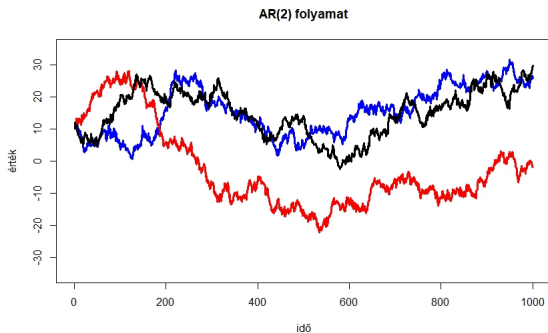
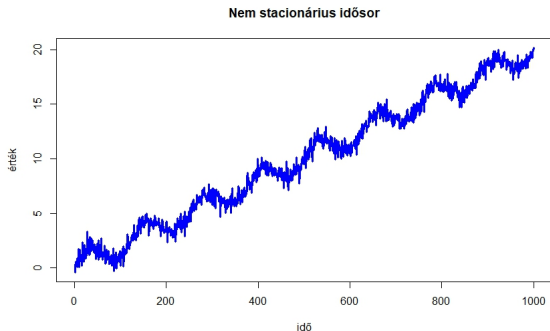


Idősorok elemzése (11. előadás)



Példák idősorra: egy másodrendű autoregressziós folyamat

Idősorok elemzése



Nem stacionárius idősor (egy lineáris tag, egy periodikus tag és egy stacionárius idősor összege)

Házi feladat május 16., kedd, 12:00-ig

A félév elején gyűjtött adatokból illesztünk lineáris modellt a sportolással töltött időre, úgy, hogy a magyarázó változók

- a) az utazással töltött idő
- b) az utazással töltött idő és a nézett sorozatok száma
- c) az utazással töltött idő, a nézett sorozatok száma és az is, hogy hányszor járnak munkába/iskolába

Melyik modell illeszkedik a legjobban? A legjobban illeszkedő modellben melyek azok a mennyiségek, amiknek az együtthatója szignifikánsan eltér 0-tól?

(Vegyük észre, hogy a feltételek nem igazán teljesülnek a kerekítések miatt, de most ettől tekintsünk el.)

Házi feladat május 16., kedd, 12:00-ig

```
> library(readxl)
```

```
> adatok <- read_excel("sstadatn.xlsx")
```

```
> summary(lm(adatok$sport~adatok$utazas))
```

```
Multiple R-squared:  0.04803, Adjusted R-squared:  0.03597
```

Házi feladat május 16., kedd, 12:00-ig

```
> library(readxl)
> adatok <- read_excel("sstadatn.xlsx")
> summary(lm(adatok$sport~adatok$utazas+adatok$sorozat))
```

Multiple R-squared: 0.0662, Adjusted R-squared: 0.02981

Mindegyik modell nagyon rosszul illeszkedik, még az első volt a legjobb a módosított R^2 szempontjából.

Házi feladat május 16., kedd, 12:00-ig

```
> library(readxl)
```

```
> adatok <- read_excel("sstadatn.xlsx")
```

```
> summary(lm(adatok$sport~adatok$utazas+adatok$sorozat+adatok$nap))
```

```
Multiple R-squared:  0.05252, Adjusted R-squared:  0.02822
```

Házi feladat május 16., kedd, 12:00-ig

```
> library(readxl)
> adatok <- read_excel("sstadatn.xlsx")
> summary(lm(adatok$sport ~ adatok$utazas))

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.3774 27.9098 2.629 0.0103 *
adatok$utazas 0.6079 0.3045 1.996 0.0493 *
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bár ez is nagyon rosszul illeszkedett, itt mindkét együttható szignifikánsan eltér a nullától.

Idősorok elemzése

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 korábbi ábrán ebből a modelltől sorsolt három folyamatot láthatunk.

Autokovariancia-függvény

Az egyes időpontokhoz tartozó valószínűségi változók közötti (lineáris) összefüggés erősségét az alábbi függvénnyel mérhetjük meg.

Definíció

Az X_1, X_2, \dots idősor autokovariancia-függvénye:

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

Autokovariancia-függvény

Az egyes időpontokhoz tartozó valószínűségi változók közötti (lineáris) összefüggés erősségét az alábbi függvénnyel mérhetjük meg.

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.

Idősorok elemzése

Az idősorok elemzésénél gyakran a következőképpen járunk el. Az idősort az alábbi három komponens összegére bontjuk (a 2. ábrán egy olyan idősor látszik, ami három ilyen tag összegeként lett előállítva):

- 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.

Ezek közül a harmadik komponens gyakran úgynevezett stacionárius folyamat.

Stacionárius folyamatok

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.

Autokorrelációs függvény

Stacionárius esetben a szórás is állandó, ezért az autokovariancia függvény mellett az autokorrelációs függvényt is gyakran használják.

Definíció

Egy gyengén stacionárius idősor **autokorrelációs függvénye**:

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

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

Az autokorrelációs függvény becslése

A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető.

Az autokorrelációs függvény becslése

A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető.

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

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

Az autokorrelációs függvény becslése

A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető.

Legyen X_0, X_1, \dots, X_{n-1} stacionárius idősről 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:

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

Az autokorrelációs függvény becslése

A várható érték a stacionárius esetben állandó, így az átlaggal torzítatlanul becsülhető.

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

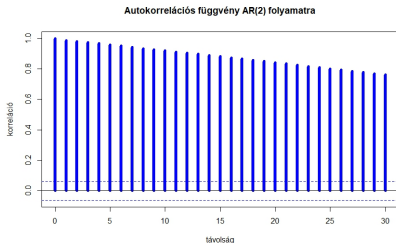
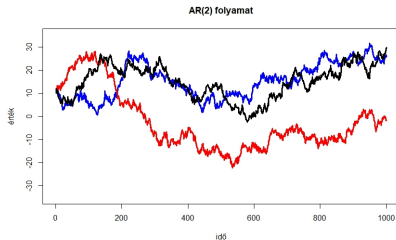
$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1} (X_j - \bar{X}) \cdot (X_{j+t} - \bar{X})}{(n-t) \cdot s_n^{*2}}.$$

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

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

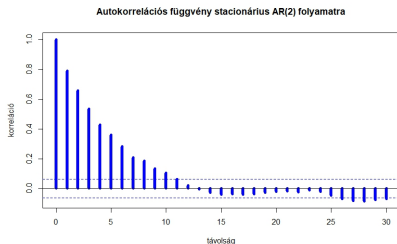
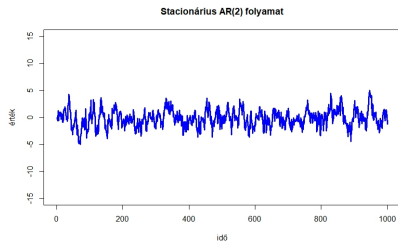
Egyik becslés sem torzítatlan $r(t)$ -re, azaz $\mathbb{E}(\hat{r}(t))$ eltér $r(t)$ -től. Ha x a megfigyelésekből álló vektor, akkor az R-ben az `acf(x)` paranccsal ábrázolható az autokorrelációs függvény becslése.

Az autokorrelációs függvény becslése



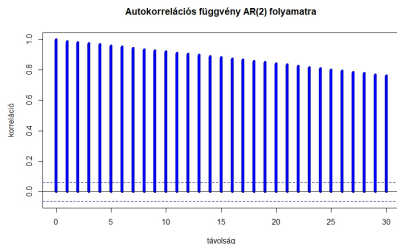
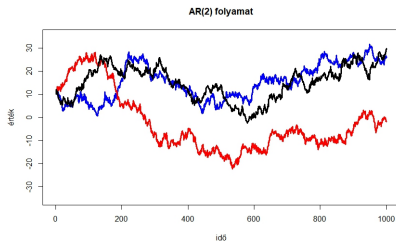
Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat három példányra, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



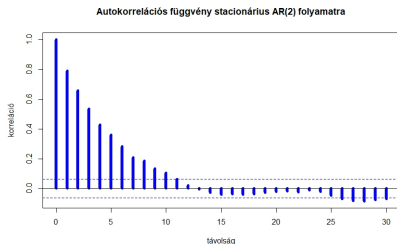
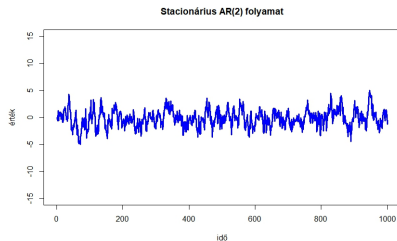
Az $X(t) = 0,7 \cdot X(t - 1) + 0,1 \cdot X(t - 2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



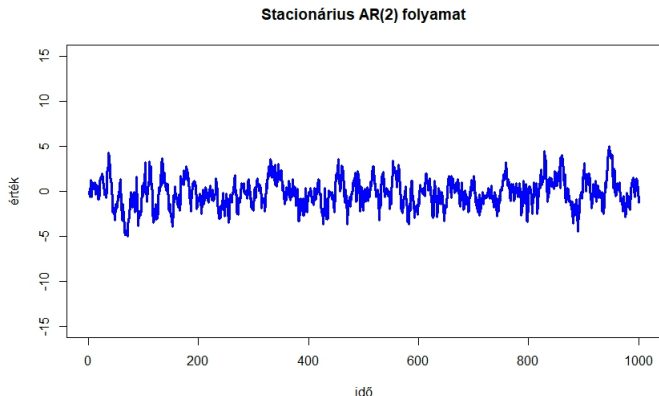
Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat három példányra, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



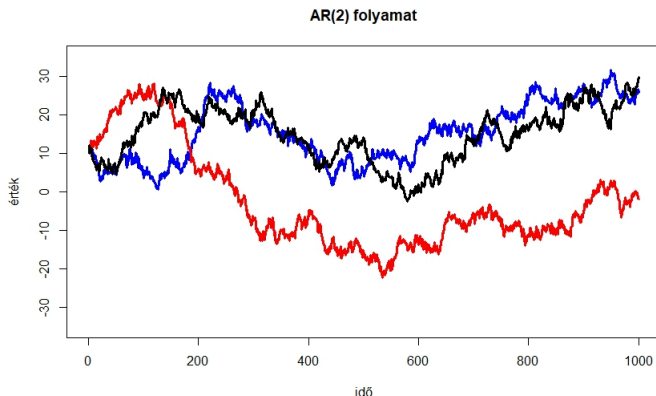
Az $X(t) = 0,7 \cdot X(t - 1) + 0,1 \cdot X(t - 2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat, illetve az autokorrelációs függvényének becslése

Autoregressziós folyamatok: stacionárius eset



$X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$ egyenletű AR(2)-folyamat: $\varepsilon(t)$ független 0 várható értékű 1 szórású valószínűségi változó $t \geq 0$ -ra (például normális eloszlásúak), és független $(X(0), \dots, X(t-1), \varepsilon(0), \dots, \varepsilon(t-1))$ -től

Autoregressziós folyamatok: nem stacionárius eset



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**

Autoregressziós folyamatok

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 $\varepsilon(t)$ minden $t \geq 0$ -ra $N(0,1)$ eloszlású valószínűségi változó, és $X(0), \dots, X(t-1)$ -től és $\varepsilon(0), \dots, \varepsilon(t-1)$ -től is független. Jelölés: $AR(p)$.

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.

Autoregressziós folyamatok stacionárius megoldása

Állítás

Az elsőrendű autoregressziós folyamatnak pontosan akkor van erősen stacionárius megoldása, ha $|\alpha_1| < 1$.

Általában, egy $AR(p)$ folyamatnak pontosan akkor van erősen stacionárius megoldása, ha az $x^p + \alpha_1 x^{p-1} + \alpha_2 x^{p-2} + \dots + \alpha_p = 0$ egyenlet minden gyökének (megoldásának) egynél kisebb az abszolút értéke.

Autoregressziós folyamatok stacionárius megoldása

Állítás

Az elsőrendű autoregressziós folyamatnak pontosan akkor van erősen stacionárius megoldása, ha $|\alpha_1| < 1$.

Általában, egy $AR(p)$ folyamatnak pontosan akkor van erősen stacionárius megoldása, ha az $x^p + \alpha_1 x^{p-1} + \alpha_2 x^{p-2} + \dots + \alpha_p = 0$ egyenlet minden gyökének (megoldásának) egynél kisebb az abszolút értéke.

A stacionárius példában: $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$

Autoregressziós folyamatok stacionárius megoldása

Állítás

Az elsőrendű autoregressziós folyamatnak pontosan akkor van erősen stacionárius megoldása, ha $|\alpha_1| < 1$.

Általában, egy $AR(p)$ folyamatnak pontosan akkor van erősen stacionárius megoldása, ha az $x^p + \alpha_1 x^{p-1} + \alpha_2 x^{p-2} + \dots + \alpha_p = 0$ egyenlet minden gyökének (megoldásának) egynél kisebb az abszolút értéke.

A stacionárius példában: $X(t) = 0,7 \cdot X(t-1) + 0,1 \cdot X(t-2) + \varepsilon(t)$

A másodfokú egyenlet: $x^2 + 0,7x + 0,1 = 0$

A megoldások:

$$\frac{-0,7 \pm \sqrt{0,7^2 - 4 \cdot 0,1}}{2} = -0,2 \text{ és } -0,5$$

Ezek egynél kisebb abszolút értékűek.

Autokorrelációs folyamatok és rövid emlékezet

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:

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

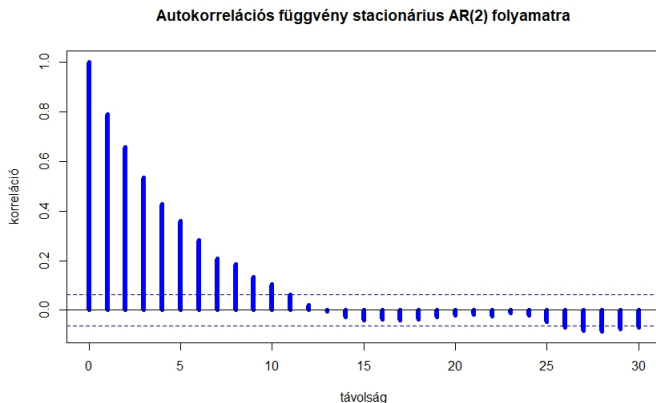
$$R(t) = \alpha_1 R(t-1) + \alpha_2 R(t-2) + \dots + \alpha_p R(t-p),$$

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

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

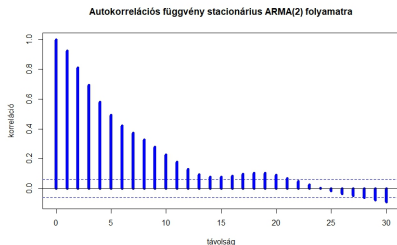
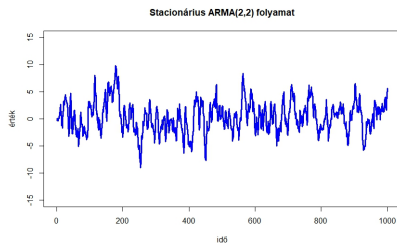
A stacionárius autoregressziós folyamatok úgynevezett rövid emlékezetű folyamatok: $\sum_{t=0}^{\infty} R(t) < \infty$, azaz $\sum_{t=0}^{\infty} r(t) < \infty$.

Autokorrelációs folyamatok és rövid emlékezet



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

ARMA-folyamatok



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

ARMA-folyamatok (általánosabb lineáris folyamatok)

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).

ARMA-folyamatok (általánosabb lineáris folyamatok)

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

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

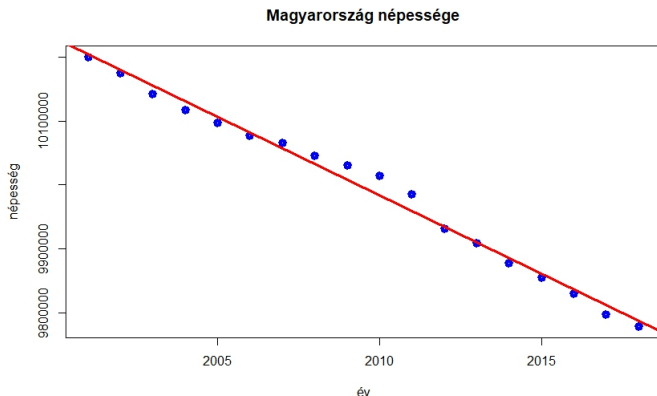
Jelölés: ARMA(p, q).

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

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

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

Autoregressziós folyamat illesztése



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

Autoregressziós folyamat illesztése és előrejelzés

- Feltételezés (itt $N(t)$ a népesség a t időpontban, és szezonális komponens nem várható):

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

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 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;

Autoregressziós folyamat illesztése és előrejelzés

- Feltételezés (itt $N(t)$ a népesség a t időpontban, és szezonális komponens nem várható):

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

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 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: az újabb hibatagokat nullának tekintjük (hiszen 0 várható értékűek, és függetlenek az előző megfigyelésektől, hibatagoktól):

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

Ebből:

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

Autoregressziós folyamat illesztése

```
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))
```

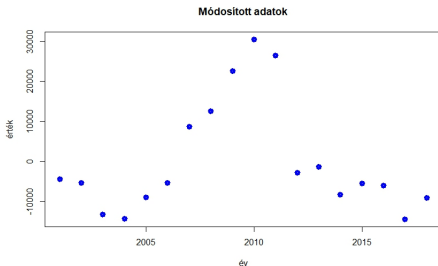
Autoregressziós folyamat illesztése a népesség adatokra

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

ahol $N(t)$ a népesség a t időpontban, a regressziós egyenes pedig $\hat{a}x + \hat{b}$ egyenletű.

```
x<-nep+24543.3*ev-59315833.4
```

```
plot(x~ev, lwd="5", col="blue", main="Módosított adatok", xlab="év",  
ylab="érték")
```



Autoregressziós folyamat illesztése a népesség adatokra

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

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.

- adott p mellett hogyan becsüljük a paramétereket?

Autoregressziós folyamat illesztése a népesség adatokra

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

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.

- adott p mellett hogyan becsüljük a paramétereket? például maximum likelihood-becsléssel
- hogyan válasszuk ki a legjobb p -t? minél nagyobb, annál több paraméter van, annál jobb lehet az illeszkedés
- ha p -t nagyra választjuk, előfordulhat a **túltanulás (overfitting)** jelensége: túl sok szabad paraméter van az adatsor méretéhez képest, és valójában nem az összefüggési struktúrát, hanem a véletlen hibákat "tanuljuk meg", ez viszont nem jó az előrejelzésnél

A rend és a paraméterek becslése

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 „bünteti” a túl sok paraméter választását)
- végül azt a p -t és azokat az együtthatókat választjuk, ahol a szorzat a legnagyobb.

Autoregressziós folyamat illesztése a népesség adatokra

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.

A népesség létszámának előrejelzése

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

$$\begin{aligned}\hat{X}(2019) &= 1,01 \cdot X(2018) - 0,33 \cdot X(2017) = \\ &= 1,01 \cdot (-9083) - 0,33 \cdot (-14436) = -4409.95\end{aligned}$$

A népesség létszámának előrejelzése

Az előrejelzés a módosított idősorban az $X(2019)$ várható értéke ($\text{predict(ar}(x), n.\text{ahead}=1)$):

$$\begin{aligned}\hat{X}(2019) &= 1,01 \cdot X(2018) - 0,33 \cdot X(2017) = \\ &= 1,01 \cdot (-9083) - 0,33 \cdot (-14436) = -4409.95\end{aligned}$$

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

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

A népesség létszámának előrejelzése

Az előrejelzés a módosított idősorban az $X(2019)$ várható értéke ($\text{predict(ar}(x), n.\text{ahead}=1)$):

$$\begin{aligned}\hat{X}(2019) &= 1,01 \cdot X(2018) - 0,33 \cdot X(2017) = \\ &= 1,01 \cdot (-9083) - 0,33 \cdot (-14436) = -4409.95\end{aligned}$$

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

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

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

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

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

Ez most még egy kicsit jobb, de nagyobb adatsorok esetében, ahol kevésbé jó a lineáris illeszkedés, az idősoros előrejelzés lényegesen jobb lehet.

Házi feladat május 23., kedd, 12:00-ig

Sorsoljunk 1000 elemű idősort az alábbi autoregressziós folyamatból (akár beépített parancs nélkül):

$$X(t) = 0,5 \cdot X(t-1) + c \cdot X(t-2) + \varepsilon(t),$$

ahol c -t mi választhatjuk. Keressünk olyan c -t, melyre van stacionárius megoldás (legyen ez c_1), és olyan c -t is, amire nincs (ez c_2).

Mind a két esetben ábrázoljuk az idősor egy-egy futását, és az autokorrelációs függvény becslését is.