

## 1. Autoregressziós (lineáris) folyamatok

Emlékeztetőül:

**1.1. 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.

**1.2. 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. Gyengén stacionárius esetben is az autokorrelációs függvényt így definiáltuk:

$$r(t) = \frac{\text{cov}(X_0, X_t)}{D^2(X_0)}.$$

Ez tetszőleges  $s$ -re megadja az  $X_s$  és  $X_{s+t}$  valószínűségi változók korrelációs együtthatóját.

Az idősorok modellezésénél az egyes tagok közötti összefüggést akarjuk modellezni. Erre egy lehetőség, hogy az időben következő tagot az előzőek egy lineáris kombinációjával, majd pedig egy véletlen hiba hozzáadásával képezzük. Például lehet  $X(0)$  és  $X(1)$  tetszőleges, majd pedig

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

ahol  $\varepsilon(t)$  az  $X(0), X(1), \dots, X(t-1)$  valószínűségi változóktól független standard normális eloszlású valószínűségi változó. Ennek egy megvalósítása látható az 1. ábrán. Az ilyen folyamatokat autoregressziós folyamatnak hívjuk.

Az egyenlet kis módosításával, a kettővel korábbi tagtól való kisebb függéssel másképp viselkedik a folyamat, ez a 2. ábrán látható.

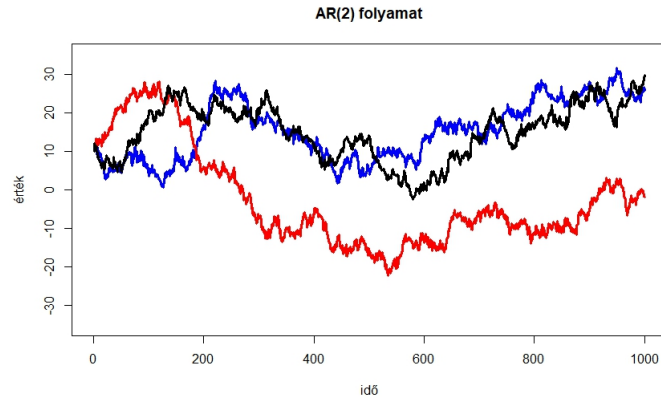
Általában az autoregressziós folyamatok esetében a korábbi tagoktól való függés időben távolabb is visszamehet, és természetesen az együtthatók is szabadon választhatók, hogy a modell kellően rugalmas legyen (ez segít a valós idősorokra való illesztésben). Így kapjuk az alábbi definíciót. A hibatagnak pedig nem kell feltétlenül normális eloszlásúnak lennie, csak azt tesszük fel, hogy 0 várható értékű és  $\sigma < \infty$  szórású.

**1.3. Definíció.** Az  $X(t)$  folyamat  **$p$  rendű autoregressziós folyamat**, ha minden  $t \geq p$ -re

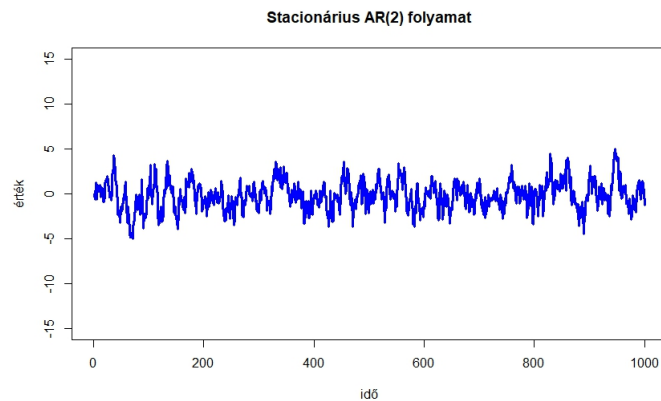
$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sigma \cdot \varepsilon(t),$$

ahol  $\varepsilon(t)$  független 0 várható értékű 1 szórású valószínűségi változó  $t \geq 0$ -ra (például normális eloszlásúak), és  $X(0), \dots, X(t-1)$ -től és  $\varepsilon(0), \dots, \varepsilon(t-1)$ -től is független. Jelölés:  $\text{AR}(p)$ .

Az előző példában tehát  $p = 2$  a rend,  $\alpha_1 = 0,7$ ,  $\alpha_2 = 0,3$  és  $\sigma = 1$ , valamint  $\varepsilon(t)$  minden  $t$ -re normális eloszlású.



1. ábra. Az  $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$  egyenletű AR(2) folyamat három trajektóriája – **ez nem stacionárius**



2. ábra. Az  $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$  egyenletű AR(2) stacionárius folyamat

**1.1. Állítás.** Az elsőrendű autoregressziós folyamatok pontosan akkor van erősen stacionárius megoldása, ha  $|\alpha_1| < 1$ .

Általában, egy AR( $p$ ) folyamatnak pontosan akkor van erősen stacionárius megoldása, ha az  $x^p + \alpha_1 x^{p-1} + \alpha_2 x^{p-2} + \dots + \alpha_p$  egyenlet minden gyökének egyénél kisebb az abszolút értéke.

A tagok közötti összefüggést többek között a korrelációs együtthatóval tudjuk mérni. Ha a folyamat stacionárius (mint a 2. ábrán látható esetben), akkor az együttes eloszlás is csak a két időpont távolságától függ, vagyis  $R(X(0), X(t))$  ugyanannyi, mint  $R(X(s), X(s+t))$  tetszőleges  $s$  egészre. Ezt a mennyiséget, vagyis a  $t$  távolságra lévő tagok korrelációs együtthatóját, amit most csak  $R(t)$ -vel jelölünk, az alábbi egyenletrendszer megoldása.

**1.2. Állítás.** Ha egy  $p$ -rendű autoregressziós folyamat gyengén stacionárius, azaz várható értéke állandó és a kovariancia csak a távolságtól függ, akkor az alábbiak teljesülnek az autokovariancia-függvényére:

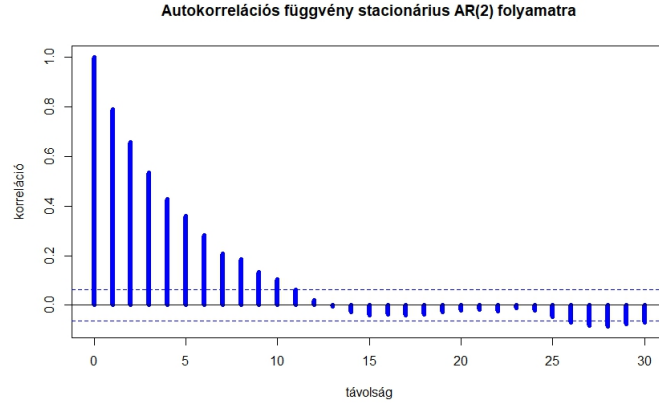
$$R(0) = \alpha_1 R(1) + \alpha_2 R(2) + \dots + \alpha_p R(p) + \sigma^2;$$

$$R(t) = \alpha_1 R(t-1) + \alpha_2 R(t-2) + \dots + \alpha_p R(t-p),$$

ahol  $t \geq 1$  tetszőleges egész. Itt  $\sigma$  a hibatag szórása. Ebből az autokorrelációs függvényre az alábbi összefüggés adódik:

$$r(t) = \alpha_1 r(t-1) + \alpha_2 r(t-2) + \dots + \alpha_p r(t-p).$$

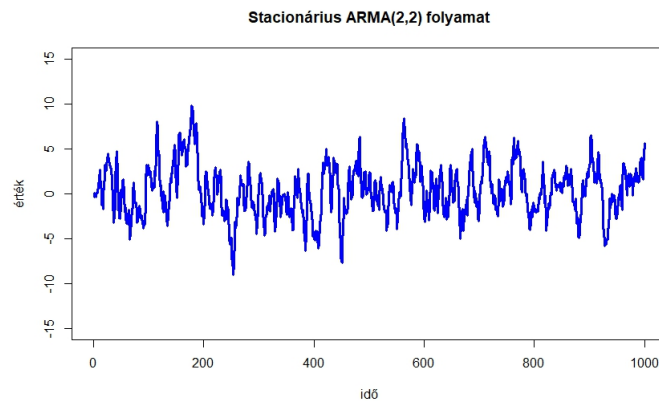
A stacionárius autoregressziós folyamatok úgynevezett rövid emlékezetű folyamatok:  $\sum_{t=0}^{\infty} R(t) < \infty$ , azaz  $\sum_{t=0}^{\infty} r(t) < \infty$ .



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

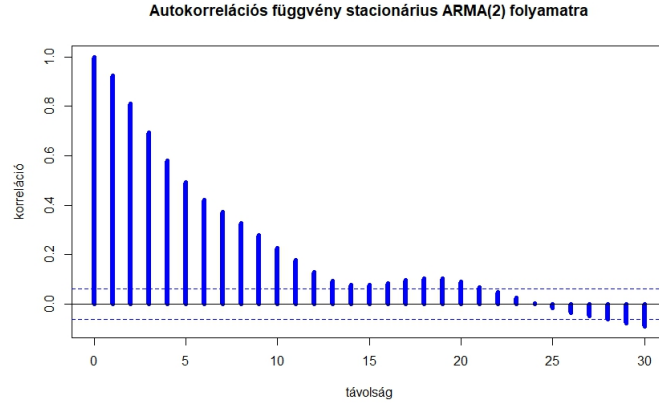
Ha a folyamatból van egy mintánk, akkor a korábban látott módon az autokorrelációs függvényt az adatokból is megbecsülhetjük. A 2. ábra másodrendű autoregressziós folyamata esetében ez a becslés a 3. ábrán látható. Ebben az esetben a korrelációs együttható tehát a távolság függvényében fokozatosan csökken, 10 távolságra pedig már csak minimális az értéke, itt feltehetően kicsi az összefüggés (bár tudjuk, hogy a korrelációs együttható nem minden összefüggést mutat ki, de ebben az esetben, ha lenne, lineáris jellegű lenne, és az látszana).

## 2. Autoregressziós mozgóátlag-folyamatok (általánosabb lineáris folyamatok)



4. ábra. Az  $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + 0,7 \cdot \varepsilon(t) + 0,2 \cdot \varepsilon(t-1) + 0,2 \cdot \varepsilon(t-2)$  egyenletű ARMA(2,2) stacionárius folyamat

Az idősorok modellezésénél eddig azt tételeztük fel, hogy a következő tag az előző néhánytól, illetve egy, a korábbiaktól független hibatagtól függ. Ez a modell gyakran még nem elég rugalmas ahhoz, hogy valós adatokra jól illeszthető legyen, ezért többféle általánosítást szokták vizsgálni. Ezek közül az egyik leggyakrabban használt az a modell, amikor a hibatagok függetlensége helyett azt tételezzük fel, hogy egy hibatag (például valamilyen külső hatás) több tagra is egyformán befolyással van. Például ha  $X(t)$  egy gazdasági mutató, akkor mondhatjuk, hogy egy, a  $t$  évben bevezetett kormányzati intézkedés az  $X(t), X(t+1), \dots, X(t+q)$  értékekre is hatással van, de természetesen, ahogy telik az idő, egyre kisebb súllyal. Ez alapján juthatunk el az alábbi definícióhoz (a korábbi tagoktól való lineáris függést megtartva, és az egyes időpontok külső hatásait függetlennek feltételezve).



5. ábra. Az  $X(t) = X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + 0,7 \cdot \varepsilon(t) + 0,2 \cdot \varepsilon(t-1) + 0,2 \cdot \varepsilon(t-2)$  egyenletű ARMA(2,2) stacionárius folyamat autokorrelációs függvényének becslése

**2.1. Definíció.** Legyenek  $\varepsilon(t)$  független 0 várható értékű 1 szórású valószínűségi változók  $t \geq 0$ -ra (például normális eloszlásúak). Az  $X(t)$  folyamat  $p, q$ -rendű autoregressziós mozgóátlag-folyamat, ha minden  $t \geq p$ -re

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sum_{m=0}^q \beta_m \varepsilon(t-m).$$

Jelölés: ARMA( $p, q$ ).

Például egy másodrendű autoregressziós ARMA(2,2) folyamat ( $\alpha_1 = 0,7, \alpha_2 = 0,3, \beta_0 = 0,7, \beta_1 = 0,2, \beta_2 = 0,2$ ):

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

Ennek egy megvalósítása látható a 4. ábrán, az ehhez tartozó autokorrelációs függvény becslése pedig az 5. ábrán.

A stacionárius ARMA-folyamatok **rövid emlékeztűek**:  $\sum_{t=1}^{\infty} R(t) < \infty$ , azaz  $\sum_{t=1}^{\infty} r(t) < \infty$ .

### 3. Lineáris folyamat illesztése és előrejelzés

Ebben a példában Magyarország népességi adatait (2001-2018-ig) elemezzük (forrás: Központi Statisztikai Hivatal). Ahogyan a 6. ábra sugallja, ez az idősor nem stacionárius, úgy tűnik, hogy a várható érték az idővel csökken. Ugyanakkor a lineáris regresszióval kapott egyenes jól illeszkedik, szezonális (periodikus) komponens pedig nem várható. Ezért a következőképpen járunk el:

- azt feltételezzük, hogy a népesség  $N(t)$  folyamata egy determinisztikus lineáris függvény és egy stacionárius folyamat összege:

$$N(t) = at + b + X(t),$$

ahol  $X(t)$  stacionárius (ebből következik, hogy az eloszlása minden  $t$ -re azonos);

- lineáris regresszióval meghatározzuk az  $a$  és  $b$  paraméterek becslését;
- az  $X(t) = N(t) - \hat{a}t - \hat{b}$  folyamatra egy autoregressziós mozgóátlag-folyamatot illesztünk:

$$X(t) = \hat{\alpha}_1 X(t-1) + \hat{\alpha}_2 X(t-2) + \dots + \hat{\alpha}_p X(t-p) + \hat{\sigma} \varepsilon(t);$$

- ebből  $N(t)$ -re is megkapjuk az illesztett modellt, a lineáris trend  $X(t)$ -hez való hozzáadásával;

- előrejelzés: ha a folyamat  $t$ -ig való megfigyelése alapján  $t + 1$ -re szeretnénk előrejelezni, akkor  $\varepsilon(t + 1)$ -et nullának tekintve (hiszen a várható értéke 0) az alábbi adódik:

$$\hat{X}(t + 1) = \hat{\alpha}_1 X(t) + \hat{\alpha}_2 X(t - 1) + \dots + \hat{\alpha}_p X(t - p + 1).$$

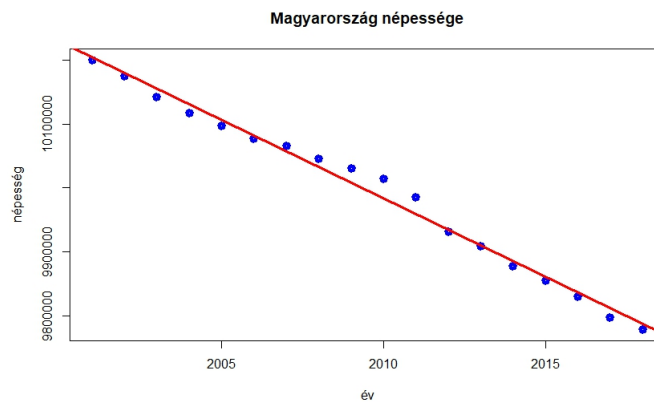
Ezt folytathatjuk is, a még nem megfigyelt időpontokban a becült értéket beírva, akár további lépéseket is tehetünk (természetesen a becslés szórása ilyenkor növekszik):

$$\hat{X}(t + 2) = \hat{\alpha}_1 \hat{X}(t + 1) + \hat{\alpha}_2 X(t - 1) + \dots + \hat{\alpha}_p X(t - p + 1).$$

Ahhoz viszont, hogy az eredeti folyamatra vonatkozó előrejelzést megkapjuk, a lineáris trendet újra figyelembe kell venni:

$$\hat{N}(t + s) = \hat{a}(t + s) + \hat{b} + \hat{X}(t + s).$$

Ez az eljárás ugyan használja a lineáris regressziót, de mivel a lineáris trendhez képest számolt hibategyek összefüggéseit is modellezzük, amikor az  $X(t)$  részt becsljük, ez a modell rugalmasabb és jobb közelítést ad, mint a lineáris modell önmagában, ahol a hibákat egymástól függetlennek tételeztük fel.



6. ábra. Magyarország népessége 2001-től 2018-ig (forrás: Központi Statisztikai Hivatal) és a regressziós egyenes

### 3.1. A lineáris trend meghatározása és eltávolítása

Magyarország népességét vizsgáljuk 2001-től 2018-ig (forrás: Központi Statisztikai Hivatal). Az első lépés tehát a regressziós egyenes meghatározása.

```
ev<-2001:2018
nep<-c(10200298, 10174853, 10142362, 10116742, 10097549, 10076581, 10066158, 10045401, 10030975,
10014324, 9985722, 9931925, 9908798, 9877365, 9855571, 9830485, 9797561, 9778371)
summary(lm(nep~ev))
```

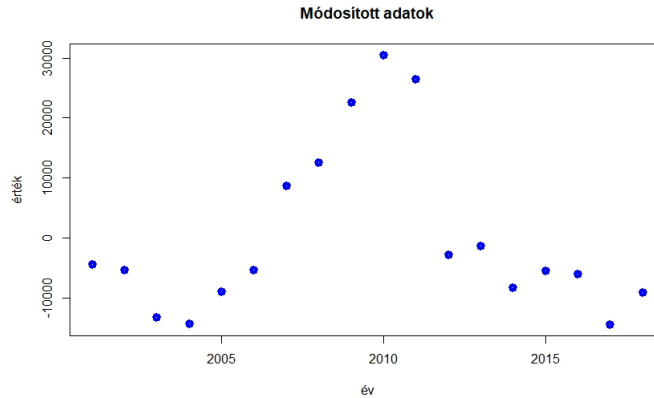
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	59315833.4	1320991.3	44.90	<2e-16 ***
ev	-24543.3	657.4	-37.34	<2e-16 ***

```
plot(nep~ev, lwd="5", col="blue", main="Magyarország népessége", xlab="év", ylab="népesség")
lines(abline(b=-24543.3, a=59315833.4, lwd="3", col="red"), xlim=c(2000, 2020))
```

A fenti gondolatmenet alapján a folyamatból a becült együtthatókkal kapott lineáris trendet eltávolítjuk, és az így kapott folyamatot vizsgáljuk majd tovább. Vagyis a lineáris regresszióval kapott függvényt kivonjuk az eredeti adatsorból:

$$X(t) = N(t) - \hat{a} \cdot t - \hat{b},$$

ahol  $N(t)$  a népesség a  $t$  időpontban, a regressziós egyenes pedig  $\hat{a}x + \hat{b}$  egyenletű. A 7. ábrán a kivonás után kapott idősor láthatjuk.



7. ábra. Magyarország népessége 2001-től 2018-ig a lineáris trend eltávolítása után

```
x<-nep+24543.3*ev-59315833.4
plot(x~ev, lwd="5", col="blue", main="Módosított adatok", xlab="év", ylab="érték")
```

### 3.2. Autoregressziós modell illesztése

Tehát ha  $N(t)$  az eredeti idősor, és  $\hat{a}, \hat{b}$  a lineáris regresszióval kapott becslések, akkor a lineáris trend elhagyása után az alábbi folyamatot kapjuk (7. ábra)

$$X(t) = N(t) - \hat{a}t - \hat{b}.$$

Erről feltételezzük, hogy stacionárius eloszlású autoregressziós folyamat.

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \dots + \alpha_p X(t-p) + \sigma \varepsilon(t),$$

ahol  $\alpha_1, \dots, \alpha_p, \sigma$  és maga  $p$  is ismeretlenek, a  $\varepsilon(t)$  valószínűségi változók pedig függetlenek, 0 várható értékűek, 1 szórásúak.

A feltételezésből viszont a folyamat rendje,  $p$  nem derül ki, és ezzel kapcsolatban a többváltozós lineáris modellhez hasonlóan előfordulhat a túltanulás („overfitting”) jelensége. Ha  $p$ -t elég nagyra választjuk, akkor rengeteg szabad paramétert állíthatunk be, így könnyen találhatunk olyan  $\alpha_1, \dots, \alpha_p$  számokat, amivel a fenti egyenlet jól illeszkedik a megfigyelt adatokra. Azonban ezzel nem a modell strukturális tulajdonságait találtuk meg, hanem, kis túlzással, minden véletlen hibához beállítottunk egy paramétert, vagyis a paraméterek valójában a véletlen hibáktól (az  $\varepsilon(t)$ -ktől) függenek. Az  $\varepsilon(t)$ -t viszont a függetlenség miatt nem tudjuk előrejelezni, így bár az illesztés jó, az előrejelzés nem lesz az.

Az autoregressziós folyamat illesztése ezért a következő módon működhet (ez az Akaike-féle információs kritérium elve, de lehetnek más módszerek is természetesen):

- többféle különböző  $p$ -t tekintünk külön-külön
- ezekre a rögzített  $p$ -re meghatározzuk, hogy melyik az  $\alpha_1, \dots, \alpha_p, \sigma$  paraméterbeállítás, amire a megfigyelt folyamat likelihood-függvénye a legnagyobb, vagyis maximumlikelihood-becslést végzünk
- minden  $p$ -re az így kapott maximális likelihood értéket megszorozzuk egy  $p$ -től függő tényezővel, ami annál kisebb, minél nagyobb  $p$  (ez a tag „büntető” a túl sok paraméter választását)
- végül azt a  $p$ -t és azokat az együtthatókat választjuk, ahol a szorzat a legnagyobb.

Tehát ha túl kevés paraméter van, akkor ugyan a büntető tényező értéke nagy, de a kevés paraméter miatt nem találunk olyan beállítást, ami jól illeszkedik, vagyis a likelihood lesz kicsi, ha pedig túl sok paraméter van, akkor a likelihood nagy, de a büntető tényező kicsi, a szorzat újra kicsi lesz. Így olyan  $p$ -t választunk végül, ami a két szempontból együttesen a legmegfelelőbb.

A példában a kiválasztott rend  $p = 2$  lesz (itt  $n = 18$ , így a paraméterek száma sem lehet 2-nél sokkal nagyobb):

```
> ar(x)      # autoregressziós modellt illesztünk
```

```
Call:  ar(x = x)
```

```
Coefficients:
```

```
      1      2  
1.0115 -0.3336
```

```
Order selected 2      sigma^2 estimated as 84281456
```

---

Tehát az Akaike-féle információs kritérium szerint illesztett autoregressziós folyamat:

$$X(t) = 1,01 \cdot X(t-1) - 0,33 \cdot X(t-2) + 9180 \cdot \varepsilon(t),$$

ahol  $\varepsilon(t)$  korrelálatlan, 0 várható értékű 1 szórású valószínűségi változók.

### 3.3. Előrejelzés

Ha kiválasztottuk  $p$ -t és az együtthatók becslését, akkor az előrejelzés a következők alapján megy. Az  $\varepsilon(t)$  várható értéke 0, és ez a tag a korábbiaktól független, így semmilyen információval nem rendelkezünk róla, a legjobb eredményt akkor kapjuk, ha az előrejelzésnél a várható értékének, azaz 0-nak tekintjük.

Előrejelzés 2019-re a módosított idősorban az  $X(2019)$  várható értéke (`predict(ar(x), n.ahead=1)`):

$$X(2019) = 1,01 \cdot X(2018) - 0,33 \cdot X(2017) = 1,01 \cdot (-9083) - 0,33 \cdot (-14436) = -4409,95$$

Ahhoz, hogy az eredeti idősorra vonatkozó előrejelzést megkapjuk, hozzá kell adni a regressziós egyenesből kapott értéket:

$$\hat{N}(2019) = \hat{a} \cdot 2019 + \hat{b} + \hat{X}(2019) = -24543,3 \cdot 2019 + 59315833,4 - 4409,95 = \mathbf{9758501}.$$

A valós adat:  $N(2019) = 9772756$ . Ez 0,15%-os relatív hibát jelent.

Ha az idősolelemzést kihagyva, csak a lineáris regresszióval számoltunk volna:

$$-24543,3 \cdot 2019 + 59315833,4 = 9762911.$$

Ez a példában még pontosabb, de hosszabb idősorok esetén érdemes lehet ezt a módszert is használni (ott lényegesen több információ áll rendelkezésre a paraméterek becslésére), főleg olyan esetekben, ahol a lineáris trend körüli ingadozás nagyobb, vagy például a lineáris trend nem is jelenik meg (0 a főegyüttható), ilyenkor ugyanis idősolelemzés nélkül lényegében csak a korábbiak átlaga lenne az előrejelzés, az utóbbi néhány taggal való összefüggést a lineáris regresszió nem tudja figyelembe venni.

---

**Házi feladat május 13., 8:15-ig** Válasszunk egy gazdasági vagy társadalmi mutatót, melyről a ksh.hu oldalon található éves adatok, legalább 15. A 2018-ig tartó adatok alapján illesszünk idősort lineáris trend és egy autoregressziós folyamat komponensekkel, készítsünk előrejelzést a 2019-es évre, majd hasonlítsuk össze az előrejelzett és a valós értéket.

---

**Házi feladat május 6., 8:15-ig** Az ismerősöket osszuk három csoportba a nézett sorozatok száma alapján (pl. 0 – 2, 3 – 5, legalább 6) úgy, hogy minden csoportba kerüljön legalább négy ismerős. Szórásanalízis segítségével vizsgáljuk meg, hogy a nézett sorozatok számának, mint faktornak (aminek most csak azt a három lehetséges értékét, szintjét tekintjük, ami a csoportosításnál létrejött), van-e szignifikáns hatása az utazási időre. Az utazási időről a feladat megoldása során feltételezhetjük, hogy normális eloszlású (holott a kerekítések miatt pontosan biztosan nem az).

A három csoport a sorozatok száma szerint: 0, 1 – 2, illetve legalább 3. Ezeknek a kódjai 0, 1, 2 lesznek. Az alábbi kódoló függvényt alkalmazzuk elemenként a sorozatok számát tartalmazó vektorra, majd utána lehet a szórásanalízist elvégezni.

```
> csoport <- function(x)if (x==0) 0 else if (x<3) 1 else 2
```

```
> szint = sapply(sorozat, csoport)
```

```
> szint
```

```
[1] 1 0 2 1 0 2 0 1 0 0 ...
```

```
> summary(aov(utazas~szint))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
szint	1	12697	12697	3.756	<b>0.0566</b>
Residuals	71	240040	3381		

---

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

A nullhipotézis az, hogy mindhárom csoportban ugyanaz az utazási idő várható értéke. Az ellenhipotézis az, hogy legalább két csoport esetében eltérő a várható érték. Mivel a  $p$ -érték **0,0566**  $>$  **0,05**, a nullhipotézist elfogadjuk, nincs szignifikáns eltérés a 0, 1 – 2 illetve legalább 3 sorozatot nézők utazási idejének várható értéke között.