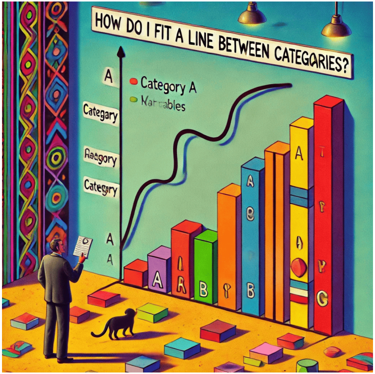
Regresszióanalízis:
Mit tegyünk, ha kategorikus a magyarázó vagy az eredmény változó?

Matematikai Statisztika
2024. október 21.



Diszkrét és kategorikus magyarázó változók

Kategorikus magyarázó változók kezelése

A lineáris regresszió tipikusan folytonos magyarázó változókkal működik, de sok esetben a magyarázó változó (X) kategorikus (diszkrét), például nemek, színek, jövedelmi kategóriák. Ha a magyarázó változó nem folytonos, akkor speciális módszereket kell alkalmaznunk, például:
Dummy változók bevezetése:

- Ha egy magyarázó változó két kategóriából áll, például bináris (0 vagy 1) – pl. férfi/nő vagy igaz/hamis –, akkor ezt a változót egyszerűen dummy változóként ábrázolhatjuk.
- A 0/1 értékek az egyes kategóriákat reprezentálják. Például, ha a változó a nemre vonatkozik, akkor:

$$\text{Gender} = \begin{cases} 1 & \text{ha nő} \\ 0 & \text{ha férfi} \end{cases}$$

- A lineáris modell ezt a változót úgy fogja kezelni, mint egy magyarázó változót, amely binárisan jelzi a kategóriákat.

Példa: Ha a magyarázó változó az "éves jövedelem kategória" (alacsony, közepes, magas), akkor minden kategóriát külön dummy változóval reprezentálhatunk: $D_1 = 1$ (alacsony), $D_2 = 1$ (közepes), és a referencia kategória lesz a magas jövedelem.

Többkategorikus magyarázó változók

Többkategorikus változók kezelése

Ha a magyarázó változó több kategóriából áll, például színek (piros, kék, zöld), vagy több jövedelmi kategória, akkor minden kategóriát külön dummy változóval ábrázolhatunk:

- Ha például három szín van: piros, kék és zöld, akkor a következőképpen lehet őket ábrázolni:

$$D_{\text{piros}} = \begin{cases} 1 & \text{ha piros} \\ 0 & \text{egyébként} \end{cases}, \quad D_{\text{kék}} = \begin{cases} 1 & \text{ha kék} \\ 0 & \text{egyébként} \end{cases}$$

- A zöld színt nem kell külön változóként bevezetni, mert az lesz az úgynevezett **referencia kategória**, amelyhez viszonyítjuk a többi kategória hatását.

Referencia kategória kiválasztása:

- Mindig ki kell választani egy referencia kategóriát, amelyet 0-val reprezentálunk, és ehhez viszonyítjuk a többi kategóriát.
- Ha a piros színre kapott paraméter pozitív, akkor azt mondhatjuk, hogy a piros szín pozitívan járul hozzá a függő változóhoz a zöld színhez képest.

Példa: A jövedelmi kategóriák esetén az alacsony és közepes jövedelmi kategóriák dummy változói mutatják, hogy azok hogyan hatnak a függő változóra a magas jövedelemhez képest (amely a referencia kategória).

Mi a logisztikus regresszió?

A logisztikus regresszió egy statisztikai módszer, amelyet akkor használunk, amikor a függő változó bináris vagy diszkrét kategóriákba esik. A logisztikus regresszió modellezni próbálja annak valószínűségét, hogy egy adott esemény bekövetkezik (pl. siker vagy kudarc, igen vagy nem), egy vagy több magyarázó változó függvényében.

Logisztikus modell

A logisztikus regresszió alapvető egyenlete a következőképpen néz ki:

$$P(Y=1) = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

ahol $P(Y=1)$ az esemény bekövetkezésének valószínűsége, α az intercept, β pedig a magyarázó változók együtthatója. A modell egy S-alakú görbét ad, amely a valószínűségeket 0 és 1 között tartja.

Alkalmazási területek

A logisztikus regressziót széles körben alkalmazzák különböző tudományterületeken, ahol a függő változó bináris:

- Orvosi kutatásokban annak valószínűségének modellezésére, hogy egy beteg meggyógyul-e egy adott kezeléstől.
- Pénzügyi elemzések során annak becslésére, hogy egy ügyfél csődbe jut-e vagy sem.
- Marketingben a vásárlók viselkedésének elemzése során, például annak becslésére, hogy egy vásárló megveszi-e egy terméket vagy sem.
- Gépi tanulásban az osztályozási problémákban, ahol két kategória közötti döntést hozunk.

A logisztikus modell előnyei és hátrányai

Előnyök

- **Könnyen értelmezhető**: A logisztikus regresszió egy jól ismert módszer, ahol az együtthatók log-odds változásként értelmezhetők.
- **Valószínűségi értelmezés**: A modell valószínűségeket ad, amelyek intuitívan értelmezhetők.
- **Nem igényli a lineáris kapcsolatot**: A logisztikus regresszió nem feltételezi, hogy a függő és a magyarázó változó között lineáris kapcsolat van, mint a lineáris regressziónál.

Hátrányok

- **Többkategorikus változók kezelése**: A logisztikus regresszió alapvetően bináris kimenetekre alkalmas. Többkategorikus kimenetek esetén más technikákat kell alkalmazni (pl. multinomiális logisztikus regresszió).
- **Outlierek érzékenysége**: A logisztikus regresszió érzékeny a kiugró adatokra, amelyek torzíthatják a modell eredményeit.
- **Nem működik jól kis minták esetén**: A logisztikus regresszió eredményei megbízhatatlanok lehetnek, ha a minta mérete túl kicsi.

Logisztikus regresszió matematikai alapjai

A log-odds

A logisztikus regresszió alapja a log-odds, amely a logaritmusa annak, hogy az esemény bekövetkezik vagy nem történik meg. Matematikailag a log-odds a következőképpen néz ki:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta X$$

A log-odds lineáris kapcsolatot feltételez a magyarázó változó (X) és az esemény bekövetkezésének valószínűsége között.

Maximum Likelihood becslés

A logisztikus regresszió paramétereit általában a maximum likelihood módszerrel becsüljük meg. A cél a modell paramétereinek olyan értékeinek megtalálása, amelyek maximalizálják annak valószínűségét, hogy a megfigyelt adatokat a modell generálta.

Miért használunk logisztikus regressziót?

A logisztikus regressziót bináris kimenetek előrejelzésére használjuk. Olyan helyzetekben alkalmazható, ahol a függő változó két kimenettel rendelkezik (pl. igen/nem, siker/kudarc). A függő változó (Y) értéke 1 vagy 0 lehet.

A logisztikus regresszió egyenlete

A logisztikus regresszió modellezésénél az alábbi egyenletet használjuk:

$$P(Y = 1) = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

ahol $P(Y = 1)$ annak valószínűsége, hogy a függő változó értéke 1, α a modell konstans tagja, β pedig a magyarázó változó (X) regressziós együtthatója.

Logisztikus regresszió - Példa

Példa: Választási eredmények előrejelzése

Vizsgáljuk egy politikai kampány során a szavazók döntését (Y), hogy egy adott jelöltre szavaznak-e ($Y = 1$), a kampányra fordított reklámkiadások (X) függvényében. A feladatunk az, hogy előre jelezzük, egy szavazó adott jelöltre fog-e szavazni, reklámkiadások alapján.

Adatok

Az alábbi táblázat tartalmazza a kampány során megfigyelt reklámkiadások (X) és a szavazók döntései (Y) közötti kapcsolatot:

Reklámköltség (X)	**Szavazó döntése (Y)**
100	1
150	0
200	1
250	1
300	0

Modell illesztése a megfigyelt adatokra

A logisztikus regresszió modellje

Az általános logisztikus regresszió modell formája:

$$P(Y = 1) = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

Az α és β paraméterek becslése történik maximum likelihood módszerrel. Az illesztés során a paraméterek optimális értékeit kapjuk meg, amelyekkel a modell a lehető legjobban leírja a megfigyelt adatokat.

Illesztett paraméterek

Az adatok alapján az illesztett paraméterek az alábbiak:

$$\hat{\alpha} = -4.5 \quad \text{és} \quad \hat{\beta} = 0.03$$

Ez azt jelenti, hogy a logisztikus regresszió modellünk az alábbi alakot veszi fel:

$$P(Y = 1) = \frac{1}{1 + e^{-(-4.5+0.03X)}}$$

Előrejelzés egy új szavazó döntésére

Cél:

Egy új szavazó esetén, aki 200 egységnyi reklámhatásnak volt kitéve, szeretnénk megjósolni, hogy adott jelöltre szavaz-e.

Előrejelzési lépés

Az illesztett modell alapján a reklámköltség ($X = 200$) behelyettesítésével kiszámítjuk a valószínűséget:

$$P(Y = 1 \mid X = 200) = \frac{1}{1 + e^{-(-4.5+0.03\cdot 200)}} = \frac{1}{1 + e^{-1.5}}$$

$$P(Y = 1 \mid X = 200) = \frac{1}{1 + 0.223} \approx 0.82$$

Tehát annak valószínűsége, hogy az új szavazó adott jelöltre fog szavazni, 82

Összegzés

Összefoglalás

A logisztikus regresszió egy hasznos eszköz bináris kimenetek modellezésére, ahol a függő változó valószínűségi értéket vesz fel 0 és 1 között. Számos alkalmazási területe van az orvostudománytól kezdve a pénzügyeken át a marketingig, és jól működik olyan helyzetekben, ahol a lineáris regresszió nem használható.