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 $X_1, \dots, X_n$ 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.

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 $(X_1, Y_1), \dots, (X_n, Y_n)$ 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!

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?

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?

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?

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

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.

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?

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.