Regresszió a sokaságban: a feladat megfogalmazása, megoldása és modellminősítés

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 9.

Tartalom

Jelölésrendszer

- Az eddigi példákból is látható, hogy van egy változó, aminek az alakulását le kívánjuk írni, amit modellezni akarunk, ennek neve **eredményváltozó** (vagy függő változó, angolul response), jele Y
- És vannak változók, amikkel le akarjuk az eredményváltozót írni, amikkel modellezünk, ezek nevei **magyarázó változók** (vagy független változók, angolul predictor), jelük X_i $(i = 1, 2, \ldots, k)$
- Az eredményváltozó a vizsgált kimenet, a magyarázó változók az azt – potenciálisan – befolyásoló tényezők (tehát a fontos, vizsgált változók és a – potenciális – confounderek egyaránt)

Kitérő: szimultaneitás

- Látszik, hogy eredményváltozóból csak egyet engedünk meg
- Ha több lenne, akkor legfeljebb külön-külön foglalkozunk mindegyikkel – mondhatjuk első ránézésre
- Ez nem igaz azonban akkor, ha változók *kölcsönösen* hatnak egymásra
- Például nem csak a rendőri erők létszáma hat a bűnözésre (jó esetben...), hanem fordítva is, hiszen a múltbeli bűnözési adatok számítanak a rendőri vezetésnek akkor, amikor határoz a rendőri erők telepítéséről
- Ez a **szimultaneitás** problémája
- Most nem foglalkozunk vele (többegyenletes ökonometria, szimultán modellek fedőnevek alatt lehet vele találkozni)

Útban a regressziós modellek felé

- Az X-ek hatnak az Y-ra... ezt kellene megragadni matematikailag!
- De hát erre ismerünk egy jó matematikai objektumot, ami pont ezt írja le:

$$Y = f\left(X_1, X_2, \ldots, X_k\right)$$

- A későbbiekben erre azt fogjuk mondani, hogy ez egy statisztikai modell
- Nehéz lenne vitatkozni ennek az általánosságával, csak épp...

Sztochasztikusság

- A fő probléma, hogy a modell azt feltételezi, hogy az Y és az X-ek kapcsolata *determinisztikus*
- Szinte teljesen mindegy is, hogy mi az Y és mik az X-ek, hogy mi a vizsgált probléma, a társadalmi-gazdasági jelenségek vizsgálata kapcsán lényegében általánosan kijelenthető, hogy ez irreális
- Egy középiskolai fizika-kísérletben ez lehet jó közelítés (megj.: igazából ott sem, mert vannak mérési hibák – legfeljebb elhanyagoljuk őket), de itt szinte kizárt, hogy *függvényszerű* módon meghatározzák a magyarázó változók az eredményváltozót
- A valódi modell **sztochasztikus** kell legyen:

$$Y = f\left(X_1, X_2, \ldots, X_k\right) + \varepsilon$$

- Rövid jelölésként az X-eket gyakran egy vektorba vonjuk össze: $Y = f\left(\underline{X}\right) + \varepsilon$
- Az ilyen f-et hívjuk (sokasági) **regressziófüggvénynek**
- ε neve: hiba

Sokaság és minta

- Ez az egyenlet egy *sokasági* modell: azt írja le, hogy a valóság hogyan működik
- Ezt persze mi nem tudhatjuk, majd mintából kell kitalálnunk (megbecsülnünk)
- Egyelőre ezzel ne törődjünk, és vizsgálódjunk tovább a sokaságban
- A nem-kísérleti jelleg miatt az az értelmes modell, ha mind az eredményváltozót, mind a magyarázó változókat – és így persze ε-t is – valószínűségi változónak vesszük, melyeknek eloszlása van (ezért használtunk eddig is nagy betűket!)

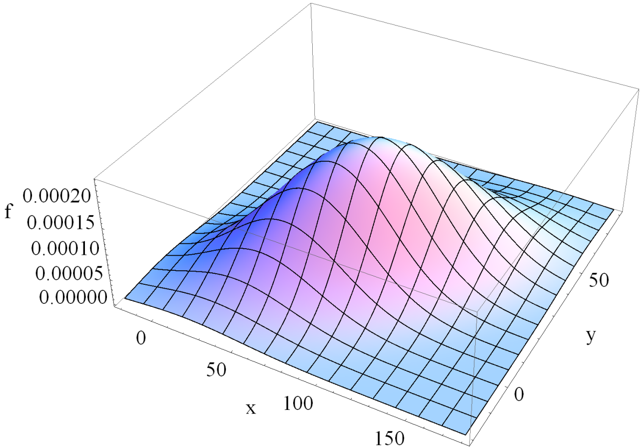
A sokaság leírása

- Most valszámos emberek leszünk: úgy vesszük mintha ismernénk a sokaságot
- (Valójában persze csak a mintán keresztül tudunk rá következtetni, de a valszámos nézőpont épp azt jelenti, hogy ezzel nem törődünk: úgy vesszük, hogy nálunk van a bölcsek köve, azaz valahonnan tudjuk, hogy mi „az” eloszlás, egyelőre nem törődve azzal, hogy ezt igazából honnan is tudhatjuk)
- Mit kell ismernünk? Nem egyszerűen Y és $X_1, X_2, \ldots, X_k$ eloszlásait (külön-külön), hanem az együttes eloszlásukat

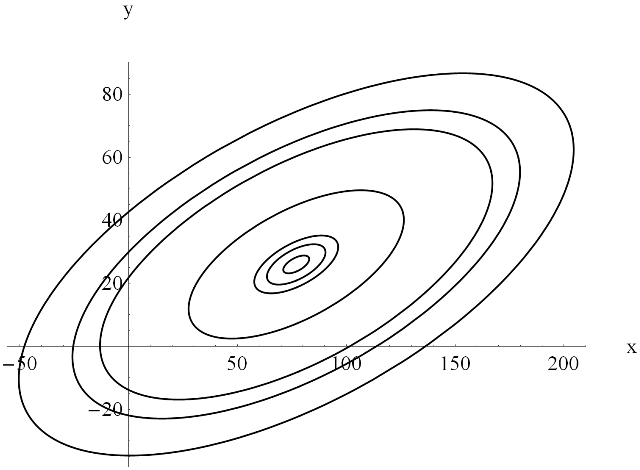
A sokaság értelme

- Ezt úgy kell elképzelnünk mint egy $k+1$ dimenziós teret: minden pont egy adott magyarázó- *és* eredményváltozó-kombináció (ami adott eloszlás szerint előállhat: van ami gyakrabban, van ami ritkábban)
- (Ha az X-eket rögzítjük, akkor egy olyan egydimenziós eloszlást kapunk, ahol a becsült érték mindenhol ugyanaz, miközben persze a – valódi – Y nem: épp ez a hiba oka)
- A tér minden pontjában valamekkora a hiba (becsült és tényleges különbsége), ennek persze az eloszlását épp az határozza meg, hogy milyen a $k+1$ dimenziós téren a sűrűségfüggvény: ha valahol kicsi, akkor az ottani hiba kis hozzájárulást fog adni az ε eloszlásához

Példa a sokaság valószínűségi leírására



Példa a sokaság valószínűségi leírására



Az optimális regressziófüggvény definiálása

- Mit nevezünk „legjobb” f-nek? Ehhez nyilván definiálni kell, hogy mit értünk jóság alatt...
- Természetes elvárás, hogy a tényleges érték (Y) és a modell szerinti érték $(f(X_1, X_2, \ldots, X_k)$, más szóval becsült vagy predikált érték) minél közelebb legyen egymáshoz, azaz, hogy ε kicsi legyen
- Az már döntés kérdése, hogy mit értünk „kicsi” alatt; tipikus választás:
    - mivel ε is egy val. változó, így a várható értékét vesszük (az már egyetlen szám, amit lehet minimalizálni)
    - és használjuk a négyzetét (hogy – egy matematikailag kényelmesen kezelhető függvénnyel – megszabaduljunk az előjelétől)
- A várható érték azért is fontos, mert jól kifejezi, hogy „ott kevésbé számít a hibázás, ami kevésbé gyakran fordul elő”

Az optimális regressziófüggvény meghatározása

- Így tehát a feladat:

$$\arg\min_f \mathbb{E}\left[Y - f\left(\underline{X}\right)\right]^2$$

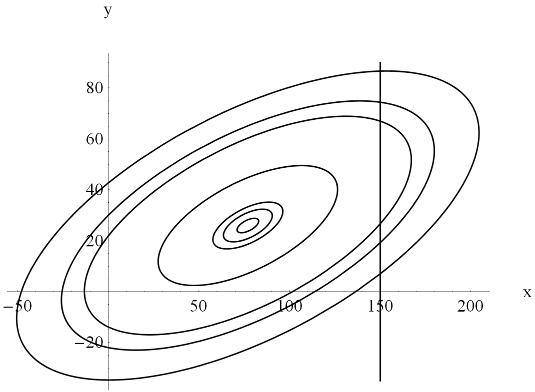
- Egészen abszurdan hangzik (az összes létező függvény körében keressünk optimumot?), de megoldható!
- A megoldás a feltételes várható érték:

$$f_{\text{opt}}\left(\mathbf{x}\right) = \mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$$

- Ez az eredmény *teljesen univerzális*, semmit nem tételeztünk fel f-ről!
- (Emlékeztetünk rá, hogy ha $\mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$ egy $f\left(\mathbf{x}\right)$ transzformációt definiál, akkor $\mathbb{E}\left(Y \mid \underline{X}\right)$ alatt $f\left(\underline{X}\right)$-et értjük – ez tehát egy valószínűségi változó)

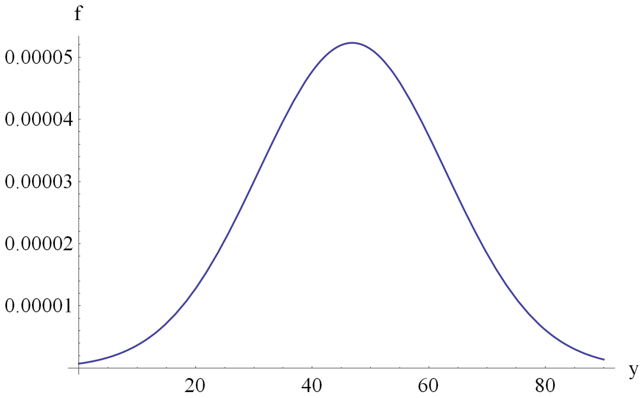
A feltételes várhatóérték – emlékeztető

Az együttes eloszlást „elmetsszük" a feltétel (például $x = 150$) pontjában:



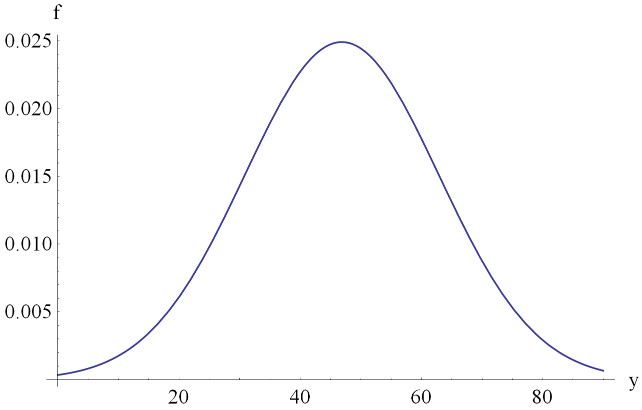
A feltételes várhatóérték – emlékeztető

Az így „kimetszett” eloszlás még nem eloszlás, mert nem 1-re normált...



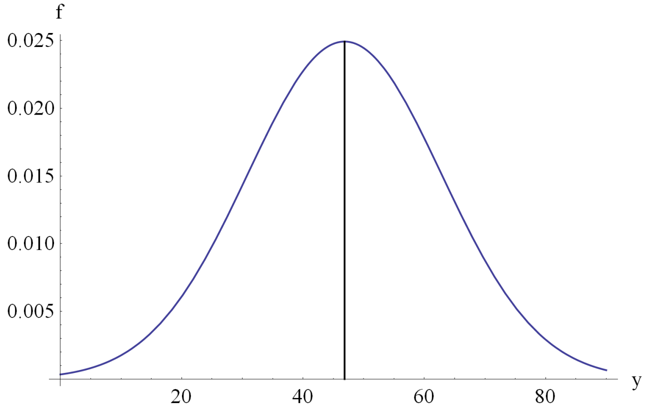
A feltételes várhatóérték – emlékeztető

... de osztva a tényleges integráljával (ami persze a peremeloszlás értéke a feltétel pontjában) kapjuk az igazi feltételes eloszlást:



A feltételes várhatóérték – emlékeztető

Ennek a várhatóértéke az adott feltétel melletti feltételes várhatóérték $\left(\mathbb{E}\left(Y \mid X = 150\right) = 46{,}9\right)$

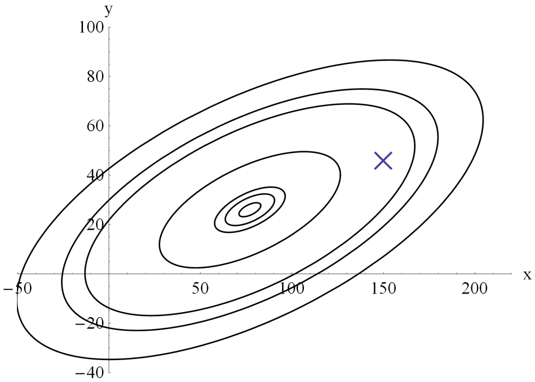


Optimális sokasági regresszió számítása

- Ez tehát – legalábbis elvileg – *pusztán* a sokasági eloszlás ismerete alapján kiszámítható, csak némi integrálást igényel
- Csakhogy: az integrál gyakorlati kiszámítása még egyszerű eloszlásokra sem feltétlenül egyszerű
- Egy nevezetes kivétel lesz, a többváltozós normális elszolás

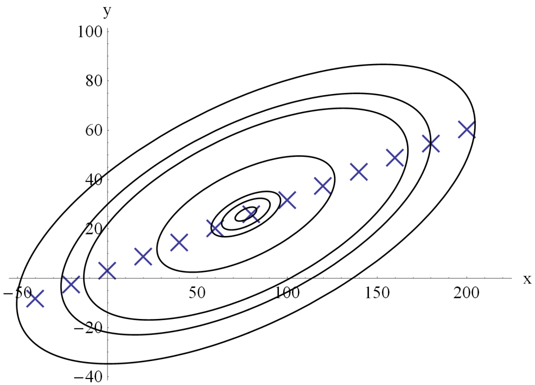
Optimális sokasági regresszió normális eloszlásnál

Az optimális becslés egy pontnál:



Optimális sokasági regresszió normális eloszlásnál

Számítsuk ki több pontra is:



Optimális sokasági regresszió normális eloszlásnál

Amit látunk, az nem véletlen:
Ha Y és $\underline{X}$ együttes eloszlása normális, akkor

$$\mathbb{E}\left(Y \mid \underline{X}\right) = \mathbb{E}Y + \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{X}\underline{X}}^{-1}\left(\underline{X} - \mathbb{E}\underline{X}\right).$$

Azaz írhatjuk, hogy

$$\mathbb{E}\left(Y \mid \underline{X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

ha bevezetjük a

$$\beta_0 = \mathbb{E}Y - \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{X}\underline{X}}^{-1}\mathbb{E}\underline{X}$$

és a

$$\begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_k\end{pmatrix}^T = \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{X}\underline{X}}^{-1}\underline{X}$$

jelöléseket.
Többváltozós normális eloszlásnál tehát speciálisan a regressziófüggvény lineáris lesz.

A hibaalak

- Általában is értelmes tehát a következő dekompozíció (a modell „error form”-ja):

$$Y = \mathbb{E}\left(Y \mid \underline{X}\right) + \varepsilon$$

- Y *mindig* felírható így! Csak majd $\mathbb{E}\left(Y \mid \underline{X}\right)$ helyébe írjuk be a mi konkért függvényformánkat, például azt, hogy $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
- Megjegyzés: amikor ilyet használunk, azaz a függvény struktúráját megadjuk, csak egy vagy több – valós szám – paramétert hagyunk ismeretlenül, akkor **paraméteres modellről** (paraméteres regresszióról) beszélünk
- Lehetne az $\mathbb{E}\left(Y \mid \underline{X}\right)$ anélkül próbálni közelíteni, hogy bármilyen konkrét függvényforma mellett elköteleződnék (nem-paraméteres modell), de ezekkel most nem fogunk foglalkozni

A hiba egy fontos tulajdonsága

- Az előbbiekből következik, hogy $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$ (hiszen $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = \mathbb{E}\left(Y - \mathbb{E}\left(Y \mid \underline{X}\right) \mid \underline{X}\right) = \mathbb{E}\left(Y \mid \underline{X}\right) - \mathbb{E}\left(Y \mid \underline{X}\right)$, a kétszeres várható érték-vétel nyilván ugyanaz, mint az egyszeres)
- Később fontos lesz, ha mindezt így fogalmazzuk meg: ha *tényleg* a jó $\mathbb{E}\left(Y \mid \underline{X}\right)$-t használjuk becslésre, *akkor* a hiba az előbbi tulajdonságú kell legyen

Modellminősítés

- Mivel $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$, így $\operatorname{cov}\left(\varepsilon, X_i\right) = 0$ és emiatt $\operatorname{cov}\left(\varepsilon, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right) = 0$ is
- Így igaz, hogy $\mathbb{D}^2 Y = \mathbb{D}^2\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right) + \mathbb{D}^2 \varepsilon$ (varianciafelbontás)

Magyarázott variancia szemlélet

- Képzeljük el, hogy látjuk az embereket, de csak a fizetésüket: az elsőnek 100 egység, a másodiknak 123, a harmadiknak 500, a negyediknek 83, és így tovább
- Nem értjük, hogy miért van ez a szóródás, ez a *variancia* ($\mathbb{D}^2 Y$)
- Megismerjük az oktatottságukat – ez *megmagyarázza* a variancia egy részét (pl. kiderül, hogy az elsőnek csak 8 általánosa van, de a másodiknak érettségije)
- Persze ez sem magyaráz mindent: lehet, hogy a negyediknek szintén 8 általánosa van, és mégsem keres 100 egységet
- Ha újabb magyarázó változókat ismerünk meg, akkor még tovább csökkenhet ez a meg nem magyarázott variancia ($\mathbb{D}^2 \varepsilon$)...

Az előbb látott felbontás tehát nem csak „statisztikai átalakítás", hanem kézzelfogható tartalom van mögötte!

Modellminősítés „magyarázott variancia hányad” elven

- Értelmes tehát azt mondani, hogy a $\mathbb{D}^2 Y$ varianciából $\mathbb{D}^2 \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right)$ az, amit „megmagyarázott” a modellünk, $\mathbb{D}^2 \varepsilon$ az, amit nem
- Ezért az
$$R^2 = \frac{\mathbb{D}^2 Y - \mathbb{D}^2 \varepsilon}{\mathbb{D}^2 Y} = 1 - \frac{\mathbb{D}^2 \varepsilon}{\mathbb{D}^2 Y}$$
a modell jóságának mutatója lesz ($0 \leq R^2 \leq 1$), a fenti „megmagyarázott variancia” értelemben, neve: **többszörös determinációs együttható**