Regressziós modellezés

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 6.

Tartalom

Tartalomjegyzék

1 Regresszió a sokaságban: a feladat megfogalmazása, megoldása és modellminősítés

1.1. Út a regressziós modellekhez

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)

Ha módunkban állna a városokba *véletlenszerű* mennyiségű rendőri állományt telepíteni, majd lemérni a bűnözési rátákat, akkor könnyen meg tudnánk határozni, hogy az előbbi hogyan hat az utóbbira. A valóságban ilyet nem tehetünk, hiszen ezt a rendőrség központilag határozza meg, ráadásul úgy – és most ez lesz a lényeg –, hogy az nem független a bűnözéstől: ahol magasabb, oda inkább vezényel több rendőrt. A kettő tehát *kölcsönösen* hat egymásra.

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:

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

- 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:

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

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

1.2. Regresszió a sokaságban

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!)

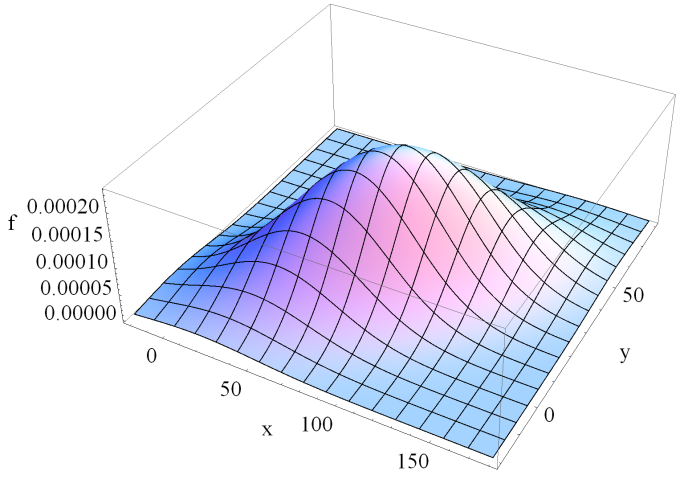
A sokaság leírása

- 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 (különkülön), 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

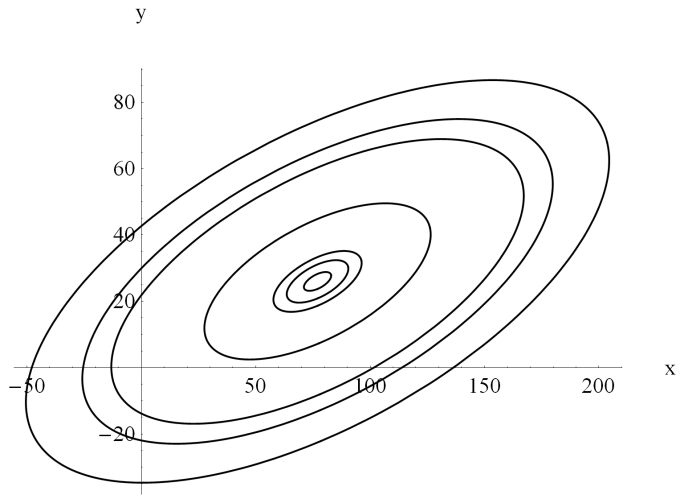
Példa a sokaság valószínűségi leírására



Ez egy kétváltozós eloszlás együttes sűrűségfüggvénye; itt az egyik változó játssza a magyarázó-, a másik az eredményváltozó szerepét. Mint sűrűségfüggvény, igaz rá, hogy tetszőleges terület felett kiszámolva a görbe alatti térfogatot (azaz kiintegrálva a függvényt), megkapjuk annak a valószínűségét, hogy a valószínűségi változó a kérdéses területre esik.

Eláruljuk, hogy a fenti eloszlás többváltozós normális (később ennek majd jelentősége lesz), $\boldsymbol{\mu} = \begin{pmatrix} 77 \\ 26 \end{pmatrix}$ várhatóérték-vektorral és $\mathbf{C} = \begin{pmatrix} 42^2 & 0{,}6 \cdot 20 \cdot 42 \\ 0{,}6 \cdot 20 \cdot 42 & 20^2 \end{pmatrix}$ kovariancia-mátrixszal.

Példa a sokaság valószínűségi leírására



Ez ugyanaz mint a fenti sűrűségfüggvény, de „szintvonalakkal” leírva (azaz különböző z magasságokban elmetszettük a sűrűségfüggvényt és a kapott metszeteket ábrázoltuk). Belátható, hogy többváltozós normális esetén ezek mindig ellipszisek. (Úgy, hogy az

ellipszis középpontját a várhatóérték-vektor adja meg, a tengelyek a kovariancia-mátrix sajátvektorainak irányába mutatnak, féltengelyeik hossza pedig a kovariancia-mátrix megfelelő sajátértékeivel arányos.) A fenti ábrát ráadásul úgy képeztük, hogy a metszetek adott valószínűségű területet határoljanak; a legnagyobb területű ellipszisről például az mondható el, hogy területére épp 95% valószínűség esik a fenti eloszlásból. Ez tehát lényegében a 0,95-ös „kvantilis-ellipszis". (A fentiek miatt az ilyen értelmű régiók többváltozós normális eloszlás esetén jól meghatározottak.) A fenti ábra ezt a 0,01, 0,05, 0,1, 0,5, 0,9, 0,95 és 0,99 valószínűségekhez tartozó ellipsziseket adja meg.

Ezt az ábrázolást szokás 'contour plot'-nak nevezni, előnye, hogy – a háromdimenziós érzékeltetéssel szemben – nem érzékeny a nézőpont megválasztására, részek nem takarnak ki másokat stb. (Ám cserében nyilván információ-vesztéssel jár, ami azzal arányos, hogy milyen sűrűn képezzük a metszeteket.)

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 $f(\mathbf{x})$ transzformációt definiál, akkor $\mathbb{E}\left(Y \mid \underline{X}\right)$ alatt $f(\underline{X})$-et értjük – ez tehát egy valószínűségi változó)

Bármilyen meglepő, de nem is olyan rettentő nehéz megoldani ezt az optimalizációs problémát. Legyen f_{opt} a feltételes várható érték, f pedig egy tetszőleges k-változós függvényt. Alakítsuk át a kritériumfüggvényt:

$$\begin{aligned}\mathbb{E}\left[Y - f(\underline{X})\right]^2 &= \mathbb{E}\left[Y - f_{\text{opt}}(\underline{X}) + f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]^2 = \\ &= \mathbb{E}\left[Y - f_{\text{opt}}(\underline{X})\right]^2 + \mathbb{E}\left\{\left[Y - f_{\text{opt}}(\underline{X})\right]\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]\right\} + \\ &+ \mathbb{E}\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]^2.\end{aligned}$$

A középső tag szerencsére nulla, ezt toronyszabállyal láthatjuk be:

$$\begin{aligned}&\mathbb{E}\left\{\left[Y - f_{\text{opt}}(\underline{X})\right]\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]\right\} = \\ &= \mathbb{E}\left\{\mathbb{E}\left\{\left[Y - f_{\text{opt}}(\underline{X})\right]\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]\right\} \mid \underline{X}\right\} = \\ &= \mathbb{E}\left\{\left[f_{\text{opt}}(\underline{X}) - f_{\text{opt}}(\underline{X})\right]\mathbb{E}\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right] \mid \underline{X}\right\} = 0,\end{aligned}$$

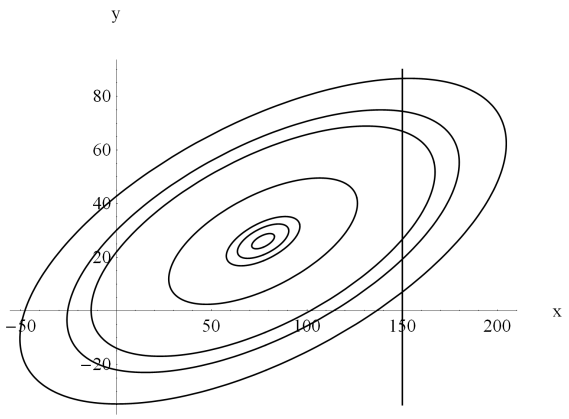
így azt kaptuk, hogy

$$\mathbb{E}\left[Y - f(\underline{X})\right]^2 = \mathbb{E}\left[Y - f_{\text{opt}}(\underline{X})\right]^2 + \mathbb{E}\left[f_{\text{opt}}(\underline{X}) - f(\underline{X})\right]^2,$$

amiből már csakugyan látható, hogy f_{opt} a legjobb választás, hiszen az első tagra nincsen ráhatásunk (mi ugye f-et állítjuk), a második tag pedig egy négyzet várható értéke, így 0-nál kisebb nem lehet, de az csakugyan elérhető, ha f-nek f_{opt}-ot választjuk.

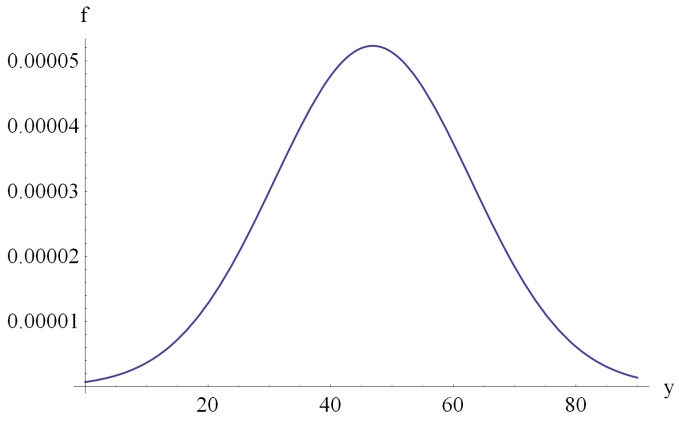
A feltételes várhatóérték – emlékeztető

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



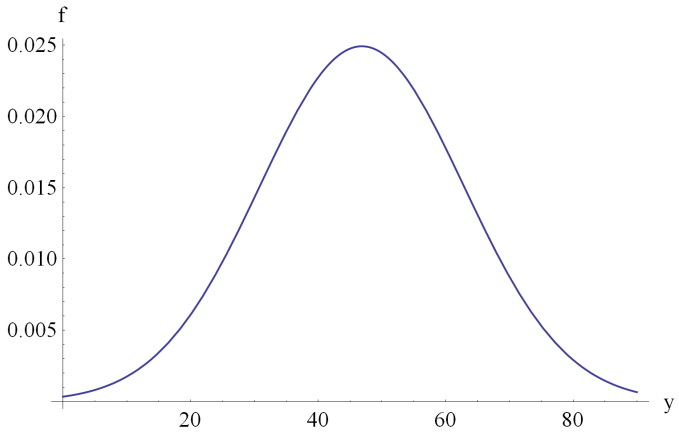
A feltételes várhatóérték – emlékeztető

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



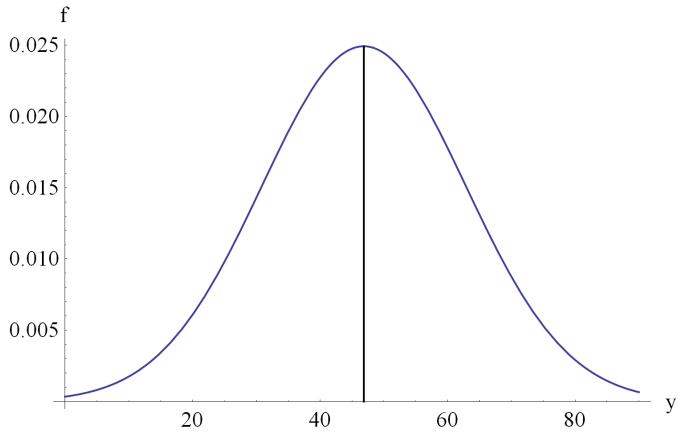
A feltételes várhatóérték – emlékeztető

… 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 ($\mathbb{E}\left(Y \mid X = 150\right) = 46{,}9$)



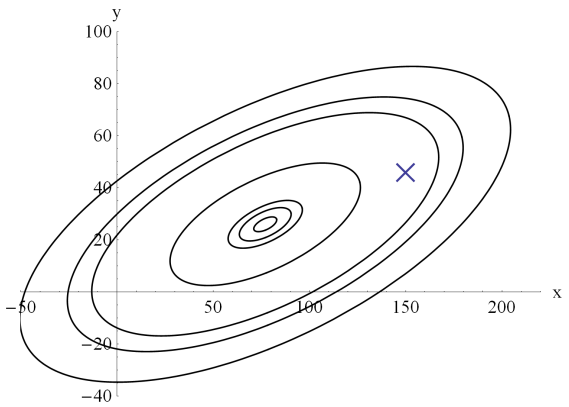
1.3. Az optimális sokasági regresszió

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

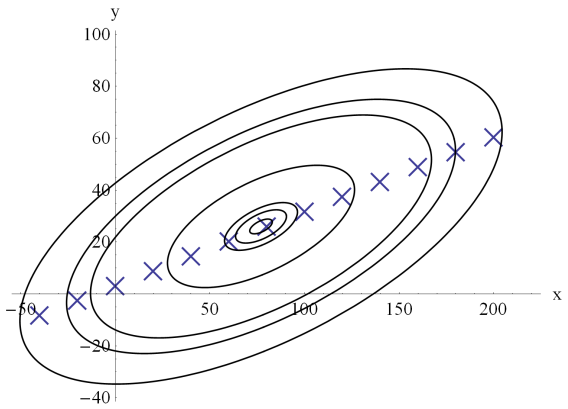
Optimális sokasági regresszió normális eloszlásnál

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.

Érdemes megfigyelni (ez kétváltozós esetben jól érzékelhető vizuálisan is), hogy a regressziófüggvény *nem* a kvantilis-ellipszis nagytengelye (tehát a korrelációs mátrix megfelelő sajátvektora) irányába mutat! (Hanem az ellipszis „vízszintesen szélső” pontjain megy át.)

Kétváltozós (X, Y) esetre: $\mathbb{E}\left(Y \mid X\right) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X} \cdot (X - \mathbb{E}X)$. Két észrevétel ennek kapcsán:

- Korreláció megjelenése: $\mathbb{E}\left(Y \mid X\right) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X}\left(X - \mathbb{E}X\right) = \mathbb{E}Y + \frac{\mathbb{D}Y}{\mathbb{D}X} \cdot \operatorname{corr}(X,Y) \cdot (X - \mathbb{E}X)$.

- A linearitás megjelenése itt: $\mathbb{E}\left(Y \mid X\right) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X}\left(X - \mathbb{E}X\right) = \left(\mathbb{E}Y - \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X} \cdot \mathbb{E}X\right) + X \cdot \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X}$ azaz $\mathbb{E}\left(Y \mid X\right) = \beta_0 + \beta_1 X$, ha $\beta_0 = \left(\mathbb{E}Y - \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X} \cdot \mathbb{E}X\right)$ és $\beta_1 = \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2 X}$.

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

Lényegében arról van szó, hogy szétbontjuk az eredményváltozó alakulását egy ’magyarázóváltozókkal elérhető legjobb becslés’ (már láttuk: a feltételes várhatóérték) és egy ’maradék hiba’ részre (ami marad). A regresszióanalízis a *feltételes* eloszlásra koncentrál! Ezért elvileg olyasmit kéne írnunk, hogy „$\left(Y \mid X\right) = \mathbb{E}\left(Y \mid \underline{X}\right) + \varepsilon$”, de ezt nem tesszük (az $\left(Y \mid \underline{X}\right)$ objektumot nem szokás definiálni), ehelyett a bal oldalra simán Y-t írunk (de ne feledjük, hogy ez *feltételes*).

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

1.4. Modellminősítés a sokasági regresszióban

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**

Érdemes észrevenni, hogy

$$\begin{aligned}
&\operatorname{cov}\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, Y\right) = \operatorname{cov}\left(Y - \varepsilon, Y\right) = \mathbb{D}^2 Y - \operatorname{cov}\left(\varepsilon, Y\right) = \\
&= \mathbb{D}^2 Y - \operatorname{cov}\left(\varepsilon, \varepsilon + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right) = \mathbb{D}^2 Y - \mathbb{D}^2 \varepsilon = \\
&= \mathbb{D}^2 \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right)
\end{aligned}$$

azaz a fent definiált R^2 nem más, mint

$$\begin{aligned}
R^2 &= \frac{\operatorname{cov}\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, Y\right)}{\mathbb{D}^2 Y} = \\
&= \frac{\left[\operatorname{cov}\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, Y\right)\right]^2}{\mathbb{D}^2 Y \cdot \mathbb{D}^2 \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right)}
\end{aligned}$$

tehát Y és $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ közti korreláció négyzete. Ennél azonban több is igaz: bebizonyítható, hogy az $X_1, X_2, \ldots, X_k$ változók *bármely* lineáris kombinációja közül szükségképp a $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$-nek lesz a legnagyobb a korrelációnégyzete az Y-nal. A lineáris regresszió tehát úgy is megfogalmazható, mint ami a magyarázó változók azon lineáris kombinációját keresi meg, melyek a legjobban korreláltak az eredményváltozóval!

2 A lineáris regresszió

2.1. A lineáris regressziós modell (a sokaságban)

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

2.2. A lineáris regressziós modell használata

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 nemlineáris kiterjesztéseknél ezt a jelenséget mélyebben meg fogjuk érteni.

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:

$$\begin{aligned}\beta_0 + \beta_1 X_1 + \ldots + \beta_l (X_l + 1) + \ldots + \beta_k X_k = \\ = (\beta_0 + \beta_1 X_1 + \ldots + \beta_l X_l + \ldots + \beta_k X_k) + \beta_l\end{aligned}$$

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

2.3. A regressziós modell használata a kauzalitás vizsgálatában

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)

Ez a megközelítés két kérdést vet fel: egyrészt, hogy vajon a valóságos jelenségeknek egyáltalán elfogadható modellje-e ez, másrészt, hogy ha valahol nem, akkor hogyan oldható fel ez a megkötés. Később mindkét kérdést részletesen is tárgyaljuk az ún. nemlineáris kiterjesztéseknél.

2.4. Az elaszticitás fogalma

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

3 A lineáris regressziós modell becslése mintából, vektoros-mátrixos formalizmus, az OLS-becslő

3.1. Az OLS-elv

Előkészületek az OLS-becsléshez

- Nem kell hozzá semmilyen regresszió, a legközönségesebb következtető statisztikai példán is elmondható
- Például: sokasági várható érték becslése normalitás esetén (legyen a szórás is ismert)
- Ami fontos: bár egy alap következtető statisztika kurzuson nem szokták mondani, de lényegében itt is az a helyzet, hogy egy *modellt* feltételezünk a sokaságra
- Jelesül $Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right)$, amit nem mellesleg úgy is írhatnánk, hogy $Y = \mu + \varepsilon$, ahol $\varepsilon \sim \mathcal{N}\left(0, \sigma_0^2\right)$
- A másik ami fontos: a modellből következik egy *becsült érték* minden mintabeli elemhez
- Jelen esetben, ha m egy feltételezett érték az ismeretlen sokasági várható értékre:

$$\widehat{y_i} = m$$

Az OLS-elv

- OLS-elvű becslés: az ismeretlen sokasági paraméterre az a becsült érték, amely mellett a tényleges mintabeli értékek, és az adott paraméter melletti, modellből származó becsült értékek közti eltérések négyzetének összege a legkisebb:

$$\widehat{\mu} = \arg\min_m \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 = \arg\min_m \sum_{i=1}^{n} \left(y_i - m\right)^2$$

- (Aminek a megoldása természetesen $\widehat{\mu} = \overline{y}$)

A mintavétel a lineáris regressziós feladatban

- Tételezzük fel, hogy az $(Y, X_1, X_2, \ldots, X_k)$ változóinkra veszünk egy n elemű mintát
- Az i-edik mintaelem: $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$
- Feltételezzük azt is, hogy a mintavétel fae (független, azonos eloszlású)

A fae feltevés keresztmetszeti esetben sokszor lehet elfogadható közelítés (bár ott sem mindig teljesül, erre egy jó példa az ún. térbeli autokorreláció – de ez túlmutat a mostani kereteken), idősoros adatoknál azonban sohasem. Ott nagyon részletesen fogunk ezzel foglalkozni.

Lineáris regresszió becslése OLS-elven

- *Hajszálpontosan ugyanaz* történik, mint az előbb, csak a sokaságra feltételezett modellünk kicsit bonyolultabb, jelesül:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- A becsült értékek adott $b_0, b_1, \ldots, b_k$ sokasági paraméterek mellett:

$$\widehat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

- A feladat tehát ugyanaz:

$$\left(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_k}\right) = \underset{b_0, b_1, b_2, \ldots, b_k}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 =$$

$$= \underset{b_0, b_1, b_2, \ldots, b_k}{\arg\min} \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik})\right]^2$$

- Annyi bonyolódottság van, hogy itt most *több* paramétert kell becsülni, de ez csak a kivitelezést nehezíti, elvileg teljesen ugyanaz a feladat

Az OLS-becslési feladat vektoros-mátrixos jelölésekkel

- A jelölések egyszerűsítése érdekében fogjuk össze mindent vektorokba és mátrixokba; egyedül a magyarázó változók nem triviálisak, mert kiegészítjük őket egy csupa 1 oszloppal (ún. design mátrix):

$$\mathbf{X}_{n\times(k+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

- Így ugyanis a feladat:
$$\arg\min_{\mathbf{b}} (\mathbf{y}-\mathbf{X}\mathbf{b})^T (\mathbf{y}-\mathbf{X}\mathbf{b})$$
- Az $(\mathbf{y}-\mathbf{X}\mathbf{b})^T (\mathbf{y}-\mathbf{X}\mathbf{b})$ hibanégyzetösszeget ESS-sel (error sum of squares) is fogjuk jelölni

Sajnos néhány irodalom az általunk használt ESS-re inkább az RSS-t (residual sum of squares) rövidítést használja, ami a jelölési zavarok legszerencsétlenebb típusa, ugyanis az RSS-t majd később mi is fogjuk használni, csak épp másra. Éppen ezért, ha ilyenekről olvasunk, mindig tisztázni kell, hogy a könyv vagy program írói mit értenek alatta.

3.2. A lineáris regresszió becslése tisztán deskriptíve

Az OLS-becslési feladat megoldása

A megoldás:

$$\arg\min_{\mathbf{b}} (\mathbf{y}-\mathbf{X}\mathbf{b})^T (\mathbf{y}-\mathbf{X}\mathbf{b}) = \arg\min_{\mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right]$$

A szélsőérték-keresést oldjuk meg többváltozós deriválással (kvadratikus felület konvex, a stacionárius pont egyértelmű globális szélsőértékhely):

$$\frac{\partial}{\partial \mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right] =$$
$$= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = 0 \Rightarrow \widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

ha $\mathbf{X}^T\mathbf{X}$ nem szinguláris

Az első lépésnél lényegében egyszerű algebrai átalakításokat végzünk (és a definíciókat használjuk), hiszen a zárójeleket felbontani, műveleteket elvégezni, mátrixokkal-vektorokkal is hasonlóan kell mint valós számokkal. (A transzponálás tagonként elvégezhető, azaz $(\mathbf{a}-\mathbf{b})^T = \mathbf{a}^T - \mathbf{b}^T$.) Egyedül annyit kell észrevenni, hogy a $\mathbf{y}^T\mathbf{X}\mathbf{b}$ egy egyszerű valós szám, ezért megegyezik a saját transzponáltjával, $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-nal. Ezért írhattunk $-(\mathbf{X}\mathbf{b})^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{b}$ helyett egyszerűen – például – $-2\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-t. (Itt mindenhol felhasználtuk, hogy a transzponálás megfordítja a szorzás sorrendjét: $(\mathbf{A}\mathbf{B})^T = \mathbf{A}^T\mathbf{B}^T$.)

Itt jelentkezik igazán a mátrixos jelölésrendszer előnye. A $\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$ lényegében egy „másodfokú kifejezés" többváltozós értelemben (az $ax^2 + bx + c$ többváltozós megfelelője), és ami igazán szép: pont ahogy az $ax^2 + bx + c$ lederiválható a változója (x) szerint (eredmény $2ax + b$), ugyanúgy ez is lederiválható a változója (azaz $\mathbf{b}$) szerint… és az eredmény az egyváltozóssal teljesen analóg lesz, ahogy fent is látható! (Ez persze bizonyítást igényel! – lásd többváltozós analízisből.) Bár ezzel átléptünk egyváltozóról többváltozóra, a többváltozós analízisbeli eredmények biztosítanak róla, hogy formálisan ugyanúgy végezhető el a deriválás. (Ezt írja le röviden a „vektor szerinti deriválás" jelölése. Egy $\mathbf{b}$ vektor szerinti derivált alatt azt a vektort értjük, melyet úgy kapunk, hogy a deriválandó kifejezést lederiváljuk $\mathbf{b}$ egyes b_i komponensei szerint

(ez ugye egyszerű skalár szerinti deriválás, ami már definiált!), majd ez eredményeket összefogjuk egy vektorba. Látható tehát, hogy a vektor szerinti derivált egy ugyanolyan dimenziós vektor, mint ami szerint deriváltunk.) Ami igazán erőteljes ebben az eredményben, az nem is egyszerűen az, hogy „több" változónk van, hanem, hogy nem is kell tudnunk, hogy mennyi – mégis, általában is működik!

Azt, hogy a megtalált stacionaritási pont tényleg minimumhely, úgy ellenőrizhetjük, hogy megvizsgáljuk a Hesse-mátrixot a pontban. A mátrixos jelölésrendszerben ennek az előállítása is egyszerű, még egyszer deriválni kell a függvényt a változó(vektor) szerint:

$$\frac{\partial^2}{\partial \mathbf{b}^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right] = \frac{\partial}{\partial \mathbf{b}}\left[-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}\right] = 2\mathbf{X}^T\mathbf{X}.$$

Az ismert tétel szerint a függvénynek akkor van egy pontban ténylegesen is (lokális, de a konvexitás miatt egyben globális) minimuma, ha ott a Hesse-mátrix pozitív definit. Esetünkben ez minden pontban teljesül. A $\mathbf{X}^T\mathbf{X}$ ugyanis pozitív szemidefinit (ez egy skalárszorzat-mátrix, más néven Gram-mátrix, amelyek mindig pozitív szemidefinitek), a kérdés tehát csak a határozott definitség. Belátható azonban, hogy ennek feltétele, hogy $\mathbf{X}^T\mathbf{X}$ ne legyen szinguláris – azaz itt is ugyanahhoz a feltételhez értünk! Megjegyezzük, hogy ez pontosan akkor valósul meg, ha az $\mathbf{X}$ teljes oszloprangú. (Erre a kérdésre a modellfeltevések tárgyalásakor még visszatérünk.)

Végül egy számítástechnikai megjegyzés: az együtthatók számításánál a fenti formula direkt követése általában nem a legjobb út, különösen ha sok megfigyelési egység és/vagy változó van. Ekkor nagyméretű mátrixot kéne invertálni, amit numerikus okokból (kerekítési hibák, numerikus instabilitás stb.) általában nem szeretünk. Ehelyett, a különféle programok igyekeznek a direkt mátrixinverziót elkerülni, tipikusan az $\mathbf{X}$ valamilyen célszerű mátrix dekompozíciójával (QR-dekompozíció, Cholesky-dekompozíció). Extrém esetekben még az is elképzelhető, hogy az egzakt, zárt alakú megoldás előállítása helyett valamilyen iteratív optimalizálási algoritmus (gradiens módszer, Newton–Raphson-módszer) alkalmazása a gyakorlatban járható út, annak ellenére is, hogy elvileg van zárt alakban megoldása.

A kapott eredmény nem más, mintha $\mathbf{X}$ Moore-Penrose pszeudoinverzével szoroznánk $\mathbf{y}$-t.

Pár további gondolat

- Az ún. reziduumok:
$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$$
- Az előrejelzések a mintánkra:
$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$
- Ez alapján vezessük be a
$$\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$
mátrixot, ezzel $\widehat{\mathbf{y}} = \mathbf{P}\mathbf{y}$

- Emiatt szokták „hat” mátrixnak is nevezni

Az OLS geometriai interpretációja

P projektormátrix lesz ($\mathbf{P}^2 = \mathbf{P}$, azaz idempotens) → út az OLS geometriai interpretációjához

Mindenekelőtt emlékeztetünk rá, hogy az $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ vektorok által kifeszített alteret azok a pontok alkotják, melyek előállnak e vektorok lineáris kombinációjaként. (E pontok mindig az eredeti vektortér – ami felett a vektorokat értelmeztük – alterét alkotják, ezért jogos az elnevezés.) Ha most vektortérnek az $\mathbb{R}^n$-et tekintjük, vektoroknak pedig az $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$ magyarázóváltozókat (és a konstanst), azaz **X** oszlopvektorait, akkor az ezek által kifeszített altér – ezt szokás egyébként az **X** mátrix oszlopterének nevezni – épp azon pontokból áll, melyek előállhatnak becsült eredményváltozó(vektor)ként valamilyen regressziós koefficiensekkel! (Hiszen a becsült eredményváltozót is e vektorok lineáris kombinációjaként állítjuk elő.) Általánosságban persze nem várható, hogy a tényleges eredményváltozó(vektor) benne legyen ebben az altérben (azaz egzaktan – értsd: minden egyes megfigyelési egységre megvalósulóan – elő lehessen állítani lineáris kombinációként), ezt fejezi ki a reziduum. Mint a tényleges és a becsült eredményváltozó(vektor) különbségvektora, a reziduum hossza megmutatja, hogy mennyire messze van a becsült és a tényleges eredményváltozó egymástól (az $\mathbb{R}^n$-ben). Mi azt szeretnénk, ha ez minimális lenne. Választva a szokásos euklideszi metrikát, visszakapjuk a legkisebb négyzetes értelmezést. A kérdés már csak az, hogy adott ponthoz (tényleges eredményváltozó) hogyan határozható meg az altér (azaz: amit lineáris regresszióval elő tudunk állítani) legközelebbi pontja… de hát ez épp a geometriai vetítés művelete! A megoldás tehát az, hogy a tényleges eredményváltozót merőlegesen rávetítjük (ortogonális projekció) a magyarázóváltozók (és a konstans) által kifeszített altérre! A vetítés eredményeként kapott pont lesz a ténylegeshez legközelebbi előállítható becsült eredményváltozó, az előállításában szereplő együtthatók pedig az optimális becsült regressziós koefficiensek. Így aztán azt is megállapíthatjuk, hogy a fenti **P** mátrix nem más, mint ami a tényleges eredményváltozót levetíti a magyarázóváltozók (és a konstans) által kifeszített altérre.

3.3. Modellminősítés tisztán deskriptíve

Modell jóságának viszonyítási pontjai

- A modell minősítése az ESS alapján? → kézenfekvő, de nem önmagában: viszonyítani kell! Két kézenfekvő alap:
  - Tökéletes (v. szaturált, perfekt modell): minden mintaelemre a pontos értéket becsüli → $\widehat{e_i} = 0 \Rightarrow ESS = 0$
  - Nullmodell: semmilyen külső (magyarázó)információt nem használ fel → minden mintaelemet az átlaggal becsül
- Egy adott regressziós modell teljes négyzetösszegének nevezzük, és TSS-sel jelöljük a hozzá tartozó (tehát ugyanazon eredményváltozóra vonatkozó) nullmodell

hibanégyzetösszegét:

$$TSS = ESS_{\text{null}} = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

Hogyan jellemezzük modellünk jóságát?

- A minősítést képezzük a „hol járunk az úton?” elven: a tökéletesen rossz modelltől a tökéletesen jó modellig vezető út mekkora részét tettük meg
- Az út „hossza” TSS $(= TSS - 0)$, amennyit „megtettünk”: $TSS - ESS$
- Egy adott regressziós modell regressziós négyzetösszegének nevezzük, és RSS-sel jelöljük a teljes négyzetösszegének és a hibanégyzetösszegének különbségét:

$$RSS = TSS - ESS.$$

Ahogy már említettük is, sajnos néhány könyv az RSS-t más néven, hogy még rosszabb legyen a helyzet, néha ESS-ként, emlegeti. (Az itteni ESS pedig épp RSS az ottani terminológiában…)

Az új mutató bevezetése

Ezzel az alkalmas modelljellemző mutató: a többszörös determinációs együttható (jele R^2):

$$R^2 = \frac{TSS - ESS}{TSS} = \frac{RSS}{TSS}.$$

Az R^2-ről bővebben

- Ha van konstans a modellben, akkor nyilván $ESS < TSS$, így minden regressziós modellre, amiben van konstans: $0 \leq R^2 \leq 1$.
- Az R^2 egy modell jóságának legszéleskörűbben használt mutatója
- Értelmezhető %-ként: a magyarázó változók ismerete mennyiben csökkentette az eredményváltozó tippelésekor a bizonytalanságunkat (ahhoz képest, mintha nem ismertünk volna egyetlen magyarázó változót sem)
- De vigyázat: nagyságának megítélése, változók száma stb.
- A belőle vont négyzetgyököt többszörös korrelációs együtthatónak szokás nevezni
- Mondani sem kell, ez az R^2 a korábban bevezetett (sokasági) R^2 mintabeli analógja

Az R^2-ről bővebben

- Ha van konstans a modellben, akkor érvényes a következő felbontás:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$$

- (Négyzetek nélkül nyilvánvaló, de négyzetekkel is!)
- Röviden tehát:

$$TSS=ESS+RSS$$

- Összevetve az előző definícióval, kapjuk, hogy

$$RSS=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$$

Egy megjegyzés a konstans szerepéről

- Az előzőek is motiválják, hogy megállapítsuk: konstanst *mindenképp* szerepeltetünk a regresszióban, ha inszignifikáns, ha nem látszik különösebb értelme stb. *akkor is*! – csak és kizárólag akkor hagyhatjuk el, ha az a modell tartalmából adódóan elméleti követelmény (erre látni fogunk nemsokára egy példát is, a standardizált regressziót)
- Ellenkező esetben (ún. konstans nélküli regresszió), a fenti felbontás nem teljesül, így a „hol járunk az úton" elven konstruált R^2 akár negatív is lehet!

Néhány könyv, az R^2 alternatív definiálása révén, a negatív esetet kizárja.

4 Az OLS becslő modellfeltevései és az OLS szolgáltatta becslések statisztikai tulajdonságai

4.1. Mintavételi helyzet

Deskriptív és következtető statisztika

- Az előbbi tárgyalás pusztán deskriptív volt: egy darab mintát tekintett, amire meghatározott egy darab regressziós függvényt és kész
- Mintha a feladat csak annyi lenne, hogy pontokra húzzunk egy rájuk jól illeszkedő görbét
- Ez a „görbeillesztési" szemlélet első ránézésre könnyen megérthető, és látszólag egyszerűsíti a helyzetet, valójában azonban rendkívül hátráltató a valódi megértésre nézve
- Nem teszi lehetővé ugyanis annak megértését, hogy a háttérben van egy sokaság, és a görbe nem univerzálisan jellemzi azt, hanem csak az adott, konkrét mintára illeszkedik legjobban, és *másik mintából másik görbét kaptunk volna*
- Azaz: figyelmen kívül hagyja a mintavételi helyzetet

A mintavételi helyzet hatásai

- Van egy elméleti regresszió a **sokaságban** ($\boldsymbol{\beta}$)
- Az adatbázisunk alapján megkaptuk a regressziós egyenest ($\widehat{\boldsymbol{\beta}}$)
- Az adatbázis azonban csak egy **minta** a sokaságából, így a $\widehat{\beta}_i$ paraméterek annak hatását *is* tükrözik, hogy konkrétan milyen mintát választottunk
- *Mintavételi ingadozás* lép fel (még akkor is, ha tökéletesen véletlen a mintavétel, ennek tehát semmi köze pl. a reprezentativitáshoz)
- Tehát: az egyes $\widehat{\beta}_i$ paraméterek „mintáról-mintára ingadoznak": minden mintából más paramétereket kapnánk
- (Természetesen reméljük, hogy az ingadozás „kellemes" tulajdonságokkal bír, például a valós érték körül történik, szorosan körülötte stb., erről később)

- Ez tehát egy becslési feladat; az OLS-nek, mint becslőfüggvénynek vizsgálhatóak a tulajdonságai

Végeredményben a mintából számolt jellemzőkben nem csak az fog tükröződni, amit szeretnék vizsgálni (azaz a tényleges – értsd: elméleti, sokasági – regresszió), hanem az is, hogy a sokaságból konkrétan hogyan vettük a mintát (azaz megjelenik a mintavételi ingadozás hatása is).

Még egy fontos megjegyzés

- Nem elég annyit mondani, hogy „jó, hát akkor a háttérben van egy sokaság is", mintha ezzel el lenne intézve ez a kérdés
- Azt is világosan látni kell, hogy az egész tárgyalás *kiindulópontja*, hogy erre *feltételezünk* egy modellt (pl. azt, hogy $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$)
- Ez egy feltételezés, mivel a sokaságot nem ismerjük, így biztosan nem tudhatjuk, hogy igaz-e (csak következtethetünk rá)
- De minden további levezetés mögött ott lesz, hogy mi mit gondoltunk, hogyan viselkedik a sokaság, mi a *sokasági modell*

Előkészület a mintavétel vizsgálatához

- Ahhoz, hogy a mintavétel hatását matematikailag tudjuk vizsgálni, az OLS-becslőt val. változókra kell ráereszteni (szemben az eddigi képlettel – $\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ – ahol konkrét számokra futtattuk)
- Pontosan ugyanúgy, ahogy az $\frac{1}{n}\sum_{i=1}^{n} x_i$-t sem tudjuk következtető statisztikailag vizsgálni (az egy szám), hanem az $\frac{1}{n}\sum_{i=1}^{n} X_i$-t nézzük
- Minket tehát $\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{Y}$ fog érdekelni!
- Ahogy az előbbi átlagos példában, így itt is igaz lesz, hogy ekkor a $\widehat{\underline{\boldsymbol{\beta}}_{\mathrm{OLS}}}$ már nem egy konkrét érték (vektor), hanem egy – vektor értékű – val. változó, tehát eloszlása van!
- Ez a mintavételi eloszlás, mi erre, ennek tulajdonságaira, a jó tulajdonságok feltételeire stb. leszünk kíváncsiak
- Előbb szimulációval nyerünk képet, aztán matematikailag is levezetjük

4.2. A mintavétel tulajdonságok szemléltetése szimulációval

Monte Carlo szimuláció használata

- Számos konkrét véletlen mintát veszünk egy előre specifikált populációból (véletlenszám-generátort használunk)
- Lényegében: empirikusan vizsgálunk egy elméleti kérdést
- Valszámos embert játszunk: ugye azt mondtuk, hogy a valszámosok úgy dolgoznak mintha valahonnan ismernék a sokasági eloszlást – hát most tényleg ismerjük!
- Példának okáért, legyen a valódi sokasági eloszlás

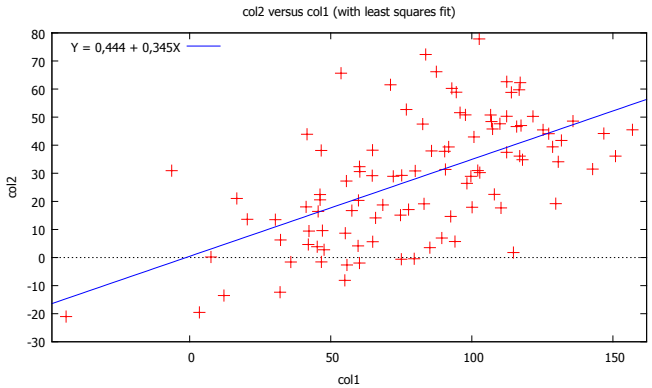
$$(X, Y) \sim \mathcal{N}\left(\begin{pmatrix}77\\26\end{pmatrix}, \begin{pmatrix}42^2 & 0{,}6 \cdot 20 \cdot 42\\ 0{,}6 \cdot 20 \cdot 42 & 20^2\end{pmatrix}\right)$$

- Ezért a *valódi* regressziós egyenes, a már látottak szerint:

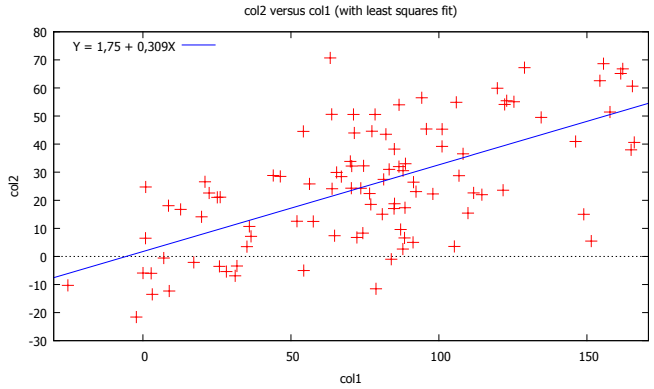
$$\mathbb{E}\left(Y \mid X\right) = 4 + \frac{12}{42}X \approx 4 + 0{,}2857X$$

- Szimulációs paraméterek: $n = 100$ elemű minta a fenti sokaságból, 1000 ismétlés

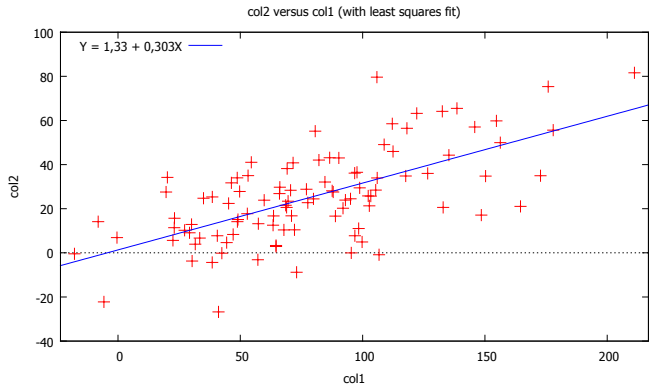
A szimuláció eredményei: 1. futtatás



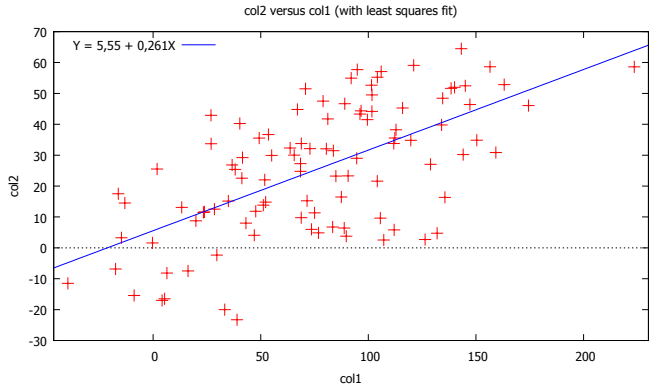
A szimuláció eredményei: 2. futtatás



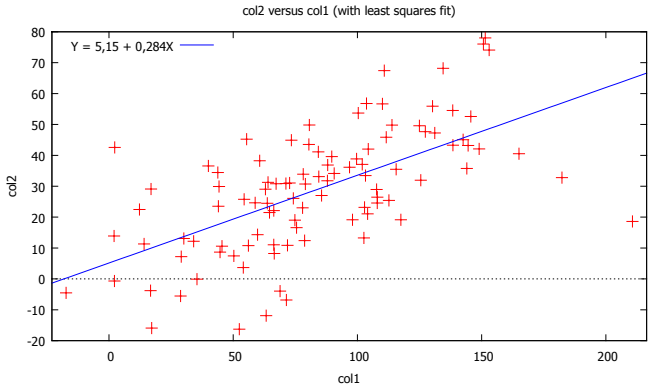
A szimuláció eredményei: 3. futtatás



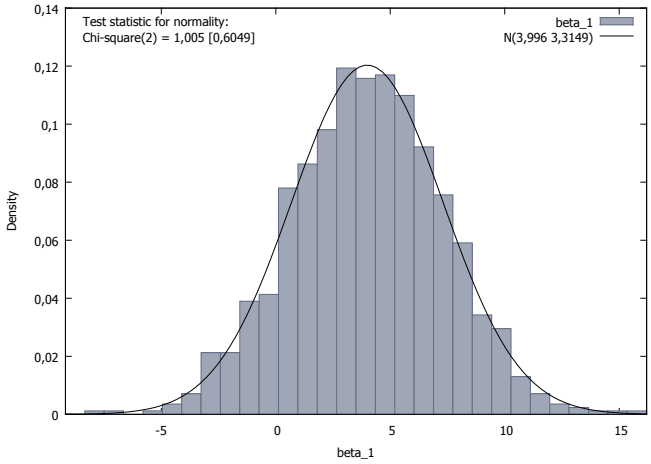
A szimuláció eredményei: 4. futtatás



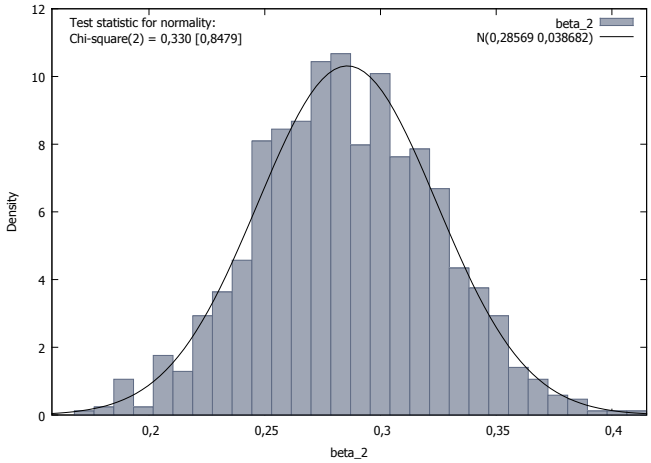
A szimuláció eredményei: 5. futtatás



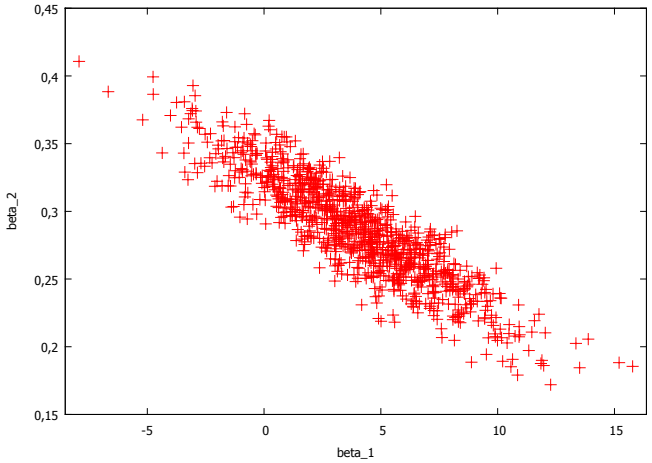
A szimuláció eredményei: konstans



A szimuláció eredményei: meredekség



A szimuláció eredményei: mindkét paraméter együtt



4.3. A mintavétel tulajdonságok matematikai levezetése

Az OLS-becslő mintavételi eloszlása

- Tudjuk, hogy $\widehat{\boldsymbol{\beta}_{\text{OLS}}} = \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{Y}$
- Valamint elfogadtuk feltételezésként, hogy a sokasági modell $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon = \underline{X}^T \boldsymbol{\beta} + \varepsilon$
  - És ez van mindegyik megfigyelési egység mögött is, tehát a mintavétel elemzéséhez ezt is írhatjuk:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

  - Röviden, értelemszerű vektorokba/mátrixokba fogással: $Y_i = \underline{X}_i^T \boldsymbol{\beta} + \varepsilon$ avagy az egész adatbázisra: $\underline{Y} = \underline{\underline{X}} \boldsymbol{\beta} + \underline{\varepsilon}$
- Na de rakjuk csak össze a kettőt:

$$\begin{aligned}\widehat{\boldsymbol{\beta}_{\text{OLS}}} &= \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{Y} = \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \left(\underline{\underline{X}} \boldsymbol{\beta} + \underline{\varepsilon}\right) = \\ &= \left(X^T X\right)^{-1} \underline{\underline{X}}^T \underline{\underline{X}} \boldsymbol{\beta} + \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{\varepsilon} = \boldsymbol{\beta} + \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{\varepsilon}\end{aligned}$$

4.4. Az OLS modellfeltevései

Az OLS standard modellfeltevési

Ahhoz, hogy az OLS-nek fennálljanak bizonyos előnyös tulajdonságai, meghatározott feltevéseknek teljesülniük kell. Az ún. standard lineáris modell feltevései:

1. Linearitás
2. Nincs egzakt multikollinearitás
3. Erős (vagy szigorú) exogenitás
4. Homoszkedaszticitás
5. Autokorrelálatlanság

Linearitás

A sokaságot *valójában* leíró modell tényleg a feltételezett, azaz fennáll, hogy

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

és ez igaz mindegyik megfigyelési egységre, és így az egész mintára is:

$$\underline{Y} = \underline{\underline{X}}\boldsymbol{\beta} + \underline{\varepsilon}$$

Nincs egzakt multikollinearitás

- Egzakt multikollinearitásnak nevezzük, ha az adatmátrix nem teljes oszloprangú
- Tehát: az oszlopok között lineáris kapcsolat van
- Azaz valamelyik változó előállítható a többi lineáris kombinációjaként
- Érezhető, hogy nem túl szerencsés: minek használjuk egyáltalán azt a változót…? (Úgyis lineáris kombinációt képezünk a többiből is!) $\rightarrow$ a hatások nem lesznek szétválaszthatóak
- Sőt: az OLS becslőfüggvényéből az is látszik, hogy ilyenkor teljesen elakadunk: $\mathbf{X}^T\mathbf{X}$ szinguláris ($\underline{\underline{X}}^T\underline{\underline{X}}$ 1 valószínűséggel szinguláris)
- Ennek feltétele: $\mathbf{X}$ ($\underline{\underline{X}}$) nem teljes oszloprangú

Nincs egzakt multikollinearitás

- A feltétel tehát: az adatmátrix 1 valószínűséggel legyen teljes oszloprangú:

$$\mathbb{P}\left(\operatorname{rank}\underline{\underline{X}} = k+1\right) = 1$$

- Ez implikálja, hogy $n \geq k+1$ (kevesebb mint $k+1$-dimenziós vektorból nincs $k+1$ független)

Érdemes megfigyelni, hogy a multikollinearitás egy mintában is elképzelhető jelenség (ezért is hivatkozhattunk rá úgy, hogy **X** nem teljes oszloprangú), de mi a tulajdonságot a sokaságban akarjuk kikötni, ezért az állítás az $\underline{\underline{X}}$-re kell vonatkozzon. Itt viszont csak annak van értelme, hogy „majdnem biztosan" (azaz 1 valószínűséggel) követeljük meg.

A másik alapvető szemléletes példa a multikollinearitásra az, ha van konstans a modellben és valamelyik magyarázó változónak nincs szórása (az összes megfigyelés ugyanaz rá). Könnyű elképzelni (pl. két dimenzióban), hogy nem lehet semmilyen regressziós egyenest húzni akkor, ha minden pontunk egymás fölött van.

Végezetül megjegyezzük, hogy gyakorlati jelentősége annak is lesz, ha ugyan nincs egzakt multikollinearitás, de van változó ami „elég jól" előállítható a többi lineáris kombinációjaként. Ezzel a kérdéskörrel fontossága miatt külön fogunk foglalkozni.

Erős exogenitás

- Minden $i = 1, 2, \ldots, n$-re
$$\mathbb{E}\left(\varepsilon_i \mid \underline{X}_i\right) = 0$$
- Tartalma: a hibák – az ún. várható érték függetlenség értelemben – függetlenek a magyarázó változóktól

Ha nem fae lenne a mintavétel, akkor azt kellene kikötni, hogy $\mathbb{E}\left(\underline{\varepsilon} \mid \underline{\underline{X}}\right) = 0$. Ha azonban a mintavétel fae, akkor az automatikusan teljesül, hogy ε_i minden $\underline{X}_j$-től független – és így várható érték független is – ha $j \neq i$.

Igazából elég azt kikötni, hogy $\mathbb{E}\left(\varepsilon_i \mid \underline{X}_i\right) = \text{konst}$, belátható, hogy ha van konstans a modellben, akkor ez egyenértékű a fentivel.

Az erős exogenitás következményei

- Toronyszabály miatt a feltétel *nélküli* várható érték is nulla:
$$\mathbb{E}\left[\mathbb{E}\left(\varepsilon_i \mid \underline{X}_i\right)\right] = \mathbb{E}\varepsilon_i = \mathbb{E}\left(0\right) = 0$$
- A várható érték függetlenség implikálja a korrelálatlanságot: $\operatorname{cov}\left(X_{ik}, \varepsilon_j\right) = 0$ avagy – ezzel egyenértékűen, hiszen $\mathbb{E}\varepsilon_i = 0$ – $\mathbb{E}\left(X_{ik}\varepsilon_j\right) = 0$
- Szokás a korrelálatlanság helyett azt is mondani, hogy a hibák ortogonálisak a magyarázóváltozókra

Az $\mathbb{E}\left(\varepsilon_i \mid \underline{X}_i\right) = 0$ *szigorúan erősebb* követelmény mint a korrelálatlanság. Ezt a fogalmat szokás várható érték függetlenségnek (mean independence) nevezni. Lássuk is ezt be, a jelölés megkönnyítése végett hívjuk egyszerűen ε-nak és X-nek a két változónkat. Emlékeztetőül $\operatorname{cov}\left(\varepsilon, X\right) = \mathbb{E}\left(\varepsilon X\right) - \mathbb{E}\varepsilon\mathbb{E}X$. Egyik oldalról

$$\mathbb{E}\left(\varepsilon X\right) = \mathbb{E}\left[\mathbb{E}\left(\varepsilon X \mid X\right)\right] = \mathbb{E}\left[X\mathbb{E}\left(\varepsilon \mid X\right)\right] = \mathbb{E}\left[X \cdot 0\right] = \mathbb{E}0 = 0,$$

így ε és X korrelálatlanok (hiszen $\mathbb{E}\varepsilon = 0$); ezzel bebizonyítottuk, hogy a várható érték függetlenség implikálja a korrelálatlanságot. Másik oldalról, tekintsük példaként az $X \sim \mathcal{N}(0,1)$ változót és egy olyan ε változót melyre $\mathbb{E}\left(\varepsilon|X\right) = X^2$. Ekkor $\mathbb{E}X = 0$, $\mathbb{E}\varepsilon = \mathbb{E}\left[\mathbb{E}\left(\varepsilon \mid X\right)\right] = \mathbb{E}\left(X^2\right) = \mathbb{D}^2X = 1$ és $\mathbb{E}\left(\varepsilon X\right) = \mathbb{E}\left[\mathbb{E}\left(\varepsilon X \mid X\right)\right] = \mathbb{E}\left[X\mathbb{E}\left(\varepsilon \mid X\right)\right] = \mathbb{E}\left(X^3\right) = 0$, így mutattunk egy példát, amikor a változók korrelálatlanok, de mégsem várható érték függetlenek.

A várható érték függetlenség tehát *szigorúan erősebb* fogalom, mint a korrelálatlanság. Érdemes azt is megjegyezni, hogy viszont *szigorúan gyengébb* mint az igazi függetlenség! (Ez könnyen belátható. Függetlenség esetén minden feltételes eloszlás ugyanaz, márpedig ekkor nyilván a várható értékük is ugyanaz. Másik oldalról tekintsünk egy origó körüli $0 < r < R$ körgyűrűre koncentrált egyenletes eloszlást. Ez nyilván várható érték független – a feltételes várható érték konstans nulla –, viszont természetesen nem független.) Lényegében arról van szó, hogy a függetlenség a feltételes eloszlások *teljes* egyezőségét követeli meg, míg a várható érték függetlenség csak annyit, hogy a feltételes eloszlások várható értéke legyen egyező. (A feltételes szórás már kapásból lehet eltérő.)

A másik fontos megjegyzés itt, hogy az OLS-nek megvan az a tulajdonsága, hogy a *reziduumok* mindig korrelálatlanok a magyarázó változókkal. Valaki megkérdezheti, hogy akkor minek ezt külön is kikötni? Vigyázat! A reziduumok (amik a mintában vannak) tényleg mindig korrelálatlanok, csakhogy itt a hibákról beszélünk, amik a sokaságban vannak! A feltétel a *sokaságról* mond valamit (amit mi nem ismerhetünk soha biztosan – épp ettől feltétel!), nem a mintáról. A mintában a reziduumok természetesen mindig korrelálatlanra lesznek állítva, de ezzel nem sokra megyünk, ha a sokaságban nem teljesül ez a feltétel, hiszen ez esetben a reziduumoknak nem sok közük lesz a hibákhoz, pont ez a probléma…

Az erős exogenitás sérülésének tipikus esetei

- Van olyan változó, ami lényeges magyarázó változó lenne (tehát valódi – sokasági – β-ja nem nulla), de mégsem szerepel a modellben, miközben legalább egy magyarázó változóval korrelál (kihagyott változó esete, „omitted variable bias") – ez épp a confounding!
- Mérési hiba magyarázó változónál (tehát a mérési változók valódi értékét nem, csak valamilyen zajjal terhelve tudjuk mérni)
- Szimultaneitás (többegyenletes modelleknél)

Az első eset szolgáltatja a talán legjellemzőbb példákat a 'korreláció nem implikál kauzalitást' statisztikai alapelvére: ez a confounding, amit már részletesen tárgyaltunk. Szemléltessük ezt egy korábban már említett példán: emberek fizetését regresszáljuk ki az oktatásban töltött éveik számával (tehát az előbbi az eredmény-, az utóbbi az – egyetlen – magyarázó változó). Ekkor a hibába vélhetően olyan tényezők fognak beleszámítani, mint a nem-oktatással összefüggő munkaalkalmasság, a munkamorál, a szakmai tapasztalat stb. Az egyszerűség kedvéért mondjuk, hogy csak a legelső adja a hibát. Ekkor a szigorú exogenitás feltétele, a várható érték függetlenség azt fogalmazza meg, hogy

a munkaalkalmasság feltételes várható értéke minden képzettség, mint feltétel mellett legyen ugyanakkora, tehát, hogy ne függjön a képzettségtől. (Amint mondtuk, konstans jelenléte esetén ez mellesleg azt jelenti, hogy nulla is legyen ez az állandó feltételes várható érték.) Baj akkor van, ha a képzettség különböző szintjei mellett a várható munkaalkalmasság *nem* állandó – tehát például a magasabb képzettségűeknek a munkaalkalmasságuk is nagyobb, azaz a nagyobb képzettséggel *egyúttal* a munkaalkalmasság is emelkedik. Ekkor megsérül a szigorú exogenitási feltétel. Épp innen kapta a feltétel a nevét: olyasmit fejez ki, hogy a magyarázó változókhoz képest exogén információ az, ami a hibákban össze van fogva. (Ez nyilván nem teljesül a fenti esetben.)

A második és harmadik kérdéskör boncolgatása meghaladja jelen kurzus kereteit.

Végül megjegyezzük, hogy idősoros esetben ez nagyon erős feltétel (hiszen például azt jelenti, hogy a magyarázó változóknak a múltbeli, a jelenbeli *és* a jövőbeli hibákra is ortogonálisnak kell lenniük!), ami sokszor nem teljesül. (Példaként gondoljunk egy egyszerű késleltetett eredményváltozós modellre.)

Az erős exogenitás sérülésének kezelése

- A problémát orvosolhatjuk a megfelelő(bb) modellspecifikációval, függően attól, hogy pontosan mi a baj oka…
- …illetve bizonyos statisztikai eszközök is a rendelkezésünkre állnak, ilyen az instrumentális változós (IV) becslés, a kétfázisú legkisebb négyzetek módszere (TSLS) stb.

E kérdések meghaladják jelen kurzus kereteit.

Homoszkedaszticitás

- A feltétel azt köti ki, hogy $\sigma_i^2 := \mathbb{D}^2\left(\varepsilon_i \mid \underline{\underline{X}}\right) = \sigma^2$ i-től függetlenül minden $i = 1, 2, \ldots, n$-re
- Tartalma: a hibák különböző megfigyelésekhez tartozó szórása állandó (nem függ attól, hogy melyik megfigyelésről van szó) avagy – másként megfogalmazva ugyanez – a becsült értékek szóródása a tényleges körül állandó
- Jellemzően keresztmetszeti adatoknál felmerülő kérdés (hamarosan foglalkozunk is vele bővebben)

Nem fae mintavételezésnél azt kellene írnunk, hogy $\mathbb{D}^2\left(\varepsilon_i \mid \underline{\underline{X}}\right) = \sigma^2$ i-től függetlenül minden $i = 1, 2, \ldots, n$-re.

A feltétellel egyenértékű, hogy $\mathbb{E}\left(\varepsilon_i^2 \mid \underline{X}_i\right) = \sigma^2$ (hiszen $\mathbb{E}\left(\varepsilon_i \mid \underline{X}_i\right) = 0$, így a szórásnégyzet a négyzet várható értéke, azaz a második momentum).

Megjegyzendő, hogy fae mintavételezésnél az mindenképp teljesül, hogy $\mathbb{D}^2\varepsilon_i$ konstans, de ez kevés: nekünk a *feltételes* szórás állandósága is kell a standard modellfeltevések között.

Autokorrelálatlanság

- Tartalma: a különböző megfigyelésekhez tartozó hibák korrelálatlanok egymással
- Fae mintavételezésnél ez tehát *automatikusan* teljesül!
- Nem fae esetben a feltétel azt köti ki, hogy $\operatorname{cov}\left(\varepsilon_i, \varepsilon_j \mid \underline{\underline{X}}\right) = 0$ minden $i, j = 1, 2, \ldots, n$, $i \neq j$-re
- Ezzel egyenértékű $\mathbb{E}\left(\varepsilon_i \varepsilon_j \mid \underline{\underline{X}}\right) = 0$ (hiszen $\mathbb{E}\varepsilon_i = 0$, így a kovariancia a két változó szorzatának várható értéke)
- Elsősorban idősoros adatok kérdésköre, most nem is foglalkozunk vele bővebben

A homoszkedaszticitás és az autokorrelálatlanság együtt

- Mindkettő felfogható úgy, mint az ε_i hibák (feltételes) kovarianciamátrixára vonatkozó megkötés
  - Homoszkedaszticitás: a kovarianciamátrix főátlójában ugyanazok az elemek (σ^2) vannak (ugye itt vannak a szórásnégyzetek)
  - Autokorrelálatlanság: a kovarianciamátrix főátlóján kívüli elemek nullák (a mátrix diagonális)
- A kettő *együtt*: a kovarianciamátrix $\sigma^2\mathbf{I}$ alakú (szokás az ilyet skalármátrixnak is nevezni)

Fae esetben $\mathbb{D}^2\underline{\varepsilon} = \mathbb{E}\left(\underline{\varepsilon}\underline{\varepsilon}^T\right) = \sigma^2\mathbf{I}$, nem fae esetben ki kell írni, hogy $\mathbb{D}^2\left(\underline{\varepsilon} \mid \underline{\underline{X}}\right) = \mathbb{E}\left(\underline{\varepsilon}\underline{\varepsilon}^T \mid \underline{\underline{X}}\right) = \sigma^2\mathbf{I}$. (Az első egyenlőségek azért állnak fenn, mert $\mathbb{E}\underline{\varepsilon} = \mathbf{0}$).
Szokás úgy is fogalmazni, hogy a hibavarianciák szferikálisak.

σ^2 becslése

Nem részletezzük, de belátható, hogy ez esetben a σ^2-re adható OLS-becslés:

$$\widehat{\sigma^2} = \frac{ESS}{n-(k+1)} = \frac{\widehat{\mathbf{e}}^T\widehat{\mathbf{e}}}{n-(k+1)}$$

Mintavételileg rögzített magyarázó változók

- Egyszerűbb tárgyalások azt feltételezik, hogy a magyarázó változók mintavételileg rögzítettek (mintha determinisztikusan megszabhatnánk az értéküket: $\underline{X}_i$ igazából $\mathbf{x}_i$)
- Ennek sok baja van:
  1. Nem annyira szép és elegáns (nyilván ez speciális esete a mi tárgyalásunknak!)

2. Nem teszi lehetővé egy sor kérdés mélyebb tárgyalását
3. Alapjában megkérdőjelezhető az alkalmazása nem-experimentális tudományokban (mint a közgazdaságtan…)

- Az előnye, hogy egyszerűsít: ekkor a hiba feltételes és feltétel nélküli eloszlása ugyanaz lesz, a '$|\ \underline{X}_i$' jellegű feltételek elhagyhatóak…
- …emiatt a modellfeltevések a következőkre egyszerűsödnek:
  - Erős exogenitás: $\mathbb{E}\varepsilon_i = 0$ minden $i = 1, 2, \ldots, n$-re
  - Homoszkedaszticitás: $\mathbb{D}^2\varepsilon_i = \sigma^2$ minden $i = 1, 2, \ldots, n$-re
  - Autokorrelálatlanság: $\mathbb{E}\left(\varepsilon_i\varepsilon_j\right) = 0$ minden $i \neq j = 1, 2, \ldots, n$

A mintavételi tulajdonságok

- Ezek lesznek a standard modellfeltevések…
- …most nekiállunk megvizsgálni, hogy a teljesülésük esetén milyen tulajdonságokkal bír az OLS-becslő

Várható érték

- Tudjuk, hogy
$$\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \boldsymbol{\beta} + \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}$$
- Ez alapján mi $\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}}$ várható értéke (várható érték-vektora)?
$$\begin{aligned}\mathbb{E}\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} &= \boldsymbol{\beta} + \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right] = \\ &= \boldsymbol{\beta} + \mathbb{E}\left\{\mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon} \mid \underline{\underline{X}}\right]\right\} = \\ &= \boldsymbol{\beta} + \mathbb{E}\left\{\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\mathbb{E}\left[\underline{\varepsilon} \mid \underline{\underline{X}}\right]\right\} = \boldsymbol{\beta}\end{aligned}$$
- Az erős exogenitás fennállása esetén tehát az OLS szolgáltatta becslések *torzítatlanok*
- Nem bizonyítjuk, de az is igaz, hogy *konzisztensek*

Az első lépésnél kihasználtuk, hogy a várható érték lineáris (összeg várható értéke a tagok várható értékeinek az összege) és hogy konstans várható értéke saját maga (ne feledjük, hogy $\boldsymbol{\beta}$ egy konstans! – nem tudjuk ugyan, hogy mennyi az értéke, de ettől még egy konstans). A második lépésnél a toronyszabályt használtuk: $\mathbb{E}X = \mathbb{E}\left[\mathbb{E}\left(X \mid Y\right)\right]$. A harmadik lépésnél pedig az erős exogenitást használtuk ki.

Kovarianciamátrix

Az előbbi ismeretében:

$$\begin{aligned}
\mathbb{D}^2\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} - \mathbb{E}\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}}\right)\cdot\left(\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} - \mathbb{E}\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}}\right)^T\right] = \\
&= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} - \boldsymbol{\beta}\right)\cdot\left(\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} - \boldsymbol{\beta}\right)^T\right] = \\
&= \mathbb{E}\left\{\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right]\cdot\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right]^T\right\} = \\
&= \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\underline{\varepsilon}^T\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\right] = \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\mathbb{E}\left(\underline{\varepsilon}\underline{\varepsilon}^T\right)\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1} = \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\cdot\sigma^2\mathbf{I}\cdot\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1} = \sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}
\end{aligned}$$

A Gauss–Markov tétel

- Ha mindegyik feltevés teljesül, akkor lineáris torzítatlan becslők körében az OLS-becslő minimális varianciájú (azaz hatásos)
- Tehát: $\mathbb{D}^2\left(\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}}\right) \leq \mathbb{D}^2\left(\widehat{\boldsymbol{\beta}'}\right)$ bármely más $\widehat{\boldsymbol{\beta}'}$ lineáris becslőre, amire $\mathbb{E}\left(\widehat{\boldsymbol{\beta}'}\right) = \boldsymbol{\beta}$ (azaz torzítatlan)

Emlékeztetőül, ha $\mathbf{A}$ és $\mathbf{B}$ négyzetes mátrixok, akkor abban az esetben mondjuk, hogy $\mathbf{A} \geq \mathbf{B}$ ha $\mathbf{A} - \mathbf{B}$ pozitív szemidefinit.

Összefoglalva

- Amennyiben a standard modellfeltevések közül teljesül a:
  - Linearitás
  - Nincs egzakt multikollinearitás
  - Erős exogenitás

  akkor az OLS szolgáltatatta becslések *torzítatlanok* és *konzisztensek*
- Ha ezen felül teljesül a:
  - Homoszkedaszticitás
  - Autokorrelálatlanság

  akkor az OLS szolgáltatta becslések *hatásosak* (minimális varianciájuk) is

BLUE-tulajdonság

Ezt röviden úgy szokták megfogalmazni, hogy ha valamennyi standard modellfeltétel teljesül, akkor az OLS szolgáltatta becslések BLUE-k:

- Best (minimális varianciájú)
- Linear (lineáris a mintaelemekben)
- Unbiased (torzítatlan)

A σ^2 és a koefficiensek kovarianciamátrixának becslői

- A σ^2-nek a $\widehat{\sigma^2} = \frac{ESS}{n-(k+1)}$ becslője torzítatlan, ha mindegyik feltétel fennáll
- A β_i koefficiensek kovarianciamátrixának $\widehat{\sigma^2}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}$ becslője szintén
- Tehát vigyázat: itt *már* a torzítatlansághoz *is* kell mindegyik feltétel (a homoszkedaszticitás és az autokorrelálatlanság is)!

A $\widehat{\beta_i}$ koefficiensek eloszlása

- Az eddigi eredmények ugyan nagyon biztatóak, de még mindig nem mondanak semmit arról, hogy konkrétan mi a becsült koefficiensek (mintavételi) eloszlása
- A $\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \boldsymbol{\beta} + \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}$ nem sok jót sejtet: ebből úgy tűnik, hogy ez $\underline{\underline{X}}$-től és ε-tól is függ, ráadásul egy elég komplexnek kinéző módon…
- Szerencsére nem ennyire rossz a helyzet!
- Van egy nevezetes speciális eset, amikor a becsült koefficiensek eloszlása egyszerű alakú, és *nem is függ* $\underline{\underline{X}}$ eloszlásától, ez pedig az, ha a hibák feltételes eloszlása normális
- Vigyázat: a hibák normalitása *nem* része a standard modellfeltevéseknek, azaz a BLUE-ság akkor is megvalósul, ha a hibák eloszlása nem normális!
- Ráadásul, még ha nem is tudjuk, hogy a normalitás teljesül, de nagy a mintánk, akkor a centrális határeloszlás-tétel miatt aszimptotikus közelítésként akkor is használhatjuk az így nyert eredményeket

Hibák normalitása

- $\underline{\varepsilon}$ feltételes eloszlása feltéve $\underline{\underline{X}}$-et többváltozós normális
- A standard modellfeltevéseket is felhasználva ez azt jelenti, hogy

$$\underline{\varepsilon} \mid \underline{\underline{X}} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

- Ez láthatóan nem függ $\underline{\underline{X}}$-től, így persze a hibák feltétel nélküli eloszlása is $\mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$

Hibanormalitás és a becsült koefficiensek eloszlása

- Ha $\underline{\varepsilon}$ eloszlása normális, akkor $\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}$-é is az
- Ez azért nagyon jó hír, mert a normális eloszláshoz csak két dolgot kell tudnunk: várható érték-vektort és kovarianciamátrixot!
- Az viszont könnyen meghatározható (az egyszerűség kedvéért a $\mid \underline{\underline{X}}$ feltételt nem írjuk ki a következőkben)
- Várható érték: $\mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right] = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\mathbb{E}\underline{\varepsilon} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\mathbf{0} = \mathbf{0}$
- Kovarianciamátrix: $\mathbb{D}^2\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\varepsilon}\right] = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\cdot\mathbb{D}^2\varepsilon\cdot\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\cdot\sigma^2\mathbf{I}\cdot\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1} = \sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}$
- Összefoglalva: $\widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\right)$

Emlékezzünk rá, hogy $\mathbb{E}\left(\mathbf{A}\underline{X}\right) = \mathbf{A}\cdot\mathbb{E}\underline{X}$ és $\mathbb{D}^2\left(\mathbf{A}\underline{X}\right) = \mathbf{A}\cdot\mathbb{D}^2\underline{X}\cdot\mathbf{A}^T$.

A fenti levezetésekben az $\underline{\underline{X}}$-ket tartalmazó kifejezések azért viselkednek úgy, mint a konstans mátrixok, mert rá feltételeztünk! Csak a rövidség kedvéért nem írtuk ki.

Konfidenciaintervallum a paraméterekre

Hibanormalitás esetén, vagy aszimptotikusan könnyen szerkeszthető konfidenciaintervallum is, $1-\alpha$ megbízhatósági szinten:

$$\widehat{\beta}_i \pm t_{n-(k+1)}^{(1-\alpha/2)}\cdot\mathrm{se}\left(\widehat{\beta}_i\right)$$

5 Hipotézisvizsgálat és intervallumbecslés lineáris modellben

5.1. Alkalmazási feltételek

Emlékeztetőül

- A most következő eredmények csak akkor egzaktak, ha a hibanormalitás is fennáll
- Ám aszimptotikusak, így közelítőleg akkor is fennálnak, ha elég nagy a mintanagyság (minél nagyobb, annál inkább)

Ez a centrális határeloszlás-tétel miatt van így.

5.2. Egy paraméter

Becsült regressziós koefficiensek mintavételi eloszlása

- A $\widehat{\beta}_i$ becsült regressziós koefficiens mintavételi ingadozását tehát a következő összefüggés írja le:

$$\frac{\widehat{\beta}_i - \beta_i}{\operatorname{se}\left(\widehat{\beta}_i\right)} \sim \mathcal{N}(0,1),$$

  ahol $\operatorname{se}\left(\widehat{\beta}_i\right) = \sqrt{\sigma^2 \left[\left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1}\right]_{kk}}$

- Sajnos ezzel a gyakorlatban nem sokra megyünk, mert σ^2-et általában nem ismerjük
- Helyettesítsük a jó tulajdonságú becslőjével, $\widehat{\sigma^2}$-tel!
- Így persze már más lesz az eloszlás, de szerencsére meghatározható, hogy mi, és nem bonyolult: $n-(k+1)$ szabadságfokú t-eloszlás

Változó relevanciája

Egy változót relevánsnak nevezünk, ha a sokasági paramétere nem nulla: $\beta_i \neq 0$.

Hipotézisvizsgálat változó relevanciájára

Ez alapján már konstruálhatunk próbát változó relevanciájának vizsgálatára:

1. $H_0 : \beta_i = 0$
2. Ekkor (azaz *ha* ez fennáll!) a $t_{\text{emp},i} = \frac{\widehat{\beta}_i}{\text{se}\left(\widehat{\beta}_i\right)}$ kifejezés $n - (k+1)$ szabadságfokú t-eloszlást követ (nulleloszlás)
3. Számítsuk ki a konkrét $t_{\text{emp},i}$-t a mintánkból és döntsük el, hogy hihető-e, hogy $t_{n-(k+1)}$-ból származik

Hipotézisvizsgálat változó relevanciájára

A hipotézisvizsgálat elvégzéséhez szükséges minden tudnivalót – a nullhipotézisen kívül – összefoglal tehát a következő kifejezés (a későbbiekben is ezt a sémát fogjuk használni hipotézisvizsgálatok megadására):

$$t_{\text{emp},i} = \frac{\widehat{\beta}_i}{\text{se}\left(\widehat{\beta}_i\right)} \overset{H_0}{\sim} t_{n-(k+1)}.$$

E próba precíz neve: változó relevanciájára irányuló (parciális) t-próba

5.3. Modell egésze

Modell egészének relevanciája

- A korábban látott t-próba azért volt „parciális”, mert egy változó irrelevanciáját vizsgálta
- Felmerül a kérdés, hogy definiálható-e a modell *egészének* irrelevanciája
- Igen, mégpedig úgy, hogy *valamennyi* magyarázó változó paramétere *együttesen is* irreleváns:
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
- (Természetesen a β_0 nincs felsorolva!)
- Rövid jelölés arra, hogy $\beta_1 = 0$ *és* $\beta_2 = 0$ stb. *és* $\beta_k = 0$ (*semmilyen más* eset jelölésére *ne* használjuk az egyenlőségláncot!)
- Figyelem: az „egyszerre nulla mindegyik” *több* mint, hogy „külön-külön nulla mindegyik”!

Modell egészének relevanciája

- A modell egészének irrelevanciájára magyarul azt jelenti, hogy a modell nem tér el lényegesen a nullmodelltől
- Implikálja, hogy minden magyarázó változó külön-külön is irreleváns (tartalmazza ezeket a hipotéziseket) → előbb teszteljük a modell egészének irrelevanciáját, és csak ennek elvetése utána teszteljük a változókat parciálisan
- A próba konkrét alakja:

$$F_{\text{emp}} = \frac{RSS/k}{ESS/\left[n-(k+1)\right]} \overset{H_0}{\sim} \mathcal{F}_{k,n-(k+1)}$$

Modell egészének relevanciája

- A tesztstatisztika átírható mint

$$\frac{RSS/k}{ESS/\left[n-(k+1)\right]} = \frac{R^2/k}{\left(1-R^2\right)/\left[n-(k+1)\right]}$$

- Persze: a „nem tér el lényegesen a nullmodelltől” úgy is megfogalmazható, hogy az „R^2 nem tér el lényegesen a nullától” ($H_0 : R^2 = 0$ is mondható lett volna)

Modell egészének relevanciája

- A próba neve: a modell egészének relevanciájára irányuló (globális) F-próba
- Szokás ANOVA-próbának is nevezni (a $TSS = ESS + RSS$ variancia-felbontáson alapszik; számlálóban és nevezőben a fokszámmal normált szórásnégyzetek vannak)
- Tipikus eredményközlés az ún. ANOVA-táblában

5.4. Tetszőleges számú paraméter

Felvezető gondolatok

- Valamennyi eddigi próba felírható úgy, hogy van egy modellünk, a nullhipotézis pedig egy megkötést jelent arra a modellre
- Azaz lényegében két modellünk van, egy megkötés nélküli és egy megkötött
- Mellesleg a megkötött modell szükségképp rosszabb, de legalábbis nem jobb (szűkebb tartományon vett optimum nem lehet jobb, mint egy bővebben vett), emiatt úgy is megfogalmazható a kérdés, hogy a különbség lényeges-e
- Az ilyen helyzetre – mint bármilyen helyzetre – többféle elven lehet tesztet konstruálni
- Wald-elv, LM-elv, LR-elv
- Az eddigi két próba Wald-elven is kihozható

Tetszőleges számú paraméter tesztelése Wald-elven

- Most felírjuk a két modellt explicite is, mert a nullhipotézis alakja szebb lesz (ez pusztán formai kérdés):
- Az egyik modell a bővebb (U – unrestricted), a másik a szűkebb (R – restricted):

$$U: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+m} X_{q+m} + \varepsilon_U$$
$$R: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R$$

- $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$, tehát megadott m darab változó *még összességében sem* bír lényeges magyarázó erővel

Tetszőleges számú paraméter tesztelése Wald-elven

A próba:

$$F_{\text{emp}} = \frac{\left(ESS_R - ESS_U\right)/m}{ESS_U/\left(n-q-m\right)} = \\ = \frac{\left(R_U^2 - R_R^2\right)/m}{\left(1-R_U^2\right)/\left(n-q-m\right)} \overset{H_0}{\sim} \mathcal{F}_{m,n-q-m}.$$

Ebből a felírásból látszik jól, hogy ez a teszt úgy is felfogható, mint ami a többszörös determinációs együtthatók különbségét ítéli meg.

Speciális esetek

- Vegyük észre, hogy ez az általános megközelítés a két, eddig látott tesztet is tartalmazza speciális esetként!
- Ha $m = 1$, akkor $F = t_j^2$: visszakaptuk a t-tesztet
  - Ám figyelem: a Wald-teszt *nem* ekvivalens a t-próba m-szeri elvégzésével (külön-külön az egyes változókra)!
- Ha $m = k$, akkor $F_{\text{Wald}} = F_{\text{ANOVA}}$: visszakaptuk a függetlenségvizsgálatot
- Logikusak, hiszen a nullhipotézisek is azonos alakúak lettek

Az első állításhoz hozzá kell még tenni, hogy az $(1, n-k)$ paraméterű F-eloszlás épp az $n-k$ szabadságfokú t-eloszlás négyzetével esik egybe.

Kitérő: a Lagrange Multiplikátor (LM)-elv

- Az LM (Lagrange Multiplikátor) próba hipotézispárja *teljesen* azonos alakú a Wald-F-teszttel:

$$U: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+m} X_{q+m} + \varepsilon_U$$
$$R: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1} + \beta_q X_q + \varepsilon_R$$

  és $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$

- A különbség a modellezés filozófiájában van (ld. később), a teszt tulajdonságai, alkalmazhatósága is eltérő
- Alapötlet: becsüljük meg a szűkebb modellt, és számítsuk ki ez alapján a becsült reziduumokat. Ha fennáll H_0, akkor ezek a reziduumok nem magyarázhatóak lényegesen sem a szűkebb modell változóival (OLS következménye), sem a vizsgált változókkal (H_0 következménye). Azaz: ha a becsült reziduumokat kiregresszáljuk az összes változóval, akkor sem tudjuk azt lényegesen magyarázni, *ha* fennáll a H_0.

Az LM-próba próbafüggvénye

- Ezen intuitív indoklás után a próbafüggvény:

$$n \cdot R^2_{\widehat{e}_R|X_1,X_2,\ldots,X_{q+m}} \overset{H_0}{\sim} \chi^2_m$$

- Itt $\widehat{e}_R$ jelölés arra utal, hogy a szűkebb (R) modellből kapott reziduumokról van szó

5.5. Lineáris megkötés(ek)

Lineáris kombináció tesztelése

- A séma:

$$r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = r$$

- Avagy röviden: $\mathbf{r}^T\boldsymbol{\beta} = r$
- Több koefficienst is érinthet, de csak egy egyenletet tartalmazhat
- Például:
  - Két koefficiens egyezik, $\beta_l = \beta_m$ (ekkor $r_l = +1$, $r_m = -1$, a többi r_i nulla és $r = 0$)
  - Egyik koefficiens c-szerese a másiknak, $\beta_l = c\beta_m$ (ekkor $r_l = +1$, $r_m = -c$, a többi r_i nulla és $r = 0$)
  - Az összes koefficiens összege épp nulla (ekkor mindegyik r_i 1 és $r = 0$)

Lineáris kombináció tesztelése

- A normális lineáris modellben erre teszt szerkeszthető
- Megvalósítás: egyik lehetőség, hogy a t-próbához hasonló alakra vezetjük vissza
- Legyen $r_1\widehat{\beta}_1 + r_2\widehat{\beta}_2 + \ldots + r_k\widehat{\beta}_k = \widehat{r}$, ekkor

$$\frac{\widehat{r} - r}{\operatorname{se}(\widehat{r})} \overset{H_0}{\sim} t_{n-(k+1)}$$

- Ez az ún. *közvetlen t-próba*
- Vizsgálható Wald-jellegű próbával is

Speciális esetek

- Ez tartalmazza speciális esetként a parciális t-próbát
- De mást nem: kettő vagy több paraméter *egyidejű* nulla mivolta több megkötést jelent
- Szerencsére az előbbi kiterjeszthető több megkötés tesztelésére is:

$$\begin{aligned}
\mathbf{r_1}^T\boldsymbol{\beta} &= r_1 \\
\mathbf{r_2}^T\boldsymbol{\beta} &= r_2 \\
&\vdots \\
\mathbf{r_m}^T\boldsymbol{\beta} &= r_m
\end{aligned}$$

- Az $\mathbf{r_i}^T$ sorvektorokat rakjuk össze egy $\mathbf{R}$ mátrixba, az r_i skalárokat egy r oszlopvektorba

Több megkötés egyidejű tesztelése

- Célszerű felírás:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

  ahol $\mathbf{R}$ $m \times k$ típusú (tehát m a megszorítások száma)

- Az erre adható teszt:

$$F_{\text{emp}} = \frac{\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)/m}{\text{ESS}/\left[n-(k+1)\right]} \overset{H_0}{\sim}$$
$$\overset{H_0}{\sim} \mathcal{F}\left[m, n-(k+1)\right]$$

Feltétel még, hogy $\mathbf{R}$ teljes sorrangú legyen ($\operatorname{rank}\mathbf{R} = m$), ami azt a kézenfekvő követelményt fogalmazza meg, hogy a megszorítások ne legyenek (lineáris értelemben) redundánsak.

Konkrét példák a fenti sémára

- Ellenőrizhető, hogy ha például…
  - …$\mathbf{R} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 0 \dots 0 \end{pmatrix}$ és $r = 0$, akkor a t-tesztet …
  - …$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$ és $\mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ akkor az ANOVA-t…
  - …$\mathbf{R} = \begin{pmatrix} \lambda_{\beta_1} & \lambda_{\beta_2} & \dots & \lambda_{\beta_k} \end{pmatrix}$ és $r = \Lambda$, akkor a lineáris kombináció tesztelését…
- …kapjuk vissza.

Speciális esetek

- Ez a képlet viszont *minden* eddig látott dolgot tartalmaz speciális esetként!
- Wald-elven

6 Kategoriális magyarázó változók

6.1. Regresszió csak minőségi változóval (ANOVA)

Minőségi változók a regresszióban

- A kérdés, ami mostani kutatásainkat motiválja: hogyan szerepeltethetünk egy *minőségi* (nominális vagy ordinális, szokás kategoriális változónak is nevezni) tulajdonságot, pl. férfi–nő, egészséges–beteg, alapfokú–középfokú–felsőfokú végzettségű stb. egy regressziós modellben
- A regresszió csak számszerű adatokat tud felhasználni → valahogy *kódolni* kell a kategoriális tulajdonság lehetséges értékeit (kimeneteit, csoportjait)
- Eddig csak mennyiségi tulajdonságokkal foglalkoztunk, aminek kódolása triviális volt: a naturáliában kifejezett értékével (m^2, eFt stb.)
- Pl. férfi = 0, nő = 1 elég kézenfekvő, de mi van az iskolai végzettséggel?
- Az alap = 0, közép = 1, felső = 2 belekódolja az adatokba, hogy a felső és a közép közti különbség *kénytelen* ugyanakkora lenni, mint a közép és alap közötti (ha felső = 3, akkor kétszer akkora stb.)
- De mi semmi ilyet nem akarunk, hiszen azt szeretnénk, hogy ezt az adatok mondják meg!

Dummy változó fogalma

- A kódolást megvalósíthatjuk olyan változóval vagy változókkal, melyek *csak* 0 vagy 1 értéket vehetnek fel
- Az ilyen változókat nevezzük dummy (bináris vagy indikátor) változónak
- Ha két kimenet van, akkor a kódolás teljesen kézenfekvő: egy dummy változóra van szükségünk, mely (például) 0 értéket vesz fel férfira, 1-et nőre
- Bonyolultabb a helyzet, ha több kimenet van
- Triviális kódolás:

	D_A	D_B	D_C
A	1	0	0
B	0	1	0
C	0	0	1

Kódolás

Ezen „kódolási tábla” alapján a kódolás (pl. X_1: jövedelem, X_2: iskolai végzettség, X_3: életkor):

$$\begin{array}{c} \begin{matrix} & X_1 & X_2 & X_3 \end{matrix} \\ \begin{pmatrix} 1 & 213 & B & 32 \\ 1 & 311 & C & 41 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 128 & B & 18 \end{pmatrix} \end{array} \rightsquigarrow \begin{array}{c} \begin{matrix} & X_1 & D_A & D_B & D_C & X_3 \end{matrix} \\ \begin{pmatrix} 1 & 213 & 0 & 1 & 0 & 32 \\ 1 & 311 & 0 & 0 & 1 & 41 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 128 & 0 & 1 & 0 & 18 \end{pmatrix} \end{array}$$

Itt már minden tisztán numerikus, működhet a regresszió

Referencia-kódolás

- …ám vegyük észre, hogy 3 csoporthoz *nem* kell 3 dummy változó, kódolható 2-vel is!
- Általában k kimenet kódolása megoldható $k-1$ dummy változóval az ún. referencia-kódolás logikájával
- Itt kiválasztunk egy kimenetet, aminél mind a $k-1$ darab dummy változó 0 értéket vesz fel (kontrollcsoport vagy referenciacsoport), és a többi $k-1$ csoportot az jelzi, hogy a $k-1$ dummy változó közül *melyik* vesz fel 1 értéket (mindig csak 1!)
- Például (3 kimenetre):

	R_A	R_B
A	1	0
B	0	1
C	0	0

- Itt C a referenciacsoport, R_A és R_B a két szükséges (ugye $k = 3$!) magyarázó változó
- Vegyük észre, hogy $R_A \equiv D_A$ és $R_B \equiv D_B$ (tehát a két kódoláshoz pontosan ugyanazon dummykra van szükség, csak a referencia-kódolásnál eldobjuk az egyiket – ez lesz a kontrollcsoport)

Dummy változó csapda

- Ha van konstans a modellben, akkor *tilos* is k csoporthoz k dummyt használni a kódoláshoz
- Ellenkező esetben egzakt multikollinearitás jön létre (gondoljuk végig, hogy a dummy változókhoz mi tartozik a design mátrixban, ld. előbb!); ez az ún. *dummy változó csapda*
- Ha k csoportot mégis k dummyval kódolunk („triviális kódolás”), akkor viszont nem szerepeltethetünk konstanst

Triviális kódolás konstans nélkül

- A két kódolási mód (k darab dummy, nincs konstans és $k-1$ darab dummy, van konstans) jól szemléltethető egy csak a nominális tulajdonsággal magyarázó regresszióval
- k darab dummy, nincs konstans:

	D_A	D_B	D_C
A	1	0	0
B	0	1	0
C	0	0	1

$$Y = \beta_A D_A + \beta_B D_B + \beta_C D_C + \varepsilon$$

- Együtthatók értelmezése: ha az A csoportban vagyunk, akkor a fenti egyenlet $Y = \beta_A + \varepsilon$ lesz $\Rightarrow$ β_A az A csoport csoportátlaga (legkisebb négyzetes elv!); hasonlóan a többi

Referencia-kódolás konstanssal

- $k-1$ darab dummy, van konstans:

	D_A	D_B
A	1	0
B	0	1
C	0	0

$$Y = \beta^* + \beta_A^* D_A + \beta_B^* D_B + \varepsilon$$

Együtthatók értelmezése referencia-kódolásnál

- Értelmezésnél egy dolgot tartsunk mindig szem előtt: ugyanarra a csoportra ugyanannak az értéknek kell kijönnie, akárhogy kódolunk!
- Így $\beta_C = \beta^*$
- Továbbá (a B csoport példáján):

$$\beta_B = \beta^* + \beta_B^* = \beta_C + \beta_B^* \Rightarrow \beta_B^* = \beta_B - \beta_C$$

- Tehát az együtthatók az *eltéréseket* jelentik a referenciacsoporttól (ami pedig a konstansba kerül)
- Vegyük észre, hogy a változónkénti szignifikanciák eltérhetnek (mert másra fognak vonatkozni), de az előrejelzése – és így a modellminősítő mutatók – nem

Fontos hipotézisvizsgálatok

- Egyrészt: szignifikáns-e egy adott csoport átlagának eltérése a referenciacsoport átlagától
- Ez itt nem más, mint β_A^* vagy β_B^* relevanciája
- Egyszerűen t-próbával ellenőrizhető!
- Másrészt: van-e egyáltalán bármilyen csoportok közötti eltérés:

$$H_0 : \beta_A^* = \beta_B^* = \ldots = 0$$
$$H_1 : \exists j : \beta_j^* \neq 0$$

- Több csoport átlaga eltér-e? De hát az az ANOVA!
- Az egyezés nem pusztán formai, teljes tartalmazi egyezés van (ez nem csak hasonló, hanem ugyanaz: az ANOVA elmondása regressziós „keretben")

Egynél több kategoriális magyarázó változó

- Ha egynél több kategoriális magyarázó változó van, akkor nem kódolható mindegyik triviálisan, ilyenkor már a konstans eltávolítása sem segít
- (Nem az lesz a baj, hogy valamelyik összege a konstans, hanem, hogy a kettő összege ugyanaz – ez elvileg is megoldhatatlan)
- Referencia-kódolás minden további nélkül használható
- A kétszempontos ANOVA megfelelője regressziós keretben!
- Természetesen feltételezhető interakció is, ez esetben a dummy-kat az összes lehetséges kombinációban szorozni kell

6.2. Regresszió minőségi és mennyiségi magyarázó változóval (ANCOVA)

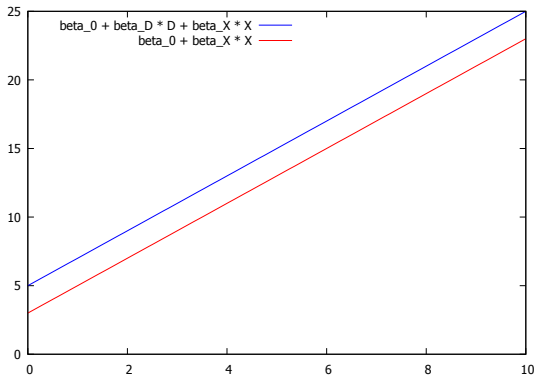
Dummyzás folytonos magyarázó változó jelenléte mellett

- Amit eddig csináltunk az lényegében az volt, amit *konstans dummyzásának* nevezhetünk: csoportonként eltérő (de konstans) értékkel becsültük az eredményváltozót
- Mi van, ha bevonunk egy magyarázó változót?
- Azaz ekkor már nem egy konstanst becsülünk az egyes csoportokra, hanem egy egyenest (a folytonos magyarázó változó függvényében)

- Dummyzással (tehát a csoporttagság szerint) eltéríthetjük az egyenesek tengelymetszetét és meredekségét is!
- Lehet csoportonként különböző
  1. +1 egység magyarázó változó hatása
  2. a 0 magyarázó változóhoz tartozó eredményváltozó
- E feladat neve: ANCOVA

Eltérő tengelymetszet

Ha csak a tengelymetszetet térítjük el (+1 egység magyarázó változó hatása ugyanaz minden csoportban, de nem ugyanannyi a 0 magyarázó változóhoz tartozó eredményváltozó):

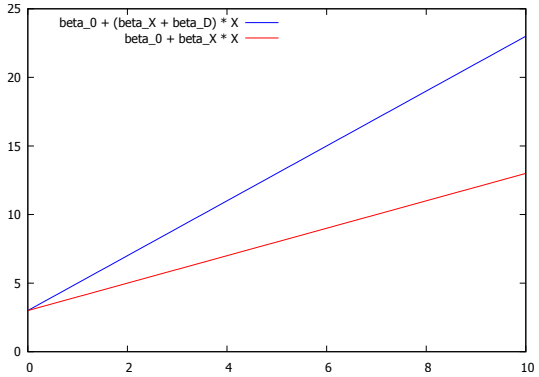


Algebrailag:

$$Y = \beta_0 + \beta_D D + \beta_X X + \varepsilon$$

Eltérő meredekség

Ha csak a meredekséget térítjük el (0 magyarázó változóhoz ugyanaz az eredményváltozó tartozik, de +1 egység magyarázó változó hatása csoportonként eltérő):



Algebrailag:

$$Y = \beta_0 + (\beta_X + \beta_D D)\, X + \varepsilon$$

Eltérő tengelymetszet és meredekség

- Akár a tengelymetszet *és* a meredekség is lehet különböző
- Ahogy előbb láttuk, csak a módszereket kell kombinálni: a konstanst és a meredekséget is megdummyzzuk:

$$Y = \beta_1 + \beta_2 X + \varepsilon,$$

de úgy, hogy $\beta_1 = \alpha + \alpha_A D_A + \alpha_B D_B$ és $\beta_2 = \gamma + \gamma_A D_A + \gamma_B D_B$

- Nagyon fontos észrevenni, hogy a meredekség dummyzása a dummy és a mennyiségi változó közti interakcióra vezet:

$$Y = \alpha + \alpha_A D_A + \alpha_B D_B + \gamma X + \gamma_A (D_A X) + \gamma_B (D_B X) + \varepsilon$$

- Logikus is: az egyik változó (folytonos) hatása eltér a szerint, hogy a másik változónak (kategoriális) mi a szintje: különböző meredekségek
- Avagy fordítva elmondva (egyenértékűen, hiszen az interakció ugye szimmetrikus): az egyik változó (kategoriális) hatása eltér a szerint, hogy a másik változónak (folytonos) mi a szintje: az egyenesek közti különbség függ attól, hogy hol nézzük

Eltérő tengelymetszet és meredekség

- De hát ez megoldható a minta szétszedésével is!
- A két módszer – természetesen – ugyanarra az eredményre vezet
- A dummyzás mégis jobb a minta szétszedésénél; vajon miért?
  - Messzemenően több lehetőségünk van a dummyzott (egybenlévő) modellel → gazdaságilag releváns hipotézisek vizsgálhatóak egyszerűen (ld. mindjárt)

Hipotézisvizsgálat a dummyzott modellben

- Pl.: van-e egyáltalán bármilyen eltérés a csoportok között? (Értsd: eltér-e a becsült egyenes (bármilyen szempontból) a csoportok között, vagy mindegyikben teljesen ugyanaz?)
- Ez az ún. *strukturális törés*, hipotézispárja: $H_0 : \alpha_A = \alpha_B = \gamma_A = \gamma_B = 0$, H_1: valamelyik ezek közül nem nulla, tehát van strukturális törés
- És most jön a szép rész: ha a fenti modellt megbecsültük (sima OLS-sel), akkor ez a hipotézis egyszerűen egy közönséges Wald- (vagy hasonló) próbát jelent!
- Hasonlóképp: nem lehet, hogy csak a tengelymetszetek eltérőek? → ez az ún. *párhuzamos ráták* hipotézise, $H_0 : \gamma_A = \gamma_B = 0$; szintén Wald-teszttel elintézhető
- Minden hasonló, gazdasági kérdés *lefordítható* ökonometriailag, például változó vagy változók relevanciájának tesztelésére

Kontraszt-kódolás

- Kontraszt-kódolás: trükkös kódolás úgy kitalálva, hogy a dummy-k együtthatója ne a referencia-csoporthoz, hanem az átlaghoz képesti eltérést jelentse
- A megoldás:

	C_A	C_B
A	1	0
B	0	1
C	-1	-1

- (A dummy változó nem 0 és 1 értéket vehet csak fel)
- Miért fog ez működni?

Kontraszt-kódolás

Mert:

$$\beta_0 + \beta_{C_A} + 0 = \overline{y}_A \tag{6.1}$$
$$\beta_0 + 0 + \beta_{C_B} = \overline{y}_B \tag{6.2}$$
$$\beta_0 - \beta_{C_A} - \beta_{C_B} = \overline{y}_C \tag{6.3}$$

És így:

- $(1)+(2)+(3) \Rightarrow 3\beta_0 = \overline{y}_A+\overline{y}_B+\overline{y}_C \Rightarrow \beta_0$ tényleg a főátlag (ha azonosak a csoportok elemszámai! különben ún. súlyozott kontraszt kellene, ahol a dummy változók már nem is feltétlenül egész értékeket vennének fel)
- $(2)+(3) \Rightarrow 2\beta_0 - \beta_{C_A} = \overline{y}_B + \overline{y}_C \Rightarrow \beta_{C_A} = 2\beta_0 - (\overline{y}_B + \overline{y}_C) = 2\beta_0 - (3\beta_0 - \overline{y}_A) \Rightarrow \beta_{C_A} = \overline{y}_A - \beta_0 \Rightarrow$ tényleg az átlagtól való eltérés (és hasonlóan a másik)

Egy terminológiai megjegyzés

- Az angol irodalomban az általunk kontrasztkódolásnak nevezett módszert nagyon gyakran „effect coding”-nak nevezik…
- …a kontraszt pedig az, amikor a csoportok tetszőleges – általunk meghatározott – lineáris kombinációját teszteljük

7 Regressziós modellek alternatív becslési lehetőségei

7.1. A maximum likelihood (ML) elv

A likelihood fogalma

- Likelihood: folytonos változónál a sűrűségfüggvény helyettesítési értéke adott ponton, diszkrétnél a valószínűség (tehát a valószínűségi súlyfüggvény értéke adott ponton)
- (Lehet többdimenziós is)
- Nem mondhatunk valószínűséget helyette, hiszen folytonos esetben nem valószínűségről van szó (csak gondoljunk bele, a sűrűség simán lehet 1-nél nagyobb)
- Részletesebben: folytonos változó minden adott konkrét értéket nulla valószínűséggel vesz fel!
- De a likelihood mégis bír kézzelfogható értelmezéssel: ezzel arányos valószínűséggel esik a valószínűségi változó az adott pont kis környezetébe
- (A '80-as években kísérleteztek a „valószerűség" szó meghonosításával erre…)

A maximum likelihood (ML) elv

- Ha ismerjük a sokasági paraméterek értékét *és* egy modellt a mintavételre, akkor meg tudjuk mondani, hogy adott minta kijövetelének mekkora a likelihood-ja
- (Hiszen a modell épp ezt írja le, csak épp függ a sokasági paraméterektől, de ha azokat is ismerjük, akkor már egy konkrét számot kapunk)
- Az ML-becslés alapgondolata: ha nem ismerjük a sokasági paramétereket, akkor válasszuk azokat becslésnek, amely mellett a lehető legnagyobb a likelihood-ja annak, hogy az a minta jöjjön ki, ami ténylegesen ki is jött
- Józan paraszti ésszel is ésszerű, de nagyon fontos felhívni a figyelmet, hogy ez *nem* ugyanaz, mint hogy azt határozzuk meg, hogy mely paramétereknek a legnagyobb a likelihoodja (fordított a feltételezés iránya!)
- (Ez az igazán kézenfekvő kérdés, csakhogy ez igényel ismeretet arról, hogy a paraméterek eloszlása milyen még mielőtt egyáltalán mintát vettünk volna – itt válik el a frekvencionista és a bayes-i statisztika)

Egyszerű várható érték becslés ML-elven

- Csináljuk meg újra ugyanazt az egyszerű példát amit az OLS-elvnél láttunk, de immár ML-elvű becsléssel!
- Emlékeztetőül a modellünk: $Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right)$ (ahol σ_0^2 ismert) és erre van egy n elemű fae mintánk
- A valódi várható érték μ, a feltételezett várható értéket, ami végigfut majd az összes lehetséges értékén, jelölje m (és a legjobb érték, amit becslésként elfogadunk $\widehat{\mu}$)
- Minta igazából most az n elem együtt, tehát a minta likelihood-ja ennek az – n-dimenziós – folytonos sűrűségfüggvénynek a helyettesítési értéke

Egyszerű várható érték becslés ML-elven

- Szerencsére fae esetben ez szétesik szorzattá:

$$L_m(\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}) = f_m(\mathbf{y}) = \prod_{i=1}^{n} f_m(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot e^{-\frac{(y_i-m)^2}{2\sigma_0^2}} =$$

$$= \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}} \cdot e^{-\sum_{i=1}^{n} \frac{(y_i-m)^2}{2\sigma_0^2}}$$

- Célszerűbb ennek a maximalizálása helyett inkább a logaritmusát maximalizálni (a logaritmus monoton transzformáció, így a logaritmált maximum ugyanott van, mint az eredeti maximuma):

$$l_m(\mathbf{y}) = \log L_m(\mathbf{y}) = -\frac{n}{2}\log\left(2\pi\sigma_0^2\right) + \sum_{i=1}^{n} -\frac{(y_i-m)^2}{2\sigma_0^2}$$

Egyszerű várható érték becslés ML-elven

- A maximalizáláshoz deriváljuk m szerint:

$$\frac{\mathrm{d}l_m(\mathbf{y})}{\mathrm{d}m} = -\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} 2(y_i - m)\cdot(-1)$$

- Tegyük egyenlővé nullával és oldjuk meg:

$$-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} 2(y_i - m)\cdot(-1) = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - m) = 0$$

$$\Rightarrow \widehat{\mu_{\mathrm{ML}}} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- (A második derivált pedig $-\frac{n}{2\sigma_0^2}$, negatív, tehát ez tényleg szélsőérték, és tényleg maximum)

Az ML-becslés általában

$$\widehat{\theta_{\mathrm{ML}}} = \arg\max_{\mathbf{t}} L_{\mathbf{t}}(\mathbf{y})$$

- Ha a mintánk fae, akkor:

$$\widehat{\theta_{\mathrm{ML}}} = \arg\max_{\mathbf{t}} \sum_{i=1}^{n} l_{\mathbf{t}}(y_i)$$

- Az ML-becslés valószínűségi modellt igényel (az OLS-becslésnél ez nem volt így, igazából elég volt az, hogy legyen egy predikált értékünk, az kijöhetett bárhogy), de cserében egy sor kellemes tulajdonsággal bír
- (*Általában*, tehát pusztán amiatt, hogy ML-becslésről van szó: függetlenül attól, hogy mi a probléma, ezeket a tulajdonságokat pusztán az ML-becslés mivolt miatt megkapjuk!)

Az ML-becslés tulajdonságai

Jelesül egy ML-becslés elég általános körülmények között

- konzisztens
- aszimptotikusan torzítatlan
- aszimptotikusan hatásos
- aszimptotikusan normális
- invariáns ($g(\theta)$ ML-becslése $g\left(\widehat{\theta_{\mathrm{ML}}}\right)$, ha g egy kölcsönösen egyértelmű függvény)

Lineáris modell becslése ML-elven

- De ezt az ML-elvet nem lehetne a sima lineáris modellre is ráereszteni (hogy megbecsüljük, de másképp mint eddig, OLS-elven)? Dehogynem!
- Egyetlen dologra van szükségünk: itt *mindenképp* fel kell tennünk valamit a hibatag eloszlásáról (különben az eredményváltozóra nem lesz eloszlásunk, anélkül pedig likelihood-unk sincs)
- Tegyük fel a normalitást (és a szferikális hibákat, tehát, hogy az $\underline{\varepsilon}$ kovarianciamátrixa $\sigma^2\mathbf{I}$)
- Ekkor ugyanis $\underline{Y}$ eloszlása is normális, mégpedig $\underline{\underline{X}}\mathbf{b}$ várhatóértékkel és $\sigma^2\mathbf{I}$ kovarianciamátrixszal (hiszen $\underline{Y} = \underline{\underline{X}}\mathbf{b} + \underline{\varepsilon}$)
- Mivel azt mondtuk, hogy $\mathbf{b}$ az (ismeretlen) sokasági paraméterek éppen feltételezett értéke; ebben fogunk majd maximalizálni
- Írhattuk volna azt is, hogy $Y_i = \underline{X}_i^T\mathbf{b} + \varepsilon_i$ és hozzátesszük, hogy a különböző megfigyelési egységek függetlenek

Lineáris modell becslése ML-elven

- A log-likelihood (a második megközelítésből felírva):

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right)+\sum_{i=1}^{n}-\frac{\left(y_i-\underline{X}_i^T\mathbf{b}\right)^2}{2\sigma^2}$$

- A log-likelihood (az első megközelítésből felírva, a többváltozós normális sűrűsége $\frac{1}{(2\pi)^{n/2}\det\mathbf{C}^{1/2}}\cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$):

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(\underline{y}-\underline{\underline{X}}\mathbf{b}\right)^T\left(\underline{y}-\underline{\underline{X}}\mathbf{b}\right)$$

- Ha ezt **b**-ben maximalizáljuk, az ugyanaz, mint $\left(\underline{y}-\underline{\underline{X}}\mathbf{b}\right)^T\left(\underline{y}-\underline{\underline{X}}\mathbf{b}\right)$-t **b**-ben minimalizálni
- …de hát ez épp az amit az OLS-nél már megoldottunk!
- Úgyhogy az eredményt már tudjuk: $\widehat{\boldsymbol{\beta}_{\mathrm{ML}}}=\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{y}$

Lineáris modell becslése ML-elven

- A lineáris regressziós modell esetén tehát az ML-becslés és az OLS-becslés egybeesik (ebből adódóan a tulajdonságaik is azonosak)
- De az ML-hez fel kellett tenni a hibanormalitást, az OLS-nél erre nem volt szükség

8 Linearitás és feloldása, nemlineáris modellek

8.1. Elöljáróban: a marginális hatás általánosabb értelmezése

A marginális hatás fogalma

- Marginális hatás: a magyarázó változó kis növelésének hatására mekkora az eredményváltozó *egységnyi magyarázóváltozó-növelésre jutó* változása
- Tipikus egyszerűsítés: a magyarázó változó egységnyi növelésének hatására mennyit változik az eredményváltozó
- (Hiszen a kettő ugyanaz, ha a változó hatása lineáris)
- Idáig az i-edik magyarázó változó ilyen módon értelmezett marginális hatása és a β_i számértéke gyakorlatilag szinonima volt

A marginális hatás precízebben

- Definíció alapján a marginális hatás: $\frac{\Delta Y}{\Delta X_j}$, ha ΔX_j kicsiny
- Ugye egyetemen vagyunk $\rightarrow$ a marginális hatás $\frac{\partial Y}{\partial X_j}$
- A többváltozós lineáris regresszió eddigi (sokasági) modelljében $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$, ezért

$$\begin{aligned}\frac{\partial Y}{\partial X_j} = \frac{\partial}{\partial X_j}\big[&\beta_0 + \beta_1 X_1 + \ldots + \\ &+ \ldots + \beta_{j-1} X_{j-1} + \beta_j X_j + \beta_{j+1} X_{j+1} + \ldots + \beta_k X_k + \varepsilon\big] = \\ &= \beta_j\end{aligned}$$

- …hát ezért tekinthettük eddig a marginális hatást és a becsült regressziós koefficienst szinonimának!

8.2. A linearitás feloldása

8.2.1. Emlékeztetőül

A linearitás következményei

- A linearitás két dolgot vont maga után:
  - Mindegy, hogy honnan indulva növelem a változót egy egységgel (a változó hatása lineáris)
  - Mindegy, hogy a többi változó milyen szinten van rögzítve (additivitás)
- E kettőt fogjuk most feloldani

8.2.2. Az additivitás feloldása: az interakció

Interakció mint a linearitás egyféle feloldása

- Eddigi modellünkben a marginális hatások a többi változó szintjétől függetlenül állandóak voltak
- Például: 1 Ft pluszjövedelem taglétszámtól függetlenül azonos többletkiadást jelent…?
- Ha nem, akkor azt mondjuk, hogy a két változó között *interakció* van: az egyik marginális hatásának *nagyságát* befolyásolja a másik *szintje*
- A kapcsolat tehát marginális hatás és szint között van (nem marginális hatás és marginális hatás vagy szint és szint között!)
- Kézenfekvő indulás: az egyik változó szintje *lineárisan* hasson a másik marginális hatására; sokaságban felírva:

$$(\beta_J + \beta_{JT}\mathrm{Tag})\,\mathrm{Jov},$$

ahol β_{JT} az interakció hatását kifejező (lineáris) együttható

Interakció

- Helyezzük ezt be a (sokasági) regresszióba:

$$Y = \beta_0 + (\beta_J + \beta_{JT}\mathrm{Tag})\,\mathrm{Jov} + \beta_T\mathrm{Tag} + \varepsilon,$$

azonban felbontva a zárójelet:

$$\begin{aligned} Y &= \beta_0 + \beta_J\mathrm{Jov} + \beta_{JT}\mathrm{Tag}\cdot\mathrm{Jov} + \beta_T\mathrm{Tag} + \varepsilon = \\ &= \beta_0 + \beta_J\mathrm{Jov} + (\beta_T + \beta_{JT}\mathrm{Jov})\,\mathrm{Tag} + \varepsilon \end{aligned}$$

- Tehát az interakció *szükségképp*, automatikusan „szimmetrikus”: ha az egyik változó szintje hat a másik marginális hatására akkor szükségképp fordítva is: a másik szintje is hatni fog az előbbi marginális hatására

Interakció

- Azaz „egyszerre” lesz igaz, hogy $(\beta_J + \beta_{JT}\text{Tag})\,\text{Jov}$ és $(\beta_T + \beta_{JT}\text{Jov})\,\text{Tag}$: attól függően, hogy milyen szempontból nézzük (melyik marginális hatását vizsgáljuk, ezt még ld. később is)
- A regresszióban így elég egyszerűen ennyit írni:

$$\beta_T\text{Tag} + \beta_J\text{Jov} + \beta_{JT}\,(\text{Jov}\cdot\text{Tag})\,.$$

- …mindkét – másik szintjétől függő – marginális hatás ebből kiadódik, függően attól, hogy hogyan bontjuk fel a zárójelet (melyik változót vizsgáljuk)
- Ez a marginális hatás pontosabb értelmezése mellett még szebben látható

A marginális hatás interakciók esetén

- Ha interakció van, például a l-edik és az m-edik tag között, akkor az l-edik marginális hatása:

$$\begin{aligned}\frac{\partial Y}{\partial X_l} = \frac{\partial}{\partial X_l}\,[&\beta_0 + \beta_1 X_1 + \ldots + \\ &+ \ldots + \beta_l X_l + \ldots + \beta_m X_m + \ldots + \beta_k X_k + \beta_{lm} X_l X_m + \varepsilon] = \\ &= \beta_l + \beta_{lm} X_m\end{aligned}$$

- Így precíz az előbbi állításunk arról, hogy ha az egyik szerint vizsgáljuk a marginális hatást, akkor az a másik szintjétől fog függeni (gondoljuk hozzá a másik szerinti deriválást is!)

8.2.3. A változónkénti linearitás feloldása

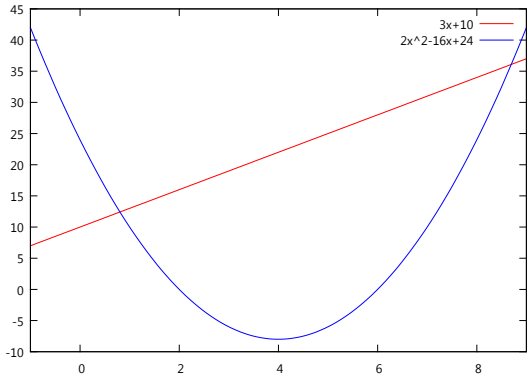
Motiváló példa: kvadratikus hatás

- Már volt: mit jelent az, ha megsértjük a „marginális hatás nem függ attól, hogy a többi magyarázó változót milyen szinten rögzítjük” következményét a linearitásnak
- És ha a „marginális hatás nem függ attól, hogy milyen szintről indulva növeljük a változót” következményt szeretnénk feloldani?
- Például: 1 évvel idősebb életkor kiinduló életkortól függetlenül azonos kiadásváltozást jelent…?
- Használjunk a lineáris függvényforma helyett mást, például négyzeteset (parabolát):

$$\frac{\partial}{\partial X_j}\left[\ldots + \beta_j X_j + \beta_{jj} X_j^2 + \ldots\right] = \beta_j + 2\beta_{jj} X_j$$

Motiváló példa: kvadratikus hatás

Szemléletesen az egy magyarázó változós esetben:

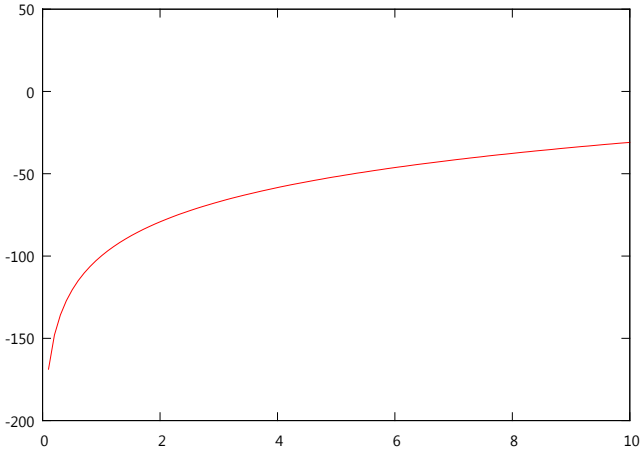


Szélsőértékhely nyilvánvaló (első derivált előjelet vált): $\beta_j + 2\beta_{jj}X_j = 0 \Rightarrow X_j = -\frac{\beta_j}{2\beta_{jj}}$

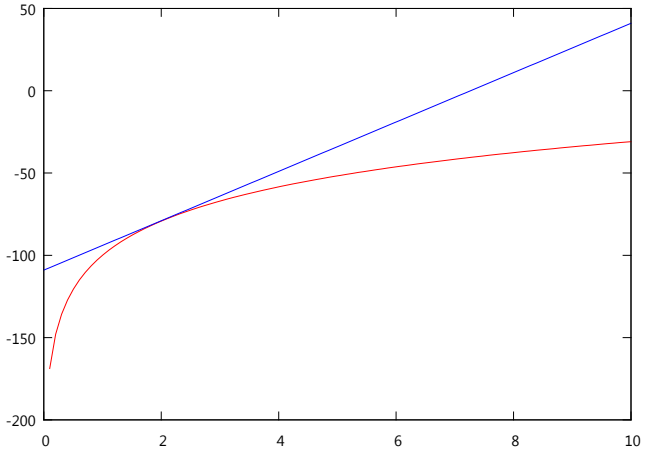
Linearitás, mint közelítés

- Az élet általában nemlineáris
- Miért használunk mégis lineáris modelleket: mert sokszor nem térnek el (nagyon) a valóságtól, de mégis sokkal könnyebben kezelhetőek matematikailag
- (Taylor-soros érvelés!)
- Ez tehát az esetek többségében egy közelítés
- Mint ilyen: vizsgálni kell az érvényességi határokat
- „Munkaponti linearizálás”

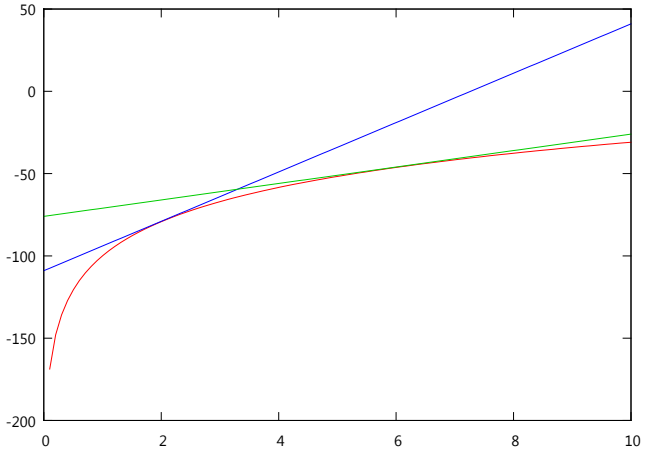
Érvényességi határok



Érvényességi határok



Érvényességi határok



Érvényességi határok

- Az érvényességi határokat az eddig látott modellekben is érdemes végiggondolni
- Azonnal kézenfekvő példa: a konstans (nagyon sok esetben)
- De sok meredekségnél is megragadható ez (fogyasztási függvény példája)
- Ez is egyfajta munkaponti linearizálás

Nemlinearitás fajtái

- Az $\beta_1 + \beta_2 X + \beta_3 X^2$ egy nemlineáris kifejezés (matematikailag)
- De figyelem: ennek ellenére minden további nélkül, tökéletesen kezelhető pusztán az eddig látott (lineáris!) eszköztárral, hiszen az OLS-nek mindegy, hogy a második magyarázó változó értékei történetesen épp az első négyzetei
- (Egészen addig nincs baj, amíg a kapcsolat nem lineáris)
- Nem úgy mint az $\beta_1 X^{\beta_2} \rightarrow$ ez nem becsülhető OLS-sel
- A megkülönböztetés végett az első esetet változójában, a másodikat paraméterében nemlineáris modellnek nevezzük
- Mi van „nemlinearitást okozó pozícióban”

Változójában nemlineáris modell

- Jellemző: továbbra is fennáll a „változók konstansokkal szorozva majd összeadva” (tehát: lineáris kombinációs) struktúra
- De elképzelhető, hogy egy változó egy „eredeti” változó transzformáltja
- Itt szükségképp nemlineáris transzformációról beszélünk!
- Vegyük észre, hogy az „eredeti” és „transzformált” közti megkülönböztetés teljesen mesterséges (csak mi tudjuk, hogy mi volt az adatbázisban bemenő adatként), az OLS-nek mindegy
- Ide tartozik a kvadratikus hatás, általában az X^a magyarázó változók, a $\log_a X$, az a^X stb., ahol a konstans
- Az előzőek miatt a becslés ugyanaz, egyedül az interpretálás igényel további tárgyalást

Paramétereiben nemlineáris modell

- Megsérti a lineáris kombináció struktúráját: paraméter nem csak szorzóként szerepel a regresszióban
- Például X^β, $\log_\beta X$ stb.
- Ez már nem becsülhető OLS-sel: az eredményváltozó nem állítható elő mátrixműveletekkel
- Más módszert fogunk használni

Interakció és kvadratikus hatás revisited

- Az előzőek fényében nyilvánvaló: a kvadratikus hatás egyfajta (igen egyszerű) változójában nemlineáris modell
- Az interakció szintén változóbeli nemlinearitás, de nem annyira kézenfekvő módon (mindenképp indokolt a külön tárgyalása)

Nemlinearitás kezelése: NLS

- Vegyük észre, hogy a $\min_{\boldsymbol{\beta}} ESS$ célfüggvény akkor is tartható, ha nemlineáris modellt specifikálunk!
- (Csak az ESS számításához szükséges $\widehat{Y}$-ok másképp jönnek ki, de ez a fenti optimalizáció szempontjából *teljesen mindegy*)
- Oldjuk meg ezt az optimalizációs feladatot!
- Ez a nem-lineáris legkisebb négyzetek (NLS, non-linear least squares) módszere

Nemlinearitás kezelése: NLS

- Sajnos mondani könnyebb, mint a gyakorlatban kivitelezni; szemben a lineáris specifikációval, a kritériumfelület nem kvadratikus, emiatt nincs egyetlen művelettel megtalálható optimum
- Van-e egyáltalán egyértelmű (globális) optimum? Mi van, ha több lokális optimum létezik?

Nemlinearitás kezelése: NLS

- Ettől el is tekintve, a konkrét optimalizáció számos gyakorlati problémát vethet fel, mivel valamilyen iteratív algoritmus kell
- Több lehetőség van, különféle előnyökkel és hátrányokkal (Gauss–Newton keresés, Levenberg–Marquardt algoritmus, konjugált gradiens keresés stb.), de mind rengeteg numerikus kérdést vet fel:
  - Meg tudjuk találni az optimumot? Biztosan? (Lehet-e baj a konvergenciával? Mi legyen a konvergencia-kritérium?)
  - Mennyi idő alatt találjuk meg?
  - Milyen kezdőértékből induljunk? (Milyen a módszer numerikus stabilitása?)
  - stb. stb. stb.

Nemlinearitás kezelése: algebrai linearizáció

- Mi a fenti (mindig alkalmazható) módszerrel szemben egy másik (könnyebb, de nem mindig alkalmazható) módszert fogunk vizsgálni: algebrai linearizálás
- Alkalmas transzformációval a nemlineáris problémát lineárissá alakítjuk, azt OLS-sel megoldjuk, majd a kapott eredményeket visszatranszformáljuk az eredeti transzformáció inverzével
- Például: $Y = \beta_1 X^{\beta_2} \varepsilon$ paramétereiben nemlineáris …
- … de mindkét oldal logaritmusát véve $\log Y = \log \beta_1 + \beta_2 \log X + \varepsilon'$ már az!
- Adatbázis logaritmálása, eredmények visszahatványozása
- (Amint mondtuk, nem mindig alkalmazható, de azért nagyon sok, gyakorlatilag fontos esetben igen)
- Természetesen itt is eltérő, specifikus értelmezések jelenthetnek meg

8.3. Néhány nevezetes, paraméterében nemlineáris modell

Log-log modell

- Például a Cobb-Douglas termelési modell:

  $$Y = \beta_1 L^{\beta_L} K^{\beta_K} \varepsilon,$$

  ahol Y a kibocsátás, L a munka, K a tőke (ill. általában a termelési tényezők) felhasználása
- Elaszticitása:

  $$\mathrm{El}_L(L,K) = \frac{\frac{\mathrm{d}Y}{Y}}{\frac{\mathrm{d}L}{L}} = \frac{\mathrm{d}Y}{\mathrm{d}L}\frac{L}{Y} = \beta_1 \beta_L L^{\beta_L - 1} K^{\beta_K} \frac{L}{\beta_1 L^{\beta_L} K^{\beta_K}} = \beta_L$$

- Ezért nevezik *konstans elaszticitású* modellnek is
- Kezelése linearizálással: mindkét oldalt logaritmáljuk

  $$\log Y = \log \beta_1 + \beta_L \log L + \beta_K \log K + \varepsilon'$$

Log-log modell

- Minden változót (eredmény és összes magyarázó is) logaritmálni kell
- Innen a modell neve
- Csak a konstans lesz logaritmálva, a többi koefficienst a transzformáció ellenére (ill. épp azért…) közvetlenül kapjuk
- Volumenhozadék (skálahozadék): $\beta_K + \beta_L$ viszonya 1-hez

Log-lin modell

- Például a jövedelem alakulása:

$$Y = e^{\beta_1+\beta_2 X+\varepsilon}$$

- Linearizálás ismét mindkét oldal logaritmálásával:

$$\log Y = \beta_1 + \beta_2 X + \varepsilon$$

- Elnevezés logikája így már látható: az eredményváltozó logaritmálva, de a magyarázó változók maradnak szintben
- Növekedési ráta: $e^{\beta_1+\beta_2(X+1)+u} = Ye^{\beta_2}$, pillanatnyi növekedési ütem: $\beta_2 = \frac{\mathrm{d}\log Y}{\mathrm{d}X} = \frac{1}{Y}\frac{\mathrm{d}Y}{\mathrm{d}X}$
- Elaszticitás: $\mathrm{El}_X(X) = \frac{\mathrm{d}Y}{\mathrm{d}X}\frac{X}{Y} = \beta_2 X$, tehát csak X-től függ

Lin-log modell – kakukktojás!

- Az előzőek alapján már világos a jelentése (pl. terület és kínálati ár összefüggése):

$$Y = \beta_1 + \beta_2 \log X + \varepsilon$$

- Miért kakukktojás?
- β_2 értelmezése:

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = \frac{\beta_2}{X} \Rightarrow \beta_2 = \frac{\mathrm{d}Y}{\mathrm{d}X/X}$$

- Elaszticitás:

$$\mathrm{El}_X(X) = \frac{\beta_2}{X}\frac{X}{Y} = \frac{\beta_2}{Y},$$

  tehát csak Y-től függ (közvetlenül)

A „közvetlenül” itt arra utal, hogy természetesen az Y helyébe beírható lenne a – csak X-eket tartalmazó – felírása.

Reciprok modell – kakukktojás

- Például keresleti modell:

$$Y = \beta_1 + \frac{\beta_2}{X} + \varepsilon$$

- Miért kakukktojás?
- Aszimptotikusan: $\lim_{X\to\infty} \mathbb{E}\left(Y \mid X\right) = \beta_1$

- Határkiadás: $\frac{\mathrm{d}Y}{\mathrm{d}X} = -\frac{\beta_2}{X^2}$
- Elaszticitás:
$$\mathrm{El}_X(X) = -\frac{\beta_2}{X^2}\frac{X}{Y} = \frac{\beta_2}{XY}$$
- Paraméterek értelmezése, β_2 előjelének jelentősége az „aszimptotikus” viselkedés szempontjából: az élvezeti cikkek példája

8.4. Specifikációs tesztek

A specifikációs tesztek

- Itt már nagyon erősen felmerül a kérdés: hogyan dönthetek a különféle függvényformák között?
- Ld. a termelési függvény példáját → megadható lineárisan és Cobb-Douglas jelleggel (eredmény nagyon nem mindegy)
- Hogyan lehet *analitikusan* dönteni?
- Az előző példára: BM-teszt, PE-teszt stb.
- Általánosságban (nem csak log/lin kérdésekre, mint az előzőek): ún. specifikációs tesztek

Egy egyszerű specifikációs teszt

- Egy egyszerű ötlet: adjuk hozzá a magyarázó változókhoz a magyarázó változók valamilyen nemlineáris transzformáltját (tipikusan négyzeteiket vagy logaritmusaikat)…
- …és nézzük meg, hogy *együttesen* szignifikánsak-e!
- Ha igen, az specifikációs hibára utal
- (Tehát figyelem, ezt alapvetően nem arra használjuk, hogy arra következtessünk, hogy egy adott konkrét változó négyzetét vagy logaritmusát hozzá kell-e adni a függvényformához, hanem összességében nézzük őket, specifikációs tesztként)
- Hátrány: sok szabadságfokot használ el, és csak elég speciális alakú nemlinearitásokkal tesztel

Ramsey RESET-je

- A modellspecifikáció *általános* tesztje; emiatt előnye: nem egy adott specifikációs kérdésre keres választ, hanem általában vizsgálja, hogy a specifikáció jó-e; hátránya, hogy ha nemleges választ ad, nem derül ki, hogy pontosan mi a specifikáció baja
- Trükk: új regressziót becsül, melynek eredményváltozója ugyanaz, de a magyarázó változókhoz hozzáadja az eredeti regresszió becsült eredményváltozójának magasabb hatványait ($\widehat{Y}^3$-ig néha $\widehat{Y}^4$-ig is):

$$Y = \beta'_0 + \beta'_1 X_1 + \beta'_2 X_2 + \ldots + \beta'_k X_k + \gamma_1 \widehat{Y}^2 + \gamma_2 \widehat{Y}^3 + \varepsilon'$$

- Amit tesztelni kell az előzőhöz hasonlóan: $H_0 : \gamma_1 = \gamma_2 = 0$ (F-próba és LM-próba is van, heteroszkedaszticitásra is robusztussá tehető)
- Ilyen módon takarékos a szabadsági fokokkal, csak 2-3-at használ el (a fenti trükkel „összesűríti" a magyarázó változókat), ráadásul általánosabb alakú nemlinearitások is belefférnek ebbe, mint a csak négyzetekkel/logaritmusokkal tesztelés
- Specifikációs teszt, tehát kihagyott változó detektálására általában nem alkalmas

9 A multikollinearitás

9.1. Multikollinearitás

A magyarázó változók körében rejlő egyéb probléma-lehetőségek

- Van egy másik oka is annak, hogy túl sok magyarázó változó használata miért lehet problémás: az, hogy a magyarázó változók a tipikus gyakorlati esetekben egymást is magyarázzák, vannak közöttük lineáris kapcsolatok
- Ezt a következő egyszerű példán mutatjuk be:

$$Y = \beta_0 + \beta_B \text{Ber} + \beta_F \text{Fo} + u,$$

- Tegyük most fel (nyilván nem igaz ilyen erősen, de nem teljesen elrugaszkodott), hogy a Bér-hez képest a Fő hozzáadása már felesleges, mégpedig azért mert „nem hordoz további információt" (ugyanazt írja le más szemszögből), mi mégis bevonjuk a modellünkbe

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becsülhetőség
- Vigyázat: *együtt* becsülhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol}\,(\text{Ber}) = 1 - R^2_{\text{Ber}|\text{Fo}}$$

Multikollinearitás leírása

- Általában: a vizsgálat magyarázó változót mennyire magyarázza a többi magyarázó változó, tehát

$$\mathrm{Tol}\,(j) = 1 - R_j^2 = 1 - R^2_{\mathbf{X_j}|\mathbf{X_1},\mathbf{X_2},\ldots,\mathbf{X_{j-1}},\mathbf{X_{j+1}},\ldots,\mathbf{X_k}}$$

- Minél nagyobb R_j^2, annál kisebb a tolerancia → intuitíve: annál kevesebb többletinformációt hoz be ez a változó a modellbe a többi magyarázó változó mellett

Multikollinearitás hatása

- Írjuk most fel egy már bent levő változó koefficiensének mintavételi varianciáját:

$$\mathbb{D}^2\left(\widehat{\beta}_j\right) = \frac{ESS/\left(n-k-1\right)}{\left(n-1\right)\mathbb{D}^2\left(X_j\right)} \cdot \frac{1}{\mathrm{Tol}\,(j)} = \frac{\widehat{\sigma^2}}{\left(n-1\right)\mathbb{D}^2\left(X_j\right)} \cdot \frac{1}{\mathrm{Tol}\,(j)}$$

- Látszik, hogy egy magyarázó változó koefficiensének a mintavételi varianciája c. p. *nő*, ahogy a tolerancia romlik (csökken); elvi minimum erre a varianciára a tolerancia = 1-nél
- Itt a c.p.-t úgy képzeljük el, mintha tudnánk *csak* a multikollinearitást változtatni
- De figyelem: a multikollinearitás, bármilyen közel is van 1-hez az R_j^2, *nem* megsértése a standard modellfeltevéseknek (hacsak nem egzakt)

A multikollinearitás mérése

- Bevezetjük a variancia infláló tényezőt (VIF):

$$\mathrm{VIF}\,(j) = \frac{1}{\mathrm{Tol}\,(j)}$$

- $\mathrm{VIF}\,(j) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\mathrm{VIF}\,(j) = 2$: a mintavételi variancia megduplázódott *pusztán a multikollinearitás miatt* (tehát amiatt, hogy a magyarázó változók egymást is magyarázzák) *ahhoz képest* mintha nem lenne multikollinearitás stb.
- A használatával kapcsolatban vannak bizonyos fenntartások!

10 A modellszelekció kérdései

10.1. Általánosítóképesség, túlilleszkedés

Pár gondolat a magyarázó változók körének kiválasztásához

- Eddig egyetlen minősítőjét láttuk egy modell jóságának: az R^2-et
- Tételmondat: új változó bevonásával R^2 értéke *mindenképp* nő (de legalábbis nem csökken), teljesen függetlenül attól, hogy mi a bevont változónk, mik vannak már a modellben stb.
- Tehát: ha az R^2-tel jellemezzük a modellünket, akkor *mindig* az összes potenciális magyarázó változó felhasználása lesz a legjobb döntés
- A valóságban azonban már nem biztos!
- Mert: az R^2 a *minta* jó leírását jellemzi, de mi a sokaságot akarjuk megragadni
- A kettő ellentmondásba kerülhet!

A tételmondat indoklásaként gondoljunk arra, hogy „legeslegrosszabb esetben" az újonnan bevont változó együtthatójára nulla mindenképp becsülhető – ekkor pedig ESS szempontjából pont ott vagyunk, mint az eredeti modell esetében!

Általánosítóképesség

- Azt, hogy a modell – a mintából kinyert információk alapján – mennyire jól tud a sokaságról (tehát a mintán kívüli világról) is számot adni, *általánosítóképességnek* nevezzük
- Igazából mi erre játszunk!
- … ennyiben (erre a célra) az R^2 nem szerencsés mutató

Az R^2 a minta jó „megjegyzését" mutatja. Ez nekünk nem öncél – gondoljunk bele: ha csak a mintát akarnánk megjegyezni, akkor kár is regressziós modellt alkotni, használhatnánk egyszerűen magát a mintát is, ami ugye a rendelkezésünkre áll…

Általánosítóképesség

- Persze az sem jó megközelítés, hogy az R^2-tel nem törődünk, hiszen ha nem szedünk ki elég információt a mintából, akkor sem várható, hogy a sokaságról jól tudunk nyilatkozni (mivel arra vonatkozóan csak a mintára támaszkodhatunk)
- Tehát: kompromisszumra van szükség a mintainformációk felhasználásában…
  - … ha túl keveset használunk fel, akkor nem nyerünk elég jó képet a sokaságról
  - … ha túl sokat használunk fel, akkor túlságosan „ráfókuszálunk” a mintára

Általánosítóképesség

- Ahogy egyre több információt nyerünk ki a mintából (egyre jobban „elköteleződünk” mellette), úgy egy pontig javul, majd ezen túl romlik az általánosítóképesség
- Tehát: nem csak nem javít a több információ, de egyenesen ront (ezért az „ellentmondás”)!

Alulilleszkedés, túlilleszkedés

- A fentiek jól értelemzhetőek a *gépi tanulás* fogalomkészletével
- Itt a tanulás információkinyerés a mintából
- Ha ezt túl kis mértékben hajtjuk végre, akkor alulilleszkedésről (alultanulásról)…
- … ha túl nagy mértékben, akkor túlilleszkedésről (túltanulásról) beszélünk
- A túltanított modell látszólag nagyon jó (a mintát jól megragadja), de valójában nem az, mert a mintán kívüli képességei gyatrák lesznek (hiszen túlságosan „ráfókuszált” a mintára)

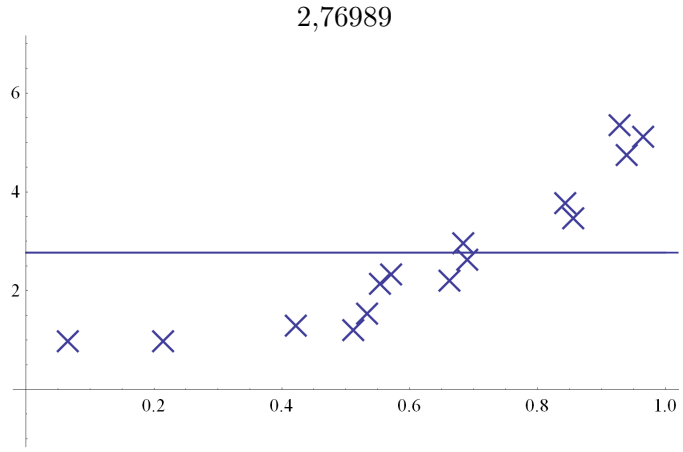
Egy példa a túlilleszkedésre

- Egyszerű kétváltozós feladat: egy magyarázó- és egy eredményváltozó
- A példánkban a tanítás fokát tehát nem a magyarázó változók számával fogjuk mérni, hanem a függvényforma bonyolultságával: $Y = \beta_1 + \beta_2 X + u$, $Y = \beta_1 + \beta_2 X + \beta_2' X^2 + u'$, $Y = \beta_1 + \beta_2 X + \beta_2' X^2 + \beta_2'' X^3 + u''$ stb.
- Tehát az eredményváltozót a magyarázó változó egyre nagyobb fokszámú polinomjával közelítjük (a polinom fokszámát jelölje p)
- (A függvényforma ilyen megválasztásával később foglalkozunk részleteiben, de most nem is ez a lényeg)

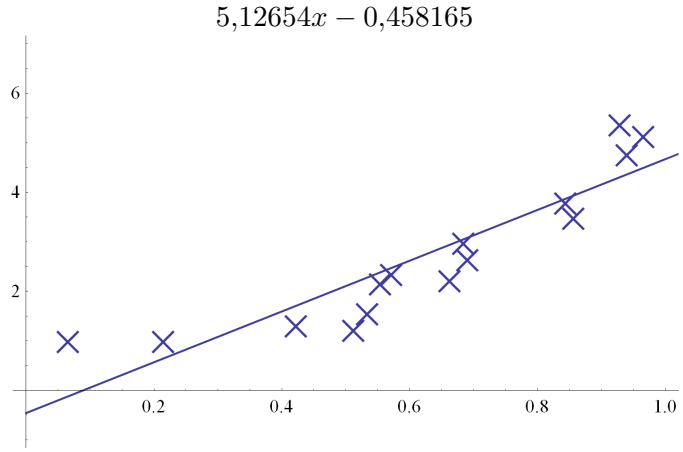
Egy példa a túlilleszkedésre

- Hogy tudjuk mi a „jól illeszkedő” modell, elárulom, hogy az adatokat valójában egy $Y = 5 \cdot X^3 + 1 + u$ modell szerint generáltam, ahol $u \sim \mathcal{N}(0; 0{,}3)$
- Tehát lényegében: „zajos harmadfokú” függvény
- A jól illeszkedő modell – ezt *most* tudjuk, általában persze nem! – a harmadfokú lenne

Alulilleszkedés: $p = 0$

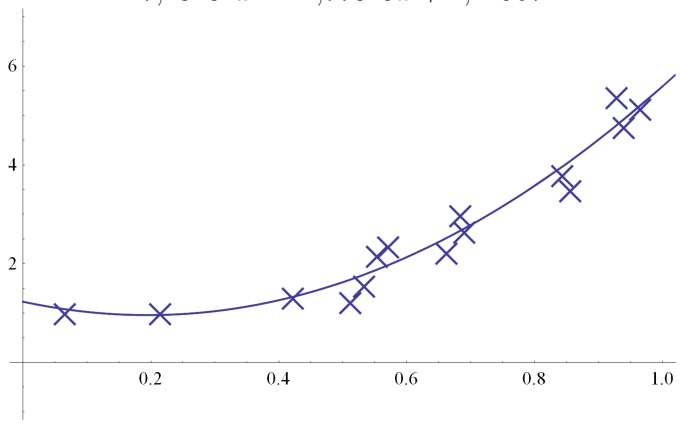


Alulilleszkedés: $p = 1$



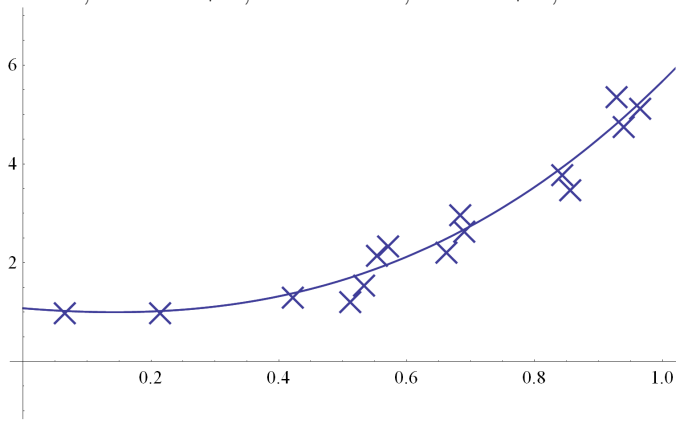
Nagyjából jó illeszkedés: $p = 2$

$$7{,}13434x^2 - 2{,}77819x + 1{,}22967$$



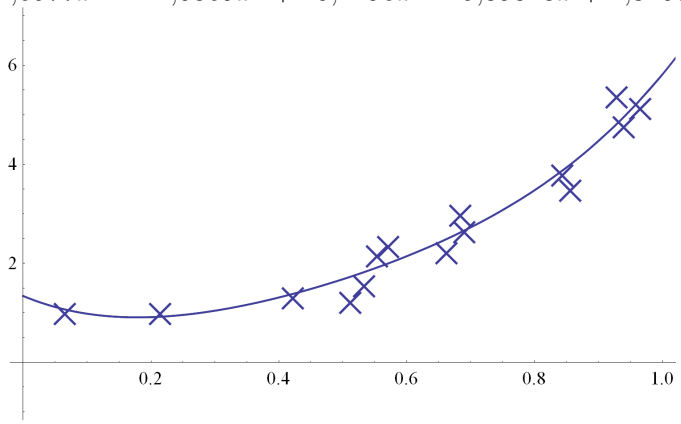
Nagyjából jó illeszkedés: $p = 3$

$$2{,}48264x^3 + 3{,}17392x^2 - 1{,}06319x + 1{,}0774$$



Nagyjából jó illeszkedés: $p = 4$

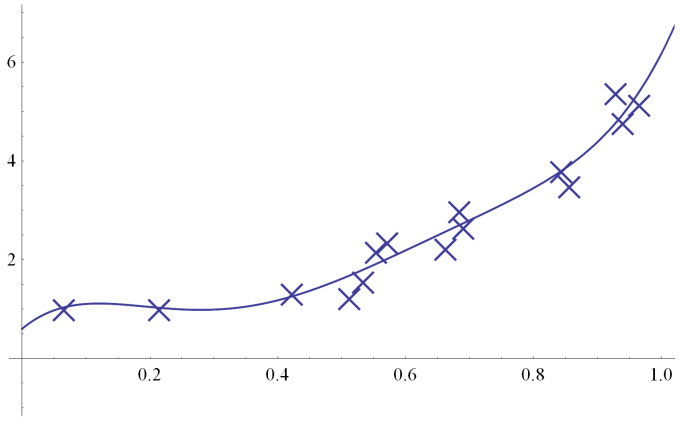
$$11{,}6577x^4 - 22{,}0369x^3 + 20{,}2496x^2 - 5{,}39823x + 1{,}34003$$



Érdemes észrevenni, hogy még ha tekintettel vagyunk az általánosítóképességre, akkor sem tudjuk mintából egyértelműen megmondani, hogy pont a $p = 3$ a jó válasz. Ez azonban nem meglepő, sőt, épp várható: egymáshoz közeli lehetőségek minta alapján nem feltétlenül különíthetőek el, ha nem elég nagy a mintanagyság. Minél kisebb a differencia, annál nagyobb mintaméret kell, hogy megbízhatóan el tudjuk dönteni, hogy a szóba jövő lehetőségek közül melyik a valódi modell – ez nyilván általános statisztikai elv. Kevés pont alapján – sajnos – elképzelhető, hogy nem eldönthető, hogy lineáris, négyzetes, vagy épp exponenciális görbe a valódi modell.

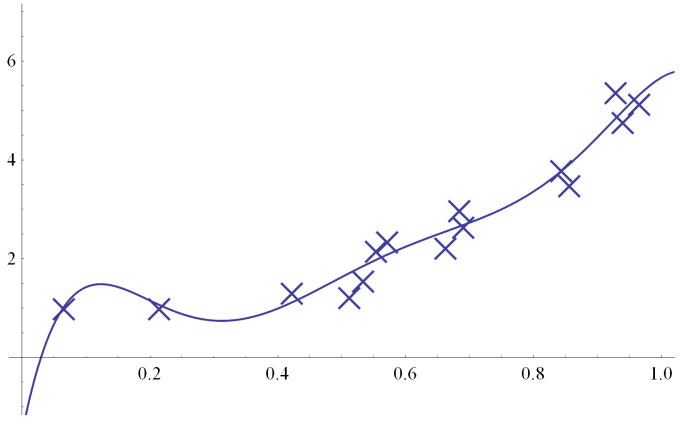
Túlilleszkedés: $p = 5$

$$94{,}7601x^5 - 236{,}514x^4 + 213{,}631x^3 - 77{,}138x^2 + 10{,}8264x + 0{,}601515$$



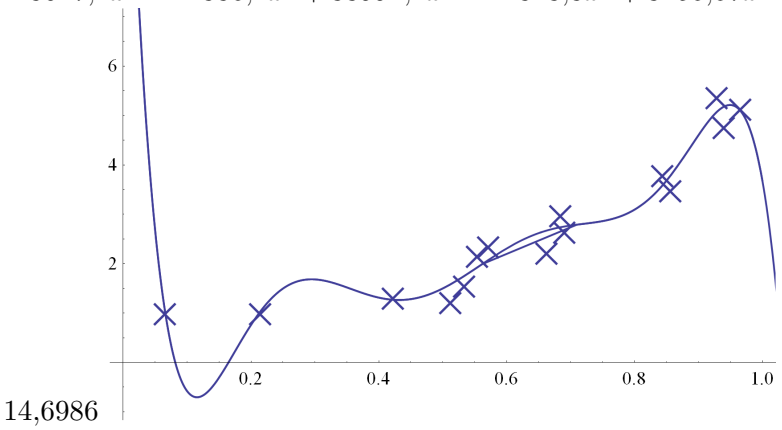
Túlilleszkedés: $p = 6$

$$-556{,}426x^6 + 1895{,}28x^5 - 2494{,}87x^4 + 1587{,}69x^3 - 489{,}325x^2 + 64{,}8299x - 1{,}52203$$



Túlilleszkedés: $p = 7$

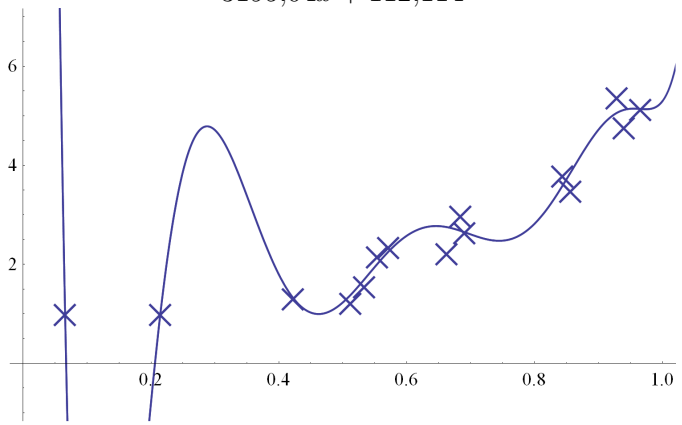
$$-7426{,}18x^7 + 28047{,}2x^6 - 42886{,}1x^5 + 33991{,}4x^4 - 14813{,}8x^3 + 3456{,}67x^2 - 380{,}286x +$$



$$14{,}6986$$

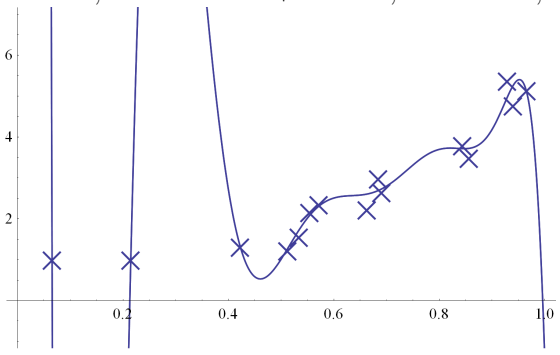
Túlilleszkedés: $p = 8$

$$59039{,}2x^8 - 282296{,}x^7 + 565254{,}x^6 - 613881{,}x^5 + 390937{,}x^4 - 146967{,}x^3 + 31001{,}6x^2 - 3195{,}04x + 112{,}114$$



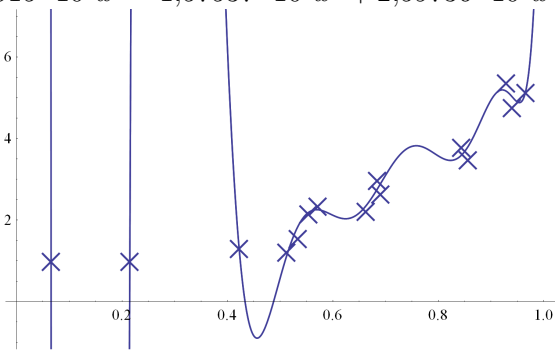
Túlilleszkedés: $p = 9$

$$-722495{,}x^9 + 3{,}85053 \cdot 10^6 x^8 - 8{,}84295 \cdot 10^6 x^7 + 1{,}1426 \cdot 10^7 x^6 - 9{,}08926 \cdot 10^6 x^5 + 4{,}57009 \cdot 10^6 x^4 - 1{,}43064 \cdot 10^6 x^3 + 262396{,}x^2 - 24485{,}1x + 807{,}137$$



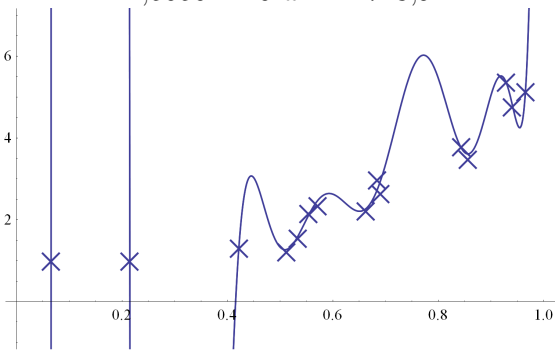
Túlilleszkedés: $p = 10$

$$8{,}61299 \cdot 10^6 x^{10} - 5{,}24999 \cdot 10^7 x^9 + 1{,}40371 \cdot 10^8 x^8 - 2{,}16006 \cdot 10^8 x^7 + 2{,}1085 \cdot 10^8 x^6 - 1{,}35546 \cdot 10^8 x^5 + 5{,}75915 \cdot 10^7 x^4 - 1{,}57537 \cdot 10^7 x^3 + 2{,}59736 \cdot 10^6 x^2 - 223991{,}x + 7044{,}46$$



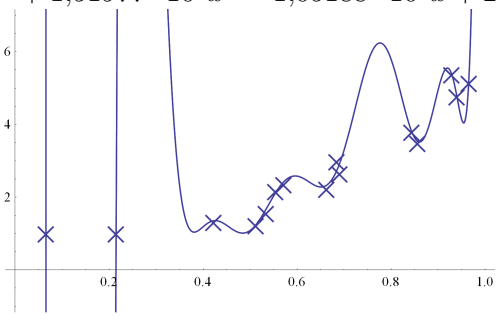
Túlilleszkedés: $p = 11$

$$9{,}81027 \cdot 10^7 x^{11} - 6{,}54761 \cdot 10^8 x^{10} + 1{,}94347 \cdot 10^9 x^9 - 3{,}37777 \cdot 10^9 x^8 + 3{,}80722 \cdot 10^9 x^7 - 2{,}91 \cdot 10^9 x^6 + 1{,}53045 \cdot 10^9 x^5 - 5{,}49469 \cdot 10^8 x^4 + 1{,}30416 \cdot 10^8 x^3 - 1{,}91189 \cdot 10^7 x^2 + 1{,}50501 \cdot 10^6 x - 44723{,}9$$



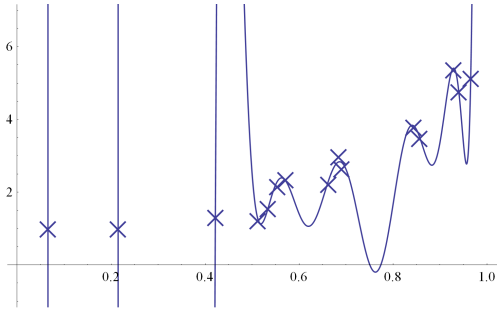
Túlilleszkedés: $p = 12$

$$1{,}97286 \cdot 10^8 x^{12} - 1{,}37728 \cdot 10^9 x^{11} + 4{,}31319 \cdot 10^9 x^{10} - 7{,}99714 \cdot 10^9 x^9 + 9{,}75531 \cdot 10^9 x^8 - 8{,}22533 \cdot 10^9 x^7 + 4{,}8983 \cdot 10^9 x^6 - 2{,}06632 \cdot 10^9 x^5 + 6{,}08915 \cdot 10^8 x^4 - 1{,}211 \cdot 10^8 x^3 + 1{,}51977 \cdot 10^7 x^2 - 1{,}05188 \cdot 10^6 x + 28665$$



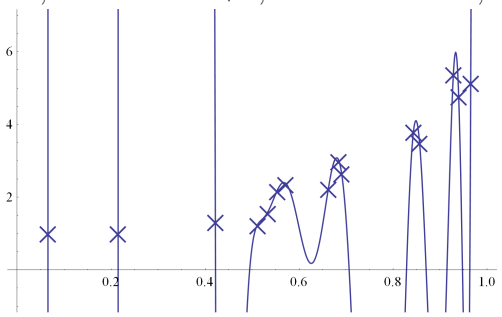
Túlilleszkedés: $p = 13$

$$1{,}33188 \cdot 10^{10}x^{13} - 1{,}09101 \cdot 10^{11}x^{12} + 4{,}06208 \cdot 10^{11}x^{11} - 9{,}08859 \cdot 10^{11}x^{10} + 1{,}36095 \cdot 10^{12}x^{9} - 1{,}43708 \cdot 10^{12}x^{8} + 1{,}0978 \cdot 10^{12}x^{7} - 6{,}12006 \cdot 10^{11}x^{6} + 2{,}4775 \cdot 10^{11}x^{5} - 7{,}14241 \cdot 10^{10}x^{4} + 1{,}41049 \cdot 10^{10}x^{3} - 1{,}77685 \cdot 10^{9}x^{2} + 1{,}24223 \cdot 10^{8}x - 3{,}41822 \cdot 10^{6}$$

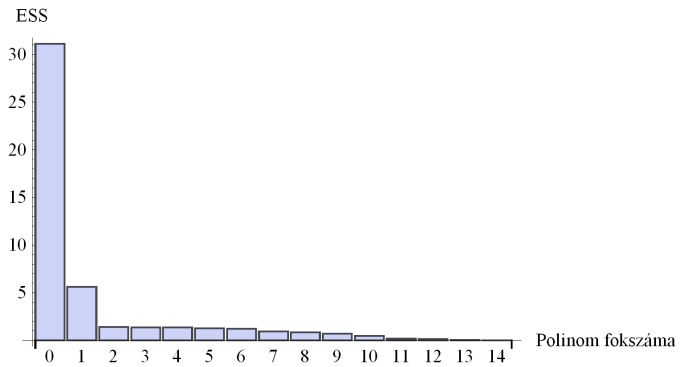


Túlilleszkedés: $p = 14$

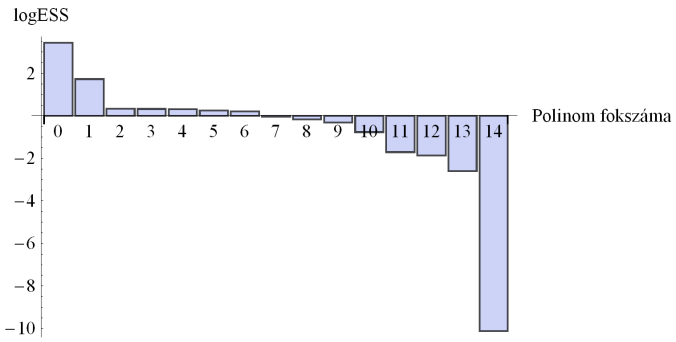
$$2{,}23808 \cdot 10^{11}x^{14} - 1{,}95447 \cdot 10^{12}x^{13} + 7{,}81606 \cdot 10^{12}x^{12} - 1{,}89512 \cdot 10^{13}x^{11} + 3{,}10833 \cdot 10^{13}x^{10} - 3{,}64245 \cdot 10^{13}x^{9} + 3{,}1386 \cdot 10^{13}x^{8} - 2{,}01508 \cdot 10^{13}x^{7} + 9{,}65479 \cdot 10^{12}x^{6} - 3{,}41996 \cdot 10^{12}x^{5} + 8{,}76076 \cdot 10^{11}x^{4} - 1{,}55904 \cdot 10^{11}x^{3} + 1{,}79489 \cdot 10^{10}x^{2} - 1{,}16536 \cdot 10^{9}x + 3{,}04682 \cdot 10^{7}$$



Hiba az egyes fokszámok mellett



Jobban láthatóan...



Itt a függőleges tengely logaritmikus beosztású, hogy a nagyon kis számok tartományában is látszódjanak a változások.

A túlilleszkedés hatása

- Itt a tanítás mértékét a polinom fokszáma jelzi
- A példa tökéletesen mutatja, hogy mi a túlilleszkedés tartalma:
  - A mintaadatokat ugyan egyre jobban megtanuljuk...
  - ... de közben a mintán kívüli világról egyre kevesebbet tudunk mondani (holott minket ez érdekelne igazából!)
- A túltanulás igazi problematikáját az adja, hogy ez utóbbi *elkerülhetetlenül* bekövetkezik, ha a tanítást túl sokáig folytatjuk (az ellentmondás a két szempont között, ugyebár)

Túlilleszkedés túl sok magyarázó változó miatt

- A magyarázó változók száma tipikus példája a tanítás fokának
- Túl kis mértékű tanítás (túl kevés magyarázó változó) esetén az alulilleszkedés miatt lesz rossz a modellünk...
- ... túl nagy mértékű tanítás (túl sok magyarázó változó) esetén a túlilleszkedés, az általánosítóképesség leromlása miatt
- Szemléletes megjelenés: a bevont magyarázó változók száma csökkenti a tesztek szabadsági fokainak számát (erre ugyanis sokszor jön elő valamilyen $n - (k + 1)$ jellegű kifejezés), így az erejüket; „leköti a szabadsági fokokat”
- Az R^2 ezt nem jellemzi, csak a mintához való illeszkedést
- Valahogy „javítani” kell; ezzel fogunk most foglalkozni

Megoldási lehetőségek I.

- Magyarázó változók számának csökkentése *csak* a bennük lévő információk alapján, tehát *nem* is nézve az eredményváltozót („blinded to the outcome")
  - A legtisztább megoldás
  - Két alapvető kivitelezési lehetőség: szakmai szempontok szerinti szűrés, vagy statisztikai alapú redundanciavizsgálat a magyarázó változók körében és redundánsak elhagyása vagy összevonása
  - Ebben segíthetnek az arra vonatkozó irányelvek, hogy adott mintanagyság mellett mennyi prediktor modellezhető
- Minden magyarázó változó felhasználása, de a regresszió regularizálása (penalizálás)
- Egyéb korszerű megoldások (pl. bayes-i modellátlagolás, BMA)

Megoldási lehetőségek II.

- Statisztikai alapú szűrés
  - Ezzel fogunk most részletesen foglalkozni
  - De vigyázat, ész nélkül nem használható, mert az *maga* is túlilleszkedéshez vezethet!
  - Ész nélkül: össze-vissza mindenféle lehetőséget megvizsgálva, hogy melyik jobb; ehelyett vezessenek minket amennyire lehet szakmai megfontolások, a próbálkozások lehetőleg legyenek pre-specifikáltak (ne az adatok sugallják őket), és ha kétség van, inkább közöljünk többféle modellt

10.2. Modellszelekció

10.2.1. A modellszelekció tartalma

A modellszelekció fogalma

- Modellszelekció alatt az optimális magyarázó változó-kör meghatározását értjük
- Ennek megfelelően foglalkozik változó bevonásának/elhagyásának hatásával…
- …de nem „mikroszkopikusan" (mi történik a többi változó becsült paramétereivel stb.), hanem „makroszkopikusan" (mi történik a modell jóságával)
- Az előbbi inkább a modellspecifikáció kérdése, később fogunk vele foglalkozni
- Továbbá: a modellspecifikációhoz szoktuk sorolni az adott magyarázó változó-kör melletti függvényforma kialakítást (de nincs egyértelmű határ a kettő között)

A modellszelekció problematikájának megoldása

- Az biztos, hogy a mintához való illeszkedés az R^2-tel jellemezhető
- Innentől két lépésben lehet továbbhaladni a modellszelekcióval:
  1. Két modell között úgy döntünk, hogy megnézzük, hogy lényeges-e köztük az R^2-beli különbség… és csak akkor választjuk a bővebbet, ha nem csak nagyobb (ez biztos), de *lényegesen* nagyobb az R^2-e (más szóval: egy modellből mindazon változókat elhagyjuk, melyek nem csökkentik *lényegesen* az R^2-et, még ha számszerűen csökkentik is)
  2. Definiálunk olyan mutatót az R^2 helyett, mely az R^2-hez hasonlóan figyelembe veszi a mintához való illeszkedést, de – azzal szemben – az ehhez szükséges magyarázó változók számát *is*
- Most e két kérdést fogjuk közelebbről is megvizsgálni

Itt (és mindenhol máshol is) a „lényeges”-et a „statisztikailag szignifikáns” szinonimájaként használjuk, tehát természetesen úgy értjük, hogy mintavételi értelemben lényeges, azaz olyan mértékű, ami nem egyeztethető össze a mintavételi ingadozással: adott szignifikanciaszinten nem hihető, hogy a változás pusztán a mintavételi ingadozásnak tudható be, ezzel szemben feltehető, hogy tényleges sokasági különbség van a hátterében.

10.2.2. Modellszelekciós tesztek

A modellszűkítésről

- Már láttuk, hogy miért akarhatunk modellt szűkíteni (változót elhagyni a modellből), még ha ezzel rontunk is az R^2-en (és még látni fogunk más okot is)
- Melyik változót lehet érdemes ezek miatt elhagyni? $\rightarrow$ *mérlegelés* a fentiekben javulás és az R^2 romlása között
- Sok vagy kevés a romlás? – a szó statisztikai értelmében lényeges-e!
- Azaz túlmutat-e a mintavételi ingadozáson: ehhez teszt kell

Nested (beágyazott) modellszelekció: a szűkebb modell minden változója benne van a bővebb modellben – elhagyhatóak anélkül, hogy a modell lényegesen romlana (azt is jelenti, hogy a két R^2 között nincs lényeges különbség)

Kitérő: modellezési filozófiák

Az LM és a Wald-teszt eltérései

- Ha ugyanazt a hipotézist vizsgálják, mi a különbség köztük?
- A nyilvánvaló: teljesen más elven épülnek fel

- Ennek konkrétabb következményei:
  1. Nem feltétlenül ugyanakkor utasítanak el; sőt, ennél több is mondható: az LM-próba *mindig* az elfogadás felé „hajlik” (olyan értelemben, hogy ha ez elutasít, akkor a Wald is, viszont ha a Wald elfogad, akkor az LM is elfogad)
  2. A Wald kismintás próba, az LM-próba nagymintás (értsd: tulajdonságai csak aszimptotikus értelemben garantáltak), de azért a gyakorlatban már néhányszor 10 mintaelemre is elég jól szokott közelíteni
  3. Belátható, hogy a Wald-teszt csak a korlátozatlan, az LM-teszt csak a korlátozott modell becslését igényli; ez utóbbi egyszerűbb (gyakorlatban számít!)

A „kismintás” természetesen nem azt jelenti, hogy a teszt csak kis mintán működik, hanem épp ellenkezőleg: azt, hogy *minden* mintanagyság mellett működik (szemben a nagymintás teszttel, ami *csak* nagy mintán működik!). Talán szerencsésebb is ezért a „véges mintás” kifejezés, mely egyúttal a különbségtétel valódi okára is utal: a kismintás tesztek tulajdonságai „$\forall n$”, míg a nagymintások „$\lim_{n\to\infty}$” értelemben garantáltak.

Az LM és a Wald-teszt eltérései

- Van egy általánosabb különbség is: más modellezési filozófiához illeszkednek
- A Wald-teszt inkább az „általánostól az egyszerűig” filozófiának (Hendry/LSE) felel meg (a korlátozatlan modellből indul, és kérdezi, hogy lépjünk-e a csökkentés irányába)
- Az LM-próba inkább az „egyszerűtől az általánosig” filozófiának felel meg (a korlátozott modellből indul, és kérdezi, hogy lépjünk-e a bővítés irányába)
- … hát ez a különbség – hiába ugyanaz *formailag* a hipotézispár!

Nem igazán lehet válaszolni arra a kérdésre, hogy melyik a „jobb” modellezési filozófia: nagyon sok, részben egymásnak ellentmondó, elméleti és gyakorlati szempont merül fel a választásnál. Ezzel a kérdéssel könyvtárnyi irodalom foglalkozik.

Az LM és a Wald-teszt eltérései

- Már most megjegyezzük, hogy az „újonnan felvett” változó nem szükségszerű, hogy még nem szereplő változó legyen: lehet egy már bent levő változó valamilyen új, nemlineáris függvényformája (pl. négyzete), vagy változók interakciója (ld. később)
- E célra általában LM-tesztet használnak, emiatt igaz az, hogy az LM-elvű teszteket kicsit általánosabban is használják az ökonometriában, más hipotézisek tesztelésére is
- … tehát: ez modellspecifikációs tesztként is felhasználható!

10.2.3. Modellszelekciós mutatók, kritériumok

Az R^2 „megjavítása”

- Ahogy láttuk az R^2 *önmagában* nem minősít egy modellt, mert csak a hibát minimalizálja, a túl sok változó káros hatásával egyáltalán nem foglalkozik („egyoldalú” mérlegelés)
- Nem lehetne ezt valahogy kijavítani? → olyan mutatót konstruálni, ami mindkét szempontra tekintettel van
- Ötlet: induljunk ki az R^2-ből, de büntessük a magyarázó változók számának növelését
- Bár máshonnan származik, de épp ennek a logikának felel meg (gondoljuk végig!) a *korrigált* R^2:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}$$

- Ez már alkalmas különböző számú magyarázó változót tartalmazó modellek összehasonlítására

A korrigált R^2 klasszikus bevezetése szerint $1 - \bar{R}^2$ megegyezik a (sokasági) hibatag és az eredményváltozó becsült varianciáinak a hányadosával. A „sima” R^2-et ugyanis úgy definiáltuk mint $R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$, ami helyett azt is írhattuk volna, hogy $R^2 = 1 - \frac{ESS/n}{TSS/n}$, ez utóbbiból még jobban látszik, hogy két szórásnégyzet hányadosától függ. A probléma, hogy belátható módon sem az ESS/n, sem a TSS/n nem torzítatlan becslője a megfelelő szórásnégyzetnek (ez utóbbi klasszikus induktív statisztikai felismerés) – az $ESS/(n-k-1)$ és a $TSS/(n-1)$ viszont az. Ezeket behelyettesítve az előbbi képletbe nyerjük a $\bar{R}^2$-et.

Ilyen szempontból – és ez is egy nagyon fontos megállapítás önmagában is! – a korrigált R^2 azért is érdekes, mert a szokásos R^2 következtető statisztikai értelemben torzított (gondoljuk végig, hogy nem is lehet más, hiszen 0 és 1 között van mindig – mi van, ha a valódi R^2 0?), a korrigált R^2 épp ezen a torzításon javít! (Sajnos nem oldja meg teljesen: igaz, hogy a nevezőre és a számlálóra is torzítatlan becslést használ, de egy hányados nem biztos, hogy torzítatatlan attól, mert a nevezője és a számlálója is torzítatlan.)

Az $\bar{R}^2$ főbb tulajdonságai

- $\bar{R}^2 \leq R^2$
- Ebből következően 1-nél nem lehet több…
- …de 0-nál lehet kisebb (ha sok magyarázó változóval is csak gyenge magyarázást (kis R^2-et) tud elérni)
- Ez már csökkenhet is új változó bevonásával (belátható, hogy ez a változó t-hányadosától függ)

- Ilyen módon már nem beágyazott modellek szelekciójára is használható…
- … de vigyázat: csak akkor, ha az eredményváltozó azért ugyanaz (különben a megmagyarázandó variancia is más lenne)

Automatikus modellszelekció

- Megadjuk a változók egy maximális halmazát (l darab potenciálisan szóba jövő magyarázó változó), és „a gép" kiválasztja, hogy melyik részhalmaza az optimális: melyeket érdemes egy modellbe bevonni, hogy az a legjobb legyen
- Jóság valamilyen célfüggvény szerint (ami ugye *nem* R^2, hogy a dolognak értelme is legyen, hanem pl. $\bar{R}^2$)
- Léteznek heurisztikus stratégiák (mind mohó algoritmus), hogy ne kelljen a 2^l kombinációt tesztelni (forward, backward, stepwise szelekció)
- Az automatikus modellszelekció használata azonban szinte *minden esetben* és *határozottan ellenjavallt*, az alkalmazásával nyert modelleknek torzítottak lesznek a regressziós koefficiensei, torzítottak lesznek a becsült standard hibái, ebből adódóan torzítottak lesznek a konfidenciaintervallumai, a szokásos p-értékek falsak lesznek, a rájuk alapozott tesztek invalidak, a t és F statisztikáknak nem t illetve F eloszlásuk lesz, torzított lesz a modell R^2-e stb.

Információs kritériumok

- Vannak további mutatók is, melyek egyszerre büntetik a magyarázó változók nagy számát és a nagy hibát, a kettő között egyensúlyt keresve, pl.
  - Akaike (AIC): $AIC = \frac{ESS}{n} e^{\frac{2(k+1)}{n}}$
  - Schwarz (SBC): $BIC = \frac{ESS}{n} n^{\frac{k+1}{n}}$
  - Hannan-Quinn (HQC): $HQC = \frac{ESS}{n} (\ln n)^{\frac{2(k+1)}{n}}$
- Teljesen más elven (információelméleti alapon) épülnek fel mint az $\bar{R}^2$
- Hiba jellegű mutatók, ezért őket *minimalizálni* akarjuk és nem maximalizálni!
- Sok van belőlük, döntsük el előre, hogy melyiket használjuk a modellszelekcióra!
- Ezekkel nem csak beágyazott modellek hasonlíthatóak össze (de azért jobbak a tulajdonságaik ilyenkor)

Modellszelekciós stratégiák

- Itt már látszik, hogy miért mondtuk az elején, hogy az ökonometriai munka iteratív
- Diagnosztizáljuk a modellt, és – ha ilyen baj van vele – szűkítjük vagy bővítjük, majd újra diagnosztizáljuk, majd…
- De vigyázat: újra fontos felhívni a figyelmet, hogy ha rengeteg ilyen iterációra kerül sor az értelemszerűen *maga is* túlilleszkedéshez vezethet (túlságosan rászabjuk a konkrét mintára a modellt)!

11 Az exogenitás és sérülése

11.1. Erős exogenitás

Emlékeztetőül

- Az erős exogenitási feltevés sérülésének három gyakorlatban tipikus esete van: kihagyott változó okozta torzítás (confounding!), mérési hiba, szimultaneitás
- Az utóbbi kettő meghaladja a mostani kereteket, így a továbbiakban az elsővel fogunk foglalkozni

Változó bevonásának hatása a modellre

Vessük össze ezt a két (demonstráció kedvéért igen kicsi) modellt az esettanulmány feladatára:

$$\widehat{\text{KiadEFt}} = \underset{(13,783)}{339,746} + \underset{(0,0064924)}{0,637354}\,\text{JovEFt}$$

$$T = 8314 \quad \bar{R}^2 = 0,5369 \quad F(1,8312) = 9637,2 \quad \hat{\sigma} = 662,02$$

(standard errors in parentheses)

$$\widehat{\text{KiadEFt}} = \underset{(16,988)}{283,172} + \underset{(0,0074136)}{0,616911}\,\text{JovEFt} + \underset{(6,0199)}{34,1727}\,\text{TLetszam}$$

$$T = 8314 \quad \bar{R}^2 = 0,5386 \quad F(2,8311) = 4852,8 \quad \hat{\sigma} = 660,78$$

(standard errors in parentheses)

Miért változott meg a jövedelem becsült koefficiense?

Változó bevonásának hatása a modellre

- Mondjuk, hogy a bővebb modell írja le a valóságos helyzetet (a gyakorlatban ezt persze soha nem tudhatjuk, filozófiai kérdés)
- Azaz a valós helyzet a második regresszió
- Az érdekes, hogy ez alapján *előre* meg tudjuk mondani, hogy az első regresszióban mi lesz a jövedelem együtthatója! (… és ebből persze a változás okát is rögtön le tudjuk olvasni)
- A jövedelem ugyanis nem csak a kiadásra hat sztochasztikusan, hanem összefügg a taglétszámmal is:

$$\widehat{\text{TLetszam}} = \underset{(0,025067)}{1,65553} + \underset{(1,1807\text{e}-005)}{0,000598206}\,\text{JovEFt}$$

$$T = 8314 \quad \bar{R}^2 = 0,2359 \quad F(1,8312) = 2566,9 \quad \hat{\sigma} = 1,2040$$

(standard errors in parentheses)

Változó bevonásának hatása a modellre

- Ebből összerakhatjuk a szűkebb regresszióban a jövedelem együtthatóját:

$$0{,}637 = 0{,}617 + 0{,}000598 \cdot 34{,}17$$

- A bővebb modellben az együttható 0,617: ennyi a jövedelem direkt hatása (ha egy egységgel nő stb.), és itt véget is ér a sztori, mert a bővebb modellben a taglétszámot állandó értéken tartjuk (v.ö.: c.p.) ezért nincs jelentősége a taglétszám és a jövedelem közti sztochasztikus kapcsolatnak
- A szűkebb modellben viszont a jövedelem egységnyi növekedése a taglétszámot is növeli tendenciájában, a növekvő taglétszám viszont (*önmagában* is!) növeli a kiadást, ez lesz az indirekt hatás
- Teljes hatás = direkt hatás + indirekt hatás(ok)

Változó bevonásának hatása a modellre

- A szűkebb regresszióban nem tudjuk *izolálni* a taglétszám hatását: ha a jövedelem nő, az a bővebb modellben nem társul a taglétszám növekedésével (v.ö. a paraméter c.p. értelmezésével), a szűkebb modellben viszont igen (hiszen ott nem endogén változó a taglétszám) → a szűkebb modellben a kihagyott változón keresztül terjedő hatások is *beépülnek* az együtthatóba
- Azaz: a bővebb regresszióval, az új változó bevonásával védekeztünk a confounding ellen (kiszűrtük a hatását: kontrolláltunk az újonnan bevont változóra)
- A gyakorlatban persze nem tudhatjuk, hogy mi a „kihagyott változó”

A specifikációs torzítás és iránya

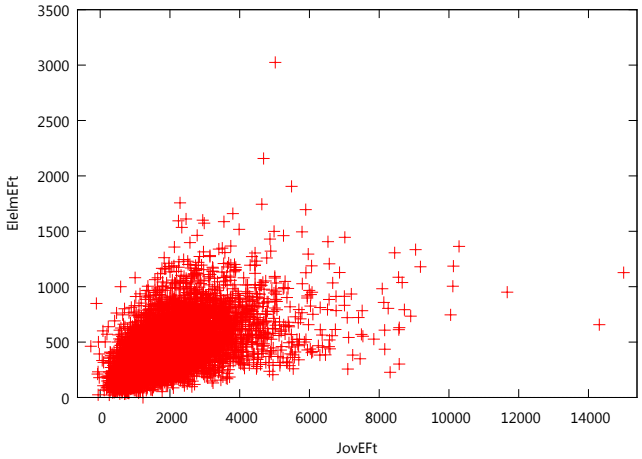
- Akkor van tehát kihagyott változó okozta torzítás, ha *egyszerre* fennáll két feltétel: a kihagyott változónak van – önmagában – hatása az eredményváltozóra (tehát a β-ja nem nulla), *és* korrelált a bennmaradt magyarázó változóval
- Ebből adódóan a torzítás iránya negatív és pozitív is lehet attól függően, hogy ez a két tényező milyen előjelű
- Belátható, hogy többváltozós esetben, ha csak egy változónak is van endogenitási baja, akkor is torzított lesz az *összes* változó becsült koefficiense

12 A homoszkedaszticitás és sérülése

12.1. A heteroszkedaszticitás és következményei

Példa a heteroszkedaszticitásra

Először próbáljunk szemléletes képet kapni a heteroszkedaszticitásról:



Emlékeztetőül

- A feltétel: $\sigma_i^2 := \mathbb{D}^2\left(\varepsilon_i \mid \underline{\underline{X}}\right) = \sigma^2$ i-től függetlenül minden $i = 1, 2, \ldots, n$
- Vagy, ezzel egyenértékűen: $\mathbb{E}\left(\varepsilon_i^2 \mid \underline{X}_i\right) = \sigma^2$

A heteroszkedaszticitás okai

A heteroszkedaszticitás oka lehet:

1. A jelenség természetes velejárója (ld. az élelmiszerfogyasztás, vagy általában a kiadások példáját: „bővülő lehetőségek az ízlés kiélésére")
2. Csoportosított adatok használatakor: például háztartásonként átlagoljuk a jövedelmet és az élelmiszerekre fordított kiadást → még ha egy háztartástag szintjén állandó σ^2 is a szórásnégyzet, a csoportosított adatokban ez σ^2/n_i lesz, ahol n_i az i-edik háztartás létszáma, ami nagyon is eltérő lesz háztartásról-háztartásra (azaz megfigyelési egységenként)

Heteroszkedaszticitás következményei

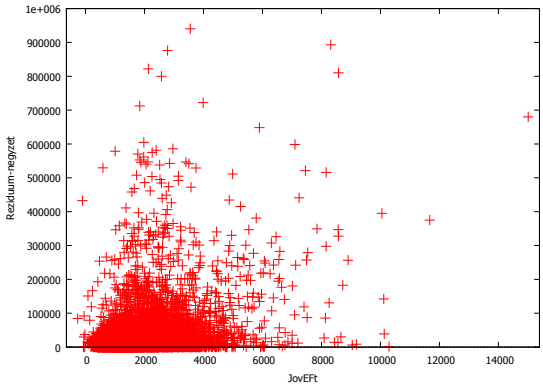
Mi történik, ha a heteroszkedaszticitással nem törődve, továbbra is a szokásos OLS-t alkalmazzuk a becslésre?

- Ahogy láttuk, az OLS szolgáltatta paraméter-becslések továbbra is torzítatlanok és konzisztensek lesznek…
- …de már nem lesz hatásos (elveszíti a BLUE-ságot) → lesz olyan lineáris torzítatlan becslés, aminek kisebb a varianciája
- Ráadásul a becsült standard hibák (illetve általában a paraméterek becsült kovariancia-mátrixa) még torzított és inkonzisztens is lesz!
- A t- és F-statisztikáknak még aszimptotikusan sem lesz t-, illetve F-eloszlásuk
- Azaz a tesztek és a paraméterekre adott konfidencia-intervallumok érvényüket vesztik
- Az előrejelzések torzítatlanok lesznek ugyan, de nem hatásosak

12.2. A heteroszkedaszticitás tesztelése

Grafikus módszerek

- Nem analitikus, de benyomás-szerzésre jó lehet
- Reziduum-négyzetek kiplottolása különféle magyarázóváltozókkal (vagy a becsült eredményváltozóval) szemben:



Azért kell különböző magyarázóváltozókat használni, mert előzetesen általában nem tudhatjuk, hogy melyik „felel” a heteroszkedaszticitásért (erre mindjárt visszatérünk még részletesebben is a tesztek kapcsán). Sőt, az is lehet, hogy nem konkrétan egy, ez korlátozza a grafikus módszer alkalmazhatóságát.

Azért a reziduum-négyzeteket célszerű vizsgálni, mert a reziduum torzítatlanul becsli a hibát (még heteroszkedaszticitás esetén is), így a hiba szórásnégyzetének – mivelhogy várható értéke nulla – ésszerű becslője a reziduum-négyzet.

Goldfeld–Quandt (GQ) próba

- Alapötlet: rendezzük a mintát azon magyarázó változó szerint, ami mentén nem állandó a feltételes szórás (az előbbi példán mondjuk a jövedelem), vágjuk szét három részre a mintát eszerint (kis, közepes és nagy értékű e változó), és hasonlítsuk össze (F-próbával) az alsó és a felső régióban a szórást
- Előnye: egyszerű, intuitív, könnyen átlátható
- Hátrányai:
  - Tudni kell, hogy mely változó mentén nem állandó a szórás (és muszáj, hogy egyetlen ilyet mutassunk)
  - Csak akkor jó, ha e változó mentén monoton módon változik a szórás
  - Gazdaságtalan, nem használ fel minden információt a mintából (a középső részt egyszerűen kidobja a kukába!)

A „monoton változás" feltétele alatt azt értjük, hogy a homoszkedaszticitást már az is elrontja, ha egy változó mentén eleinte nő, aztán csökken a feltételes szórás, viszont egy ilyen helyzetet a GQ-próba nem feltétlenül fog észrevenni (kis szórást hasonlít kis szórással, azt fogja hinni, hogy minden rendben).

LM-próbák

- Mi volna, ha a modell paraméterének tekintenénk, hogy az i-edik megfigyelési egységnél mekkora a σ_i^2 feltételes variancia?
- Önmagában nyilván rossz ötlet: lehetetlen lesz megbecsülni, hiszen minden paraméterre csak egyetlen megfigyelésünk (az $\widehat{u}_i^2$ reziduum-négyzet) lesz
- De: egyszerűsítsük a struktúráját! Azaz: feltételezzünk egy kevesebb paraméterre redukált formát, mely meghatározza a feltételes szórást

LM-próbák

- Tehát: van elképzelés, hogy mely változók „felelősek" potenciálisan a heteroszkedaszticitásért, melyek mozgatják a hibatag szórását → rakjunk erre egy (lineáris regressziós) modellt; pár lehetséges példa erre:

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$
$$\sigma_i = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$
$$\ln\left(\sigma_i^2\right) = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$

- Itt Z_i-k ismert változók, melyek körét mi határozzuk meg, mint amik „felelhetnek" a nem-állandó szórásért (ezek természetesen részben vagy egészben magyarázó változók is lehetnek az eredeti regresszióban)
- A σ_i feltételes szórás helyébe annak a becslőjét, az $|\widehat{u}_i|$ reziduumot írjuk a segédregresszióban

LM-próbák

- Akkor nincs heteroszkedaszticitás, ha a segédregresszióban teljesül a $H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_P = 0$ nullhipotézis (hiszen ekkor a σ_i speciálisan állandó lesz, nekünk épp ez kellett)
- Ezt ún. LM-elven vizsgáljuk (részletesen lásd később), a lényeg, hogy az erre irányuló próba:
$$LM_{\text{emp}} = nR^2 \overset{H_0}{\sim} \chi^2_{p-1},$$
ahol R^2 természetesen a *segéd*regresszió többszörös determinációs együtthatója
- A fenti modellekhez tartozó próbák nevei rendre: Breusch–Pagan-próba, Glejser-próba, Harvey–Godfrey-próba (ún. multiplikatív heteroszkedaszticitásra)
- (Valójában a próbák eredetileg kicsit más alakúak, de nagy mintán egységesen a fenti formára hozhatóak)

Az ún. Park-próba a Harvey–Godfrey-próba speciális esete.

LM-próbák

- Előnyeik:
  - Ezek már minden információt felhasználnak
  - Nem muszáj, hogy a heteroszkedaszticitásért egyetlen változó legyen felelős
- Hátrányaik:
  - Továbbra is nekünk kell tudnunk, hogy mely változó(k) felelős(ek) a nem-állandó szórásért
  - Hibanormalitást igényelnek és erre érzékenyek

Breusch–Pagan (BP) próba

- Még egyszer, a segédregressziója:
$$\widehat{u}_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP} + e_i$$
- (Valójában a jobb oldal helyett egy $f\left(\alpha_1 + \alpha_2 Z_{i2} + \ldots + \alpha_P Z_{iP}\right)$ transzformáltat is tekinthetnénk valamilyen f függvénnyel, a próba végeredménye ugyanis belátható-an ugyanaz lesz, ez tehát erre általánosodik)
- A hibanormalitásra robusztusabb változata: Koenker-próba

White–teszt

- Az összes eddigi tesztnek még mindig hátránya, hogy tudni kell, hogy mi mozgatja a heteroszkedaszticitást
- A White-próba ötlete: ha nincs ötletünk, használjunk „mindent”, ami a változóinkból kinyerhető (a valódi ok inkább az, hogy a homoszkedaszticitási feltétel gyengíthető arra, hogy az interakciókkal és a kvadratikus hatásokkal nincs összefüggése ε^2-nek)
- Minden: az összes magyarázó változó, az összes interakció, az összes kvadratikus hatás (persze csak ahol van értelme)
- Innentől olyan, mint a BP-próba
- Nagy mintás
- További előnye, hogy a hibanormalitásra sem annyira érzékeny
- Hátránya: itt is érvényesül a „minél kevesebb előfeltevésre épít egy próba, annál gyengébb” elv (itt szemléletesen: nagyon megnő a segédregresszióban a magyarázó változók száma) → ha van *a priori* információnk, használjuk! (itt: ha ismerjük, mi felel a heteroszkedaszticitásért, erősebb a BP-próba)

12.3. A heteroszkedaszticitás kezelése

Modellspecifikáció változtatása

- Ötlet: úgy módosítjuk a modellspecifikációt, hogy az új specifikációban szereplő hiba már ne legyen heteroszkedasztikus
- Nem-statisztikai jellegű korrekció: szakmai ismeretet (is) igényel arról, hogy vajon mi a jó módosított specifikáció
- Nem is univerzális (nem feltétlenül alkalmazható minden esetben)
- Például:
  - Logaritmálás (ld. a nemlinearitásokról szóló részt a 6. fejezetben)
  - „Deflálás” (áttérés valamilyen méret-jellegű mutatóra leosztott változóra, pl. népességről népsűrűségre)

Heteroscedasticity Consistent Covariance Matrix, HCCM

- Ötlet: a becsült értékek torzítatlanok, azokat hagyjuk békén: maradjanak ugyanazok, mint a heteroszkedaszticitás figyelmen kívül hagyásával becsült modellben
- A standard hibákkal kéne valamit kezdeni

- HCCM nevű eljárás képes ezeket korrigálni: robusztus (vagy Huber–White–Eicker) standard hibák
- Matematikai részletekkel nem törődünk

Heteroscedasticity Consistent Covariance Matrix, HCCM

- Univerzálisan működőképes, nem igényel semmilyen feltevést a heteroszkedaszticitás struktúrájáról (legalábbis nagy mintán: itt emiatt gyakran automatikusan robusztus standard hibát adnak meg, esetleg mindkét standard hibát)
- Viszont ha fennáll a homoszkedaszticitás, akkor jobban járunk a szokásos standard hibával, mert annak a kismintás viselkedése is garantált (már csak ezért is érdekes a tesztelés)
- Alapozható rá más teszt is, nem csak a t-próba

Általánosított legkisebb négyzetek módszere (GLS)

- Ez már a teljes modellt újrabecsüli: a becsült koefficiensek is mások lesznek
- Alapötlet: a hibák kovarianciamátrixa nem skalármátrix $\rightarrow$ semmi baj, feltételezzünk egy általánosabb mátrixot, és számoljuk azzal végig a legkisebb négyzetes becslést
- Matematikai részletek nélkül a végeredmény:

$$\widehat{\boldsymbol{\beta}_{\mathrm{GLS}}} = \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{y},$$

  ahol $\boldsymbol{\Omega}$ a – teljesen általános – feltételes kovarianciamátrix: $\boldsymbol{\Omega} = \mathbb{E}\left(\underline{UU}^T \mid \underline{\underline{X}}\right)$
- A baj, hogy ez önmagában csak akkor alkalmazható, ha ismerjük ezt a $\boldsymbol{\Omega}$ mátrixot, azaz az egyes – feltételes – szórásokat

Érdemes észrevenni, hogy $\boldsymbol{\Omega}$-ra most már semmilyen kikötést nem tettünk (teljesen általános mátrix), így a GLS ugyanúgy alkalmazható autokorreláció jelenléte esetén is a becslésre! (Persze ugyanazzal a limitációval, hogy ti. ismerni kell ezt a mátrixot.)

Súlyozott legkisebb négyzetek módszere (WLS)

- Van gyakorlati példa arra, amikor – legalábbis konstans szorzó erejéig – ismerjük a feltételes szórásokat: ha tudjuk, hogy azok mely változóval arányosak
- Például: fogyasztási egységek számával arányos a feltételes szórás: $\sigma_i = \sigma \cdot F_i$ (a fogyasztási egységek F_i száma minden megfigyelési egységre ismert)

- Ennyi (tehát a konstans szorzó erejéig ismert feltételes szórás) már elég: ha

$$Y_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

heteroszkedasztikus is, a

$$\frac{Y_i}{\sqrt{F_i}} = \beta_1 \frac{1}{\sqrt{F_i}} + \beta_2 \frac{X_{i2}}{\sqrt{F_i}} + \ldots + \beta_k \frac{X_{ik}}{\sqrt{F_i}} + \frac{\varepsilon_i}{\sqrt{F_i}}$$

könnyen belátható, hogy nem lesz az

Súlyozott legkisebb négyzetek módszere (WLS)

- Ez az ún. súlyozott legkisebb négyzetek módszere (WLS, weighted least squares); nem keverendő össze a megfigyelési egységek súlyozásával (erre is látni fogunk példát)
- Fontos, hogy a súlyok *nem* becslésből származtak, hanem ismertek voltak (ld. a fogyasztási egységek példáját)
- Természetesen egy változónál több is felhasználható (fogyasztási egységek száma *és* település stb.), hogy leírjuk a σ_i-t és a függvényforma is lehet akármilyen bonyolult (fogyasztási egységek négyzetével arányos feltételes szórás stb.), a lényeg, hogy olyan kifejezést konstruáljunk *kizárólag* ismert változókból, mellyel egyenesen arányos lesz a feltételes szórás
- A kulcs az, hogy visszaredukáljuk egyetlen ismeretlen paraméterre a feltételes szórásokat (ugyanúgy, ahogy homoszkedaszticitásnál lenne)

Kivitelezhető általánosított legkisebb négyzetek módszere (FGLS)

- Ha nem ismert a heteroszkedaszticitás struktúrája, akkor más megoldás kell, hogy a gyakorlatban alkalmazható legyen a GLS
- Egy segédregresszióban az eredeti regresszió reziduumainak négyzeteit regresszáljuk ki a White-tesztnél látott módon, innen kapjuk a hiba becsült varianciáit
- Ezek felhasználásával egy súlyozott regressziót számítunk, amivel újra közelítjük a hibák varianciáit
- Így nyerünk – ismert struktúra nélkül is – becslést a feltételes szórásra, amit az alapregresszióban a WLS-hez hasonló módon alkalmazhatunk a heteroszkedaszticitás korrigálására
- (A segédregresszióban reziduum-négyzet helyett mást is alkalmazhattunk volna, ahogy a heteroszkedaszticitásra irányuló LM-próbáknál is volt)

13 Kategoriális eredményváltozó modellezése: a logisztikus regresszió és változatai

13.1. Általános gondolatok

Kvalitatív változó eredményváltozó pozíciójában

- Például a feladat egy „csődbe megy-e vagy sem” jellegű változó modellezése
- Ez bináris változó → mint az eddig tárgyalt dummy változók, csak ezúttal eredményváltozóként
- Jelent ez módosulást? (Hiszen például magyarázó változóként mindegy volt, hogy egy változó bináris, az OLS-t nem zavarta, hogy történetesen csak 0 és 1 értékeket vesz csak fel)
- Most drasztikusan más a helyzet: Y *nem modellezhető* OLS-sel

OLS és a bináris eredményváltozó

- Matematikai részletekbe nem megyünk bele
- Intuitíve: gondoljunk arra, hogy az OLS – elvileg – bármilyen értéket becsülhet $-\infty$ és ∞ között → egy ilyen hogyan lenne értelmezhető egy „csődbe megy-e vagy sem” kérdés válaszaként?!
- De: mégis lineáris struktúrában fogjuk megoldani a problémát… csak trükkösebben alkalmazzuk: bináris Y helyett egy transzformált változóra
- Avagy – fordítva megfogalmazva – megtartjuk a lineáris kombinációt, de annak az eredményét áteresztjük egy olyan függvényen, ami a $(-\infty, \infty)$-t a $[0, 1]$-re képezi le

A mostani feladat általánosabban

- Tegyük fel, hogy elkészült a bináris Y-ra adott modellünk, és azt előrejelzésre használjuk

- Vegyük észre, hogy az Y szerinti érték egyfajta *csoporttagságot* jelent: becsődölő, működő
- Az előrejelzés ebben a kontextusban lényegében besorolás egy csoportba!
- Tehát mégegyszer: a megfigyelési egység két csoport valamelyikébe tartozik, mi a csoporttagságával összefüggő adatok alapján tippeljük meg a csoporttagságot
- Ezt a feladatot általában *osztályozásnak* (klasszifikáció) nevezik
- A klasszifikáció hatalmas gyakorlati jelentőségű feladat: melyik cég megy csődbe (a mérlegadatai alapján), melyik beteg fog meghalni (a laboreredmények alapján), kit vesznek fel adott munkahelyre (egyéni jellemzők alapján) stb. stb.

13.2. Alapfogalmak bevezetése

A feladat átalakítása

- Hogy a kérdést a magyarázó változók lineáris kombinációjával tudjuk kezelni, áttérünk más változóra
- Először is: nem az 1-es csoportba tartozás tényét, hanem annak $\mathbb{P}_{\underline{X}}$ feltételes valószínűségét fogjuk modellezni
- Az alsó index értelme: az 1-es csoportba tartozás valószínűsége, *feltéve*, hogy a magyarázó változók $\underline{X}$ értékűek, azaz precízen: $\mathbb{P}_{\underline{X}} = \mathbb{P}\left(Y = 1 | \underline{X}\right)$
- Ezzel a $\{0, 1\}$ változó helyett egy $[0, 1]$-on lévőt kell modellezni
- Vegyük észre, hogy ezzel még nem léptünk ki az eddigi regressziós keretből, sőt, teljesen megfelelünk neki, hiszen egy bináris (0-1) változóra ez a feltételes valószínűség épp a feltételes várható érték!
- Azt fogjuk mondani, ez a későbbiek szempontjából lesz fontos, hogy az eredményváltozó eloszlása Bernoulli (p valószínűséggel vesz fel 1-et, $1 - p$ valószínűséggel 0-t), és ennek a feltételes várható értékét modellezzük

A feladat további átalakítása

- Ez persze még mindig kevés, ezért újabb transzformációt alkalmazunk
- Odds (esély) fogalma: az 1-es csoportba tartozás valószínűsége a 0-s csoportba tartozás valószínűségéhez viszonyítva, jelen esetben valószínűség osztva 1-valószínűséggel
- Azaz

$$\text{odds}_{\underline{X}} = \frac{\mathbb{P}_{\underline{X}}}{1 - \mathbb{P}_{\underline{X}}}$$

- Könnyen megoldható $\mathbb{P}_{\underline{X}}$-re:

$$\mathbb{P}_{\underline{X}} = \frac{\text{odds}_{\underline{X}}}{1 + \text{odds}_{\underline{X}}}$$

És még egy átalakítás

- Az odds már a $[0, \infty)$ intervallumon van
- Majdnem jó, egy utolsó trükk: bevezetjük a *logit* fogalmát, mint log-odds:

$$\text{logit}_{\underline{X}} = \ln \text{odds}_{\underline{X}}$$

- És ez már a $(-\infty, \infty)$-n van (és szimmetrikussá is tettük a siker és kudarc eloszlását rajta)!
- Na, ezt fogjuk lineáris struktúrával modellezni!

$$\text{logit}_{\underline{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k = \underline{X}^T \boldsymbol{\beta}$$

- A módszer neve: logit regresszió, vagy logisztikus regresszió

Figyeljük meg, hogy itt nincs hibatag, hiszen az ingadozás abban fejeződik ki, hogy ez csak a feltételes várható érték, amire „rárakódik" az eredményváltozó Bernoulli eloszlása. Igazából a lineáris regresszió is leírható lett volna így: a feltételes várható értéket modellezzük lineárisan, majd rákeverünk egy $\mathcal{N}\left(0, \sigma^2\right)$ eloszlást. Itt ugyanez történik, csak a lineáris kombinációt még meg is transzformáljuk, és nem normálisat, hanem Bernoulli-t keverünk rá. Erre a gondolatra később még visszatérünk.

A logisztikus regresszió visszafejtése

- Játszuk el mindezt visszafelé, feltéve, hogy β-k már ismert:

$$\begin{aligned}
\text{logit}_{\underline{X}} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \\
\text{odds}_{\underline{X}} &= e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k} \\
\mathbb{P}_{\underline{X}} &= \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}} = \frac{e^{\underline{X}^T \boldsymbol{\beta}}}{1 + e^{\underline{X}^T \boldsymbol{\beta}}}
\end{aligned}$$

- Az utolsó lépésben kapott $f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$ épp a korábban emlegetett, $(-\infty, \infty)$-t a $[0, 1]$-be képező függvény!
- $\boldsymbol{\beta}$ ismeretében egyszerű algebrai műveletekkel kapjuk a siker valószínűségeit
- És az utolsó lépés: hogy becsüljük meg $\boldsymbol{\beta}$-t?
- Sajnos az OLS – ahogy már mondtuk – nem jó, új módszer kell: maximum likelihood (ML) becslés

13.3. A logisztikus regresszió becslése és alkalmazása

A logisztikus regressziós modell becslése

- Minden $\mathbf{b}$ választáshoz meghatározható a minta (itt: adatbázisunk) likelihood-ja (precízen: adott $\mathbf{b}$ mellett mekkora likelihood-dal jött volna ki a mintánk)
- Ezt fogjuk a $\mathbf{b}$-ban maximalizálni, és így kapjuk $\widehat{\boldsymbol{\beta}_{\mathrm{ML}}}$-t
- Kérdés: hogyan kapjuk a minta likelihood-ját?
- Annyira nem nehéz, hiszen *egy* mintaelemre a kijövetelének valószínűsége $\mathbb{P}_{\underline{X}}$ (ha az eredményváltozója 1) illetve $1 - \mathbb{P}_{\underline{X}}$ (ha eredményváltozója 0), mely értékek kiszámíthatóak adott $\mathbf{b}$ mellett (már láttuk is)
- Már csak az egész mintára (nem egyes mintaelemekre) kell kiszámítani, itt függetlenség feltételezésével élünk

A logisztikus regressziós modell becslése

- Az egész minta likelihood-ja így:

$$L(b_0, b_1, \ldots, b_k) = \prod_{Y_i=1} \mathbb{P}_{\underline{X}_i} \prod_{Y_i=0} \left(1 - \mathbb{P}_{\underline{X}_i}\right) = \prod_{i=1}^{n} \mathbb{P}_{\underline{X}_i}^{Y_i} \left(1 - \mathbb{P}_{\underline{X}_i}\right)^{1-Y_i} =$$
$$= \prod_{i=1}^{n} \left(\frac{e^{\underline{X}_i^T \mathbf{b}}}{1 + e^{\underline{X}_i^T \mathbf{b}}}\right)^{Y_i} \left[1 - \left(\frac{e^{\underline{X}_i^T \mathbf{b}}}{1 + e^{\underline{X}_i^T \mathbf{b}}}\right)\right]^{1-Y_i}$$

- Ezzel a megoldandó feladat:

$$\max_{b_0, b_1, \ldots, b_k} L(b_0, b_1, \ldots, b_k)$$

A logisztikus regressziós modell becslése

- E helyett a gyakorlatban inkább a vele ekvivalens

$$\min_{b_0, b_1, \ldots, b_k} -2 \ln L(b_0, b_1, \ldots, b_k)$$

  feladatot oldjuk meg (nem csak numerikus okokból)
- Fontos különbség, hogy míg lineáris regresszió esetén volt zárt alakú megoldás, itt általában nincs, numerikus eljárást kell használni

Alkalmazás: elemzés

- Értelmezzük az együtthatókat:

$$\frac{\text{odds}_{X_1,\ldots,X_{l-1},X_l+1,X_{l+1},\ldots,X_k}}{\text{odds}_{X_1,\ldots,X_{l-1},X_l,X_{l+1},\ldots,X_k}} = \frac{e^{X_1,\ldots,X_{l-1},X_l+1,X_{l+1},\ldots,X_k}}{e^{X_1,\ldots,X_{l-1},X_l,X_{l+1},\ldots,X_k}} = e^{\beta_l}$$

- Ezért az e^{β_l}-kat is meg szokták adni a programok, a nevük esélyhányados (odds ratio, OR)

Alkalmazás: előrejelzés

- Még egy megfontolást kell tenni: csak csődvalószínűséget kaptunk… de az előrejelzésben konkrét kimenet kell! Mikor soroljuk becsődölőbe? Ha ez a valószínűség 0,5-nél nagyobb? 0,1-nél? 0,99-nél…?

- Jelölje ezt a határt C (cut-off point, cut value):

$$\widehat{Y} = 1 \Leftrightarrow \mathbb{P}_{\underline{X}} > C$$

- Ekkor különböző C-khez különböző konkrét klasszifikációk tartoznak

A klasszifikáció jóságának mérése

- Legalapvetőbb eszköz a klasszifikációs mátrix:

	$\widehat{y}=1$	$\widehat{y}=0$
$y=1$	6	1
$y=0$	5	38

- Főátlóban a helyes osztályozások, ezek aránya a helyes osztályozási ráta (itt $\frac{6+38}{6+1+5+38} = 0{,}88$)

- Mellékátlóban: első- és másodfajú hibák (specificitás, szenzitivitás)

- Gondoljuk végig, hogyan változik ezek aránya C növelésére, ill. csökkentésére

- Szenzitivitás az (1-specificitás) függvényeben különböző C-kre: ROC-görbe (terület alatta: AUC)

C megválasztása veszteség-függvény alapján

- Ha tudjuk, hogy az egyes hibák milyen „költséget” jelentenek, akkor analitikusan választhatunk optimális C-t
- Veszteség-mátrix:

	$\widehat{y}=1$	$\widehat{y}=0$
$y=1$	0	1
$y=0$	0,2	-0,2

- Ezzel az előző klasszifikációs mátrix költsége:

$$6 \cdot 0 + 1 \cdot 1 + 5 \cdot 0{,}2 + 38 \cdot (-0{,}2) = -5{,}6$$

- Azt a C-t választjuk, aminél ez minimális!

C megválasztása veszteség-függvény nélkül

- C korrekt megválasztása *csak* veszteség-függvény ismeretében lehetséges: ha nem tudjuk, hogy milyen súlyú a kétféle hibázás, akkor honnan tudhatnánk egyáltalán megmondani, hogy mi az, hogy „jó” választás?
- Néha azonban mégis rákényszerülünk a veszteségek ismerete nélküli döntésre
- Klasszikus (nem ROC-görbére támaszkodó) heurisztikák:
  - Fix 0,5-ös cutoff
  - A cutoff legyen az 1-esek mintabeli aránya
  - A cutoff legyen olyan, hogy azzal a predikált 1-esek aránya megegyezzen az 1-esek mintabeli arányával
- Optimalizálás a ROC-görbe alapján:
  - A specificitás és a szenzitivitás összege legyen maximális (Youden-szabály)
  - A bal felső – optimális – ponthoz legközelebbi pont választása (azaz $(1-Se)^2 + (1-Sp)^2$ legyen minimális)

Modelljellemzés pszeudo-R^2 mutatóval

- Az OLS-nél látott R^2-hez hasonló elvű („hol járunk az úton?”) mutató szeretnénk LR-re is
- Az ESS helyett itt a $-2\ln L$ jellemzi a modellt
- Mi a tökéletes modell? $\rightarrow \mathbb{P}_{\underline{X}} = 1$ ha $Y=1$ és $\mathbb{P}_{\underline{X}} = 0$ ha $Y=0$ $\rightarrow$ mennyi ennek a likelihoodja?

- Épp 1, $-2\ln L = 0$
- Az üres – semmilyen magyarázó változót nem tartalmazó modell – $-2\ln L$-je analitikusan meghatározható (analóg a helyzet az OLS-sel)
- Az alapján a McFadden-féle pszeudo-R^2:

$$R^2 = \frac{(-2\ln L_{\text{null}}) - \left(-2\ln L_{\text{targy}}\right)}{-2\ln L_{\text{null}}}$$

- Sok fenntartás van az ilyen mutatókkal kapcsolatban!

Modellszelekció

- Nested modellszelekció,

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$$

- Ha nagy mintánk van, akkor rendkívül kényelmesen vizsgálható egy új próbakészítési elvvel, az ún. likelihood-hányados (LR) elven konstruált teszttel:

$$\left(-2\ln \widehat{L}_{H_0}\right) - \left(-2\ln \widehat{L}_{H_1}\right) \sim \chi^2_m$$

- Üres modelltől való szignifikáns különbözés tesztelése: függetlenségvizsgálat
- Szaturált modelltől van szignifikáns különbözés tesztelése: illeszkedésvizsgálat

14 Az általánosított lineáris modell (GLM)

14.1. Az általánosított lineáris modell (GLM)

A lineáris és a logisztikus regresszió közös keretben

- Vegyük észre a hasonlóságokat!
  1. Van valamilyen eredményváltozó-eloszlás
     - Lineárisnál normális, logisztikusnál Bernoulli
  2. A feltételes várhatóérték valamilyen transzformáltját modellezzük: $g\left[\mathbb{E}\left(Y|\underline{X}\right)\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
     - Lineárisnál az identitás, logisztikusnál a korábban látott f (pontosabban szólva annak az inverze)
  3. Elvileg valamit mondani kellhet a varianciáról is
     - Lineárisnál azt, hogy $\sigma_i^2 = \sigma_0^2$, logisztikusnál megspóroltuk ezt, mert a várható értéke meghatározta a szórást is (egy paramétere volt az eloszlásnak)

Az általánosított lineáris modell (GLM)

- A fenti komponensek határozzák meg az ún. általánosított lineáris modellt (generalized linear modell, GLM)
- Az eredményváltozó eloszlása legyen exponenciális eloszláscsaládból származó
- A g függvény neve: link függvény
- Becslés maximum likelihood-dal
- A lineáris és logisztikus regresszió mind speciális esete ennek (alkalmasan választott eredményváltozó eloszlással, link függvénnyel és szórás-függvénnyel)
- Sok minden más is ide tartozik, lássunk még egy példát

Poisson regresszió

- Mi van, ha az eredményváltozó valamilyen darabszám, események száma jellegű változó (count data)?
- Ilyenekre tipikusan feltételezett eloszlás első közelítésben: Poisson-eloszlás
- Ez exponenciális családbeli
- Várható értéke itt is épp a paramétere
- Tipikus link függvény választás: a log
- Összerakva mindezeket a modellünk:

$$Y \sim \text{Poi}(\lambda)$$
$$\log\left[\mathbb{E}\left(Y|\underline{X}\right)\right] = \log\lambda = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$