

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**. És mely c értékre lesz $h_2(c) := \mathbb{E}[(c - X)^2]$ minimális? A válasz: az $\mathbb{E}X$ **várható értékre** (Steiner-tétel, tanultuk.)

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

- Akkor lesz a $h_1(f) := \mathbb{E}|f(X) - Y|$ hiba minimális, ha $\mathbb{E}[|f(x) - Y| \mid X = x]$ -et minimalizáljuk minden rögzített x -re, azaz $f(x)$ az Y **feltételes mediánja** az $X = x$ feltétel mellett.
- Akkor lesz a $h_2(f) := \mathbb{E}[(f(X) - Y)^2]$ hiba minimális, ha $\mathbb{E}[(f(x) - Y)^2 \mid X = x]$ -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 $\mathbb{E}[(g(Y) - X)^2]$ 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 $\mathbb{E} \sum_{i=1}^n (X_i - \mu(X))^2 / (n - 1) = \text{Var}(X)$, ha utánaszámolunk.

Regressziós feladatok

- 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).
 - Határozzuk meg az X és Y minták átlagait, szórásait, mediánjait, és korrelációs együtthatójukat!
 - 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?
 - 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?
- Legyen X a Duna mai bécsi vízállása, Y pedig legyen a holnaputáni budapesti vízállás. Statisztikai megfigyelések alapján (X, Y) együttes sűrűségfüggvénye $f(x, y) = \frac{6}{5}(x + (y - 1)^2)$ ha $0 < x < 1, 0 < y < 1$, egyébként 0. A mért bécsi vízállás ismeretében mi a legjobb tipp a holnaputáni budapesti vízállásra ha a négyzetes eltérés várható értékét akarjuk minimalizálni? Mi a helyzet akkor, ha csak lineáris függvényt használhatunk a becsléshez? (Az adatok nem valóságosak, továbbá feltesszük, hogy a vízállást egy 0 és 1 közötti szám jellemzi)
- 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,
 - ha feltesszük, hogy $k(y)$ lineáris,
 - ha $k(y)$ tetszőleges valós lehet?
- Kétszer dobtunk egy kockával, a dobások összege 10. Mi az első dobás feltételes várható értéke? És mit tippelünk az első dobásra?
 - Legyen X két dobás összege, Y pedig az első dobás. Határozzuk meg a regressziós egyenest!
 - Tízszor 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?
- Legyenek X_1, X_2 független $\text{RAND}()$ számok, minimumuk X , maximumuk Y . Határozzuk meg az Y feltételes mediánját és várható értékét az $X = x$ feltétel mellett, és az első regressziós egyenest.
- Legyen az (X, Y) kétdimenziós valószínűségi változó együttes sűrűségfüggvénye:
 - $$f(x, y) = \begin{cases} 2 \exp(-(x + 2y)), & \text{ha } 0 \leq x, y; \\ 0, & \text{egyébként.} \end{cases}$$

(b)

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

(c)

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

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

7. (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 $\mathbb{E}[Y | X = x]$ feltételes várható érték és az $\sqrt{\mathbb{E}[(Y - \mathbb{E}[Y | X = x])^2 | X = x]}$ 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?