Villamosmérnök A4

11. gyakorlat (2012. nov. 26-27.)

Regressziók

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**.

Hasonlóképpen, ha az $X_1, \ldots, X_n$ 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** $m(X_1, \ldots, X_n)$ **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** $\mu(X_1, \ldots, X_n) := (X_1 + \cdots + X_n)/n$, és ez egy jó becslés lesz a valószínűségi változó várható értékére.

Ez akkor kezd izgalmassá válni, amikor egy kétdimenziós (X, Y) valószínűségi változóból látunk független kísérleteket, és ezek alapján meg akarjuk érteni, 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ adathalmazból.

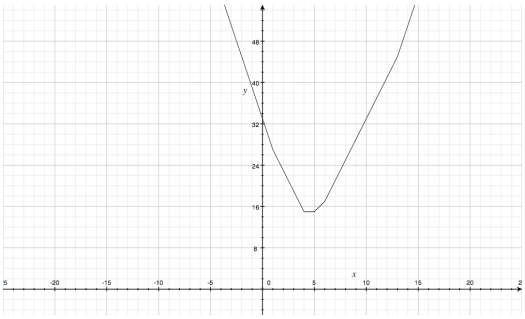
Láttuk korábban, hogy **kétdimenziós normális** eloszlásokra a feltételes medián és a feltételes várható érték függvény is megegyezik a regressziós egyenessel.

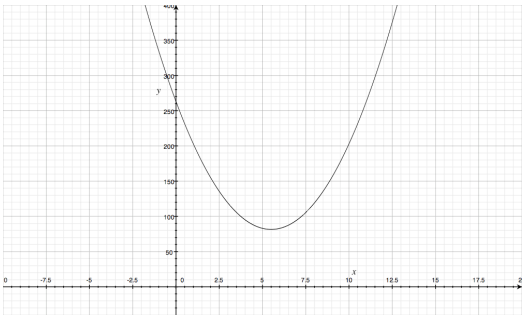
1. Vegyük a $4, 6, 1, 4, 13, 5$ adathalmazt (más néven mintát).

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

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

   Megoldás:

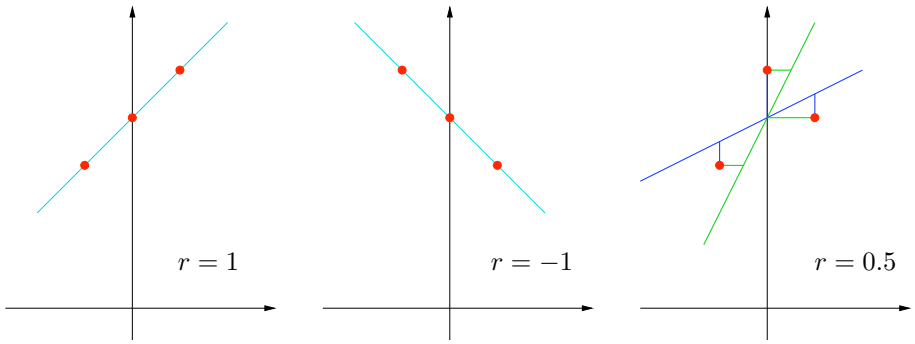




   Az (a) részben egy törtlineáris konvex függvényt kapunk, melynek minimuma a 4 és 5 között vétetik föl, ott konstans 15. A mediánok a 4 és 5 közötti számok. A (b) részben egy pozitív állású parabolát kapunk, aminek minimumhelye 5.5, ez a mintaátlag.

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 minta-mediánok, -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 $y = 2x + 4$.



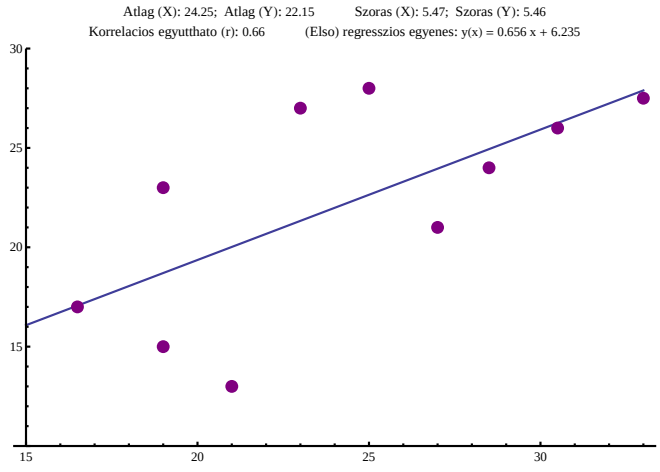
3. Egy tízfős A4 csoportban, az i-edik diák első hét röpZH eredményének összegét jelölje X_i, első nagyZH-jának eredményét pedig Y_i. Az eredmények: $(21, 13)$, $(25, 28)$, $(19, 23)$, $(30.5, 26)$, $(28.5, 24)$, $(19, 15)$, $(27, 21)$, $(23, 27)$, $(33, 27.5)$, $(16.5, 17)$.

   (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. (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 $k=1,2,\dots,6$-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.

5. Legyenek $X_1, \dots, X_k$ 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,

   (a) $k=2$-re;

   (b) általános k-ra.

   Megoldás: (a) Az együttes sűrűségfgv: $f(x,y)\Delta x\Delta y \approx 2\Delta x\Delta y \mathbf{1}_{\{0\le x\le y\le 1\}}$, egyenletes eloszlás egy háromszögben. Tehát $f_{2|1}(y|x) = \mathbf{1}_{\{x\le y\le 1\}}/(1-x)$, egyenletes eloszlás az $[x,1]$ szakaszon. A feltételes medián és várható érték is $(x+1)/2$, és mivel ez lineáris x-ben, ez az első regressziós egyenes.

   (b) Az együttes sűrűségfgv: $f(x,y)\Delta x\Delta y \approx k(k-1)\,(y-x)^{k-2}\Delta x\Delta y \mathbf{1}_{\{0\le x\le y\le 1\}}$. Mivel $f_{2|1}(y|x) = f(x,y)/f_1(x)$, szükségünk van az első marginálisra: $f_1(x) = \int_x^1 k(k-1)\,(y-x)^{k-2}\,dy = k\,(1-x)^{k-1}$. Tehát $f_{2|1}(y|x) = (k-1)\,\frac{(y-x)^{k-2}}{(1-x)^{k-1}}\,\mathbf{1}_{\{x\le y\le 1\}}$. A feltételes várható érték $\int_x^1 y\, f_{2|1}(y|x)\,dy = \frac{x+k-1}{k}$. Ez megint csak lineáris x-ben, így ez az első regressziós egyenes. A feltételes medián akkor $m = m(x)$, ha $\int_x^m (y-x)^{k-2}\,dy = \frac{(1-x)^{k-1}}{2(k-1)}$, azaz $(m-x)^{k-1} = (1-x)^{k-1}/2$, azaz $m = x + (1-x)/2^{1/(k-1)}$.

6. 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: (a) $f(x,y) = f_X(x)f_Y(y)$, 1 illetve 2 paraméterű független exponenciálisok. Így az $X=x$ feltétel mellett, tetszőleges x-re, egy 2 paraméterű exponenciálist látunk. A regressziós egyenes vízszintes, azaz konstans.

7. Egy hivatalban minden ügyfél kiszolgálása 10 perc várható értékű exponenciális valószínűségi változót vesz igénybe, egymástól függetlenül. Ha k ügyfelet összesen x idő alatt szolgálnak ki, akkor a legelső ügyfél kiszolgálásának mi a feltételes mediánja és várható értéke, mi a regressziós egyenes?

8. (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.

(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?

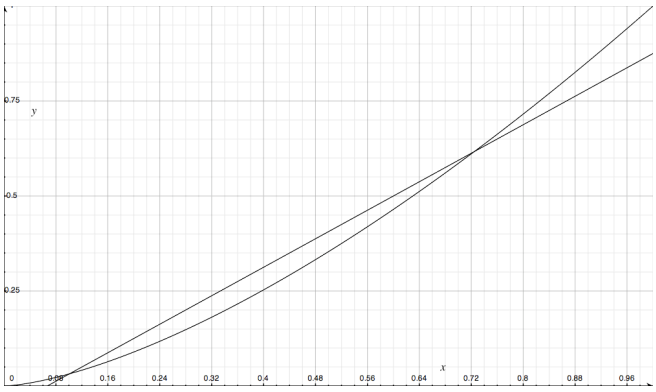
Megoldás: $\mathbb{E}X = \mathbb{E}U^2 = \int_0^1 u^2 du = 1/3$. $\mathbb{E}X^2 = \mathbb{E}U^4 = \int_0^1 u^4 du = 1/5$. $\mathbb{V}\mathrm{ar}X = 1/5 - 1/9 = 4/45$.

$\mathbb{E}Y = \mathbb{E}U^3 = \int_0^1 u^3 du = 1/4$. $\mathbb{E}Y^2 = \mathbb{E}U^6 = \int_0^1 u^6 du = 1/7$. $\mathbb{V}\mathrm{ar}Y = 1/7 - 1/16 = 9/112$.

$\mathbb{E}XY = \mathbb{E}U^5 = \int_0^1 u^5 du = 1/6$. $r(X,Y) = \frac{(1/6-1/12)\cdot 3\sqrt{5}\cdot 4\sqrt{7}}{2\cdot 3} = \sqrt{35/36}$.

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: $1/4 + \sqrt{35/36}(x - 1/3) \cdot 3 \cdot 3\sqrt{5}/(4\sqrt{7} \cdot 2) = 1/4 + (x - 1/3)15/16$



9. Magyarországon a felnőtt férfiak testmagassága átlagosan 178 cm, 9 cm szórással, míg testsúlyuk 85 kg, 10 kg szórással. A korrelációs együttható 0.7, azaz minél magasabb valaki, annál súlyosabb is.

   (a) Átlagosan mekkora súlyú egy 190 cm magas férfi?

   (b) Átlagosan milyen magas egy 94.3 kg-os férfi?

   (c) Hasonlítsuk össze e két eredményt.

   Megoldás: (a) $85 + 0.7 \cdot (190 - 178) \cdot 10/9 \approx 94.3$ kg.

   (b) $178 + 0.7 \cdot (94.3 - 85) \cdot 9/10 \approx 183.9$ cm.

   (c) A 183.9 cm sokkal közelebb van a 178-as átlagmagassághoz, mint a 190 cm, amiből indultunk. A 190 cm az $12/9 = 4/3$ szórásnyira van az átlagtól, a 94.3 kg már csak 0.93 szórásnyira, az 183.9 cm pedig csak 0.66 szórásnyira. Ennek az az oka, hogy azon faktorok, melyek egyszerre tesznek valakit magasabbá és nehezebbé, csak részben felelősek a 190 cm-ért, és vannak olyan faktorok is, melyek csak magassá teszik. Ha látjuk a 190 cm-es végeredményt, természetesebb arra gondolni, hogy mindkét fajta hatás kissé átlag fölött produkált, mintsem arra, hogy az egyik hatás a felelős teljesen, a másik pedig átlagosan viselkedett. Tehát a testsúlyra is ható faktorok alighanem csak kb fele annyira voltak átlag felettiek, mint amennyire a 190 cm az átlag felett van, és így a testsúly tipikusan már nem lesz olyan kiemelkedő.