

# Lineáris modell (11. előadás)

## 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.

# Becslések a lineáris modellben

## 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. 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:

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

# Konfidenciaintervallumok

$1 - \alpha$  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.

# Konfidenciaintervallumok

$1 - \alpha$  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:

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

## Az $a = 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$ . Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól?

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

## Az $a = 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$ . Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól?

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

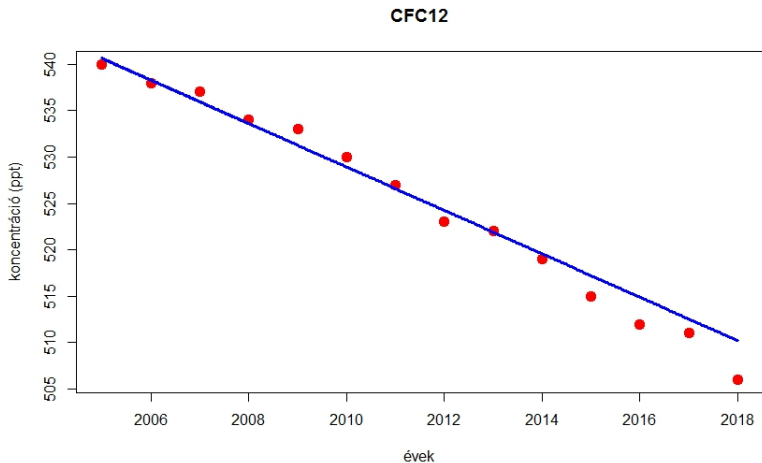
Kétoldali  $t$ -próbát végezhetünk az alábbi próbastatisztikával és  $f = n - 2$  szabadsági fokkal:

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

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.

# Lineáris regresszió



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

## Az $a = 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$        $H_1: a \neq 0$

## Az $a = 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$        $H_1: a \neq 0$

Kétoldali  $t$ -próbát végezhetünk az alábbi próbastatisztikával:

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

A példában

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

## Az $a = 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$        $H_1: a \neq 0$

Kétoldali  $t$ -próbát végezhetünk az alábbi próbastatisztikával:

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

A példában

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

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

## Az $a \leq 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$ . Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál?

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

## Az $a \leq 0$ hipotézis ellenőrzése

Lineáris modell:  $Y_i = aX_i + b + \varepsilon_i$ . Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál?

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

Egyoldali  $t$ -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  $t > \bar{t}_{n-2, \alpha}$ , 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.

## Lineáris modell: példa R-ben

```
> cfc12<-c(540, 538, 537, 534, 533, 530, 527, 523, 522, 519, 515, 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
```

## Lineáris modell: példa R-ben (folytatás)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5807.73626	159.19290	36.48	1.15e-13 ***
ev	-2.62637	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

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

## Többváltozós lineáris regresszió (multiple linear regression)

Az  $Y$  változót fejezzük ki az  $X_1, \dots, X_p$  valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük ( $X_{i,p} \equiv b$  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  $\varepsilon_i$  független  $N(0, \sigma^2)$  normális eloszlású valószínűségi változók.

**Például:**  $X_{i,1}$  az év,  $X_{i,2}$  a CFC-12 kibocsátás,  $Y$  a koncentráció. Ekkor a lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \varepsilon_i.$$

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

## Többsváltozós lineáris regresszió (multiple linear regression)

Az  $Y$  változót fejezzük ki az  $X_1, \dots, X_p$  valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük ( $X_{i,p} \equiv b$  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  $\varepsilon_i$  független  $N(0, \sigma^2)$  normális eloszlású valószínűségi változók.

**Például:**  $X_{i,1}$  az év,  $X_{i,2}$  a CFC-12 kibocsátás,  $Y$  a koncentráció. Ekkor a lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \varepsilon_i.$$

Vektoros formában:  $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ , 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):

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

Ekkor is megfelelő próbastatisztikával  $t$ -próbával tesztelhetők az  $a_i = 0$  hipotézisek, vagyis ellenőrizhető, hogy az  $Y$  mely mennyiségektől függ szignifikánsan.

# Hipotézisvizsgálat a lineáris modellben

Többváltozós lineáris modell:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i, \text{ azaz } \underline{Y} = X\beta + \varepsilon.$$

Legyen  $H$  olyan  $r \times p$  méretű mátrix, aminek a rangja  $r$  (itt  $r < p$ ). Ekkor a nullhipotézis:  $H_0 : H\beta = 0$ , és  $H_1 : H\beta \neq 0$ . (Például: ha  $H$  egy sora a  $j$ . egységvektor, az  $a_j = 0$ -t jelenti.)

A valószínűséghányados próba 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  $\beta^*$  a  $\beta$  becslése a  $H\beta = 0$  feltétel mellett a redukált lineáris modellben (például: bizonyos  $X$ -ek együtthatója 0, ezeket nem használhatjuk).

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.

# Szórásanalízis

	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

Néhány évi középhőmérséklet (forrás: Országos Meteorológiai Szolgálat), különböző évekből

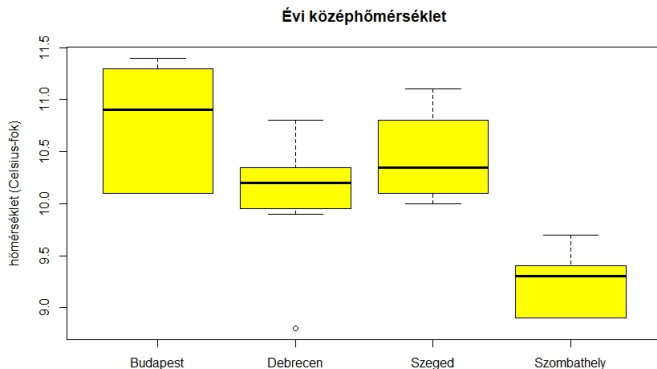
# Szórásanalízis

	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

Néhány évi középhőmérséklet (forrás: Országos Meteorológiai Szolgálat), különböző évekből

Igaz-e, hogy az egyes városokban az évi középhőmérséklet várható értéke megegyezik, vagy szignifikáns különbség látható?

# Szórásanalízis



A városok évi középhőmérséklet adatai különböző évekből. Van-e szignifikáns eltérés a várható értékek között?

# Szórásanalízis (analysis of variance, ANOVA)

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  $\mu_i$ , szórása  $\sigma$ .

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

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

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

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

Másképpen:

$H_0$ : a faktornak nincs szignifikáns hatása

$H_1$ : a faktornak szignifikáns hatása van.

# Szórásanalízis (ANOVA)

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  $\mu_i$ , szórása  $\sigma$ .

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

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

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

- normális eloszlások várható értékére vonatkozó próba
- a **kétmintás párosítatlan** Student-féle  $t$ -próba általánosításának is tekinthető, most nem kettő, hanem több csoport van, a szórások mindenhol megegyeznek
- a lineáris regresszió speciális esete  $\beta = (\mu_1, \mu_2, \dots, \mu_k)$ -val, ahol a magyarázó változók értéke 0 vagy 1, mert ezt  $\beta$ -val megszorozva kapjuk valamelyik  $\mu_j$ -t, és ehhez adódik hozzá a hiba.
- a nullhipotézis  $H\beta = 0$  alakú, ezért  $F$ -próbát végezhetünk.

# Szórásanalízis (analysis of variance, ANOVA)

$X_{ij}$  valószínűségi változók,  $i = 1, \dots, k$ ,  $j = 1, \dots, n_i$ . Vagyis  $k$  csoport van, és az  $i$ -ben  $n_i$  darab megfigyelés van. Jelölések:

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

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

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

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

# Szórásanalízis

	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 ( $\bar{X}_{j.}$ )	10,8	10,1	10,5	9,2	$\bar{\bar{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

Teljes szóródás = csoportokon belüli + csoportok közötti:

$$S = S_e + S_t = 5,43 + 7,15 = 12,49.$$

# Szórásanalízis

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)},$$

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

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

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

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

## Szórásanalízis

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

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)} = \frac{7,15 \cdot 20}{5,43 \cdot 3} = 8,77,$$

ahol  $n$  a megfigyelések száma,  $k$  a csoportok száma, és a csoportokon belüli szóródás (hiba):  $S_e = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 = 5,43$ , 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  $\alpha = 0,05$  terjedelem mellett:  $c_{\text{krit}} = 3,86$ .

## Szórásanalízis

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

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_e(k - 1)} = \frac{7,15 \cdot 20}{5,43 \cdot 3} = 8,77,$$

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

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

Mivel  $F = 7,15 > c_{\text{krit}} = 3,86$ , akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van.

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

# 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.

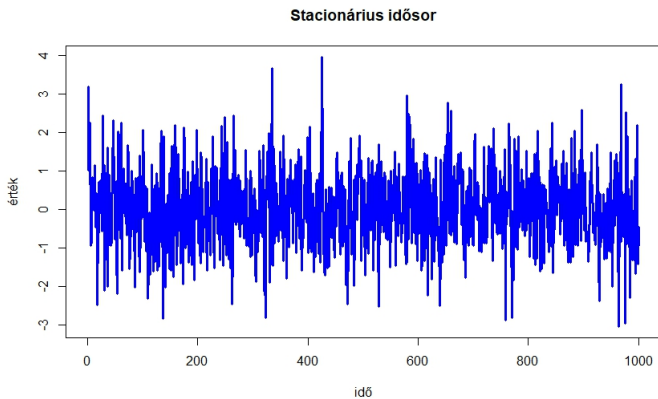
Az idősorok általában **nem független** valószínűségi változókból állnak. Az összefüggéseket jellemzi például az autokovariancia-függvény.

## Definíció

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

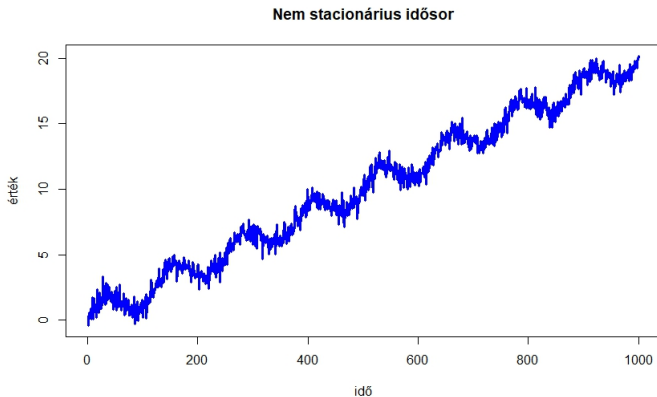
$$R(s, t) = \text{cov}(X_s, X_t) = \mathbb{E}(X_s X_t) - \mathbb{E}(X_s)\mathbb{E}(X_t).$$

# Stacionárius idősor



Példa stacionárius idősorra

# Nem stacionárius idősor



Példa nem stacionárius idősorra (egy lineáris tag, egy periodikus tag és egy stacionárius idősor összege)

## Házi feladat május 15., 9:00-ig

A húszelemű mintában osszuk fel a megfigyeléseket három csoportra aszerint, hogy az egyes emberek cipőmérete legfeljebb 39, 40 és 43 között van, vagy legalább 44 (lehet más csoportokat is választani).

Készítsünk szórásanalízist: állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszinten, hogy a cipőméret mint faktor szignifikáns hatással van a testmagasságra? Mennyi a  $p$ -érték ebben a hipotézisvizsgálati feladatban?

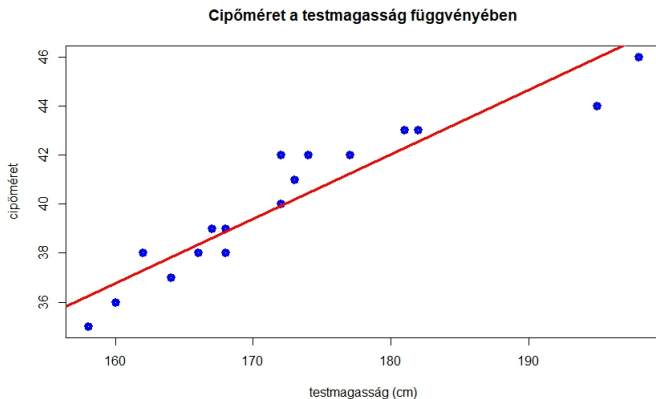
## Házi feladat május 8-ig, megoldás

Tekintsük a húszelemű mintában a cipőméreteket ( $Y_j$ ) a testmagasság ( $X_j$ ) függvényében.

- 1 Határozzuk meg a regressziós egyenes egyenletét és a megmagyarázott ingadozás részarányát.
- 2 Ábrázoljuk a cipőméretet a testmagasság függvényében a regressziós egyenessel együtt.

```
> magassag<-c(160, 158, 173, 162, 177, 168, 168, 182, 172, 174, 158,
172, 172, 195, 166, 164, 198, 181, 168, 167)
> cipo<-c(36, 35, 41, 38, 42, 39, 39, 43, 42, 42, 35, 40, 40, 44, 38,
37, 46, 43, 38, 39)
> plot(cipo magassag, lwd="5", col="blue", main="Cipőméret a testmagasság
függvényében", xlab="testmagasság (cm)", ylab="cipőméret")
> lines(abline(b=0.2627, a=-5.274, lwd="3", col="red"), xlim=c(155,195))
```

## Házi feladat május 8-ig, megoldás



A cipőméret a testmagasság függvényében és a regressziós egyenes:

$$y = 0,2627x - 5,274 \text{ (itt } R^2 = 0,88\text{)}$$

## Házi feladat május 8-ig, megoldás

```
> summary(lm(cipo ~ magassag))
```

Call: lm(formula = cipo ~ magassag) Residuals:

Min	1Q	Median	3Q	Max
-1.9585	-0.7756	0.1098	0.7137	2.0843

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	<b>-5.27420</b>	3.75379	-1.405	0.177
magassag	<b>0.26273</b>	0.02182	12.044	<b>4.76e-10 ***</b>

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

Residual standard error: 1.029 on 18 degrees of freedom

Multiple R-squared: 0.8896, Adjusted R-squared: **0.8835**

F-statistic: 145 on 1 and 18 DF, p-value: 4.76e-10

Az egyenes meredeksége szignifikánsan pozitív,  $p = 2,38 \cdot 10^{-10}$ .