alpha}$ , azaz  $p < \alpha$ , akkor elutasítjuk  $H_0$ -t, az egyenes meredeksége szignifikánsan eltér 0-tól (itt  $t_{n-2,\alpha}$  az  $\alpha$  szignifikanciaszintű  $f = n - 2$  szabadsági fokú kétoldali  $t$ -próba kritikus értéke).

Ha  $|t| \leq t_{n-2,\alpha}$ , azaz  $p \geq \alpha$ , akkor elfogadjuk  $H_0$ -t, az egyenes meredeksége nem tér el szignifikánsan 0-tól.

A korábbi példában (az 1. ábra):

$$t = -33,19; \quad \alpha = 0,05; \quad n = 14; \quad f = n - 2 = 12; \quad c_{\text{krit}} = 2,19.$$

Mivel  $|t| = 33,19 > c_{\text{krit}} = 2,19$ , elutasítjuk a nullhipotézist, az egyenes meredeksége **szignifikánsan eltér 0-tól**. A  $p$ -érték:  $p = 3,6 \cdot 10^{-13} < 0,05 = \alpha$ .

A  $t$  és a  $p$  is kiolvasható az R-kódból (1.1. példa).

Egy másik kérdés: állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál? A modellen belül ez jelenti azt, hogy a vizsgált mennyiségek között pozitív irányú összefüggés van,

minél nagyobb az  $X$ , annál nagyobb az  $Y$  értéke is (természetesen a fordított irányú összefüggés is hasonlóképpen tesztelhető lenne).

$$H_0: a \leq 0 \quad H_1: a > 0$$

Továbbra is használhatjuk, hogy  $a = 0$  esetén az alábbi próbat statisztika  $t$ -eloszlású  $f = n - 2$  szabadsági fokkal.

Egyoldali  $t$ -próbát végezhetünk az alábbi próbat statisztikával és  $f = n - 2$  szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha  $t > \bar{t}_{n-2, \alpha}$ , azaz  $p < \alpha$ , akkor elutasítjuk  $H_0$ -t, az egyenes meredeksége szignifikánsan több 0-nál (itt  $\bar{t}_{n-2, \alpha}$  az  $\alpha$  terjedelmű  $f = n - 2$  szabadsági fokú egyoldali  $t$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett).

Ha  $t \leq \bar{t}_{n-2, \alpha}$ , azaz  $p \geq \alpha$ , akkor elfogadjuk  $H_0$ -t, az egyenes meredeksége nem szignifikánsan pozitív.

---

**Házi feladat május 5., szerda, 9:00-ig** Válasszunk egy, népeiséggel kapcsolatos adatsort (például a <https://ourworldindata.org/world-population-growth> oldalon a világ, az egyes földrészek, országok adatsorai is elérhetők), ahol elérhető az elmúlt 20 év adatai. Az első 10 évre illesszünk lineáris modellt, becsüljük meg az együtthatókat.

- Állíthatjuk-e a 10 év adatsora alapján, hogy a főegyüttható szignifikánsan különbözik 0-tól?
- Mennyire jó a lineáris modell illeszkedése?
- Az első 10 év alapján készítsünk előrejelzést a második tíz évre, és hasonlítsuk össze az előrejelzett és a valós értékeket.