A lineáris regresszió

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

Tartalom

A linearitás és jelentősége

- Ha a háttéreloszlás normális, akkor $\mathbb{E}\left(Y \mid \underline{X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ és így $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$
- A továbbiakban általában is ebben az ún. **lineáris modellben** fogunk gondolkozni, függetlenül attól, hogy mit tudunk a háttéreloszlásról, ugyanis:
    1. Többváltozós normalitásnál *egzaktan* ez a helyzet
    2. Más esetekben csak *közelítés*, de cserében nagyon kellemesek a tulajdonságai, különösen ami az interpretációt illeti
    3. Ráadásul az is elmondható, hogy – a Taylor-sorfejtés logikáját követve – bármi más is a jó függvényforma, legalábbis lokálisan ez is jó közelítés kell legyen
    4. Végezetül pedig: majd látni fogjuk, hogy szerencsés módon egy sor nemlineáris kiterjesztés is könnyen kezelhető ugyanebben a keretben
- Azt fogjuk mondani, hogy ezt a modell *feltételezzük* a sokaságra (hogy aztán ezt jól tettük-e, azt majd különböző szempontokból persze vizsgáljuk)

A lineáris regressziós modell

- A modellünk tehát:
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$
- Egyelőre még semmilyen feltételt nem kötöttünk ki, bár annyit már láttunk, hogy ha ez jó modell, akkor $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$ igaz kell legyen (ez persze csak szükséges feltétel, arról még semmit nem tudunk, hogy elégséges-e) – erre a kérdésre később térünk vissza

A modellünk használata: előrejelzés

- Teljesen kézenfekvő, csak egy dolgot kell megbeszélni: előrejelzésnél ε helyébe 0-t írunk
- (Hiszen a feltételes várható értékre lövünk)
- Azaz
$$\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

A modellünk használata: elemzés

A paraméterek értelmezésével elemezhetjük a modellünket; kérdéseket válaszolhatunk meg a modellezett jelenségről.

A modellünk használata: elemzés (tengelymetszet)

- A β_0 konstans értelmezése: ha valamennyi magyarázó változó nulla értékű, akkor modellünk szerint várhatóan mekkora az eredményváltozó
- Ha a minden magyarázó változó nulla kombináció kívül esik az értelmes tartományon, akkor ennek lehet, hogy nincs tárgyi értelme (ilyenkor: egyszerűen az illeszkedést javító paraméter)

A modellünk használata: elemzés (meredekség)

- A meredekségek egyszerű értelmezése: ha a vizsgált magyarázó változó egy egységnyivel nagyobb lenne úgy, hogy minden más változót rögzített értéken tartunk (ceteris paribus, röviden c. p.), akkor modellünk szerint várhatóan hány egységnyit változna az eredményváltozó
- Hiszen:

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_l \left(X_l + 1\right) + \ldots + \beta_k X_k =$$
$$= \left(\beta_0 + \beta_1 X_1 + \ldots + \beta_l X_l + \ldots + \beta_k X_k\right) + \beta_l$$

- Figyelem:
  - Ceteris paribus
  - Mindegyik változót a saját egységében mérve
  - Abszolút változásokat kapcsol össze
- Később precízebben is értelmezzük a meredekséget

Kauzalitás és a regressziós modellek

- Két dolgot már részletesen láttunk: a kauzalitás kutatásának problémáját, ha csak megfigyeléses adataink vannak, és a regressziós modellek alapjait
- Na de mi köze a kettőnek egymáshoz?
- Azonnal világossá válik, ha az elemzésnél látottakra gondolunk – ceteris paribus!
- A β_l együttható úgy értendő, hogy az X_l növekedésének hatása... ha *minden más változatlan marad*!
- Ez *épp* a confounding kiszűrése, hiszen ott pont az a probléma, hogy ha X_l nő, akkor vele együtt más is változik!
- Voilá – megoldottuk a problémát

Visszatérve a példákra

- Az oktatás β-ja a magasabb iskolai végzettség hatása, *miközben* minden mást (így a nem oktatással összefüggő munkaalkalmasságot is!) rögzítetten tartottuk – azaz kiszűrtük a confound-oló hatását...
- Az előadáslátogatás β-ja a több előadáslátogatás hatása, *miközben* minden mást (így a motivációt is!) rögzítetten tartottuk – azaz kiszűrtük a confound-oló hatását...
- és így tovább, és így tovább... (érdemes végiggondolni a többi példára is!)

Limitációk

- Az előbbi kijelentés persze valójában túl optimista volt
- A legfontosabb probléma: valójában nem tudunk „minden másra" kontrollálni – csak amit beleraktunk a modellbe!
- De mi van, ha valamit nem tudunk jól lemérni? Még jobb: mi van, ha valamiről eszünkbe sem jut, hogy confounder? (Ez a kísérlet hatalmas előnye!)
- Másrészt a regressziós modelleknek vannak előfeltevéseik (részletesen fogunk vele foglalkozni), melyeknek teljesülniük kell, hogy valós eredményt kapjunk
- Csak a példa kedvéért: a lineáris specifikáció kényelmes, de cserében kiad dolgokat a modell változóira nézve

A lineáris specifikáció hatása

- Eddigi definíció a meredekségre: a többi változót rögzítjük, a vizsgált egy egységgel nagyobb... de: milyen szinten rögzítjük a többit? milyen szintről indulva nő egy egységgel a vizsgált?
- A linearitás fontos következménye, hogy *mindkettő mindegy*!
  - Mindegy milyen szinten rögzítjük a többi változót...
  - Mindegy milyen szintről indulva növeljük eggyel a vizsgált változót...
- ...mindenképp *ugyanannyi* lesz a növelés hatása az eredményváltozóra!
- Szemléletes tartalom: gondoljunk az egyenesre (illetve síkra)

A modellünk használata: elemzés (rugalmasság, elaszticitás)

- A meredekséghez hasonló mutatót szeretnénk, de úgy, hogy ne abszolút, hanem relatív változásokat kössön össze
- Tehát: ha a vizsgált magyarázó változó 1 %-nyival nagyobb lenne c. p., akkor modellünk szerint várhatóan hány %-nyit változna az eredményváltozó
- Számítás:

$$\mathrm{El}_l\left(\underline{X}\right) = \frac{\beta_l / Y}{1/X_l} = \beta_l \cdot \frac{X_l}{Y} = \beta_l \cdot \frac{X_l}{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}$$

- Figyelem:
  - Ceteris paribus
  - Minden elmozdulást relatíve (%-osan) mérve
- Ami új: az érték függ attól, hogy milyen pontban vagyunk, tehát, hogy az *összes* magyarázó változó milyen értékű (ezt tükrözi a jelölés is); teljesen logikus módon