Programtervező informatikus BSc, C szakirány
Valószínűségszámítás és statisztika gyakorlat

1. (10-11 hét) Nemparaméteres próbák, egyszerű lineáris regresszió

Elmélet

Nemparaméteres próbák:

Diszkrét illeszkedésvizsgálat

Legyen $X_1, \dots, X_n$ egy n elemű minta és tegyük fel, hogy a mintaelemek r különböző x_j $(j = 1, \dots r)$ értéket vehetnek fel. Továbbá jelölje $\nu_j (j = 1, \dots r)$ az egyes értékek megfigyelt gyakoriságát, azaz n független megfigyelést osztályozunk valamilyen szempont szerint, r páronként diszjunkt osztályba. Az egyes osztályok feltételezett valószínűségei rendre $p_1, \dots p_r$.

Osztályok	1	2	...	r	Összesen
Értékek	x_1	x_2	...	x_r	
Gyakoriságok	ν_1	ν_2	...	ν_r	n
Valószínűségek	p_1	p_2	...	p_r	1

Azt vizsgáljuk, hogy a minta eloszlása megegyezik-e a feltételezett eloszlással. Ismert eloszlás esetén tiszta illeszkedésvizsgálatot végzünk. Ha viszont az eloszlás paraméteres és csak az eloszláscsaládot ismerjük, a paraméter(eke)t viszont nem (pl. az a kérdés, hogy származhatnak-e az adatok p paraméterű binomiális eloszlásból), akkor becsléses illeszkedésvizsgálatot végzünk.

Tiszta illeszkedésvizsgálat:

$H_0 : P(X_i = x_j) = p_j \quad j = 1, \dots, r$
$H_1 : \exists$ legalább egy j melyre $P(X_i = x_j) \neq p_j$

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat:

Legyen ϑ egy s dimenziós paramétervektor, valamint legyen $\hat{\vartheta}$ a ϑ paramétervektor ML-becslése, és legyen $\hat{p}_j = p_j(\hat{\vartheta})$.

$H_0 : P(X_i = x_j) = \hat{p}_j \quad j = 1, \dots, r$
$H_1 : \exists$ legalább egy j melyre $P(X_i = x_j) \neq \hat{p}_j$

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - n\hat{p}_j)^2}{n\hat{p}_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-s-1}$ Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-s-1,1-\alpha}\}$

Megjegyzés: Mivel a próba aszimptotikus, vigyáznunk kell arra, hogy a minta elemszáma elég nagy legyen. Ökölszabályként meg szokás követelni, hogy az ún. elméleti gyakoriság (np_j) legalább 5 legyen minden i-re. Ha ez nem teljesül, akkor a kis várt gyakoriságokkal rendelkező eseményeket összevonjuk.

Függetlenségvizsgálat

n független megfigyelést két szempont szerint osztályozunk, az 1. szempont szerint r osztály, míg a 2. szempont szerint s osztály van. Annak a valószínűsége, hogy egy megfigyelést az 1. szempont szerint az i-edik, a második szerint pedig a j-edik osztályba sorolunk, p_{ij}. Az ilyen tulajdonságú megfigyelések számát pedig ν_{ij}-vel jelöljük. Az osztályozási eljárás eredményét ún. kontingenciatábla formájában szokás megadni:

		2. szempont					
		1	...	j	...	s	Sorösszegek
1. szempont	1	ν_{11}	...	ν_{1j}	...	ν_{1s}	$\nu_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	i	ν_{i1}	...	ν_{ij}	...	ν_{is}	$\nu_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	ν_{r1}	...	ν_{rj}	...	ν_{rs}	$\nu_{r\bullet}$
Oszlopösszegek		$\nu_{\bullet 1}$	...	$\nu_{\bullet j}$	...	$\nu_{\bullet s}$	n

ν_{ij} = megfigyelések gyakorisága az (i, j) osztályban

$\nu_{i\bullet} = \sum_{j=1}^{s} \nu_{ij} \quad \nu_{\bullet j} = \sum_{i=1}^{r} \nu_{ij}$

Hasonlóan $p_{i\bullet}$ ill. $p_{\bullet j}$ a marginális eloszlást jelölik, tehát a $[p_{ij}]$ mátrix sor-, illetve oszlopösszegei: $p_{i\bullet} = \sum_{j=1}^{s} p_{ij} \quad p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$

H_0: a két szempont független egymástól, azaz $p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \;\; 1 \le i \le r,\, 1 \le j \le s$
H_1: a két szempont nem független, azaz $p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}$ legalább egy (i,j) párra

Próbastatisztika: $T_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{(r-1)(s-1)}$
Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Megjegyzés: Ha $r = s = 2$, akkor a próbastatisztika a következőképpen leegyszerűsödik:
$T_n = n \cdot \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_1.$

Homogenitásvizsgálat

Van két független mintánk (adatsorunk) az egyikben n, a másikban m megfigyeléssel. Valamilyen szempont szerint r, páronként diszjunkt osztályba soroljuk a megfigyeléseket. Az i-edik osztály valószínűsége p_i az 1. minta és q_i a 2. minta esetén $(i = 1, 2, ..., r)$. Legyenek az egyes osztályok gyakoriságai $\nu_1, \ldots, \nu_r$ az 1. minta és $\mu_1, \ldots, \mu_r$ a 2. minta esetén.

Osztályok	1	2	...	r	Összesen
1. minta					
Gyakoriságok	ν_1	ν_2	...	ν_r	n
Valószínűségek	p_1	p_2	...	p_r	1
2. minta					
Gyakoriságok	μ_1	μ_2	...	μ_r	m
Valószínűségek	q_1	q_2	...	q_r	1

Azt vizsgáljuk, hogy a két minta ugyanolyan eloszlás szerint sorolódik-e be az egyes osztályokba:

H_0 : a két eloszlás megegyezik, azaz $p_i = q_i \; i = 1, \ldots r$
H_1 : a két eloszlás nem megegyezik meg, azaz $\exists$ legalább egy i, hogy $p_i \neq q_i$

Próbastatisztika: $T_{n,m} = nm \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_{n,m}(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Korreláció becslése:

Legyenek $X_1, \ldots, X_n$ és $Y_1, \ldots, Y_n$ n elemű minták. A korreláció becslése a minták alapján:

Tapasztalati korrelációs együttható: $r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$

Egyszerű lineáris regresszió:

Adott $(x_1, y_1), \ldots, (x_n, y_n)$ számpárokra szeretnénk egyenest illeszteni.
Modell: $y_i = a + bx_i + \varepsilon_i,$ ahol $E\varepsilon_i = 0$ és $D^2\varepsilon_i = \sigma^2 < \infty \quad (i = 1, \ldots, n)$
Cél: a és b becslése
Módszer: legkisebb négyzetek: $\min \sum_{i=1}^{n} (y_i - (a + bx_i))^2$
Megoldás: $\hat{b} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$, ez torzítatlan b-re és a szórásnégyzete: $D^2(\hat{b}) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$

$\hat{a} = \overline{y} - \hat{b}\overline{x}$, ez torzítatlan a-ra és a szórásnégyzete: $D^2(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right)$

Reziduálisok: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i) \quad (i = 1, \ldots, n)$
Reziduális szórásnégyzet becslése: $\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$

Determinációs együttható: $R^2 = 1 - \dfrac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} = r^2$ Az R^2 mutatja meg, hogy X változékonysága mennyire magyarázza Y változékonyságát. Értéke 0 és 1 között lehet, minél nagyobb, annál jobban teljesít a model, azaz annál jobban illeszkedik az egyenes.

Feladatok

1.1. Feladat. Egy gyárban egy termék minőségét 4 elemű mintákat véve ellenőrzik, havonta 300 mintavétellel. Megszámolták, hogy a legutóbbi hónapban hányszor volt selejtes a minta, melynek eredményét az alábbi táblázat tartalmazza:

Selejtesek száma	0	1	2	3	4
Darabszám	80	113	77	27	3

Modellezhető a mintákban levő selejtesek száma
a) $(4; 0,25)$, ill.
b) $(4; p)$ paraméterű binomiális eloszlással $(\alpha = 0,05)$? $\left(\chi^2_{3;0,95} = 7,81,\ \chi^2_{2;0,95} = 5,99\right)$

Megoldás

a) Tiszta illeszkedésvizsgálat

H_0: $X_i \sim Bin(4;\ 0,25)$.
H_1: X_i nem ilyen eloszlású

Vegyük észre, hogy az utólsó oszlopra vonatkozóan $np_5 = 300 \cdot \binom{4}{4} \cdot 0,25^4 \cdot 0,75^0 = 1,2 < 5$, így az utolsó két oszlopban levő eseményeket vonjuk össze. A módosított táblázat a következő:

Selejtesek száma	0	1	2	3 vagy 4
Esetszám	80	113	77	30

$r = 4$, $n = 300$

Határozzuk meg az egyes selejtszámokra vonatkozó valószínűségeket, illetve ezek alapján a várt gyakoriságokat:
$p_1 = P(X_j = 0) = \binom{4}{0} \cdot 0,25^0 \cdot 0,75^4 = 0,3164 \Rightarrow n \cdot p_1 = 300 \cdot 0,3164 = 94,9$
$p_2 = P(X_j = 1) = \binom{4}{1} \cdot 0,25^1 \cdot 0,75^3 = 0,4219 \Rightarrow n \cdot p_2 = 300 \cdot 0,4219 = 126,6$
$p_3 = P(X_j = 2) = \binom{4}{2} \cdot 0,25^2 \cdot 0,75^2 = 0,2109 \Rightarrow n \cdot p_3 = 300 \cdot 0,2109 = 63,3$
$p_4 = P(X_j \geq 3) = 1 - p_1 - p_2 - p_3 = 0,0508 \Rightarrow n \cdot p_4 = 300 - 94,9 - 126,6 - 63,3 = 15,2$

Selejtesek száma	0	1	2	3 vagy 4
Esetszám (gyakoriság) (ν_j)	80	113	77	30
Valószínűségek (p_j)	0,3164	0,4219	0,2109	0,0508
Elméleti (várt) gyakoriságok (np_j)	94,9	126,6	63,3	15,2

Próbastatisztika: $T_n = \sum_{j=1}^{r} \dfrac{(\nu_j - np_j)^2}{np_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$, melynek értéke

$\frac{(80-94,9)^2}{94,9} + \frac{(113-126,6)^2}{126,6} + \frac{(77-63,3)^2}{63,3} + \frac{(30-15,2)^2}{15,2} = 2,339 + 1,461 + 2,965 + 14,411 = 21,176$

A próbastatisztika mely értékeire utasítjuk el H_0-t?
Mivel a szabadsági fok $r - 1 = 3$, így $\chi^2_{3;0,95} = 7,81$, azaz a kritikus tartomány $= \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 7,81\}$. Mivel most $T_n = 21,176 > 7,81$, így elutasítjuk H_0-t, azaz mondhatjuk, hogy a selejtes termékek száma nem $Bin(4;\ 0,25)$ eloszlást követ.

A hipotézist a p-érték α-val való összehasonlításával is tesztelhetjük:
p-érték $= P(\chi^2_3 > 21,176) = 0,0001 < 0,05$, így elutasítjuk H_0-t.

b) Becsléses illeszkedésvizsgálat

H_0: $X_i \sim Bin(4; p)$ valamilyen p-re
H_1: X_i nem ilyen eloszlású

Először meg kell becsülni az ismeretlen p paramétert ML-módszerrel. (Egy paramétert becslünk, így $s = 1$.) A 3.6 a) feladat alapján tudjuk, hogy $Bin(m, p)$ eloszlású minta esetén (m ismert) a p ML-becslése $\hat{p} = \frac{\overline{X}}{m}$. Mivel $\overline{x} = \frac{0 \cdot 80 + 1 \cdot 113 + 2 \cdot 77 + 3 \cdot 27 + 4 \cdot 3}{300} = \frac{360}{300} = 1,2$, így $\hat{p} = \frac{1,2}{4} = 0,3$. Vegyük észre, hogy az utolsó oszlopra vonatkozóan $np_5 = 300 \cdot \binom{4}{4} \cdot 0,3^4 \cdot 0,7^0 = 2,43 < 5$, így az utolsó két oszlopban levő eseményeket vonjuk össze. A módosított táblázat a következő:

Selejtesek száma	0	1	2	3 vagy 4
Esetszám	80	113	77	30

$r = 4$, $n = 300$

Határozzuk meg az egyes selejtszámokra vonatkozó valószínűségeket, illetve ezek alapján gyakoriságokat:

$\hat{p_1} = \hat{P}(X_j = 0) = \binom{4}{0} \cdot 0,3^0 \cdot 0,7^4 = 0,2401 \Rightarrow n \cdot \hat{p_1} = 300 \cdot 0,2401 = 72$
$\hat{p_2} = \hat{P}(X_j = 1) = \binom{4}{1} \cdot 0,3^1 \cdot 0,7^3 = 0,4116 \Rightarrow n \cdot \hat{p_2} = 300 \cdot 0,4116 = 123,5$
$\hat{p_3} = \hat{P}(X_j = 2) = \binom{4}{2} \cdot 0,3^2 \cdot 0,7^2 = 0,2646 \Rightarrow n \cdot \hat{p_3} = 300 \cdot 0,2646 = 79,4$
$\hat{p_4} = \hat{P}(X_j \geq 3) = 1 - \hat{p_1} - \hat{p_2} - \hat{p_3} = 0,0837 \Rightarrow n \cdot \hat{p_4} = 300 - 72 - 123,5 - 79,4 = 25,1$

Selejtesek száma	0	1	2	3 vagy 4
gyakoriságok (ν_j)	80	113	77	30
Valószínűségek $(\hat{p_j})$	0,2401	0,4116	0,2646	0,0837
Elméleti gyakoriságok $(n\hat{p_j})$	72	123,5	79,4	25,1

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - n\hat{p_j})^2}{n\hat{p_j}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-s-1}$, melynek értéke
$\frac{(80-72)^2}{72} + \frac{(113-123,5)^2}{123,5} + \frac{(77-79,4)^2}{79,4} + \frac{(30-25,1)^2}{25,1} = 0,889 + 0,893 + 0,073 + 0,957 = 2,812$

A próbastatisztika mely értékeire utasítjuk el H_0-t?

Mivel 1 paramétert becsültünk ($s = 1$), a szabadsági fok $r - s - 1 = 2$, így $\chi^2_{2;0,95} = 5,99$, azaz a kritikus tartomány $= \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-s-1,1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 5,99\}$. Mivel most $T_n = 2,812 < 5,99$, így nem utasítjuk el H_0-t, tehát tekinthetjük a selejtes termékek számát $Bin(4, p)$ eloszlásúnak.

1.2. Feladat. Az alábbi kontingencia-táblázat mutatja, hogy egy 100 éves időszakban egy adott hónapban a csapadék mennyisége és az átlaghőmérséklet hogyan alakult:

Hőmérséklet | Csapadék	kevés	átlagos	sok
hűvös	15	10	5
átlagos	10	10	20
meleg	5	20	5

A cellákban az egyes esetek gyakoriságai találhatóak. $\alpha = 0,05$ mellett tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek? $(\chi^2_{4;0,95} = 9,49)$

Megoldás

Függetlenségvizsgálat

H_0: a csapadék és a hőmérséklet függetlenek
H_1: nem függetlenek

Egészítsük ki a táblázatot egy "összesen" sorral és oszloppal:

Hőmérséklet | Csapadék	kevés	átlagos	sok	Összesen
hűvös	15	10	5	$\nu_{1\bullet} = 30$
átlagos	10	10	20	$\nu_{2\bullet} = 40$
meleg	5	20	5	$\nu_{3\bullet} = 30$
Összesen	$\nu_{\bullet 1} = 30$	$\nu_{\bullet 2} = 40$	$\nu_{\bullet 3} = 30$	$n = 100$

A várt gyakoriságok $\hat{\nu}_{ij} = n \cdot \frac{\nu_{i\bullet}}{n} \cdot \frac{\nu_{\bullet j}}{n} = \frac{\nu_{i\bullet} \cdot \nu_{\bullet j}}{n}$ táblázatban:

Hőmérséklet | Csapadék	kevés	átlagos	sok	Összesen
hűvös	$\frac{30 \cdot 30}{100} = 9$	$\frac{40 \cdot 30}{100} = 12$	$\frac{30 \cdot 30}{100} = 9$	$\nu_{1\bullet} = 30$
átlagos	$\frac{30 \cdot 40}{100} = 12$	$\frac{40 \cdot 40}{100} = 16$	$\frac{30 \cdot 40}{100} = 12$	$\nu_{2\bullet} = 40$
meleg	$\frac{30 \cdot 30}{100} = 9$	$\frac{30 \cdot 40}{100} = 12$	$\frac{30 \cdot 30}{100} = 9$	$\nu_{3\bullet} = 30$
Összesen	$\nu_{\bullet 1} = 30$	$\nu_{\bullet 2} = 40$	$\nu_{\bullet 3} = 30$	$n = 100$

Próbastatisztika: $T_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{(r-1)(s-1)}$ (r az oszlopok, s a sorok száma), melynek értéke
$\frac{(15-9)^2}{9} + \frac{(10-12)^2}{12} + \frac{(5-9)^2}{9} + \frac{(10-12)^2}{12} + \frac{(10-16)^2}{16} + \frac{(20-12)^2}{12} + \frac{(5-9)^2}{9} + \frac{(20-12)^2}{12} + \frac{(5-9)^2}{9} =$
$= 4 + 0,333 + 1,778 + 0,333 + 2,25 + 5,333 + 1,778 + 5,333 + 1,778 = 22,916$

A próbastatisztika mely értékeire utasítjuk el H_0-t?

Mivel a szabadsági fok $(r-1)(s-1) = 2 \cdot 2 = 4$, így $\chi^2_{4;0,95} = 9,49$, azaz a kritikus tartomány $=$
$\{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 9,49\}$. Mivel most $T_n = 22,916 > 9,49$, így elutasítjuk H_0-t, tehát állíthatjuk, hogy a csapadék és a hőmérséklet nem függetlenek.

A hipotézist a p-érték α-val való összehasonlításával is tesztelhetjük:

p-érték $= P(\chi^2_4 > 22,916) = 0,0001 < 0,05$, így elutasítjuk H_0-t.

1.3. Feladat. Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

Dobások	1	2	3	4	5	6
1. kocka	27	24	26	23	18	32
2. kocka	18	12	15	21	14	20

$\alpha = 0,05$ mellett döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak! $\left(\chi^2_{5;0,95} = 11,1\right)$

Megoldás

Homogenitásvizsgálat

H_0 a két eloszlás megegyezik
H_1 a két eloszlás nem egyezik meg

Egészítsük ki a táblázatot egy „összesen" oszlopplal:

Dobások	1	2	3	4	5	6	Összesen
1. kocka (ν_i)	27	24	26	23	18	32	$n = 150$
2. kocka (μ_i)	18	12	15	21	14	20	$m = 100$

$r = 6$

Próbastatisztika: $T_{n,m} = nm \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ melynek értéke

$$T_{150,100} = 150 \cdot 100 \left(\frac{\left(\frac{27}{150} - \frac{18}{100}\right)^2}{27+18} + \frac{\left(\frac{24}{150} - \frac{12}{100}\right)^2}{24+12} + \frac{\left(\frac{26}{150} - \frac{15}{100}\right)^2}{26+15} + \frac{\left(\frac{23}{150} - \frac{21}{100}\right)^2}{23+21} + \frac{\left(\frac{18}{150} - \frac{14}{100}\right)^2}{18+14} + \frac{\left(\frac{32}{150} - \frac{20}{100}\right)^2}{32+20} \right) =$$
$$= 0 + 0,67 + 0,20 + 1,09 + 0,19 + 0,05 = 2,2$$

A próbastatisztika mely értékeire utasítjuk el H_0-t?

Mivel a szabadsági fok $r - 1 = 5$, így $\chi^2_{5;0,95} = 11,1$, azaz a kritikus tartomány $= \{\mathbf{x} : T_{n,m}(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\} =$
$= \{\mathbf{x} : T_{n,m}(\mathbf{x}) > 11,1\}$. Mivel most $T_{150,100} = 2,2 < 11,1$, így nem utasítjuk el H_0-t, ami nem mutat elentmondást a két eloszlás azonosságával.

1.4. Feladat. A következő feladatot csak R-rel kell megoldani, a számolások tájékoztató jellegűek, ha valakit érdekel, nem szerepelnek a számonkérésen. Adottak a következő $(\mathbf{x}, \mathbf{y})$ párok:

$\mathbf{x}$	0	1	6	5	3
$\mathbf{y}$	4	3	0	1	2

a) Határozzuk meg és ábrázoljuk is az $a + bx$ alakú regressziós egyenest!
b) Mi az $x = 2$-höz tartozó előrejelzett y érték?
c) Számoljuk ki a reziduálisokat és becsüljük meg a hiba szórásnégyzetét, valamint a becsléseink szórásnégyzetét!

Megoldás

$\overline{x} = \frac{0+1+6+5+3}{5} = 3$; $\quad \overline{y} = \frac{4+3+0+1+2}{5} = 2$; $\quad \sum (x_i - \overline{x})^2 = (-3)^2 + (-2)^2 + 3^2 + 2^2 + 0^2 = 26$

A paraméterek becslésének meghatározásához célszerű egy táblázatot készíteni:

	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$\hat{y}_i = \hat{a}x_i + \hat{b}$	$\hat{\varepsilon}_i = y_i - \hat{y}_i$
	0	4	-3	2	$\frac{50}{13} \approx 3,85$	$\frac{2}{13} \approx 0,15$
	1	3	-2	1	$\frac{42}{13} \approx 3,23$	$-\frac{3}{13} \approx -0,23$
	6	0	3	-2	$\frac{2}{13} \approx 0,15$	$-\frac{2}{13} \approx -0,15$
	5	1	2	-1	$\frac{10}{13} \approx 0,77$	$\frac{3}{13} \approx 0,23$
	3	2	0	0	$\frac{26}{13} \approx 2$	0
Összesen	15	10	0	0	-	0

a) A táblázat első négy oszlopából megkaphatjuk a képletek alapján a keresett együtthatókat:

$$\hat{b} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{-6 - 2 - 6 - 2}{9 + 4 + 9 + 4} = -\frac{8}{13}; \qquad \hat{a} = \overline{y} - \hat{b}\overline{x} = 2 - \left(-\frac{8}{13}\right) \cdot 3 = \frac{50}{13}$$

Tehát a regressziós egyenes: $\frac{50}{13} - \frac{8}{13}x = 3,846 - 0,615x$

b) $x = 2$-höz tartozó előrejelzett y érték $\hat{y} = \frac{50}{13} - \frac{8}{13}2 = 3,846 - 0,615 \cdot 2 = 2,616$

c) A reziduálisok meghatározásához az előbbi táblázat utolsó előtti oszlopa használható, a reziduálisokat a táblázat utolsó oszlopa tartalmazza (mindkettő a képlet alapján könnyen számolható).

Reziduális szórásnégyzet (ill. hiba) becslése:

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\left(\frac{2}{13}\right)^2 + \left(-\frac{3}{13}\right)^2 + \left(-\frac{2}{13}\right)^2 + \left(\frac{3}{13}\right)^2}{3} = \frac{\frac{2}{13}}{3} = \frac{2}{39} \approx 0,051$$

Az együtthatók becsléseinek szórásnégyzetét a következő képletekkel becsülhetjük:

$$\hat{D}^2(\hat{b}) = \frac{\hat{\sigma}^2}{\sum(x_i - \overline{x})^2} = \frac{\frac{2}{39}}{26} \approx 0,00197$$

$$\hat{D}^2(\hat{a}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right) = \frac{2}{39}\left(\frac{1}{5} + \frac{3^2}{26}\right) \approx 0,028$$