

Többváltozós lineáris modell (11. előadás)

A lineáris modellben több magyarázó változót is bevezethetünk.

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:

$$Y_1 = a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1;$$

$$Y_2 = a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2;$$

...

$$Y_n = a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n.$$

Többváltozós lineáris modell (11. előadás)

A lineáris modellben több magyarázó változót is bevezethetünk.

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:

$$Y_1 = a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1;$$

$$Y_2 = a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2;$$

...

$$Y_n = a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n.$$

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

$$\underline{Y} = X\beta + \underline{\varepsilon}.$$

Az együtthatók becslése

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

$$\underline{Y} = X\beta + \underline{\varepsilon}.$$

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

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

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 megfelelő illeszkedés ellenőrzése

A megmagyarázott ingadozás részaránya:

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

- érzékeny a kiugró értékekre
- nem veszi figyelembe a becsült paraméterek számát
- elég sok paraméterrel megfigyelhető a **túltanulás (overfitting)** jelensége: valójában nem modellt illesztünk, hanem a véletlen hibákat külön-külön tanuljuk meg

A megfelelő illeszkedés ellenőrzése

A megmagyarázott ingadozás részaránya:

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

- érzékeny a kiugró értékekre
- nem veszi figyelembe a becsült paraméterek számát
- elég sok paraméterrel megfigyelhető a **túltanulás (overfitting)** jelensége: valójában nem modellt illesztünk, hanem a véletlen hibákat külön-külön tanuljuk meg

Ezért az R^2 -nek az alábbi módosított (adjusted) változata is gyakran használt:

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

Ha $p = 0$, visszakapjuk az eredetit (persze ez nem egy valódi modell).

Például: $n = 100$, $p = 10$ esetén jelzi, hogy a mintaelemszámhoz képest túl sok a paraméter.

Hipotézisvizsgálat a lineáris modellben, egyváltozós eset

Egyváltozós eset: $Y_i = aX_i + b + \varepsilon_i$

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

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 α 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.

Hipotézisvizsgálat a lineáris modellben

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

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

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

$$H_0 : H\beta = 0$$

$$H_1 : H\beta \neq 0.$$

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.

Hipotézisvizsgálat a lineáris modellben

$$H_0 : H\beta = 0$$

$$H_1 : H\beta \neq 0.$$

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

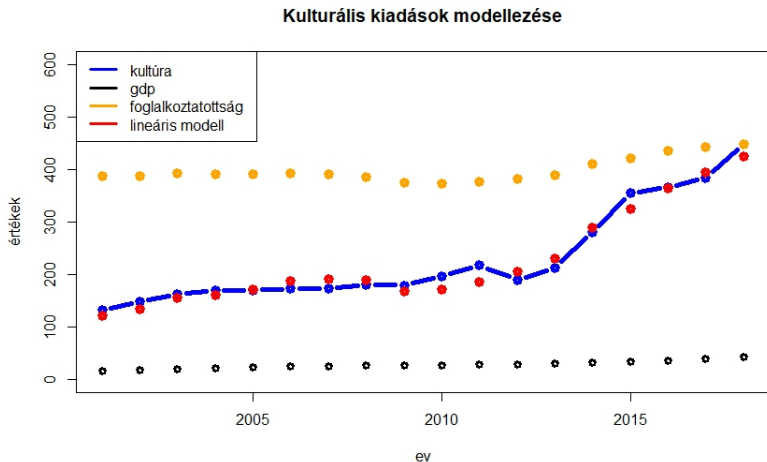
$$F = \frac{(\underline{Y} - X\beta^*)^T(\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})},$$

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). Itt tehát a számláló első tagja a nullhipotézis esetén a maximumlikelihood-becslés, míg a nevezőben ugyanúgy maximumlikelihood-becslés szerepel, de a nullhipotézis feltétele nélkül, a teljes paramétertartományon.

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.

Többváltozós lineáris modell: példa



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; forrás: KSH)

Többváltozós lineáris regresszió: példa

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.

```
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)
```

Többsváltozós lineáris modell: példa

```
> 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 ***  
--
```

Többváltozós lineáris modell: példa

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:

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

Többváltozós lineáris modell: példa

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:

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

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:

Többváltozós lineáris modell: példa

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:

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

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~fogl))
```

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.

Szórásanalízis (analysis of variance, ANOVA)

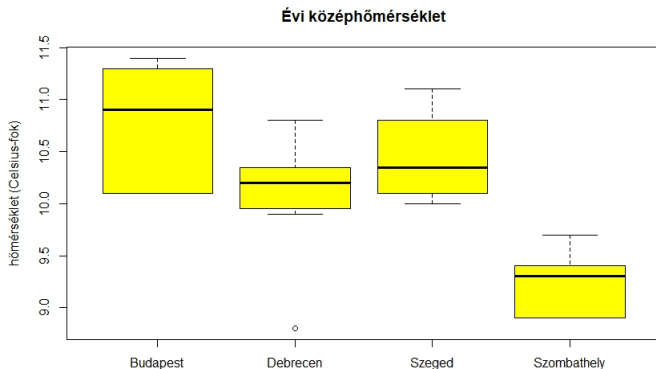
- hipotézisvizsgálati eljárás
- az egyedek különböző csoportokba soroljuk
- egy **normális eloszlású** mennyiséget vizsgálunk: igaz-e, hogy ennek a várható értéke az egyes csoportokban azonos?
- **feltesszük, hogy a vizsgált mennyiség szórása az egyes csoportokban azonos**
- az egyes csoportokon belül és közöttük is függetlenek a megfigyelések
- két csoport esetén ez a kétmintás, Student-féle t -próba
- általánosabban is a többváltozós lineáris modell
- 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

Szórásanalízis (analysis of variance, ANOVA)

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? 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

Szórásanalízis (analysis of variance, ANOVA)



Boxplot ábra az egyes városok éves középhőmérséklet adataiból

Feltevésék 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 : a faktornak nincs szignifikáns hatása

H_1 : a faktornak szignifikáns hatása van.

Kapcsolat a lineáris modellel

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.

Kapcsolat a lineáris modellel

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.

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.

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

A próbastatisztika:

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

A szórásanalízis eljárása

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

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

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.

Szórásanalízis: példa

A korábbi példára visszatérve feltételezzük, hogy a szórások 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 ($\bar{X}_{j\cdot}$)	10,8	10,1	10,5	9,2	$\bar{\bar{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

Szórásanalízis: példa

A csoportokon belüli szóródás kiszámítása:

$$\begin{aligned} S_g &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \\ &= ((10,8 - 10,8)^2 + (10,1 - 10,8)^2 + \dots + (10,1 - 10,8)^2) + \\ &\quad + ((8,8 - 10,1)^2 + (9,9 - 10,1)^2 + \dots + (10,3 - 10,1)^2) + \\ &\quad + ((11,1 - 10,5)^2 + (10,8 - 10,5)^2 + \dots + (10,3 - 10,5)^2) + \\ &\quad + ((8,9 - 9,2)^2 + (9,4 - 9,2)^2 + \dots + (9,7 - 9,2)^2) = 5,34. \end{aligned}$$

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 átlagától vett eltérések négyzetét adjuk össze, majd hasonlóképpen az $i = 3$ és $i = 4$ esetekben is.

A csoportok közötti szóródás kiszámítása:

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

Szórásanalízis: példa

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: $S_t = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X})^2 = 7,15$.

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.

Házi feladat május 7., kedd, 12:00-ig

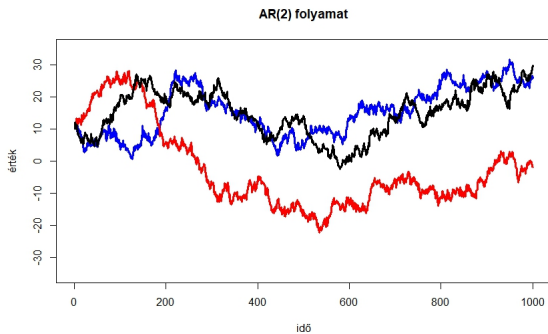
A félév elején gyűjtött adatokból illesztünk lineáris modellt a sportolással töltött időre, úgy, hogy a magyarázó változók

- az utazással töltött idő és az, hogy hányszor járnak munkába egy héten
- az utazással töltött idő, az, hányszor járnak munkába/iskolába és hogy hány sorozatot néztek
- az utazással töltött idő, a nézett sorozatok száma, az, hogy hányszor járnak munkába, és hogy van-e kutyájuk

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

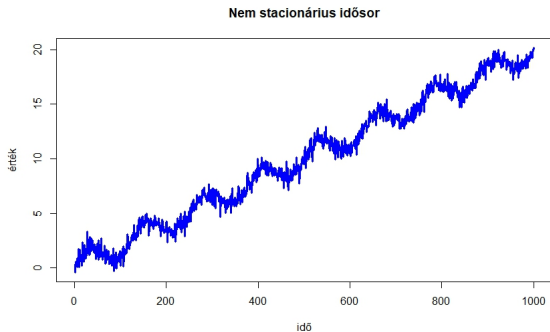
(Vegyük észre, hogy a feltételek nem igazán teljesülnek a kerekítések miatt, de most ettől tekintsünk el.)

Idősorok elemzése



Példák idősorra: egy másodrendű autoregressziós folyamat

Idősorok elemzése



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

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 24. 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:

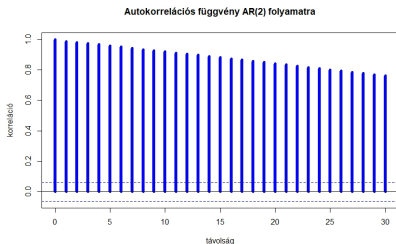
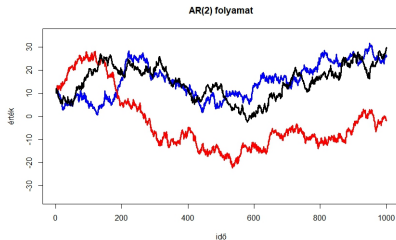
$$\hat{r}(t) = \frac{\sum_{j=0}^{n-t-1} (X_j - \bar{X}) \cdot (X_{j+t} - \bar{X})}{(n-t) \cdot s_n^{*2}}.$$

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

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

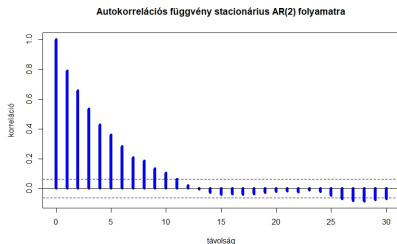
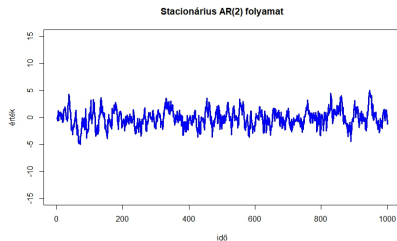
Egyik becslés sem torzítatlan $r(t)$ -re, azaz $\mathbb{E}(\hat{r}(t))$ eltér $r(t)$ -től. Ha x a megfigyelésekből álló vektor, akkor az R-ben az `acf(x)` paranccsal ábrázolható az autokorrelációs függvény becslése.

Az autokorrelációs függvény becslése



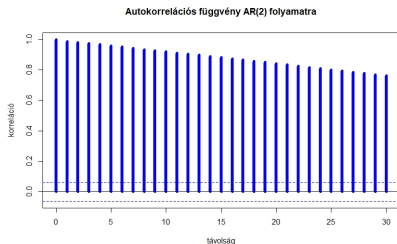
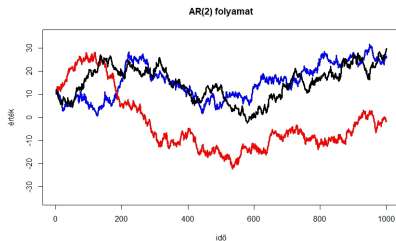
Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat három példányra, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



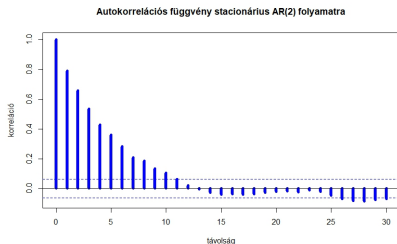
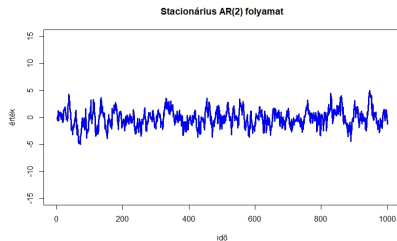
Az $X(t) = 0,7 \cdot X(t - 1) + 0,1 \cdot X(t - 2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



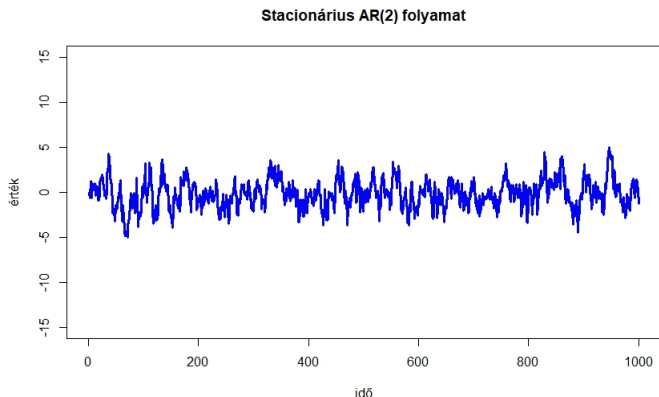
Az $X(t) = 0,7 \cdot X(t-1) + 0,3 \cdot X(t-2) + \varepsilon(t)$ folyamat három példányra, illetve az autokorrelációs függvényének becslése

Az autokorrelációs függvény becslése



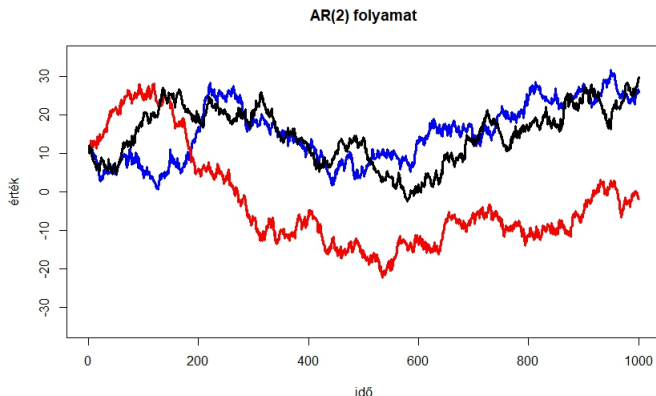
Az $X(t) = 0,7 \cdot X(t - 1) + 0,1 \cdot X(t - 2) + \varepsilon(t)$ egyenletű stacionárius AR(2) folyamat, illetve az autokorrelációs függvényének becslése

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.

Autoregressziós 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:

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

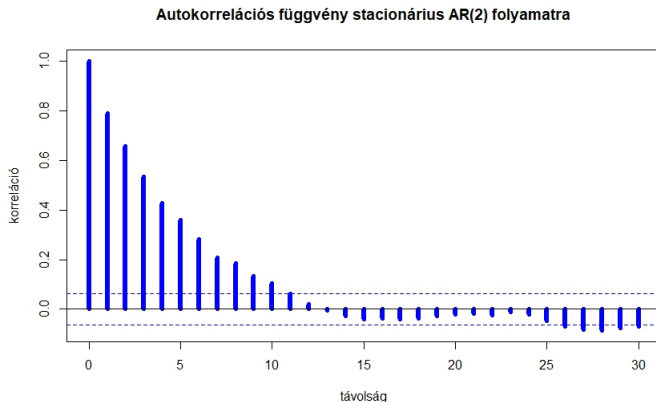
$$R(t) = \alpha_1 R(t-1) + \alpha_2 R(t-2) + \dots + \alpha_p R(t-p),$$

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 $\sum_{t=0}^{\infty} r(t) < \infty$.

Autoregressziós 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