A4 12. hét - Megoldások

Regressziók összefoglalás

Adott egy X valószínűségi változó, mely c érték(ek)re lesz a $h_1(c) := \mathbb{E}|c-X|$ **hiba minimális**? A válasz: az $m(X)$ **mediánra** (amiből lehet több is, diszkrét változó esetén). És mely c értékre lesz $h_2(c) := \mathbb{E}\big[(c-X)^2\big]$ minimális? A válasz: az $\mathbb{E}X$ **várható értékre** (Steiner-tétel, tanultuk.)

Hasonlóképpen, ha az $X_1, \dots, X_n$ független kísérleteket látjuk egy ismeretlen eloszlású valószínűségi változóra, és ezek alapján akarjuk becsülni a változót, akkor:

(a) az a c, amire $h_1(c) := \sum_{i=1}^n |c-X_i|$ a minimális, az a **minta** $m(X_1, \dots, X_n)$ **mediánja**, azaz a nagyságra középső érték a kísérletekből (páros n esetén a két középső érték között bármi), és persze nagy n esetén ez egy jó becslés lesz a valószínűségi változó mediánjára; illetve

(b) az a c, amire $h_2(c) := \sum_{i=1}^n (c-X_i)^2$ a minimális, az a **mintaátlag** $\mu(X_1, \dots, X_n) := (X_1 + \dots + X_n)/n$, és ez egy jó becslés lesz a valószínűségi változó várható értékére.

Izgalmasabb, amikor egy kétdimenziós (X, Y) valószínűségi változóból látunk független kísérleteket, és ezek alapján értenénk meg, hogyan függ Y az X-től; pontosabban, X **milyen függvényével tudnánk Y-t a legjobban becsülni**? A válasz függ attól, hogyan mérjük, milyen jó a becslésünk:

(a) Akkor lesz a $h_1(f) := \mathbb{E}|f(X)-Y|$ hiba minimális, ha $\mathbb{E}\Big[|f(x)-Y| \,\Big|\, X = x\Big]$-et minimalizáljuk minden rögzített x-re, azaz $f(x)$ az Y **feltételes mediánja** az $X = x$ feltétel mellett.

(b) Akkor lesz a $h_2(f) := \mathbb{E}\big[(f(X)-Y)^2\big]$ hiba minimális, ha $\mathbb{E}\Big[(f(x)-Y)^2 \,\Big|\, X = x\Big]$-et minimalizáljuk, azaz $f(x)$ az Y **feltételes várható értéke** az $X = x$ feltétel mellett.

Ha csak **lineáris** f függvényeket engedünk meg, akkor a $h_2(f)$ négyzetes hibát az **első regressziós egyenes** minimalizálja: $f(x) = \mu_2 + r\,(x-\mu_1)\,\sigma_2/\sigma_1$, ahol μ_1 és σ_1 az X várható értéke és szórása, μ_2 és σ_2 az Y-éi, r pedig X és Y korrelációs együtthatója. A **második regressziós egyenes** pedig azon g lineáris függvény, mely az $\mathbb{E}\big[(g(Y)-X)^2\big]$ hibát minimalizálja: $g(y) = \mu_1 + r\,(y-\mu_2)\,\sigma_1/\sigma_2$.

Ezek nem csak azért fontosak, mert a lineáris összefüggéseket fogadja be a legkönnyebben az értelmünk, hanem mert a μ_i, σ_i, r értékeket természetes módon becsülhetjük egy $(X_1, Y_1), \dots, (X_n, Y_n)$ adathalmazból. Mégpedig: a $\mu(X)$ mintaátlagot már feljebb definiáltuk, a **minta varianciája** pedig $\sum_{i=1}^n (X_i - \mu(X))^2/(n-1)$; kovariancia hasonlóan. Az $n-1$-gyel osztás n helyett nem nyomdahiba, hanem így lesz $\mathbb{E}\sum_{i=1}^n (X_i - \mu(X))^2/(n-1) = \mathbb{V}\mathrm{ar}(X)$, ha utánaszámolunk.

Láttuk az előzőhéten, hogy **kétdimenziós normális** eloszlásokra a feltételes eloszlás normális, így mediánja és várható értéke megegyezik, ráadásul lineáris függvénye a feltételnek, így megegyezik a regressziós egyenessel.

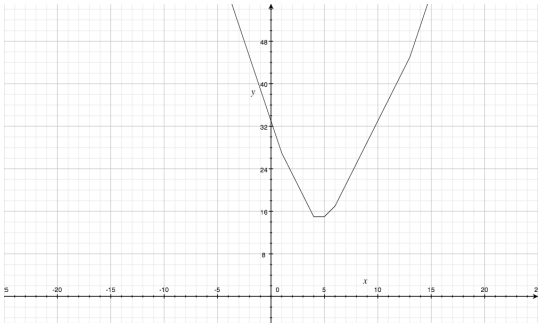
Regressziós feladatok

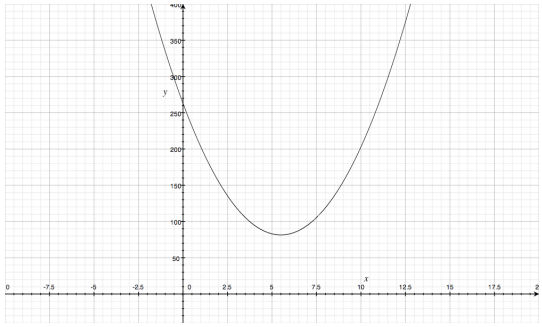
1. Vegyük a $4, 6, 1, 4, 13, 5$ adathalmazt (más néven mintát).

   (a) Határozzuk meg a $h_1(c)$ hibafüggvényt és a minta mediánjait!

   (b) Határozzuk meg a $h_2(c)$ hibafüggvényt és a mintaátlagot!

   Megoldás:





   Az (a) részben egy törtlineáris konvex függvényt kapunk, melynek minimuma a 4 és 5 között vétetik föl, ott konstans 15. A mediánok a 4 és 5 közötti számok. A (b) részben egy pozitív állású parabolát kapunk, aminek minimumhelye 5.5, ez a mintaátlag.

2. Egy kétdimenziós háromelemű mintánk első koordinátái $-1, 0, 1$, második koordinátái $3, 4, 5$, valamilyen sorrendben. Világos, hogy $3! = 6$-féleképpen lehet összepárosítani a koordinátákat. A koordinátákkénti mintamediánok, -átlagok, és -szórások persze nem függenek a párosítástól. Mik ezek a koordinátákkénti értékek? És mi a korrelációs együttható a 6 lehetséges párosításban?

Megoldás: Az első koordináta mediánja és átlaga is 0, szórása $((1-0)^2+(0-0)^2+(-1-0)^2)/2 = 1$. A második koordináta mediánja és átlaga is 4, szórása megint 1. A korrelációs együttható 1 és -1 az azonosan rendezett illetve az ellentétetesen rendezett esetben. A $(-1,3)$, $(0,5)$, $(1,4)$ esetben pedig, például, $((-1)(-1)+0\cdot 1+1\cdot 0)/2 = 0.5$ a kovariancia és a korrelációs együttható is. Nem volt kérdés, de az első regressziós egyenes az $y = 4 + x/2$, a második regressziós egyenes az $x = 0 + (y-4)/2$, azaz $y = 2x + 4$.

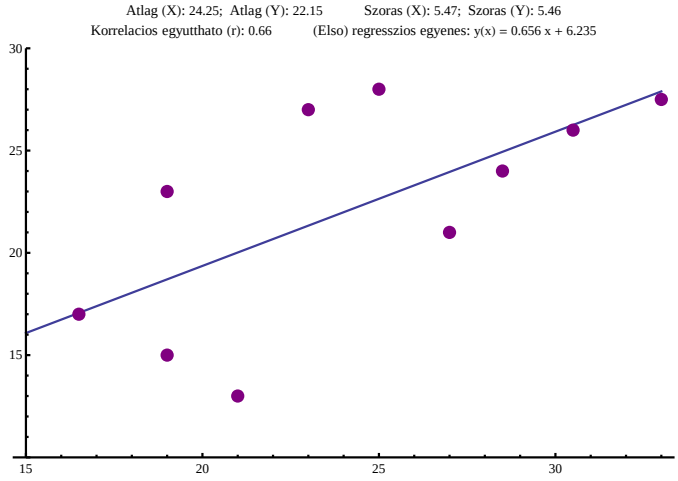
3. Egy tízfős A4 csoportban, az i-edik diák első hét röpZH eredményének összegét jelölje X_i, első nagyZH-jának eredményét pedig Y_i. Az eredmények: $(21,13)$, $(25,28)$, $(19,23)$, $(30.5,26)$, $(28.5,24)$, $(19,15)$, $(27,21)$, $(23,27)$, $(33,27.5)$, $(16.5,17)$.

   (a) Határozzuk meg az X és Y minták átlagait, szórásait, mediánjait, és korrelációs együtthatójukat!

   (b) Írjuk föl a minta két regressziós egyenesét! Mennyire tűnik jónak az adatok alapján a lineáris közelítés, és mennyire gondoljuk, hogy elvileg lineárisnak kellene lennie az összefüggésnek?

   (c) Kiderül, hogy volt még egy láthatatlan diák is a csoportban, akinek a nagyZH-ja 25 pontos lett. Milyen röpZH összpontszámot tippelünk neki? És ha az derült volna ki, hogy a röpZH összpontszáma 26, akkor milyen nagyZH pontszámot tippelnénk?

   Megoldás:



   A mintaátlag $(24.25, 22.15)$. X mediánjai a $[23,25]$ intervallum, Y mediánjai a $[23,24]$ intervallum. A szórások, $\sum_{i=1}^{10}(X_i-\mu(X))^2/9$-cel számolva, 5.47 és 5.46. A korrelációs együttható, szintén 9-cel osztva, 0.66. A második regressziós egyenes $y \mapsto 24.25 + 0.66(y-22.15)5.47/5.46$, ami az $y = 25$ helyen $x = 26.13$-at ad. Az első regressziós egyenes pedig $x \mapsto 22.15 + 0.66(y-24.25)5.46/5.47$, ami az $x = 26$ helyen $y = 23.3$-at ad, szóval nem kaptuk vissza az $y = 25$ indulóértéket, ugyanis a két regressziós egyenes különbözik egymástól.

   Az elég nagy korrelációs együttható látszik az adathalmazon, de aközött különbséget tenni, hogy lineáris összefüggés van elég nagy véletlen hibával, vagy pedig egy rendkívül bonyolult összefüggés véletlen nélkül, azt biztosan eldönteni nem lehet.

4. Legyen X a Duna mai bécsi vízállása, Y pedig legyen a holnaputáni budapesti vízállás. Statisztikai megfigyelések alapján (X,Y) együttes sűrűségfüggvénye $f(x,y) = \frac{6}{5}\left(x + (y-1)^2\right)$ ha $0 < x < 1, 0 < y < 1$, egyébként 0. A mért bécsi vízállás ismeretében mi a legjobb tipp a holnaputáni budapesi vízállásra ha a négyzetes eltérés várható értékét akarjuk minimalizálni? Mi a helyzet akkor, ha csak lineáris függvényt használhatunk a becsléshez? Mi a helyzet akkor ha az abszolút eltérés várható értékét akarjuk minimalizálni? (Az adatok nem valósak, továbbá feltesszük, hogy a vízállást egy 0 és 1 közötti szám jellemzi)

   Megoldás:

   Ha a négyzetes eltérés várható értékét akarjuk minimalizálni, akkor a legjobb függvény ere a $k(x) = E(Y|X = x)$. A VIII/12-es feladatban meghatároztuk az $Y|X = x$ eloszlás $f_{2|1}(y|x)$ feltételes sűrűségfüggvényét. Ilyenkor x egy paraméter, amit látunk az y-ban sűrűségfüggvény. Nincs más dolgunk, mint kiszámolni a várható értéket a megszokott módon:

$$k(x) = E(Y|X=x) = \int_{-\infty}^{\infty} y \cdot f_{2|1}(y|x)dy = \int_0^1 y \cdot \frac{x+(y-1)^2}{x+\frac{1}{3}}dy = \frac{1+6x}{4+12x}$$

   Ha továbbra is a négyzetes eltérés várható értékét akarjuk minimalizálni, de csak lineáris függvényt használhatunk, akkor a legjobb becslő egyenes a regressziós egyenes, vagyis az $y = l(x) = m_Y + (x-m_X)r\frac{\sigma_Y}{\sigma_X}$ egyenes. Ehhez meg kell határoznunk a várható értékeket, szórásokat és a korrelációt. Az együttes sűrűségfüggvényből mindezek számolhatóak. A számolás nem nehéz, mert egyszerű kétváltozós polinómokat kell integrálni, de egy kicsit időigényes:

$$E(X) = \int_0^1 \int_0^1 x\frac{6}{5}(x+(y-1)^2)dydx = \frac{3}{5}$$

$$E(Y) = \int_0^1 \int_0^1 y\frac{6}{5}(x+(y-1)^2)dydx = \frac{2}{5}$$

$$E(XY) = \int_0^1 \int_0^1 xy\frac{6}{5}(x+(y-1)^2)dydx = \frac{1}{4}$$

$$E(X^2) = \int_0^1 \int_0^1 x^2\frac{6}{5}(x+(y-1)^2)dydx = \frac{13}{30}$$

$$E(Y^2) = \int_0^1 \int_0^1 y^2\frac{6}{5}(x+(y-1)^2)dydx = \frac{6}{25}$$

A fenti integrálok értékéből, már felírhatóak a szórások és a korreláció is, az adódó regressziós egyenes képlete:

$$y = l(x) = \frac{2}{5} + (x - \frac{1}{3})0,13)\frac{\sqrt{\frac{2}{25}}}{\sqrt{\frac{11}{150}}}.$$

Ha pedig az abszolút eltérés várható értékét akarjuk minimalizálni, akkor a legjobb függvény az $a(x) = Med(Y|X = x)$ vagyis az $Y|X = x$ feltételes eloszlás mediánja. Ez ebben az esetben egy harmadfokú egyenletre vezet, így ezt nem számoljuk ki. Megjegyzem, hogy a 12. heti előadáson számoltatok feltételes mediánt.

5. Egy kétdimenziós valószínűségi változó sűrűségfüggvénye $\frac{1}{6}xy$ $(0 < x < 2, x < y < 2x)$.

   Milyen $k(y)$ függvénnyel érdemes a második koordinátából az elsőt tippelni, ha az a célunk, hogy a tippelésnél elkövetett hiba négyzetének átlagos értéke sok kísérlet esetén minél kisebb legyen,

   (a) ha feltesszük, hogy $k(y)$ lineáris,

   (b) ha $k(y)$ tetszőleges valós lehet?

   Megoldás: Az (a) részben képlet szerint $x = l(y) = m_X + (y - m_Y)r\frac{\sigma_X}{\sigma_Y}$ a második regressziós egyenes képlete. Az előző feladattal való hasonlóság miatt jelöltük ez a függvényt $l(y)$-al. Az előző feladat mintájára minden számolható a képletben. A kiszámolásra fordítható idő hiányában mellőzzük a végeredményt.

   A (b) részben pedig $k(y) = E(X|Y = y)$ függvényt kell kiszámolni. Itt van egy technikai nehézség, az $f_2(y)$ marginális sűrűségfüggvényt máshogy kell számolni az $0 < y < 2$ és az $2 < y < 4$ esetekben (ha felrajzoljátok $f(x, y)$ tartóját akkor világos, hogy miért). Mindez a lenti képletek nevezőjében látszik.

$$f_{1|2}(x|y) = \frac{f(x,y)}{\int\limits_{y/2}^{y} f(x,y)dx} = \frac{\frac{1}{6}xy}{\int\limits_{y/2}^{y} \frac{1}{6}xydx} = \frac{xy}{\frac{3}{8}y^3} = \frac{8x}{3y^2} \quad , ha \quad y < 2.$$

$$f_{1|2}(x|y) = \frac{f(x,y)}{\int\limits_{y/2}^{2} f(x,y)dx} = \frac{\frac{1}{6}xy}{\int\limits_{y/2}^{2} \frac{1}{6}xydx} = \frac{xy}{2y - \frac{1}{8}y^3} = \frac{x}{2 - \frac{1}{8}y^2} \quad , ha \quad y > 2.$$

$$k(y) = E(X|Y = y) = \int\limits_{y/2}^{y} x \cdot f_{1|2}(x|y)dx = \int\limits_{y/2}^{y} x \cdot \frac{8x}{3y^2}dx = \frac{7}{9}y \quad , ha \quad y < 2.$$

$$k(y) = E(X|Y = y) = \int\limits_{y/2}^{2} x \cdot f_{1|2}(x|y)dx = \int\limits_{y/2}^{2} x \cdot \frac{x}{2 - \frac{1}{8}y^2}dx = \frac{y}{3} + \frac{16}{12 + 3y} \quad , ha \quad y > 2.$$

6. (a) Kétszer dobtunk egy kockával, a dobások összege 10. Mi az első dobás feltételes várható értéke? És mit tippelünk az első dobásra?

   (b) Legyen X két dobás összege, Y pedig az első dobás. Határozzuk meg a regressziós egyenest!

   (c) Tízszer dobtunk egy kockával, a dobások összege 50. Most mi az első dobás feltételes várható értéke? És mit tippelünk az első dobásra? És mi a regressziós egyenes?

   Megoldás: (a) Legyen Z a második dobás. Világos, hogy $\mathbb{E}[Y|X = x] = \mathbb{E}[Z|X = x]$, a szimmetria miatt. Viszont $\mathbb{E}[Y|X = x] + \mathbb{E}[Z|X = x] = \mathbb{E}[Y + Z|X = x] = x$, így $\mathbb{E}[Y|X = x] = x/2$. Ha $x = 10$, akkor ez 5. Mi a feltételes eloszlás? Háromféle képpen kaphatunk $X = Y + Z = 10$-et: $4 + 6$, $5 + 5$, $6 + 4$. Azaz $\mathbb{P}[Y = 4|X = 10] = P[Y = 5|X = 10] = P[Y = 6|X = 10] = 1/3$, nincs igazán jó tipp.

   (b) Mivel a fent meghatározott feltételes várható érték $x \mapsto x/2$ lineáris, ez a regressziós egyenes is.

   (c) A feltételes várható érték ugyanúgy 5, mint az előbb. Mi a feltételes eloszlás? Legyen X a 10 dobás összege, Z pedig az utolsó 9 dobásé. $\mathbb{P}[Y = k|X = 50] = \mathbb{P}[Y = k, Z = 50-k]\mathbb{P}[X = 50] = \mathbb{P}[Z = 50-k]/(6\mathbb{P}[X = 50])$. Azaz azt keressük, mely $k = 1, 2, \ldots, 6$-ra lesz $\mathbb{P}[Z = 50 - k]$ maximális. A Z várható értéke 31.5, módusza is

ekörül van. Intuitíven világos, hogy $k = 6$-nál lesz a minket érdeklő maximum, azaz a 6-osra érdemes tippelni. Ezt be is lehet bizonyítani, a legegyszerűbben Excellel.

X szórásnégyzete 10-szer az Y-é, azaz $10 \cdot 2.917 = 29.17$, szórása 5.40. A kovariancia: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y + Z, Y) = \mathbb{V}\mathrm{ar}(Y) = 2.917$, hiszen Y és Z függetlenek. Így a korrelációs együttható $r = 2.917/\sqrt{29.17 \cdot 2.917} = 1/\sqrt{10} \approx 0.316$. A regressziós egyenes: $y = 3.5 + (x - 35)/\sqrt{10} \cdot \sqrt{2.917}/\sqrt{29.17} = 3.5 + (x - 35)/10 = x/10$. Ja, ezt tudtuk is, hiszen $x/10$ volt az $\mathbb{E}[Y|X = x]$, ami lineáris. Akkor ez fölösleges munka volt.

7. Legyenek X_1, X_2 független RAND() számok, minimumuk X, maximumuk Y. Határozzuk meg az Y feltételes mediánját és várható értékét az $X = x$ feltétel mellett, és az első regressziós egyenest.

   Megoldás: A feladat megoldásában nincs szükségünk az együttes sűrűségre. Mivel tanulságos, ezért később kiszámoljuk. De előbb koncentráljunk a kérdésre. Szükségünk van a $Y|X = x$ eloszlás mediánjára és várható értékére. Nézzük meg jobban az $Y|X = x$ eloszlást. Tudjuk, hogy a két RND minimuma x, ezen feltétel mellett mi a maximum eloszlása ez a kérdés. Lehet érezni hogy az $X = x$ feltétel mellett Y eloszlása egyenletes lesz az $(x, 1)$ szakaszon. Vagyis $Y|X = x$ egyenletes eloszlású az $(x, 1)$ szakaszon. Az egyenletes eloszlás mediánja és várható értéke egybeesik, így a válasz $E(Y|X = x) = Med(Y|X = x) = \frac{x+1}{2}$. Szerencsénkre ez lineáris függvénye x-nek, így a regressziós egyenes is $y = \frac{x+1}{2}$.

   Ha ki akarjuk számolni az együttes sűrűségfüggvényt, akkor a következőt tehetjük. A tartóját könnyű látni: X biztosan 0 és 1 között van, továbbá ha tudjuk, hogy $X = x$ akkro Y x és 1 között mozoghat. Így a tartó $0 < x < y < 1$ háromszög. A képletet legegyszerűbb az $f(x, y) = f_1(x)f_{2|1}(y|x)$ azonosságon keresztül meghatározni. Vegyük észre, hogy $f_{2|1}(y|x) = \frac{1}{1-x}$ ha $x < y < 1$ egybként 0, mert a korábbiak szerint $Y|X = x$ egyenletes az $(x, 1)$ szakaszon. Már csak $f_1(x)$-t kell meghatároznunk. Határozzuk meg X $F_1(x)$ eloszlásfüggvényét, majd abból deriválással adódik $f_1(x)$. Legyen x 0 és 1 között (ez az érdekes eset)

$$\begin{aligned} F_1(x) &= \mathbb{P}(X < x) = \mathbb{P}(\text{két független RND minimuma kisebb, mint x}) \\ &= \mathbb{P}(\text{mindkettő kisebb, mint x}) + \mathbb{P}(\text{pontosan az egyik kisebb, mint x}) = x^2 + 2x(1 - x). \end{aligned} \tag{1}$$

   A fenti képlet deriválásából adódik, hogy $f_1(x) = 2 - 2x$ ha $0 < x < 1$. Mindent összerakva pedig kapjuk hogy $f(x, y) = 2$ ha az $0 < x < y < 1$ háromszögön vagyunk, egyébként pedig 0. Vagyis végeredményben azt kaptuk, hogy (X, Y) eloszlása egyenletes az $0 < x < y < 1$ háromszögön.

8. Legyen az (X, Y) kétdimenziós valószínűségi változó együttes sűrűségfüggvénye:

   (a)
$$f(x, y) = \begin{cases} 2\exp(-(x + 2y)), & \text{ha } 0 \le x, y; \\ 0, & \text{egyébként.} \end{cases}$$

   (b)
$$f(x, y) = \begin{cases} x + y, & \text{ha } 0 < x < 1;\ 0 < y < 1; \\ 0, & \text{egyébként.} \end{cases}$$

   (c)
$$f(x, y) = \begin{cases} 24xy, & \text{ha } 0 \le x;\ 0 \le y \text{ és } 0 \le x + y \le 1; \\ 0, & \text{egyébként.} \end{cases}$$

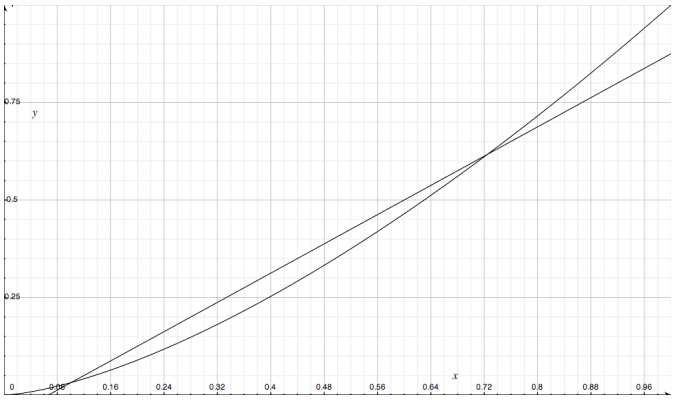
   Határozzuk meg az Y feltételes mediánját és várható értékét az $X = x$ feltétel mellett, és az első regressziós egyenest.

   Megoldás: Az (a) részben a fenti képleten látszik, hogy X marginális eloszlása 1 paraméterű exponenciális, míg Y marginális eloszlása 2 paraméterű exponenciális, továbbá X és Y függetlenek. A függetlenség miatt az $Y|X = x$ eloszlás megegyezik Y marginális eloszlásával, így az $E(Y|X = x) = E(Y) = \frac{1}{2}$, továbbá $MED(Y|X = x) = MED(Y) = \frac{\ln(2)}{2}$ (lásd VII/2-es feladat). A regressziós egyenes képletében a függetlenség miatt $r = 0$, ezért ez is egyszerűen az $y = \frac{1}{2}$ egyenes lesz. Megjegyezzük, hogy most minden kiszámolt becslő függvény konstans a függetlenség miatt, nem tudjuk használni X értékét Y megbecslésére.

   A (b) és (c) rész megoldása HF.

9. (a) Legyen U egy egyenletes véletlen szám a $[0,1]$ intervallumból, $X = U^2$ és $Y = U^3$. Mi X és Y korrelációs együtthatója? Mi az $\mathbb{E}[Y \mid X = x]$ feltételes várható érték és az $\sqrt{\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X = x])^2 \,\big|\, X = x\big]}$ feltételes szórás? Határozzuk meg az első regressziós egyenest.

Megoldás: $\mathbb{E}X = \mathbb{E}U^2 = \int_0^1 u^2 du = 1/3$. $\mathbb{E}X^2 = \mathbb{E}U^4 = \int_0^1 u^4 du = 1/5$. $\mathbb{V}\mathrm{ar}X = 1/5 - 1/9 = 4/45$.
$\mathbb{E}Y = \mathbb{E}U^3 = \int_0^1 u^3 du = 1/4$. $\mathbb{E}Y^2 = \mathbb{E}U^6 = \int_0^1 u^6 du = 1/7$. $\mathbb{V}\mathrm{ar}Y = 1/7 - 1/16 = 9/112$.
$\mathbb{E}XY = \mathbb{E}U^5 = \int_0^1 u^5 du = 1/6$. $r(X,Y) = \frac{(1/6-1/12)\cdot 3\sqrt{5}\cdot 4\sqrt{7}}{2\cdot 3} = \sqrt{35/36}$.
Ha $X = x$, akkor $U = \sqrt{x}$ és $Y = x^{3/2}$. Ez a feltételes várható érték, a feltételes szórás pedig 0.
A regressziós egyenes: $1/4 + \sqrt{35/36}(x - 1/3) \cdot 3 \cdot 3\sqrt{5}/(4\sqrt{7} \cdot 2) = 1/4 + (x - 1/3)15/16$
A következő ábrán látható a feltételes várható érték (legjobb becslés várható négyzetes eltérés értelemben) és a regressziós egyenes (legjobb lineáris becslés várható négyzetes eltérés értelemben).



(b) Most legyen U egy egyenletes véletlen szám a $[0,2]$ intervallumból, és, mint az előbb, $X = U^2$ és $Y = U^3$. Változott-e a korrelációs együttható? *Megoldás*: HF