

Lineáris modell többváltozós esetben; szórásanalízis; idősorok bevezetés

1. Többváltozós lineáris regresszió (multiple linear regression)

Természetesen az is elképzelhető, hogy az Y mennyiség nem egy, hanem több változónak a lineáris függvénye, valamilyen hiba hozzáadásával.

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük ($X_{i,p} \equiv 1$ lehet a konstans tag):

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

ahol ε_i független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók.

Például: $X_{i,1}$ az év, $X_{i,2}$ a CFC-12 kibocsátás az i . mérésnél, és $X_{i,3} = b$ egy konstans tag (vagyis az $X_{i,1}$ évben Y a koncentráció, ami az időnek és a kibocsátásnak is a függvénye). Ekkor a lineáris modell:

$$\begin{aligned} Y_1 &= a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1; \\ Y_2 &= a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2; \\ &\dots \\ Y_n &= a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n. \end{aligned}$$

Itt a_1 az, hogy milyen együtthatóval számít az év, a_2 az, hogy milyen együtthatóval számít a kibocsátás, b a konstans tag, ε pedig a véletlen hiba, amely évről évre független, azonos eloszlású.

Vektoros formában, visszatérve az általános esetre: $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, ahol X az $X_{i,j}$ megfigyelésekből készített mátrix, és $\underline{\beta} = (a_1, a_2, \dots, a_p)^T$ az együtthatók oszlopvektora.

Ezután az a_1, \dots, a_p együtthatók becslése (torzítatlan, és ugyanaz a legkisebb négyzetek módszerével és maximumlikelihood-módszerrel):

$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{Y}.$$

Az egyváltozós esethez képest most a konstans tagot másképpen vettük figyelembe, ezt is egy valószínűségi változónak tekintettük, az X vektor része. Ennek következménye, hogy az együtthatók becslésében nem kell levonni az átlagot. A konstans tag nélkül (vagyis ha $b = 0$ lenne) ugyanazt kapnánk vissza, ha $p = 1$, hiszen ekkor $X^T X = \sum_{j=1}^n X_j^2$, és $X^T Y = \sum_{j=1}^n X_j Y_j$.

A megmagyarázott ingadozás részaránya:

$$R^2 = \frac{(X^T X)^{-1} X^T \underline{Y}}{\underline{Y}^T \underline{Y}}.$$

Ez a mennyiség azonban nem csak például a kiugró értékekre érzékeny, hanem nem veszi megfelelően figyelembe a p -től való függést, vagyis a becsült paraméterek számát. Ez azért okoz problémát, mert túl sok becsült paraméter esetén gyakran megfigyelhető a túltanulás (overfitting) jelensége, amikor a paraméterek becslései jobban függnek a megfigyelések véletlen hibáiból adódó komponensétől, mint a megfigyelt rendszer valódi szerkezetétől, és ezért valójában nem lesz jó a modellillesztés és az előrejelzés sem. Ezért az R^2 -nek az alábbi módosított (adjusted) változata is gyakran használt:

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}.$$

Itt tehát $p = 0$ (ez nem is egy valódi modell) esetén az eredeti R^2 -et kapjuk vissza, és ha a megfigyelések száma nagy, p pedig kicsi, akkor szintén közel van a módosított érték az eredetihez. Ha azonban például $n = 100$ -as mintaelemszám mellett $p = 10$ -et használunk, akkor az R^2 -nek az 1-től való eltérése, amit figyelni szoktunk, nagyjából 10%-kal megnő, jelezve, hogy a mintaelemszámhoz képest túl sok lehet a paraméter.

1.1. Hipotézisvizsgálat a lineáris modellben

Ekkor is megfelelő próbastatisztikával t -próbával tesztelhetők az $a_i = 0$ hipotézisek. Ennek jelentősége a következő. A modell eredeti felírásakor kiválasztottunk p mennyiséget, melyről feltételeztük, hogy ezek olyan értelemben meghatározzák Y viselkedését, hogy egy lineáris függvényükhöz már csak egy véletlen hiba adódik hozzá. Az a_i együttható mondja meg, hogy az i . mennyiség milyen súllyal szerepel, vagyis ha például

$$Y_j = 5X_{j,1} + 3X_{j,2} + 0,02X_{j,3} + \varepsilon_j,$$

és az $X_{j,i}$ valószínűségi változók várható értéke és szórása nagyjából megegyezik, akkor Y_j -re az első mennyiség hatása a legjelentősebb, a második is ugyanennyire fontos, ugyanakkor a harmadik mennyiség sokkal kisebb súllyal szerepel, felmerülhet, hogy ezt ne is vegyük figyelembe a modellezés során. Vagyis megtehetjük, hogy az elsőként felépített lineáris modellt leszűkítjük csak azokra a változókra, amiknél az együttható szignifikánsan különbözik 0-tól, vagyis aminek jelentős hatása van a megfigyelt Y mennyiségre, újra megbecsüljük a paramétereket, és csak ezzel a leszűkített modellel számolunk tovább, feltéve, hogy ott is még jó illeszkedés kapható. Ennek az az előnye, hogy elkerülhetjük az előző részben említett túltanulás jelenségét, amikor túl sok a becsült paraméter, és nem az illesztés nem tükrözi a megfigyelt rendszer valódi szerkezetét.

A fent megfogalmazott hipotézisnél általánosabb feladatot oldunk meg, így több együttható 0 volta is egyszerre tesztelhető például.

Többváltozós lineáris modell ($X_{i,p}$ lehet a konstans tag):

$$Y_i = a_1X_{i,1} + a_2X_{i,2} + \dots + a_pX_{i,p} + \varepsilon_i, \text{ azaz } \underline{Y} = X\beta + \varepsilon.$$

Legyen H olyan $r \times p$ méretű mátrix, aminek a rangja r (itt $r < p$). Ekkor az alábbi hipotézisvizsgálati feladatot tekintjük:

$$H_0 : H\beta = 0 \qquad H_1 : H\beta \neq 0.$$

Ha például $r = 3$, akkor a nullhipotézis három olyan típusú egyenletet jelent, hogy $5a_1 + 3a_2 - 2a_3 = 0$, vagyis az együtthatók valamely lineáris kombinációja 0.

Ha például H egy sora a j . egységvektor, akkor βH egy eleme az a_j együttható, a nullhipotézis az $a_j = 0$ -t jelenti. Ha H -t különböző egységvektorokból állítjuk össze, akkor tudjuk több együttható 0 voltát egyszerre tesztelni.

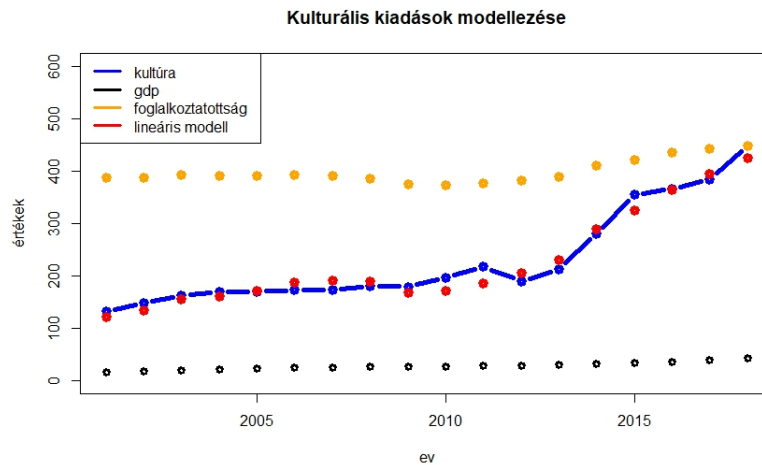
A valószínűséghányados próba (ami a Neyman–Pearson-lemmában szerepelt) próbastatisztikája:

$$F = \frac{(\underline{Y} - X\beta^*)^T(\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})},$$

ahol β^* a β becslése a $H\beta = 0$ feltétel mellett a redukált lineáris modellben (a fenti példában ez annak felel meg, amikor bizonyos magyarázó változókat nem használhatunk).

Ha H_0 igaz, akkor $F \cdot (n - p)/r$ eloszlása F -eloszlás $(r, n - p)$ szabadsági fokkal. Ezért H_0 -t elutasítjuk, ha F értéke nagyobb ennek az F -próbának a kritikus értékénél, különben elfogadjuk H_0 -t. Ha $r = 1$ és $p = 2$, valamint a próbastatisztikából gyököt vonunk, akkor az egyváltozós eset próbastatisztikáját és egy t -eloszlás abszolút értékét kapjuk, így lesz ez a korábban látott módszer általánosítása.

1.2. Többváltozós lineáris regresszió: példa



1. ábra. A költségvetés kultúrára szánt kiadásai és lineáris modell a gdp, a foglalkoztatottság és az évszám figyelembevételével (az ábrán minden mennyiség valamilyen konstansszorosra látható, a valódi nagyságrendek eltérőek)

Az alábbi adatsorok (forrás: KSH) Magyarország kultúrára fordított költségvetési összegeit (milliárd forintban), gdp-jét (milliárd forintban), illetve a Magyarországon foglalkoztatottak számát (ezer fő) mutatják 2001-2018-ig. Olyan lineáris modellt építünk, ahol Y a kultúrára fordított éves kiadás, legyen X_1 az évszám, X_2 a gdp, X_3 a foglalkoztatottak száma, $X_4 \equiv 1$ a konstans tag:

$$Y = a_1X_1 + a_2X_2 + a_3X_3 + a_4 + \varepsilon,$$

ahol $\varepsilon \sim N(0, \sigma^2)$ normális eloszlású hiba. A magyarázó változók nagyságrendje eltérő, de a becslés szempontjából ez nem baj, ha valamelyik változót átskálázzuk, a becslt együttható is átskálázódik, de a például R^2 változatlan marad.

```
kultura<-c(132, 148, 163, 170, 170, 173, 173, 181, 179, 197, 217, 190, 213, 281,
355, 366, 384, 448)
```

```
ev<-2001:2018
```

```
gdp<-c(15399, 17434, 19134, 21078, 22549, 24316, 25701, 27217, 26458, 27269, 28371,
28848, 30290, 32694, 34785, 35896, 38835, 42662)
```

```
fogl<-c(3868, 3871, 3922, 3900, 3902, 3928, 3902, 3848, 3749, 3732, 3759, 3827, 3893,
4101, 4211, 4352, 4421, 4470)
```

```
summary(lm(kultura ev + gdp + fogl))
```

```
Call: lm(formula = kultura ev + gdp + fogl)
```

```
Residuals:
```

```
Min 1Q Median 3Q Max
```

```
-22.1858 -14.4101 0.9424 10.3284 27.5662
```

```
Coefficients:
```

```
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -8.394e+03 9.580e+03 -0.876 0.396
```

```
ev 3.801e+00 4.788e+00 0.794 0.441
```

```
gdp 3.939e-03 3.896e-03 1.011 0.329
fogl 2.201e-01 3.351e-02 6.568 1.25e-05 ***
```

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

Residual standard error: 18.46 on 14 degrees of freedom

Multiple R-squared: 0.9683, Adjusted R-squared: 0.9615

F-statistic: 142.6 on 3 and 14 DF, p-value: 9.94e-11

A becslések alapján az illesztett modell (ez látható az 1. ábrán):

$$Y = 3,8X_1 + 0,0039X_2 + 0,22X_3 - 8394 + \varepsilon.$$

Az R^2 értéke 1-hez viszonylag közeli, mondhatjuk, hogy jól illeszkedik a modell. A t -próba egyedül a foglalkoztatottak számánál mutat 0-tól való szignifikáns eltérést. Ha most csak ezt a változót tartjuk meg, és így illesztünk modellt:

```
> summary(lm(kultura focl))
```

Ekkor az illesztett modell ez lenne: $Y = 0,37X_3 - 1261$, és $\tilde{R}^2 = 0,83$, ez tehát kevésbé jó illeszkedést jelent az előzőhöz képest.

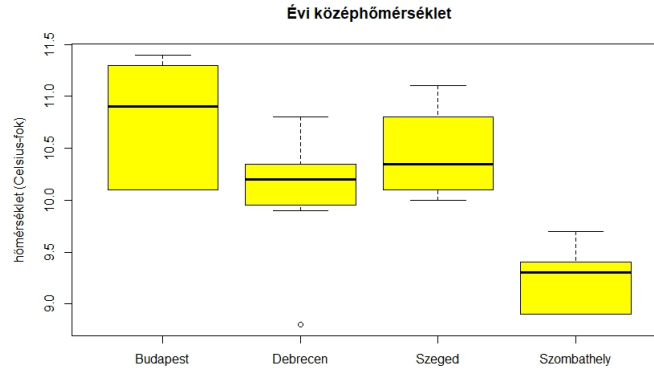
2. Szórásanalízis (analysis of variance, ANOVA)

A szórásanalízis olyan hipotézisvizsgálati eljárás, melynél ugyanazt a mennyiséget vizsgáljuk különböző csoportokba sorolt egyedek esetében, és azt szeretnénk eldönteni, hogy ennek a mennyiségnek az egyes csoportokra jellemző eloszlásának ugyanaz-e a várható értéke. Másképpen fogalmazva, igaz-e, hogy a vizsgált mennyiség várható értékére nincs hatása annak, hogy a megfigyelés melyik csoportból származik. Amint látni fogjuk, több csoport esetén ez a feladat a többváltozós lineáris modellhez is kapcsolódik. Ha viszont csak két csoport van, és feltételezzük, hogy a megfigyelések normális eloszlásúak, akkor a t -próba feladatát kapjuk vissza.

Azt, hogy a megfigyelések különböző csoportokból származnak, úgy is szokták fogalmazni, hogy a mérés egy faktor különböző szintjein történik, és az a kérdés, hogy a faktornak van-e szignifikáns hatása a várható értékre.

Példa. Az alábbi táblázat néhány éves középhőmérséklet érték (forrás: Országos Meteorológiai Szolgálat), különböző évekből, különböző helyszínekről. A kérdés: elfogadható-e, hogy az egyes városokban az évi középhőmérséklet várható értéke megegyezik, vagy szignifikáns különbség mutatható ki? A 2. ábra az adatokból készült boxplot ábrát mutatja. Ebben a példában a „faktor” a helyszín, és ennek négy „szintje” van.

	Budapest	Debrecen	Szeged	Szombathely
	10,8	8,8	11,1	8,9
	10,1	9,9	10,8	9,4
	11,4	10,0	10,1	8,9
	11,3	10,2	10,0	9,3
	11,0	10,4	10,4	9,7
	10,1	10,8	10,3	
		10,3		
átlag (\bar{X})	10,8	10,1	10,5	9,2
szórás (s_n^*)	0,57	0,63	0,42	0,34



2. ábra. Boxplot ábra az egyes városok éves középhőmérséklet adataiból

2.1. Feltevések és kapcsolat a lineáris modellel

Legyenek X_{ij} független normális eloszlású valószínűségi változók, $i = 1, \dots, k$ és $j = 1, \dots, n_i$. Az X_{ij} valószínűségi változó várható értéke μ_i , szórása σ .

$$X_{ij} \sim N(\mu_i, \sigma) \quad (j = 1, 2, \dots, n_i).$$

Vagyis: k csoport van, és a k . csoportban μ_i a várható érték. Másképpen: egy faktor különböző szintjein történik mérés, az i . csoportban a faktor i . szintjének hatása μ_i .

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k.$$

$$H_1 : \mu_1 = \mu_2 = \dots = \mu_k \text{ nem teljesül.}$$

Másképpen:

$$H_0: \text{a faktornak nincs szignifikáns hatása}$$

$$H_1: \text{a faktornak szignifikáns hatása van.}$$

Összefoglalva:

- normális eloszlások várható értékére vonatkozó próba: feltettük, hogy a megfigyelések normális eloszlásúak, és a hipotézisek a várható értékre vonatkoznak
- feltettük, hogy a megfigyelések szórása minden esetben azonos, σ
- a **kétmintás párosítatlan** Student-féle t -próba általánosításának is tekinthető: most nem kettő, hanem több csoport van, a szórások mindenhol megegyeznek

Ezt a feladatot a lineáris modell egy speciális esetének is tekinthetjük. A lineáris modell ez volt:

$$Y_j = a_1 X_{j,1} + a_2 X_{j,2} + \dots + a_k X_{j,k} + \varepsilon_j,$$

ahol $\varepsilon_j \sim N(0, \sigma^2)$ független normális eloszlású valószínűségi változók.

Most tegyük fel, hogy az $X_{j,i}$ valószínűségi változók értéke csak 0 vagy 1 lehet, sőt, hogy ezek közül mindig pontosan egy lesz 1, a többi 0 (a lineáris modellben a magyarázó változók függetlenségét nem kellett feltenni).

Ekkor ha $a_i = \mu_i$ (minden $i = 1, 2, \dots, k$ esetén), és az Y_j esetében, vagyis a j . mérésnél a k_j . valószínűségi változó 1, a többi 0, akkor $Y_j = \mu_{k_j} + \varepsilon_j$, azaz Y_j normális eloszlású μ_{k_j} várható

értékkel és σ szórással. Vagyis az Y_j -ket aszerint csoportosítva, hogy melyik $X_{j,k}$ értéke 1, éppen a p csoporthoz tartozó méréseket kapjuk vissza.

A többváltozós lineáris modellben $H\beta = 0$ alakú nullhipotéziseket tudunk tesztelni, ahol β az együtthatók vektora. Most tehát $\beta = (\mu_1, \dots, \mu_k)$, és lehet

$$H = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \dots & & \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}.$$

Ekkor $H\beta = (\mu_1 - \mu_2, \mu_2 - \mu_3, \dots, \mu_{k-1} - \mu_k)^T$, így $H\beta = 0$ éppen azzal ekvivalens, hogy minden μ_j megegyezik, ami a szórásanalízis nullhipotézise volt.

A többváltozós lineáris modell esetében a megadott próbastatisztika F -eloszlású volt a nullhipotézis mellett és az F -próba kritikus értékeit használhattuk. Mivel tehát a szórásanalízis egy speciális eset, most is hasonlóképpen járhatunk el, a próbastatisztika pedig szintén megegyezik az ott látottal, bár most más alakban írjuk fel.

2.2. A szórásanalízis eljárása

X_{ij} valószínűségi változók, $i = 1, \dots, k$, $j = 1, \dots, n_i$. Vagyis k csoport van, és az i -ben n_i darab megfigyelés van. A szórásanalízis elvégzéséhez az alábbi mennyiségekre lesz szükség.

Csoporton belüli átlagok: $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$.

Az összes megfigyelés száma: $n = n_1 + \dots + n_k$.

Teljes átlag: $\bar{X} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}$.

Csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$.

Csoportok közötti szóródás: $S_t = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$.

Teljes szóródás: $S = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = S_t + S_g$.

A próbastatisztika:

$$F = \frac{S_t(n-k)}{S_g(k-1)}.$$

Legyen c_{krit} az $f_1 = k - 1$ és $f_2 = n - k$ szabadsági fokú F -próba kritikus értéke α terjedelem mellett.

Ha $F > c_{\text{krit}}$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van legalább egy pár esetében.

Ha $F < c_{\text{krit}}$, akkor **elfogadjuk a nullhipotézist**, a várható értékek között nincs szignifikáns eltérés.

2.3. Szórásanalízis: példa

A korábbi példára visszatérve kiszámíthatjuk a csoporton belüli és csoportok közötti szóródásokat. Most feltételezzük, hogy a szórássok az egyes városok esetében megegyeznek, és hogy a középhőmérséklet normális eloszlású, az egyes helyszínek esetében egymástól független (ez utóbbi nagyjából helyes is, mert az adatok mind különböző évekből származnak).

	Budapest	Debrecen	Szeged	Szombathely	összesen
	10,8	8,8	11,1	8,9	
	10,1	9,9	10,8	9,4	
	11,4	10,0	10,1	8,9	
	11,3	10,2	10,0	9,3	
	11,0	10,4	10,4	9,7	
	10,1	10,8	10,3		
		10,3			
átlag (\bar{X}_i)	10,8	10,1	10,5	9,2	$\bar{\bar{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

A csoportokon belüli szóródás kiszámítása:

$$S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = ((10,8 - 10,8)^2 + (10,1 - 10,8)^2 + \dots + (10,1 - 10,8)^2) +$$

$$+ ((8,8 - 10,1)^2 + (9,9 - 10,1)^2 + \dots + (10,3 - 10,1)^2) +$$

$$+ ((11,1 - 10,5)^2 + (10,8 - 10,5)^2 + \dots + (10,3 - 10,5)^2) +$$

$$+ ((8,9 - 9,2)^2 + (9,4 - 9,2)^2 + \dots + (9,7 - 9,2)^2) = 5,34.$$

Itt az első sor Budapestnek (az $i = 1$ esetnek) felel meg, minden mérésnél a budapesti mérések átlagától vett különbség négyzetét számítjuk ki, és ezeket adjuk össze. A második sor, $i = 2$, Debrecen, ekkor az itteni átlagtól vett eltérések négyzetét adjuk össze, majd hasonlóképpen az $i = 3$ és $i = 4$ esetekben is.

A csoportok közötti szóródás kiszámítása:

$$S_t = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2 = 6 \cdot (10,8 - 10,17)^2 + 7 \cdot (10,1 - 10,17)^2 + 6 \cdot (10,5 - 10,17)^2 + 5 \cdot (9,2 - 10,17)^2 = 7,15.$$

Itt minden csoportra az átlagnak a teljes átlagtól vett eltérését emeljük négyzetre, majd ezt a csoport mintaelemszámával szorozzuk, és így adjuk össze az egyes csoportokra.

Teljes szóródás = csoportokon belüli + csoportok közötti:

$$S = S_g + S_t = 5,34 + 7,15 = 12,49.$$

Az előző példában: $n = 24$ a megfigyelések száma, $k = 4$ az osztályok száma.

A próbastatisztika:

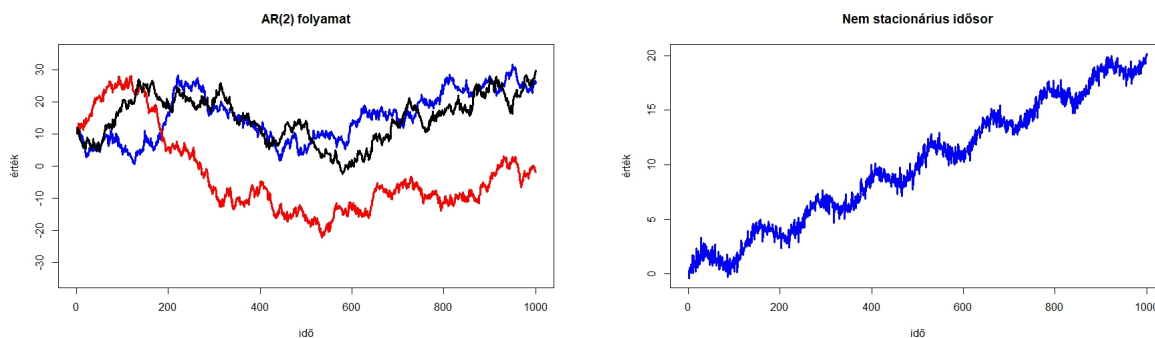
$$F = \frac{S_t(n - k)}{S_g(k - 1)} = \frac{7,15 \cdot 20}{5,34 \cdot 3} = 8,77,$$

ahol n a megfigyelések száma, k a csoportok száma, és a csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = 5,43$, a csoportok közötti szóródás: $S_t = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2 = 7,15$.

Az $f_1 = k - 1 = 3$ és $f_2 = n - k = 20$ szabadsági fokú F -próba kritikus értéke $\alpha = 0,05$ terjedelemben: $c_{krit} = 3,86$.

Mivel $F = 8,77 > c_{krit} = 3,86$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van.

Vagyis a helynek mint faktornak (tényezőnek) **szignifikáns hatása** van az évi középhőmérsékletre.



3. ábra. Példák idősorra: egy másodrendű autoregressziós folyamat (balra), illetve egy nem stacionárius idősor (egy lineáris tag, egy periodikus tag és egy stacionárius idősor összege, jobbra)

3. Idősorok elemzése

Az idősorok fogalma minden olyan elemzésben előjöhethet, ahol egy mennyiség (pl. egy ország gdp-je vagy népessége, munkanélküliségi ráta, infláció, más pénzügyi vagy gazdasági mutatók) időbeli függését szeretnénk megérteni. A lineáris modellben $Y(t) = at + b + \varepsilon(t)$ alakú idősorokat tudunk vizsgálni, amik egy lineáris függvényből és egy hozzáadott véletlen hibából állnak, de természetesen a valós folyamatok modellezésére ez a legtöbb esetben nem elég rugalmas.

3.1. Definíció. Az

$$X_0, X_1, X_2, X_3, \dots, X_t, \dots$$

valószínűségi változók sorozata idősor, ha az indexparaméter (sorszám) időpontként is értelmezhető.

Az idősorok általában **nem független** valószínűségi változókból állnak. Sőt, a következő értéket gyakran az előzőekből, egy véletlen hiba hozzáadásával számítjuk ki. Például lehet $X(1) = 10, X(2) = 12$, ezután pedig

$$X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t) \quad t = 3, 4, \dots \quad (1)$$

ahol $\varepsilon(3), \varepsilon(4), \dots$ egymástól és az korábbi X -ektől független standard normális eloszlású valószínűségi változók. A 3. ábrán ebből a modelltől sorsolt három folyamatot láthatunk.

Az összefüggéseket jellemzi például az autokovariancia-függvény. Ezt azt mondja meg, hogy az s és t időpontokban mért értékek között mennyi a kovariancia, azaz nagyjából azt, hogy ezek között mennyire erős a lineáris jellegű kapcsolat. A kovariancia függ például attól, hogy milyen mértékegységben tekintjük a mennyiségeket, viszont ha ezt az idősoron belül már nem változtatjuk, akkor az autokovariancia-függvény értékeit egymással már össze tudjuk hasonlítani.

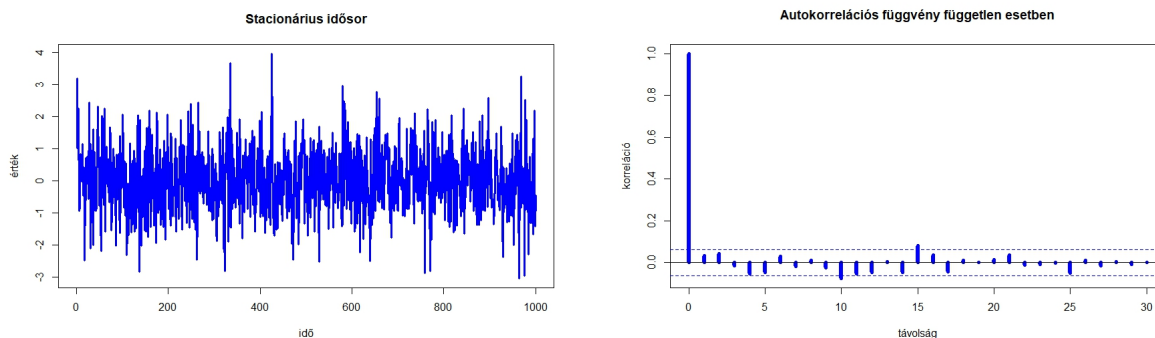
3.2. Definíció. Az X_1, X_2, \dots idősor autokovariancia-függvénye:

$$R(s, t) = \text{cov}(X_s, X_t) = \mathbb{E}(X_s X_t) - \mathbb{E}(X_s)\mathbb{E}(X_t).$$

Itt $R(t, t) = \mathbb{E}(X_t^2) - \mathbb{E}(X_t)^2 = D^2(X_t)$ a t időpontban vett szórásnégyzet. Ha viszont s és t távolságát növeljük, akkor az X_s és X_t egyre távolabbi időpontokhoz tartoznak, így sok esetben annál gyengébb közöttük az összefüggés, annál kisebb a kovariancia értéke.

3.1. Stacionárius folyamatok

Az idősorok elemzésénél gyakran a következőképpen járunk el. Az idősort az alábbi három komponens összegére bontjuk (a 3. ábrán egy olyan idősor látszik, ami három ilyen tag összegeként lett előállítva):



4. ábra. Független azonos eloszlású valószínűségi változók, mint idősor, és ennek a becslt autokorrelációs függvénye

- lineáris trend: $at + b$ alakú determinisztikus lineáris függvény;
- szezonális komponens: $f(t)$ determinisztikus periodikus függvény, melyre valamilyen h periódussal az igaz, hogy $f(t + h) = f(t)$ teljesül minden t -re;
- egy olyan X_t véletlen tag, melynek az eloszlása már t -től minél kevésbé függ, például a várható értéke és a szórása időben állandó, sőt például az X_s, X_t együttes eloszlása is csak attól függ, hogy s és t egymástól milyen messze vannak.

Például ha egy szálloda havi bevételeit szeretnénk elemezni tízéves megfigyelések alapján, akkor abban lehet egy növekvő trend (akár csak az infláció miatt is), a szezonális komponens adódhat abból, hogy például februárban feltehetően a legtöbb esetben alacsonyabb a bevétel, mint júliusban, és még ehhez adhatunk hozzá egy véletlen hibát. Ilyenkor, ha t a hónap sorszámát jelöli ($t = 1, 2, \dots, 120$), akkor a periodikus részben $f(t)$ olyan érték, ami csak t tizenkettes maradékától függ, vagyis attól, hogy ez a hónap január, február stb.

Az alábbi definíciók arra vonatkoznak, hogy a harmadik komponense, vagyis az időben állandó eloszlású véletlen részre pontosan milyen feltételeket írunk elő.

3.3. Definíció. Az X_0, X_1, X_2, \dots idősor **gyengén stacionárius**, ha

- várható értéke állandó: $\mathbb{E}(X_t) = \mathbb{E}(X_0)$ minden t -re;
- a kovariancia csak az időpontok távolságától függ:

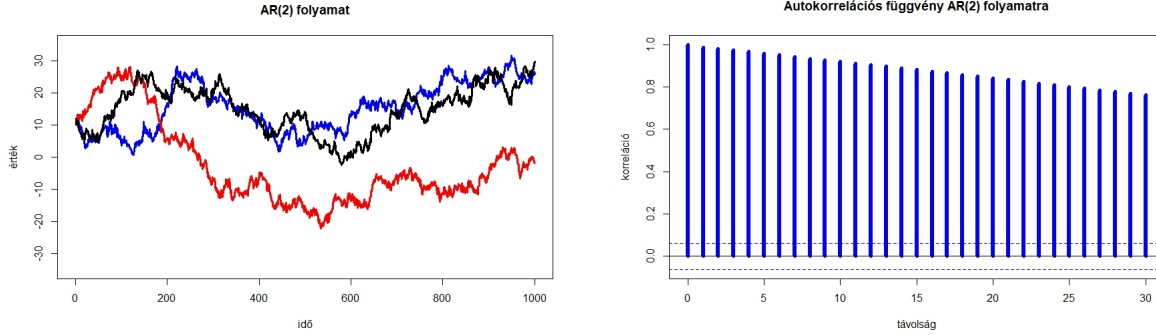
$$R(s, t) = \text{cov}(X_s, X_t) = \text{cov}(X_0, X_{t-s}) = R(0, t - s).$$

Az X_0, X_1, X_2, \dots idősor **erősen stacionárius**, ha tetszőleges n, t_1, t_2, \dots, t_n és h nemnegatív egészek esetén az

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \text{ és } (X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h})$$

valószínűségi vektorváltozók eloszlása megegyezik.

Egy erősen stacionárius idősor gyengén stacionárius, fordítva nem feltétlenül. A gyengén stacionárius esetben nem csak a kovariancia, hanem a korrelációs együttható is jól használható az adott távolságra lévő tagok közötti „lineáris jellegű” kapcsolat erősségének mérésére, hiszen a szórás is időben állandó.



5. ábra. Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat három példányá, illetve az autokorrelációs függvényének becslése

3.4. Definíció. Egy gyengén stacionárius idősor **autokorrelációs függvénye**:

$$r(t) = \frac{R(0, t)}{R(0, 0)} = \text{corr}(X_s, X_{s+t}) = \frac{\text{cov}(X_s, X_{s+t})}{D(X_s)^2} = \frac{\mathbb{E}((X_s - \mathbb{E}(X_s))(X_{s+t} - \mathbb{E}(X_{s+t})))}{D^2(X_s)},$$

ahol $s \geq 0$ tetszőlegesen választható a gyenge stacionaritás tulajdonsága miatt, és corr a két valószínűségi változó korrelációs együtthatóját jelöli.

3.2. Az autokorrelációs függvény becslése

Általában az idősort leíró modellt nem ismerjük, és ezért a várható értéket, szórást, korrelációkat sem. A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető. Az autokorrelációs függvény becslésére az alábbi módszerek szokásosak.

Legyen X_0, X_1, \dots, X_{n-1} stacionárius időorból származó n elemű minta. Az autokorrelációs függvény becslése:

$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1} (X_j - \bar{X}) \cdot (X_{j+t} - \bar{X})}{(n-t) \cdot s_n^{*2}}.$$

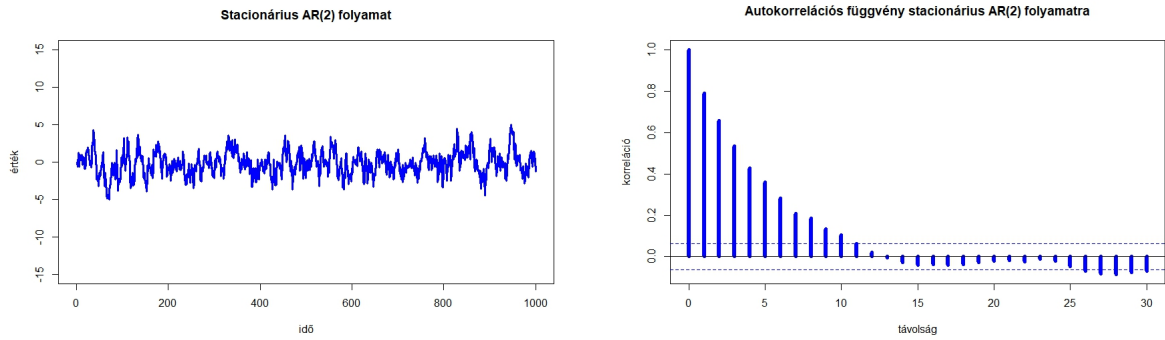
Egy másik lehetőség, hogy a tagok száma helyett n -nel osztunk:

$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1} (X_j - \bar{X}) \cdot (X_{j+t} - \bar{X})}{n \cdot s_n^{*2}}.$$

Egyik becslés sem torzítatlan $r(t)$ -re, azaz $\mathbb{E}(\hat{r}(t))$ eltér $r(t)$ -től. Ha \mathbf{x} a megfigyelésekből álló vektor, akkor az R-ben az `acf(x)` paranccsal ábrázolható az autokorrelációs függvény becslése.

A 4. ábrán egy egyszerű, független, standard normális eloszlású valószínűségi változókból álló idősort látunk. Itt valójában $r(0) = 1$ és $r(1) = r(2) = \dots = 0$. A 4. ábra jobb oldalán ennek az autokorrelációs függvénynek az adatokból kapott becslése látható, itt azonnal látszik, hogy a különböző tagok korrelációja nagyon kicsi.

Az 5. ábrán az (1) egyenlettel leírt modellnek (melynek folyamata az 5. ábra bal oldalán látható) egy megvalósításából számolt autokorreláció becslése. Itt a távolság növelésével csak lassan csökken a korreláció, távoli tagok között is erős összefüggés látszik. Ez a folyamat valójában nem is stacionárius, a becslés itt nem is jól alkalmazható. Ha a második együtthatót $0,3$ -ról $0,1$ -re csökkentjük, akkor a folyamat stacionáriussá válik (a 6. ábra), és az autokorrelációs függvény becslésén is látszik, hogy a távolság növelésével a korreláció gyorsabban csökken (a 6. ábra).



6. ábra. Az $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat, illetve az autokorrelációs függvényének becslése

Házi feladat május 12., szerda, 9:00-ig Az ismerősöktől gyűjtött adatok alapján vizsgáljuk meg, hogy ha a közösségi médiával töltött időt írjuk fel a szabadban töltött idő, a nézett sorozatok száma és az olvasott könyvek számának egy lineáris modelljeként, akkor mely együtthatók térnek el szignifikánsan a nullától, és hogy mennyire jól illeszkedik a modell.