Többszörös lineáris regresszió
Outliers, overfitting és egyéb dolgok

Matematikai statisztika
2024. október 28.

Outlierek – Hatásuk a regresszióra

Definíció: Outlier

Az **outlier** egy olyan adatpont, amely jelentősen eltér a többi adatponttól. Az outlierek felismerése fontos, mert jelentős hatást gyakorolhatnak a regressziós modell eredményeire és az illeszkedés minőségére.

Outlierek típusai

- **Távoli outlierek**: Adatok, amelyek messze helyezkednek el az átlagos értékektől, befolyásolva az átlagot és a szórást.
- **Befolyásoló pontok**: Adatok, amelyek jelentősen hatnak a regressziós együtthatók becsléseire, így a modell eredményére.

Példa

Outlierek azonosítása

Hat mátrix használata outlierekhez

A hat mátrix diagonális elemeit (h_{ii}) használhatjuk annak mérésére, hogy egy adatpont mennyire befolyásolja a regressziós becslést:

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}X^T$$

- Átlagos hatás: $h_{ii} \approx \frac{k+1}{n}$
- Jelentős hatás: $h_{ii} > 2\frac{k+1}{n}$
- Outlier: $h_{ii} - \frac{1}{n} \geq 0.5$

További azonosítási módszerek

- **Box plot**: Az interkvartilis távolság segítségével kimutathatók a szélsőséges értékek.
- **Grubbs-teszt**: A legszélsőségesebb outlier azonosítására szolgál.
- **Cook-távolság**: Megmutatja, mely adatpontok befolyásolják leginkább az

23

Grubbs-teszt

A Grubbs-teszt egy statisztikai teszt, amely az adatsor legszélsőségesebb értékének outlierként való azonosítására szolgál. A teszt azt vizsgálja, hogy a legnagyobb vagy legkisebb adatpont szignifikánsan eltér-e a minta többi elemétől.

- **Alkalmazás**: A Grubbs-teszt különösen hasznos, ha egyetlen kiugró értéket szeretnénk azonosítani egy normál eloszlású adatsorban.
- **Számítási mód**: A teszt a legnagyobb abszolút értékű különbséget számolja a mintaátlaghoz képest:

$$G = \frac{\max |x_i - \bar{x}|}{s},$$

ahol $\bar{x}$ az adatok átlaga, s pedig a minta szórása.

Döntési folyamat

A Grubbs-teszt döntési folyamata:

1. **Nullhipotézis megfogalmazása** (H_0): A legszélsőségesebb érték nem kiugró, vagyis nincs szignifikáns eltérés a többi adattól.
2. **Alternatív hipotézis** (H_1): A legszélsőségesebb érték kiugró, és szignifikánsan eltér a többi adattól.
3. **Szignifikanciaszint** (α) kiválasztása: Általában 5
4. **Kritikus érték meghatározása**: A kritikus G_{kritikus} értéket a Grubbs-teszt táblázataiból vagy számítógépes szoftver segítségével határozzuk meg.
5. **Összehasonlítás és döntés**: Ha $G > G_{\text{kritikus}}$, akkor elutasítjuk a nullhipotézist, és az értéket kiugrónak tekintjük. Ha $G \leq G_{\text{kritikus}}$, akkor a nullhipotézist elfogadjuk, és az érték nem kiugró.

Grubbs-teszt előnyei és korlátai

A Grubbs-teszt erős módszer egyetlen outlier azonosítására, azonban nem alkalmas több

Cook-távolság

A Cook-távolság egy diagnosztikai eszköz, amely azt méri, hogy egy-egy adatpont milyen mértékben befolyásolja a regressziós modell illeszkedését. Ezzel a módszerrel megállapítható, mely adatpontok torzíthatják leginkább a modell eredményeit.

- **Számítási mód**: A Cook-távolság az i-edik adatpont eltávolításának hatását méri a modellre:

$$D_i = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_{j(-i)})^2}{p \cdot MSE},$$

ahol $\hat{y}_{j(-i)}$ a modell jóslata az i-edik adatpont nélkül, p a paraméterek száma, és MSE az átlagos négyzetes hiba.

További azonosítási módszerek - Cook-távolság döntési folyamata

Döntési folyamat

A Cook-távolság alapján a döntési folyamat a következő:

1. **Küszöbérték meghatározása**: Általánosan elfogadott küszöbérték Cook-távolság $> 4/n$, ahol n az adatpontok száma.
2. **Adatpontok elemzése**: Minden adatpont Cook-távolságát kiszámoljuk, és összevetjük a küszöbértékkel.
3. **Döntés**:
   - Ha $D_i > 4/n$, akkor az i-edik adatpont jelentős befolyást gyakorol a modellre, és külön figyelmet igényel.
   - Ha $D_i \leq 4/n$, akkor az i-edik adatpont nem gyakorol jelentős hatást a modell illeszkedésére.

Cook-távolság előnyei

A Cook-távolság lehetővé teszi, hogy azonosítsuk a modellre legnagyobb befolyással bíró adatokat, és segít felismerni azokat a pontokat, amelyek esetlegesen torzíthatják a modell illeszkedését.

Outlierek kezelése

Outlierek hatása és kezelési lehetőségek

Az outlierek kezelésekor el kell dönteni, hogy az adatpont valóban hibás vagy extrém, de érvényes érték. A kezelési lehetőségek közé tartozik:

- **Eltávolítás**: Ha az outlier hibás adat, az eltávolítása szükséges lehet.
- **Transzformációk**: Logaritmus vagy más nemlineáris transzformáció csökkentheti az outlierek hatását, különösen hosszú farkú eloszlások esetén.
- **Súlyozott regresszió**: Jelentős hatású outlierek kisebb súlyt kaphatnak a modellben.

Outlierek figyelmen kívül hagyásának kockázatai

Az outlierek elhanyagolása torzíthatja a modell eredményeit és ronthatja az általánosíthatóságot. Az outlierek forrásának vizsgálata fontos ahhoz, hogy megállapítsuk, valós jelenségeket vagy hibás adatokat tükröznek.

Overfitting jelensége

Mi az Overfitting?

Az **overfitting** akkor fordul elő, amikor a modell túlzottan illeszkedik a tanulási adathalmazra, de csökkent az általánosíthatósága, így nem teljesít jól új, ismeretlen adatokon.

Miért probléma az Overfitting?

Az overfitting miatt a modell érzékennyé válik az adathalmazban lévő véletlenszerű zajokra és mintázatokra, amelyek nem tükrözik az alapvető összefüggéseket.

Overfitting felismerése - Tanulási és teszthiba összehasonlítása

Tanulási és teszthiba fogalma

Az overfitting felismerésének egyik alapvető módszere a tanulási (train) és a teszthiba (test error) összehasonlítása. Ezt gyakran az **RMSE** (gyök négyzetes középérték) mutatóval mérjük.

- **RMSE a tanulási adathalmazon**: Az RMSE (Root Mean Squared Error) kiszámításával megmérjük a modell illeszkedésének hibáját a tanulási adatokon.

$$\mathrm{RMSE}_{\mathrm{train}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **RMSE a teszt adathalmazon**: Ugyanezt a hibamértéket a teszt adathalmazon is kiszámítjuk, hogy megvizsgáljuk, hogyan teljesít a modell új adatokon.

$$\mathrm{RMSE}_{\mathrm{test}} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{y}_j)^2}$$

Az Overfitting elkerülése

Rendszeresítés (Regularizáció)

- **Ridge regresszió**: Büntetési tagot adunk a modellhez, ami csökkenti az együtthatók nagyságát, mérsékelve a modell komplexitását.
- **Lasso regresszió**: A Ridge regresszióhoz hasonlóan működik, de az abszolút értékek minimalizálásával akár nullára is csökkentheti bizonyos együtthatókat, egyszerűsítve a modellt.

Keresztvalidáció

A modell kiértékelése különböző adathalmazokon történik. A keresztvalidáció során a modellt új részhalmazokon teszteljük, hogy lássuk, hogyan teljesít idegen adatokon.

Példa - Keresztvalidáció

AIC – Akaike Information Criterion alapjai

AIC célja

Az AIC a modellek összehasonlítását segíti az illeszkedés és a komplexitás figyelembevételével. Célja a legjobb előrejelzési képességgel rendelkező, de nem túlzottan bonyolult modell kiválasztása.

AIC képlete

Az AIC-t a következő képlet határozza meg:

$$\mathrm{AIC} = 2k - 2\ln(L),$$

ahol:

- k: a modell paramétereinek száma,
- L: a modell maximális likelihood értéke.

AIC – Értelmezés és használat

AIC interpretációja

Az AIC értéke minél kisebb, annál jobb a modell illeszkedése az adatokra, figyelembe véve a komplexitást is. Az AIC összehasonlítható különböző modellek között, de csak azonos adathalmazra illesztett modellek esetén.

Miért fontos a modell komplexitása?

A túlzottan komplex modell jól illeszkedik a tanulási adathalmazra, de rosszul teljesíthet új adatokon (overfitting). Az AIC bünteti a túl sok paramétert, így elősegíti a generalizálhatóbb modell kiválasztását.

AIC – Példa alkalmazás

Példa - Két modell összehasonlítása

Tegyük fel, hogy van két modellünk:

- **Modell A**: Egyszerűbb modell kevesebb paraméterrel,
- **Modell B**: Összetettebb modell több paraméterrel.

AIC számítása mindkét modell esetén

Mindkét modellre kiszámítjuk az AIC-t, és összehasonlítjuk az értékeket. Az alacsonyabb AIC-értékkel rendelkező modellt választjuk, mivel az nagyobb valószínűséggel generalizálható jobban.

BIC – Bayesian Information Criterion alapjai

BIC célja

A BIC is a modellek összehasonlítására szolgál, de nagyobb hangsúlyt fektet a modell komplexitásának korlátozására, mint az AIC. A BIC segítségével a modellek valószínűségi bizonyíték alapján értékelhetők.

BIC képlete

A BIC képlete:

$$\mathrm{BIC} = k \ln(n) - 2 \ln(L),$$

ahol:

- k: a modell paramétereinek száma,
- n: a megfigyelések száma,
- L: a modell maximális likelihood értéke.

BIC – Értelmezés és használat

BIC interpretációja

Az AIC-hez hasonlóan a BIC kisebb értéke jobb illeszkedést jelez. Ugyanakkor a BIC szigorúbb a modell komplexitásával, ezért nagyobb adathalmaz esetén a BIC nagyobb büntetést ad a paraméterek számának növekedésekor.

A BIC használata előnyben részesíti az egyszerűbb modelleket

Mivel a BIC büntetése a minta méretétől függ, nagyobb mintaszám esetén a BIC hajlamos az egyszerűbb modelleket választani. Ez azzal jár, hogy kisebb modellekkel is megfelelő predikciót biztosít.

BIC – Példa alkalmazás

Példa - Több modell közül a legjobb kiválasztása

Tegyük fel, hogy három modellünk van, különböző paraméterszámmal. Mindegyik modellre kiszámítjuk a BIC értéket, és az alacsonyabb BIC értékkel rendelkezőt választjuk.

BIC és minta méretének kapcsolata

Nagyobb minták esetén a BIC erősebben bünteti a túlkomplikált modelleket, míg kisebb minták esetén kevésbé szigorú. Így a BIC inkább a nagy mintaszámú elemzések során hasznos.

AIC és BIC összehasonlítása

Különbségek az AIC és BIC között

- **AIC**: A modell illeszkedésére és a paraméterek számára figyel, célja az előrejelzési pontosság maximalizálása.
- **BIC**: Nagyobb büntetést alkalmaz a paraméterekre, célja a valószínűségi alapú modell választása, különösen nagy minták esetén.

AIC és BIC – Előnyök és hátrányok

AIC előnyei és hátrányai

- **Előny**: Hatékony az előrejelzési pontosság maximalizálásában.
- **Hátrány**: Hajlamos lehet túlzottan bonyolult modelleket választani, mivel kevésbé bünteti a paraméterszámot.

BIC előnyei és hátrányai

- **Előny**: Nagy minták esetén szigorúbban szabályozza a modell komplexitását.
- **Hátrány**: Hajlamos az egyszerűbb modellek választására, ami kisebb minták esetén túlzott egyszerűsítést okozhat.

AIC és BIC – Használati javaslatok

Mikor használjuk az AIC-t?

Az AIC javasolt, ha a cél az előrejelzés pontosságának maximalizálása, és a modell összetettsége kevésbé fontos. Általában kisebb mintaszám esetén vagy prediktív modellekhez ideális.

Mikor használjuk a BIC-t?

A BIC előnyös, ha a modell komplexitásának csökkentése a cél, például nagy mintaszám esetén, ahol a valószínűségi alapú modellkiválasztás relevánsabb.

Kombinált használat

AIC és BIC együttes használata segíthet a megfelelő modell kiválasztásában, különösen akkor, ha több modell is szóba jöhet az adathalmaz alapján.

Mit jelent az interakciós hatás?

Az interakciós hatás akkor fordul elő, amikor két vagy több független változó közötti kölcsönhatás befolyásolja a célváltozót. Ilyen esetekben egy változó hatása a célváltozóra függhet egy másik változó jelenlététől vagy értékétől.

- **Alapgondolat**: Az interakció hatása azt jelenti, hogy a független változók nemcsak önmagukban, hanem együtt is befolyásolják a célváltozót.
- **Matematikai kifejezés**: Az interakciós hatásokat az alábbi módon lehet beépíteni a regressziós modellbe:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \varepsilon,$$

ahol a β_3 együttható azt méri, hogy a X_1 és X_2 közötti interakció milyen hatással van a Y célváltozóra.

Interakciós hatások vizsgálatának fontossága

Az interakciók felismerése és modellezése javíthatja a regressziós modell pontosságát, mivel figyelembe veszi, hogyan változnak a független változók hatásai más változók

Példa interakciós hatásra - Életkor és munkaórák hatása a stressz szintjére

Példa felállítása

Tegyük fel, hogy egy vállalatnál vizsgáljuk a dolgozók stressz szintjét (Y), amelyet befolyásolhat a dolgozó **életkora** (X_1) és a **munkaórák száma** (X_2).

- **Alaphatások**: Feltételezhetjük, hogy a stressz szintje önmagában növekszik a munkaórák számával (X_2) és csökken az életkorral (X_1), mivel az idősebb dolgozók gyakran jobban kezelik a stresszt.
- **Interakciós hatás**: Az életkor és a munkaórák kölcsönhatása is befolyásolhatja a stressz szintjét. Például a fiatalabb dolgozóknál a hosszabb munkaórák nagyobb stresszt okozhatnak, mint az idősebb dolgozóknál.

Interakciós modell felírása

Az interakciós hatást figyelembe vevő modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \varepsilon,$$

ahol a β_3 együttható mérése jelzi, hogy az életkor és munkaórák kölcsönhatása miként

Interakciós hatások vizsgálata és értelmezése - Eredmények

Az interakciós együttható értelmezése

A regressziós modell futtatása után megkapjuk az interakciós együtthatót (β_3), amely segít eldönteni, hogy a független változók kölcsönhatása jelentős hatással van-e a célváltozóra.

- **Pozitív interakciós hatás** ($\beta_3 > 0$): Az életkor növelheti a munkaórák hatását a stresszre. Például, ha a stressz növekedése munkaóránként nagyobb az idősebb dolgozóknál, ez pozitív interakciós hatásra utalhat.
- **Negatív interakciós hatás** ($\beta_3 < 0$): Az életkor csökkenti a munkaórák hatását a stresszre. Például, ha a fiatalabb dolgozóknál a munkaórák jobban növelik a stresszt, akkor a negatív β_3 azt jelzi, hogy az életkor mérsékli ezt a hatást.

Az interakciós hatás jelentőségének ellenőrzése

A modellben az interakciós hatás szignifikanciáját statisztikai teszttel is ellenőrizhetjük. Ha az interakciós együttható szignifikáns, akkor az interakció valóban befolyásolja a célváltozót.