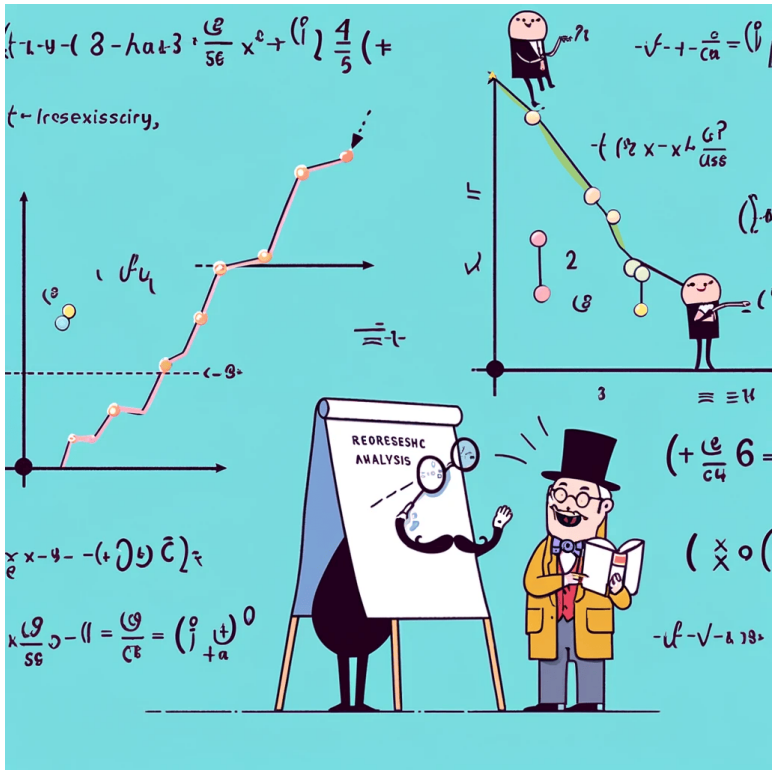
Regresszióanalízis: elméleti regresszió

Matematikai Statisztika
2024. október 21.



Regresszióanalízis - Elméleti regresszió

Alapfeladat

A **regresszióanalízis** célja egy függő változó Y és egy vagy több független változó $X_1, \ldots, X_p$ közötti kapcsolat leírása. Az alapvető kérdés az, hogy hogyan függ össze ezek között a változók között egy valószínűségi függvény.
Az elméleti megközelítésben a célunk az, hogy megtaláljuk azt a modellt, amely becslést ad a **függő változó várható értékére** a független változók függvényében.
Matematikailag ez azt jelenti, hogy a **feltételes várható érték függvényt** próbáljuk megbecsülni:

$$f(X_1, \ldots, X_p) = \mathbb{E}(Y|X_1, \ldots, X_p).$$

Ez a függvény leírja, hogy hogyan változik a függő változó Y, ha a független változók $X_1, \ldots, X_p$ értékei változnak.

Probléma a feltételes várható értékkel

A feltételes várható érték **megoldja a feladatot**, hiszen ez a "legjobb" függvény, amely a függő változó várható értékét adja meg a független változók függvényében. Azonban ennek a függvénynek a pontos kiszámítása gyakran nehézkes, mivel összetett és bonyolult függvényformákhoz vezethet.

Alternatív megoldás - Könnyebben kezelhető függvények

Egyszerűbb modellek keresése

A **feltételes várható érték** meghatározása a legjobb megoldás lenne, de gyakran **túl bonyolult** és nehezen kezelhető. Emiatt olyan **egyszerűbb modelleket** keresünk, amelyek:

- jól közelítik az elméleti modellt,
- könnyebben kezelhetők és érthetőek,
- általában a gyakorlati feladatok megoldására alkalmasak.

Ezek a modellek **nem pontosan oldják meg** az eredeti feladatot, de elegendően jó közelítést adnak, amelyeket könnyebb alkalmazni és interpretálni.

Hogyan választunk függvénycsaládot?

A célunk, hogy egy adott **függvénycsaládból** válasszuk ki azt a függvényt, amely a legjobban közelíti a függő változó várható értékét a független változók függvényében.
Matematikailag ez azt jelenti, hogy keressük azt a függvényt, amelyre a következő **kifejezés minimális**:

$$\mathbb{E}((Y - f(X_1, \ldots, X_p))^2)$$

vagyis a függő változó (Y) és a becsült érték ($f(X_1, \ldots, X_p)$) közötti különbség négyzetének **várható értékét** minimalizáljuk.

Informálisan: olyan függvényt keresünk, amely a legkisebb átlagos négyzetes eltérést adja a tényleges és a becsült értékek között.

Lineáris függvények mint gyakori megoldás

Az egyik leggyakrabban alkalmazott modell a **lineáris regresszió**, amely a következő formában írható fel:

$$f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

ahol:

- β_0 a konstans (tengelymetszet),
- $\beta_1, \ldots, \beta_p$ a magyarázó változókhoz tartozó együtthatók.

A lineáris függvények **egyszerűek** és **könnyen kezelhetők**, ezért gyakran választják őket, még ha nem is adják a legpontosabb megoldást. Azonban sok esetben elegendően jó becslést adnak a valós kapcsolatokra.

Egyszerű lineáris regresszió

Mi az egyszerű lineáris regresszió?

Az **egyszerű lineáris regresszió** a legegyszerűbb regresszióanalízis példa, ahol egyetlen magyarázó változó (X) van. Célunk, hogy megtaláljuk azt a lineáris kapcsolatot, amely a legjobban közelíti a függő változó (Y) értékeit.

A következő formát keresünk:

$$Y = a + bX$$

ahol a az Y tengely metszéspontja, b pedig a meredekség.

Cél: Minimális eltérés

A regresszió célja, hogy minimalizáljuk az **eltéréseket** a függő változó (Y) és a modell ($a + bX$) között. Ezt az alábbi képlet segítségével érjük el:

$$h(a, b) = \mathbb{E}((Y - (a + bX))^2),$$

vagyis minimalizáljuk az eltérés négyzetének **várható értékét**.

A regressziós egyenes levezetése

Matematikai megoldás

A célunk tehát, hogy megkeressük azokat az a és b értékeket, amelyek minimalizálják $h(a, b)$-t. Ehhez **deriválunk** a és b szerint:

- Az a szerinti derivált:
$$\frac{\partial h(a,b)}{\partial a} = -2\mathbb{E}(Y - a - bX) = 0,$$
- A b szerinti derivált:
$$\frac{\partial h(a,b)}{\partial b} = -2\mathbb{E}((Y - a - bX)X) = 0.$$

Ezek az egyenletek határozzák meg a regressziós egyenes paramétereit.

Az egyenletrendszer megoldása

Megoldás: Regressziós paraméterek

A két egyenlet megoldásával a következő kifejezéseket kapjuk:

- **Meredekség** (b):

$$b = \frac{\text{cov}(X, Y)}{\sigma_X^2} = \text{corr}(X, Y)\frac{\sigma_Y}{\sigma_X},$$

- **Y tengelymetszet** (a):

$$a = \mathbb{E}(Y) - b\mathbb{E}(X).$$

A meredekség (b) mutatja meg, hogyan változik Y az X változására, míg az a az Y tengelyen való metszéspontot adja meg.

A regressziós egyenes végső formája

Regressziós egyenlet

A végső formája az **egyszerű lineáris regresszió** egyenletének:

$$Y = a + bX = \text{corr}(X, Y)\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}(X)) + \mathbb{E}(Y),$$

Ez az egyenlet adja meg a legjobban illeszkedő egyenest X és Y között, minimalizálva a predikciók és a tényleges értékek közötti különbséget.

Összegzés

Az **egyszerű lineáris regresszió** alapját képezi a statisztikai modellezésnek, mivel egyszerűsége miatt jól értelmezhető, és könnyen alkalmazható különböző területeken. Az egyenlet megadja, hogyan változik Y az X értékeinek változására.