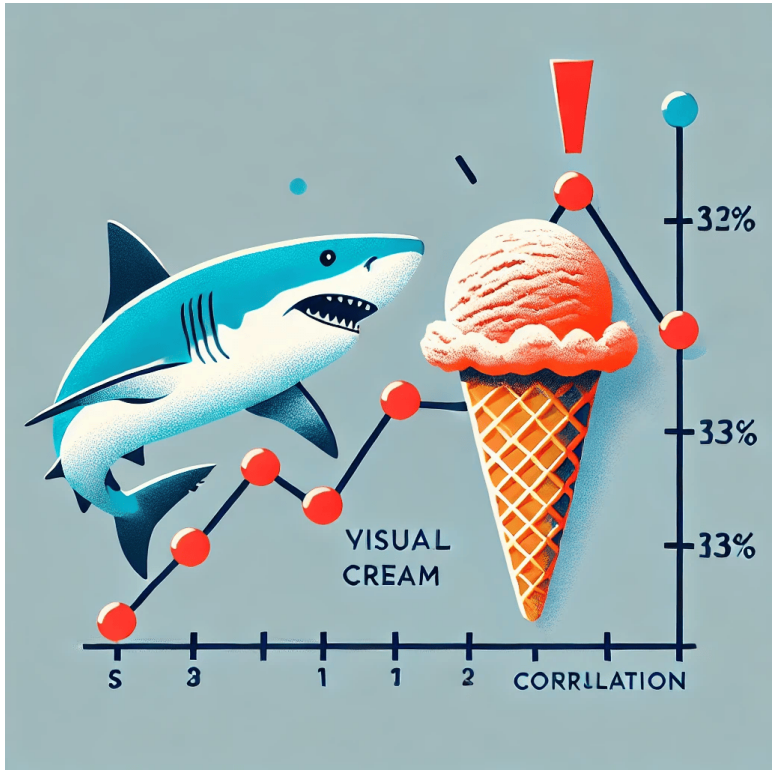
Regresszióanalízis: korreláció

Matematikai Statisztika
2024. szeptember 16.



Első lépés – Kovariancia fogalma

(elméleti) Kovariancia fogalma

A **kovariancia** egy mérőszám, amely két **valószínűségi változó** közötti **kapcsolat erősségét** írja le a következő értelemben: azt "mutatja meg", hogy ha az egyik változó növekszik, hogyan változik a másik: nő-e vagy csökken.

Kovariancia esetén:

- **Pozitív kovariancia**: Ha az egyik változó nő, a másik is hajlamos növekedni.
- **Negatív kovariancia**: Ha az egyik változó nő, a másik hajlamos csökkenni.
- **Kovariancia = 0**: A változók között "nincs kapcsolat".

Definíció

Két **valószínűségi változó** X és Y kovarianciája:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Azaz X és Y változók eltérésének szorzatának várható értéke.

Kovariancia – Tulajdonságok és problematika

Tulajdonságok

A kovariancia néhány fontos tulajdonsága:

- $\text{cov}(X, X) = \sigma_X^2$, azaz egy valószínűségi változó kovarianciája saját magával a szórásnégyzet.
- $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$
- $|\text{cov}(X, Y)| \leq \sigma_X \sigma_Y$, azaz a kovariancia abszolút értéke kisebb vagy egyenlő a szórások szorzatával.

A kovariancia problémája

A **kovariancia** egy fontos mérőszám, azonban van egy alapvető probléma vele: **skálaérzékeny**. Ez azt jelenti, hogy a kovariancia értéke attól függ, milyen mértékegységben mérjük a változókat. **Alacsony kovariancia** két különböző okból is adódhat:

- A változók **értékei kicsik**, így a kovariancia is kisebb lesz.
- A változók között **nincs lineáris kapcsolat**, tehát a kovariancia értéke ennek megfelelően nulla közeli lehet.

Emiatt a kovariancia értelmezése nehéz lehet, mivel nem világos, hogy a kis érték valódi gyenge kapcsolatot jelent-e, vagy csak a változók mértékegysége okozza a kis kovarianciát. A korrelációs együttható ebben segít, mivel normálja a kovarianciát a változók szórásaival.

Korrelációs együttható - fogalma és tulajdonságok

Korreláció fogalma

A kovariancia finomított változata a **korrelációs együttható**, amelyet úgy kapunk, hogy a kovarianciát a két valószínűségi változó szórásával normáljuk. Ezáltal a korreláció értéke mindig -1 és 1 között lesz. Ezt nevezzük **elméleti korrelációnak**.

A korrelációs együttható:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Tulajdonságok

- $|\text{corr}(X, Y)| \leq 1$
- Ha $\text{corr}(X, Y) = 1$, akkor és akkor, ha X és Y között pozitív lineáris kapcsolat van.
- Ha $\text{corr}(X, Y) = -1$, akkor és csak akkor, ha X és Y között negatív lineáris kapcsolat van.
- Ha $\text{corr}(X, Y) = 0$, nincs lineáris kapcsolat X és Y között.
- 0 korrelációból nem (feltétlen) következik függetlenség!!

Empirikus kovariancia

Ha adott egy statisztikai minta $(X_1, Y_1), \ldots, (X_n, Y_n)$ az X és Y eloszlásból, akkor az **empirikus kovariancia** az alábbi módon számítható:

$$C_n(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

ahol $\overline{X}$ és $\overline{Y}$ az X és Y **mintaátlagai**.
Az empirikus kovariancia célja, hogy hasonlóképpen vizsgálja a két változó közötti kapcsolatot, mint ahogy az elméleti kovariancia teszi, de adatsorokból, hiszen a való életben **adatsoraink** vannak, amelyekből szeretnénk következtetéseket levonni.

Empirikus korrelációs együttható

Az **empirikus kovariancia** alapján számítjuk ki az **empirikus korrelációs együtthatót**, amelyet a minták szórásával normálunk:

$$r_n(X, Y) = \frac{C_n(X, Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

ahol s_X és s_Y az X és Y **empirikus szórásai**.

Az **empirikus korreláció** segít jobban megérteni az adatok közötti **kapcsolatot**, és az eredmények alapján dönthetünk arról, hogy milyen kapcsolat van a változók között.

Guess the Correlation – Játékok

- Guess the Correlation – Melbourne University játék
- Guess the Correlation – Online játék

Empirikus korreláció tesztelése

Hipotézisvizsgálat az empirikus korrelációra

Lehetséges, hogy a kapott empirikus korreláció valójában csak véletlen eredmény. Ezért szükség van statisztikai tesztre annak ellenőrzésére, hogy a korreláció tényleg szignifikáns-e.

Hipotézisek:

- H_0: $r_n(X, Y) = 0$ (Nincs korreláció)
- H_1: $r_n(X, Y) \neq 0$ (Van korreláció)

Próbastatisztika

A teszteléshez használt próbastatisztika:

$$t = r_n(X, Y)\sqrt{\frac{n-1}{1 - r_n^2(X, Y)}}$$

amely t-eloszlású $n - 2$ szabadságfokkal.

Empirikus korrelációs együttható – Feladat

Feladat

Tekintsünk egy három elemű statisztikai mintát az (X, Y) háttérváltozókhoz. A minta az alábbi:

X	-1	2	5
Y	5	4	6

Feladat: Számítsuk ki az **empirikus kovarianciát** és **korrelációs együtthatót**, majd teszteljük le, hogy $\varepsilon = 0.05$ szignifikancia szinten a korreláció szignifikáns-e.

Megoldás

- **Mintaátlagok:** $\overline{X} = 2$, $\overline{Y} = 5$
- **Empirikus kovariancia:** $C_3(X, Y) = 1$
- **Empirikus korrelációs együttható:** $r_3(X, Y) = 0.5$
- **Próbastatisztika:** $t = \sqrt{\frac{2}{3}}$, ami kisebb a kritikus értéknél, így nem szignifikáns.

Korreláció vs. Ok-okozati kapcsolat

Mi az a korreláció?

A **korreláció** egy statisztikai mérőszám, amely megmutatja, hogy két változó hogyan mozog együtt. Ha az egyik változó nő, a másik is hasonlóan változik, de ez nem feltétlenül jelenti, hogy egyik okozza a másikat.

Fontos: A korreláció önmagában nem bizonyítja az ok-okozati kapcsolatot, csak azt jelzi, hogy ESETLEG van valamiféle statisztikai összefüggés.

Korreláció és ok-okozat közötti különbség

A korreláció és az ok-okozati kapcsolat közötti különbség megértése fontos a statisztikai következtetések helyes értelmezéséhez. Példa erre a jégkrém eladások és a cápatámadások korrelációja: mindkettő növekszik a nyári hónapokban, de a jégkrém fogyasztása nem okoz cápatámadásokat. Az igazi ok itt a hőmérséklet, ami mindkettőt befolyásolja.

Hibás következtetés a korrelációból

Jégkrém és cápatámadások

Tegyük fel, hogy egy kutatás kimutatta, hogy minél több jégkrémet adnak el, annál több cápatámadás történik. A statisztikai adatok szerint erős pozitív korreláció van a jégkrém eladások és a cápatámadások száma között.

De vajon a jégkrém okozza a cápatámadásokat? Vagy fordítva? Természetesen nem! A rejtett változó itt a hőmérséklet, amely mindkét jelenségre hatással van: a meleg nyári napokon több ember megy a strandra, ahol nő a cápatámadások esélye, és ekkor több jégkrémet is fogyasztanak.

Összegzés

Fontos: A korreláció felfedezése érdekes, de nem jelent ok-okozati kapcsolatot. A statisztikai elemzés során mindig érdemes további vizsgálatokkal ellenőrizni, hogy valójában mi okozza a kapcsolatot, és figyelembe venni a rejtett változókat (pl. hőmérséklet), amelyek torzíthatják az eredményeket.