BUDAPESTI CORVINUS EGYETEM
MATEMATIKAI KÖZGAZDASÁGTAN ÉS GAZDASÁGELEMZÉS TANSZÉK

AZ ÖKONOMETRIA ALAPJAI:

Többváltozós lineáris regresszió és kiterjesztései

VINCZE JÁNOS

Budapest, 2018. december

Tartalomjegyzék

1 Bevezetés

Ez a jegyzet a Gazdaság és Pénzügymatematikai Elemző osztatlan szak "Bevezetés az ökonometriába" tárgyához készült, amit a szak hallgatói a II. év második szemeszterében hallgatnak. Kiegészíti *R. Ramanathan: Bevezetés az ökonometriába alkalmazásokkal* (2003, PANEM, Budapest) című tankönyvét a szak speciális igényeinek és tulajdonságainak megfelelően. A szak hallgatói, amikor ökonometriát tanulnak, már a szokásos közgazdasági képzésekhez képest magasabb szintű analízis, algebra és valószínűségszámítási ismeretekkel rendelkeznek. A jegyzet arra alapoz, hogy emiatt "absztraktabb" ökonometriai ismereteket lehet nyújtani nekik, mint amit a nevezett tankönyv lehetővé tesz. Másfelől, míg a Ramanathan könyv gyakorlati szempontból a Gretl statisztikai programcsomag használatára alapul, ez az anyag az R nyelv használatán keresztül mutatja be az ökonometriai számításokat.

A jegyzet bizonyos, a továbbiakban fontos, matematikai fogalmakat és állításokat tartalmazó fejezettel indul. Ezt követi a klasszikus statisztikai módszereket ismertető fejezet. Egy nem szorosan a tananyaghoz tartozó fejezet megismertet a bayes-i ökonometria alapfogalmaival. Az anyag fő része egy meglehetősen hosszú fejezet a lineáris regressziós modellről, ami több hét tananyagát foglalja magában. Külön fejezetek foglalkoznak a heteroszkedaszticitással, az instrumentális becslésekkel és a kvalitatív változókat tartalmazó modellekkel. Végül egy a hagyományos statisztikai gondolkodásól némileg eltérő filozófiát használó megközelítésről is szól egy - ismét nem szorosan véve tananyag - fejezet. A tankönyvhöz képest a legfontosabb hiány az idősorelemzés, ami teljesen kimarad, de evvel a témával külön jegyzet foglalkozik. A fejezetek végén találhatók összegyűjtve a témákhoz tartozó R programok és példák, magyarázatokkal ellátva.

2 Matematikai alapok

2.1 Valószínűség-számítás

2.1.1 Valós valószínűségi változó

Létezik (Ω, Φ, P) valószínűségi mező, és $X : \Omega \to R$ P-mérhető függvény. Ekkor X-et valós valószínűségi változónak nevezzük. Ennek eloszlásfüggvénye:

$$F(x) = P\left\{\omega \mid X(\omega) < x\right\}.$$

Ha

$$F'(x) = f(x)$$

létezik, akkor a változót folytonosnak nevezzük.

2.1.2 Többdimenziós eloszlás

Létezik (Ω, Φ, P) valószínűségi mező, és $\mathbf{X} : \Omega \to R^n$ P-mérhető függvény, akkor $\mathbf{X}$ n-dimenziós valós valószínűségi vektor változó. A többdimenziós eloszlásfüggvény:

$$F(\mathbf{x}) = P\left\{\omega \mid \mathbf{X}(\omega) < \mathbf{x}\right\}$$

ahol $\mathbf{x} \in R^N$.
Speciálisan kétdimenziós változókra, ha folytonosak:

$$F(x_1, x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f(x_1, x_2) ds_1 ds_2.$$

Definiálható a peremeloszlás függvény:

$$F_1(x_1) = \int_{-\infty}^{+\infty} F(x_1, s_2) ds_2,$$

és a peremsűrűség függvény:

$$F_1(x_1) = \int_{-\infty}^{x_1}\int_{-\infty}^{+\infty} f(x_1, x_2) ds_2 = \int_{-\infty}^{x_1} f_1(s_1) ds_1.$$

Az x_1 és x_2 valószínűségi változók függetlenek, ha

$$F(x_1, x_2) = F_1(x_1) F_2(x_2)$$

$$f(x_1, x_2) = f_1(x_1) f_2(x_2).$$

A feltételes eloszlásfüggvény definíciója:

$$F(x_1 \mid x_2) = \frac{F(x_1, x_2)}{F_2(x_2)}.$$

Míg a feltételes sűrűségfüggvényé:

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}.$$

2.1.3 Egy valószínűségi változó momentumai

Egy X valószínűségi változó, melynek eloszlásfüggvénye F, elsőrendű momentuma, vagy várható értéke:

$$E(X) = \int x dF.$$

Ha létezik f sűrűségfüggvény, akkor

$$E(X) = \int x f(x) dx.$$

Amennyiben az eloszlás diszkrét, akkor a formula

$$E(X) = \sum p_i x_i,$$

ahol p_i az x_i valószínűsége.
Ha $Y = g(X)$ is valószínűségi változó, akkor a megfelelő formulák:

$$\begin{aligned} E(Y) &= \int g(x) dF \\ E(Y) &= \int g(x) f(x) dx \\ E(Y) &= \sum p_i g(x_i). \end{aligned}$$

Állítás: a várható érték lineáris operátor.

$$\begin{aligned} E(a + bX) &= a + bE(X) \\ E(\sum \alpha_i X_i) &= \sum \alpha_i E(X_i). \end{aligned}$$

A centrális másodrendű momentum (variancia) definíciója:

$$var(X) = E(X - E(X))^2 = E(X^2) - E(X)^2.$$

Láthatólag ez is egy várható érték: az $(X - E(X))^2$ változó várható értéke.
A variancia fontos tulajdonsága:

$$var(a + bX) = b^2 var(X).$$

A kovariancia a variancia általánosítása, két valószínűségi változóhoz rendel egy értéket (vegyes másodrendű momentum):

$$cov(X,Y) = E((X - E(X)(y - E(Y)) = E(XY) - E(X)E(Y).$$

Egyszerű számolással igazolható, hogy

$$var(\sum_i \alpha_i X_i) = \sum_i \alpha_i^2 var(X_i) + 2\sum_i \sum_{j \neq i} \alpha_i \alpha_j cov(X_i, X_j).$$

Belátható, hogy amennyiben X és Y függetlenek, akkor $cov(X,Y) = 0$.

A korreláció a mértékegység függetlenné tett kovariancia:

$$-1 \leq cor(X,Y) = \frac{cov(X,Y)}{\sqrt{varX}\sqrt{varY}} \leq 1.$$

Többdimenziós eloszlásokra, ha $\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu}$ és $\mathbf{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ vektor elemeinek kovarianciáiból felépített mátrix ($\boldsymbol{\Sigma}_{ij} = cov(x_i, x_j)$), akkor $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ esetén:

$$\begin{aligned} \mathbf{E}(\mathbf{y}) &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\ \mathbf{cov}(\mathbf{y}) &= \mathbf{A} \qquad \boldsymbol{\Sigma}\mathbf{A}'. \end{aligned}$$

2.1.4 A normális eloszlás

Normális eloszlású valószínűségi változó sűrűség és eloszlásfüggvénye:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ F(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp(-\frac{(t-\mu)^2}{2\sigma^2}) dt. \end{aligned}$$

A várható értéke és a varianciája:

$$\begin{aligned} E(x) &= \mu \\ var(x) &= \sigma^2. \end{aligned}$$

Speciális szerepe van a standard normális eloszlásnak (z), amelyre $\mu = 0$,és $\sigma^2 = 1$.

Állítás: Ha X normális, akkor $Y = aX + b$ is normális és

$$\begin{aligned} E(Y) &= aE(X) + b \\ var(Y) &= a^2 var(X). \end{aligned}$$

Következmény: Ha X véletlen változó $N(\mu, \sigma)$ (jelölési szokás: $X \sim N(\mu, \sigma)$)

$$f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

sűrűségfüggvénnyel, akkor az $Y = \frac{X-\mu}{\sigma}$ (standardizált véletlen változó, azaz $X \sim N(0,1)$.)

A lognormális eloszlást olyan változókra értelmezhetjük, amelyek csak pozitív értékeket vehetnek fel. Azt mondjuk, hogy X lognormális, ha $Y = \ln X$ normális.

$$Y = \ln X \tilde{} N(\mu, \sigma).$$

Ekkor $X = \exp(Y)$ és

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp(-\frac{(\ln x - \mu)^2}{2\sigma^2})$$

$$\begin{aligned} E(x) &= \exp(\mu + \frac{\sigma^2}{2}) \\ var(x) &= \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1). \end{aligned}$$

A többdimenziós normális eloszlás A sűrűségfüggvény

$$f(\mathbf{x}) = (2\pi)^{-n/2} (\det \Sigma)^{-1/2} \exp((-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

alakú, ahol $\boldsymbol{\mu}$ a várható érték vektor, és $\boldsymbol{\Sigma}$ a kovariancia mátrix.
Ha $\mathbf{x}$ normális eloszlású, akkor $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ is normális eloszlású és

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tilde{} N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

Minden normális vektort előállíthatunk az n dimenziós standard normális eloszlásból, amelynek sűrűségfüggvénye

$$f(\mathbf{x}) = (2\pi)^{-n/2} \exp(-\frac{1}{2}\mathbf{x}'\mathbf{x}),$$

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{0} \quad , \\ \boldsymbol{\Sigma} &= \mathbf{I} \quad . \end{aligned}$$

Mivel

$$\mathbf{A}\mathbf{A}' = \mathbf{C} = \mathbf{C}^{1/2}\mathbf{C}^{1/2},$$

ha $\mathbf{x}$ standard normális, akkor a kovariancia mátrixa és inverze is $\mathbf{I}$, és

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{b} \tilde{} N(\mathbf{b}, \mathbf{C}) \\ & \quad \mathbf{C}^{1/2}\mathbf{x} + \mathbf{b} \tilde{} N(\mathbf{b}, \mathbf{C}). \end{aligned}$$

(Itt kihasználjuk, hogy szigorúan pozitív definit mátrixoknak létezik négyzetgyökük.)

Ha $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ n dimenziós, akkor

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$$

is n dimenziós standard normális.

Állítás Normális együttes eloszlás peremeloszlásai és feltételes eloszlásai is normálisak.

$$\mathbf{E}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

$$\mathbf{cov}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{x}_1 \tilde{} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

$$\mathbf{x}_2 \tilde{} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

$$(\mathbf{x}_1 | \mathbf{x}_2) \tilde{} N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}),$$

ahol

$$\begin{aligned} \boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

2.1.5 Fontos eloszlások a matematikai statisztikában

A **Khi-négyzet eloszlás**: χ_k^2

A k szabadságfokú Khi-négyzet eloszlás k darab független standard normális eloszlás négyzetének az összege.

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2.$$

Várható értéke k, és varianciája $2k$.

Példa: Mint láttuk, ha $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ n dimenziós, akkor $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{Y}$ $\sim \mathbf{N}(\mathbf{0}, \mathbf{I})$. Tehát

$$\begin{aligned} \mathbf{y}'\mathbf{y} &= (\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}))'\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}) = \\ (\mathbf{x}-\boldsymbol{\mu}))'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &\sim \chi^2_n. \end{aligned}$$

A **Student-féle t eloszlás**: t_k.
Ha Z standard normális és $U \sim \chi^2_k$ függetlenek, akkor

$$\frac{Z}{\sqrt{\frac{U}{k}}} \sim t_k.$$

k szabadságfokú t eloszlás. Ekkor $E(t_k) = 0$,és $var(t_k) = \frac{k}{k-2}$.Az utóbbi formulából sejthető, hogy ha $k \to \infty$, akkor a t_k varianciája tart 1-hez. Több is belátható: a t_k-k eloszlásfüggvényei pontonként konvergálnak a standard normális eloszlásfüggvényhez.

F eloszlás: $F_{k,n}$.
Független $U_1 \sim \chi^2_k$ és $U_2 \sim \chi^2_n$ eloszlásokból képezzük a (k, n) szabadságfokú F eloszlást.

$$\frac{\frac{U_1}{k}}{\frac{U_2}{n}} \sim F_{k,n}.$$

A várható érték csak a nevező szabadságfokától függ: $E(F_{k,n}) = \frac{n}{n-2}$.
Idempotens mátrixok:
P idempotens, ha

$$\mathbf{P}^2 = \mathbf{P}.$$

Egy idempotens mátrix minden sajátértéke 0 vagy 1, és $rank(P) = tr(P)$.
A nyom ciklikus permutáció tulajdonsága:
Bármely A, B, C mátrixra, ahol a szorzás elvégezhető:

$$tr(ABC) = tr(CAB) = tr(BCA).$$

Ha M pozitív definit idempotens mátrix, $rank(M) = n-k$, és $\xi \sim N(0, I_n)$, akkor $\xi' M \xi \sim \chi^2_{n-k}$.
Cochrane Tétel
Legyen $Z_1, ...Z_n$ iid és normális. Továbbá $U_i = \mathbf{Z}'\mathbf{Q}_i\mathbf{Z}, i = 1, .., L$, és $\sum_{i=1}^{L} r_i = n$, ahol r_i az U_i-t definiáló kvadratikus forma rangja. Ekkor minden $U_i \sim \chi^2_{r_i}$.

2.1.6 Valószínűségi változók sorozatainak határértéke

Nagy Számok (gyenge) Törvénye Ha X_i $(i = 1, 2, ...)$ azonos eloszlású és kölcsönösen független valószínűségi változókból álló sorozat (iid), akkor minden $\epsilon > 0$ számra

$$\lim_{n\to\infty} P(\mid \overline{X_n} - \mu \mid > \epsilon) = 0,$$

ahol $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$és $E(X_i) = \mu$.
Ezt úgy is jelöljük, hogy

$$\begin{aligned} E(\overline{X_n}) &\rightarrow {}^{P}\mu, \\ p\lim \overline{X_n} &= \mu, \\ n &\rightarrow \infty. \end{aligned}$$

Központi Határeloszlás Tétel Legyen

$$\begin{aligned} p\lim \overline{X_n} &= \mu, \\ n &\rightarrow \infty. \end{aligned}$$

Ekkor

$$\sqrt{n}(\overline{X_n} - \mu) \rightarrow^{d} N(0, \sigma^2),$$

ami azt jelenti, hogy ha F_n-nel jelöljük a $\sqrt{n}(\overline{X_n} - \mu)$ valószínűségi változó eloszlásfüggvényét, akkor az F_n függvénysorozat pontonként konvergál a 0 várható értékű és σ^2 varianciájú normális eloszlás eloszlásfüggvényéhez, ahol σ^2 az X_i változók (amelyek nem feltétlenül normális eloszlásúak) varianciája.

Aszimptotikus normalitás definíciója

Valamely $\mathbf{b}_n$ (vektor) valószínűségi változó sorozatot aszimptotikusan normálisnak nevezzük, ha

- $p\lim \mathbf{b}_n = \mathbf{b}$ (vagyis valószínűségben konvergens), és
- $\sqrt{n}(\mathbf{b}_n - \mathbf{b}) \rightarrow^{d} N(0, \mathbf{C})$) (azaz a $\sqrt{n}(\mathbf{b}_n - \mathbf{b})$ vektorváltozó eloszlásfüggvénye pontonként konvergál egy $\mathbf{0}$ várható értékű normális vektorváltozóhoz).

Slutsky Tétel

Ha a_n változó sorozat eloszlásban konvergál és b_n változó sorozat valószínűségben konvergál egy b számhoz, akkor

$a_n + b_n$ ugyanahhoz az eloszláshoz konvergál, mint $a_n + b$.

Ha a_n változó sorozat eloszlásban konvergál és b_n változó sorozat valószínűségben konvergál egy b számhoz, akkor

$a_n b_n$ ugyanahhoz az eloszláshoz konvergál, mint $a_n b$.

Folytonos Leképezés Tétel

Ha $p\lim b_n = b$,és h folytonos a b helyen, akkor $p\lim h(b_n) = h(b)$.

A Delta Módszer

Legyen $\mathbf{h} : R^k \rightarrow R^m$ differenciálható. Ha $\mathbf{b}_n$ aszimptotikusan normális ($p\lim \mathbf{b}_n = \mathbf{b}$, $\sqrt{n}(\mathbf{b}_n - \mathbf{b}) \rightarrow^{d} N(0, \mathbf{C})$), akkor $\lim \mathbf{h}(\mathbf{b}_n) = \mathbf{h}(\mathbf{b})$, és $\sqrt{n}(\mathbf{h}(\mathbf{b}_n) - \mathbf{h}(\mathbf{b})) \rightarrow^{d} N(0, \nabla\mathbf{h}(\mathbf{b})'\mathbf{C}\nabla\mathbf{h}(\mathbf{b}))$, vagyis $\mathbf{h}(\mathbf{b}_n)$ is aszimptotikusan normális $\nabla\mathbf{h}(\mathbf{b})'\mathbf{C}\nabla\mathbf{h}(\mathbf{b})$ kovariancia mátrixszal.

3 Klasszikus statisztika

3.1 Minta és statisztika

Tételezzük fel, hogy n objektumról vannak megfigyeléseink. Ezek a megfigyelések mindegyik objektumról egy véges számegyüttessel jellemezhetőek. A klasszikus statisztikai megközelítés feltételezi, hogy létezik egy olyan valószínűségi mező amelyen valószínűségi változókat definiálhatunk, és amely valószínűségi változók egy realizációi a megfigyelések.

A legegyszerűbb feltevés az, hogy létezik $X_1....X_n$ skalár valószínűségi változók, amelyek mindegyikének ugyanaz az eloszlásfüggvénye ($F_i = F$), és egymástól páronként függetlenek, vagyis a minta eloszlásfüggvénye:

$$F_n(x_1, x_2, ...x_n) = F_1(x_1)F_2(x_2)...F_n(x_n) = F(x_1)F(x_1)...F(x_n)$$

alakban írható. Az ilyen mintát a továbbiakban iid-nek (independently, identically distributed) nevezzük. Mintának szokás nevezni mind az $(X_1, X_2, ...X_n)$ valószínűségi változó vektort, mind pedig annak egy $(x_1, x_2, ...x_n)$ realizációját. A mintaeloszlás mellett szokás populáció eloszlásról is beszélni, ami az X_i változók (elméleti) eloszlása, és amiről nem tudunk mindent, amennyiben statisztikailag érdekes problémánk van. Nemcsak iid mintákkal foglalozik a matematikai statisztika, de ebben a jegyzetben kizárólag ilyenekről lesz szó. Figyeljük meg, hogy a minta eloszlásfüggvénye mindig többdimenziós, ha $n > 1$.

A minta valamely mérhető függvényét statisztikának nevezzük. Tehát a statisztika is valószínűségi változó. Itt is figyeljünk arra, hogy mikor van szó a statisztikáról, mint valószínűségi változóról, és mikor annak egy konkrét realizációjáról.

3.1.1 A (klasszikus) matematikai statisztika problémaköre

Feltételezéseket teszünk a minta eloszlásáról, anélkül, hogy egyértelműen meghatároznánk az eloszlást, azaz az eloszlásfüggvények egy halmazában gondolkodunk. Definiálunk statisztikákat, amelyek alapján következtetéseket vonunk le arról, hogy melyik az "igazi" eloszlás.

Példa: megfigyeljük 2000 ma élő magyar állampolgár testmagasságát. Feltevésünk az, hogy minden egyes megfigyelt egyén testmagassága egy ugyanolyan F^{mm} eloszlás egymástól független realizációja. Tehát

$$F^{mm}_{2000}(X_1, X_2, ...X_n) = F^{mm}(X_1)F^{mm}(X_2)...F^{mm}(X_n).$$

A populáció eloszlásról például feltesszük, hogy létezik $E(X_i) = \mu, var(X_i) = \sigma^2$.

A statisztikai problémát informálisan úgy fogalmazhatjuk meg, hogy megpróbáljunk olyan statisztikákat (a megfigyelések függvényeit) definiálni, amelyekből következtethetünk az ismeretlen μ és σ^2 paraméterekre!

Mint látni fogjuk a megfigyelések számtani átlagából "jól" lehet következtetni a várható értékre. A kérdés, hogy mit is jelent pontosan ez a "jól". Formális

definíciókat fogunk adni a becslés (becslőfüggvény) tulajdonságairól, de a végső (nem-matematikai) kérdés az, hogy mi is a célunk pontosan a statisztikai elemzéssel.

Az alapsokaság (populáció) két interpretációja Az egyik lehetséges megközelítés az, hogy véges alapsokaságban gondolkodunk: a minta a jelenleg létező majdnem 10 millió magyar populációjából "véletlen" és független (azaz visszatevéses) kiválasztással történt. A 10 millió magyar aktuális átlagmagasságát akarjuk becsülni, amit megkaphatnánk úgyis, hogy a mintát 10 millióra növeljük.

A másik lehetséges megközelítésnél végtelen alapsokaságban gondolkodunk. A "potenciális" magyarok végtelen populációjából választottuk ki a mintát "véletlenül" és függetlenül (nem kell visszatevéses mintavételről beszélnünk). A "potenciális" magyar magasság várható értékét és varianciáját akarjuk becsülni.

Ebben az esetben a két interpretáció ugyanarra a matematikai modellre vezet, de nem mindegy, hogy mire akarjuk használni az eredményeket. Az első esetben megelégszünk azzal, hogy a jelen pillanatban érvényes átlagmagasságot kívánjuk megbecsülni anélkül, hogy minden magyar állampolgárról mintát vennénk. A második interpretációnál magasabbra is tehetjük a lécet, és megkísérelhetünk egy későbbi vagy (korábbi) időpontra előre(hátra)-jelzéseket tenni a magyar lakosság magasságának eloszlásáról. Ez azonban már nem matematikai jellegű kérdés. Az ökonometria általában a végtelen populáció interpretációjából indul ki.

3.2 Pontbecslés

Pontbecslésnél az alapsokaság egy ismeretlen paraméterét kívánjuk minél pontosabban "belőni" a minta alapján.

3.2.1 A várható érték és a variancia pontbecslése iid mintában

Az egyik legegyszerűbb, de már érdekes, statisztikai probléma az, hogy hogyan lehet egy iid minta esetén a várható értéket és a varianciát "jól" megbecsülni? Egy ötlet: vegyünk egy mintarealizációt $(x_1, ...x_n)$ és keressük azt az m-et, amely a

$$\min_m \sum_{i=1}^{n} (x_i - m)^2$$

feladat megoldása. (Ez a legkisebb négyzetek általános elvének az alkalmazása.

Deriválás és az elsőrendű feltétel megoldása után

$$m = \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

adódik, vagyis a mintaelemek számtani átlaga. Tetszőleges mintára a számtani átlag adja meg az eltérés négyzetek összegét minimalizáló számot, az $m = \overline{X}_n$ statisztikát, azaz véletlen függvényt.

3.2.2 A mintaátlag, mint becslőfüggvény tulajdonságai

Definíció: Torzítatlanság Egy T statisztikát egy t paraméter torzítatlan becsésének nevezünk, ha $E(T) = t$.

A torzítatlanság megkövetelése hasznos tulajdonságnak tűnik. Azt jelenti, hogy a becslésünk nem vezet szisztematikus tévedéshez, ha sok becslést hajtanánk végre ugyanabból a populációból vett sok mintából, akkor átlagosan pontos eredményt kapnánk.

Állítás: $E(\overline{X_n}) = \mu$ (a mintaátlag torzítatlan becslése a populáció várható értékének). Bizonyítás:

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu.$$

Nagyon sok torzítatlan becslés van. Minden $\boldsymbol{\lambda}$ vektorra, amelyre $\sum \lambda_i = 1$

$$\overline{X}_{n\lambda} = \sum \lambda_i X_i$$

is torzítatlan becslést ad. Feltehetjük azt a kérdést is, hogy ezen torzítatlan becsléseknek mekkora a varianciája, azaz

$$var(\overline{X}_{n\lambda}) = E(\overline{X}_{n\lambda} - \mu)^2.$$

Definíció: Valamely paraméter torzítatlan becslései közül a minimális varianciával rendelkezőt hatásosnak nevezzük.

Állítás: A mintaátlag hatásos becslése μ-nek, és $var(\overline{X}_n) = \frac{\sigma^2}{n}$. Bizonyítás: $var(\overline{X}_n) = \frac{\sigma^2}{n}$következik abból, hogy $var(\frac{X_i}{n}) = \frac{\sigma^2}{n^2}$, és a függetlenségi feltevésből. Legyen $\left[\frac{1}{n}\right]$ egy olyan n dimenziós vektor, amely minden eleme $\frac{1}{n}$. Ekkor a várható érték minden torzítatlan becslése

$$\mathbf{m}_c\mathbf{x} = (\left[\frac{1}{n}\right] + \mathbf{c})\mathbf{x}$$

alakban írható, ahol $[1]'\,\mathbf{c}' = 0$. Írjuk fel az $E((\mathbf{m}_c'\mathbf{x} - \boldsymbol{\mu})^2)$ várható négyzetes hibát és vonjuk ki belőle az $E((\left[\frac{1}{n}\right]'\mathbf{x} - \boldsymbol{\mu})^2)$ értéket (a számtani átlag várható négyzetes hibáját.) Azt találjuk, hogy a különbség

$$E(\mathbf{c}'(\mathbf{x}\mathbf{x}')\mathbf{c})$$

vagyis a $\mathbf{c}'\mathbf{x}$ valószínűségi változó négyzetének a várható értéke. Ez viszont $(\sigma^2 + \mu^2)\sum c_i^2 > 0$.

Definíció: Egy T statisztikát a t paraméter konzisztens becslésének nevezzük, ha $\lim_{n\to\infty} T \to^{\Pr} t$.

Állítás: A mintaátlag konzisztens becslése μ-nek. Bizonyítás: A Nagy Számok Törvényéből közvetlenül.

3.2.3 A variancia becslése

Kézenfekvő ötlet, hogy a varianciát a minta varianciájával (empirikus szórásnégyzet) becsüljük.

$$s_{en}^2 = \sum_i \frac{1}{n}(X_i - \overline{X_n})^2.$$

Állítás: $E(s_{en}^2) = \frac{n-1}{n}\sigma^2$**,vagyis az empirikus szórásnégyzet "lefelé" torzított becslése a varianciának.** Bizonyítás:

$$\begin{aligned} E(s_{en}^2) &= E(\sum_i \frac{1}{n}(X_i - \frac{\sum_j X_j}{n})^2) \\ &= E(\sum_i \frac{1}{n}((X_i - \mu) + (\mu - \frac{\sum_j X_j}{n}))^2). \end{aligned}$$

Ennek kifejtése, majd az $E(X_i) = \mu, E(X_i^2) = \sigma^2 + \mu^2$ és $E(X_i X_j) = \mu^2 (i \neq j)$ összefüggések felhasználása.

A torzítás azonban egyszerűen korrigálható, legyen

$$s_n^2 = \frac{n}{n-1} s_{en}^2 = \frac{1}{n-1} \sum (X_i - \overline{X}_n)^2.$$

a korrigált empirikus szórásnégyzet.

Állítás:

$$E(s_n^2) = \sigma^2,$$

vagyis s_n^2 torzítatlan becslése σ^2-nek.

Állítás: Mind s_n^2, mind s_{en}^2 konzisztens becslése a σ^2-nek. Bizonyítás: Az állítás ismét a Nagy Számok Törvénye következménye, ha feltesszük, hogy az alapsokaság eloszlásának létezik negyedik momentuma.

Normális eloszlású alapsokaság Ha az iid mintáról további feltevéseket teszünk, akkor a mintaátlag és a (korrigált) empirikus szórásnégyzet további tulajdonságait fedezhetjük fel. A leggyakoribb feltevés az, hogy a minta normális eloszlású. Ekkor egy n elemű véletlen minta sűrűségfüggvénye:

$$f(X|\,\boldsymbol{\mu},\boldsymbol{\sigma}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2n\sigma^2}\right].$$

3.2.4 A Maximum Likelihood elv

Példa Legyen n független kísérletünk egy karakterisztikus változóra. Hogyan becsüljük a p-t? Egy kézenfekvő ötlet az, hogy keressük meg azt a valószínűséget, amely mellett a minta a legnagyobb "valószerűséggel" (likelihood) realizálódott. Tudjuk, hogy azon minták, amelyekben k-szor fordul elő a p valószínűségű esemény ugyanolyan "valószerűek" az előfordulás sorrendjétől függetlenül ($p^k(1-p)^{n-k}$). Tehát

$$P(k \mid p,n) = \binom{n}{k} p^k (1-p)^{n-k},$$

másként kifejezve k binomiális valószínűségi változó n és p paraméterekkel. Itt n–t ismerjük, tehát maximáljuk ezt a valószínűséget p szerint adott k-ra és n-re. Vegyük $P(k \mid p,n)$ logaritmusát és deriváljuk p szerint, majd tegyük egyenlővé 0-val a deriváltat:

$$\ln(P(k \mid p,n)) = \ln\binom{n}{k} + k\ln p + (n-k)\ln(1-p),$$

$$\begin{aligned} \frac{k}{p} &= \frac{n-k}{1-p}, \\ p^{ml} &= \frac{k}{n} \end{aligned}$$

adódik. Azt mondjuk, hogy p^{ml} a p paraméter maximum likelihood becslése.

Miért jó az ML becslés? Az ML becslés nagyon hasznos tulajdonsága az invariancia. Ha t egy ML becslése θ-nak, akkor $r(t)$ is ML becslése $r(\theta)$nak, ahol r egy tetszőleges függvény. Továbbá belátható, hogy amennyiben egy ML becslés torzítatlan, akkor hatásos is, azaz a legkisebb varianciájú az összes torzítatlan becslés között. Ez az úgynevezett Cramér-Rao egyenlőtlenség következménye.

Cramér-Rao egyenlőtlenség Legyen

$$I(\theta) = -E(\frac{\partial^2 \log L(\theta)}{\partial \theta^2})$$

a Fisher-információ. Ekkor igaz, hogy bármely torzítatlan becslés varianciája legalább $\frac{1}{I(\theta)}$.

Bizonyos regularitási feltevések teljesülése esetén az alábbi állítások is igazak.

1. Az ML becslések konzisztensek.
2. Az ML becslések aszimptotikusan normálisak.
3. Az ML becslések aszimptotikus varianciája a legkisebb az összes konzisztens becslés között.

Az ML becslés normalitási feltevés mellett Tekintsük most μ-t és σ^2-et változónak és rögzítsük $\mathbf{x}$-et. Válasszuk a várható érték és a variancia becslésének azokat a számokat, amelyek adott $\mathbf{x}$ megfigyelés mellett maximalizálják ezt az úgynevezett likelihood függvényt.

$$\max_{m,s} L(m, \mathbf{s} \mid \mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - m)^2}{2ns^2} \right].$$

Állítás: A várható érték maximum likelihood becslése újra a mintaátlag, és a variancia becslése az empirikus szórásnégyzet.

$$m^{ML} = \overline{X}_n$$

$$s^{2ML} = s^2_{en} = \frac{1}{n}\sum (X_i - \overline{X_n})^2.$$

Bizonyítás:

Maximáljuk a likelihood függvény logaritmusát:

$$\log L(m, \mathbf{s}^2 | \mathbf{x}) = -\frac{n}{2}(\log(2\pi) + \log s^2)) - \left[-\frac{\sum_{i=1}^{n}(X_i - m)^2}{2ns^2} \right],$$

Csak az utolsó tag függ m-től, és annak maximalizálása ekvivalens az eltérés négyzetösszeg minimalizálásával. A log-likelihood függvényt deriválva s^2 szerint, és a deriváltat egyenlővé téve 0-val megkapjuk a variancia ML becslését is.

Állítás: Ha a minta normális iid, akkor a mintaátlag normális, $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. Ezért

$$z = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim N(0,1).$$

Bizonyítás: Normális eloszlású valószínűségi változók lineáris kombinációja is normális.

3.3 Intervallum becslés

Intervallumbecslésről beszélünk, amikor a keresett paraméterre egy intervallumot, és nem csupán egy pontot adunk meg. Amikor a populáció várható értékére számolunk konfidencia intervallumot, akkor egy olyan (véletlen végpontú, nemelfajult) intervallumot próbálunk megadni, amelybe nagy valószínűséggel beleesik a várható érték. A várható értékre normális eloszlás feltevése esetén meg tudunk adni ilyen intervallumokat.

3.3.1 Konfidencia intervallum: Normális alapsokaság várható értéke, ismert σ esete

Mint láttuk

$$z = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

standard normális. Legyen $1-\alpha$ az úgynevezett konfidencia szint. Jelöljük F-fel a standard normális eloszlásfüggvényt. Definiáljuk $z_{\alpha/2}$-t implicite az alábbi egyenletből:

$$F(z_{\alpha/2}) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z^\alpha} \exp(-\frac{x^2}{2})dx = 1 - \frac{\alpha}{2}.$$

Ekkor

$$P\left[-z_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq z_{\alpha/2}\right] = 1-\alpha.$$

amiből

$$P\left[\overline{X}_n - \frac{1}{\sqrt{n}}\sigma z_{\alpha/2} \leq \mu \leq \overline{X}_n + \frac{1}{\sqrt{n}}\sigma z_{\alpha/2}\right] = 1-\alpha.$$

Állítás: Ha

$$\begin{aligned} T_2 &= \overline{X}_n + \frac{1}{\sqrt{n}}\sigma z_{\alpha/2}, \\ T_1 &= \overline{X}_n - \frac{1}{\sqrt{n}}\sigma z_{\alpha/2}, \end{aligned}$$

akkor $1-\alpha$ valószínűséggel a várható érték benne van a (T_1, T_2) véletlen végpontú konfidencia intervallumban.

Figyeljünk a pontos jelentésre: $1-\alpha$ a valószínűsége annak, hogy a várható érték beleesik ebbe a véletlen végpontú intervallumba. Vigyázat, a várható érték nem véletlen változó, az intervallum végpontjai viszont azok, ezek különböző mintákban különböző értékeket vesznek fel.

3.3.2 Konfidencia intervallum: normális alapsokaság várható értéke, ismeretlen σ esete

Mi történik, ha nem ismerjük σ^2-et? Kézenfekvő, hogy annak torzítatlan becslésével helyettesítsük, de kérdés, hogy a

$$\sqrt{n}\frac{\overline{X}_n - \mu}{s_n}$$

eloszlást ismerjük-e?

Állítás: Ha a minta iid és normális, akkor a Cochrane Tétel alapján

$$v = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

továbbá $z = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$ és v függetlenek.

Állítás: Ha a minta iid és normális, akkor

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \sim t_{n-1}.$$

Bizonyítás:

$$\begin{aligned} z &= \frac{\overline{X}_n - \mu}{\sigma^2} \\ t_{n-1} &= \sqrt{n-1}\frac{z}{\sqrt{v}} = \sqrt{n}\frac{\overline{X}_n - \mu}{s_n}. \end{aligned}$$

Definiáljuk $t_{n-1,\alpha/2}$−et implicite a következő összefüggésből:

$$F_{t_{n-1}}(t_{n-1,\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Ekkor a t_{n-1} eloszlás szimmetriája miatt

$$F_{t_{n-1}}(-t_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

és

$$P(-t_{n-1,\alpha/2} \leq \sqrt{n}\frac{-\mu}{s_n} \leq t_{n-1,\alpha/2}) = 1-\alpha.$$
$$P(\overline{X_n} - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}) = 1-\alpha$$

Vagyis s_n-nel helyettesítve σ-t t_{n-1} eloszlást kapunk, és ezzel számolhatunk a mintaátlagra konfidencia intervallumot, azaz intervallum becslést $1-\alpha$ konfidencia szinten.

Állítás: A $\left[\overline{X_n} - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}, \overline{X_n} + \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}\right]$**intervallum** $1-\alpha$ **valószínűséggel tartalmazza** μ**-t.** A különbség a z és t eloszlásokkal számolt konfidencia intervallumok között az, hogy az utóbbiak hossza is véletlen változó, míg az előzőeknél a hossz állandó.

3.4 Hipotézis vizsgálat

A statisztikai modellben a feltevések "soha nem vitatott" részét fenntartott hipotézisnek nevezzük. Például feltehetjük, hogy a minta iid és normális, és ezt soha nem vonjuk kétségbe az elkövetkező vizsgálat során.

3.4.1 Hipotézisek tesztelése (statisztikai próbák)

A fenntartott hipotézisen belül elkülöníthetünk egy null hipotézist, és annak a komplementerét alternatív hipotézisnek nevezzük. Például a fenntartott hipotézisnek megfeleltethetjük a $(\mu \in R, \sigma^2 = \sigma_0^2)$ egyenest. Ekkor a nullhipotézis lehet $\mu_0 = 1$. Geometrialilag a nullhipotézis egy pont az egyenesen, az alternatív hipotézis pedig ennek a pontnak a komplementere, A nullhipotézis név onnan származik, hogy nagyon gyakran $\mu_0 = 0$ a nullhipotézis, de ez egyáltalán nem szükségszerű.

A klasszikus tesztelmélet alapgondolata az, hogy találunk egy S statisztikát, amelynek a nullhipotézis fennállása esetén ismerjük az eloszlását. Megadunk egy Ω halmazt és egy α valószínűséget, amelyekre igaz az, hogy

$$P(s \in \Omega) = 1-\alpha.$$

Kiszámítjuk a mintából a S statisztika realizált s értékét, és a következő döntési szabályt követjük:

Ha $s \in \Omega$, akkor elfogadjuk a nullhipotézist, ha nem, akkor elutasítjuk.

A döntési szabály alkalmazásával kétféleképpen tévedhetünk. Elsőfajú hibát követünk el, ha elvetjük a nullhipotézist, pedig igaz. Az elsőfajú hiba valószínűsége α, ezt szignifikancia szintnek is nevezzük.

Másodfajú hibát követünk el, ha elfogadjuk a nullhipotézist, pedig az nem igaz. A másodfajú hiba valószínűsége általában egy függvény, annak a függvénye, hogy az alternatív hipotézis melyik "eleme" az "igazi" paraméter. A másodfajú hiba meghatározásához ismernünk kell a S statisztika eloszlását olyankor,

amikor az alternatív hipotézis egyes elemei érvényesek. A másodrendű hiba valószínűségét kivonva 1-ből megkapjuk a teszt (próba) erejét. Vagyis az erős próba az, amikor nem kell félnünk attól, hogy a nullhipotézist tévedésből fogadjuk el.

3.4.2 Várható érték hipotézis tesztelése normális esetben, amikor a szórás ismert

Ha a nullhipotézis $\mu = \mu_0$, és $\sigma = \sigma_0$, és α szignifikancia szinten (azaz ekkora elsőfajú hibával) tesztelünk, akkor az

$$P(abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}) > z_{\alpha/2}) = \alpha$$

egyenlet meghatározza $z_{\alpha/2}$-t, ahogy a konfidencia intervallum meghatározásánál láttuk. Ekkor a $z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$ változóra:

$$P\left[-z_{\alpha/2} \leq z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma} \leq z_{\alpha/2}\right] = 1 - \alpha,$$

Tehát, ha

$$abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}) \leq z_{\alpha/2},$$

akkor a $\mu = \mu_0$ $\sigma = \sigma_0$ hipotézist elfogadjuk, egyébként elvetjük. A hibás elutasítás valószínűsége α, ezért igyekszünk α-t kicsire választani. ($\alpha = 0,1$ vagy $0,05$ vagy $0,001$ gyakori választások.)

Az elfogadási kritérium ekvivalens módon megfogalmazható úgy is, hogy a nullhipotézist akkor és csakis akkor fogadjuk el, ha

$$\overline{X}_n \in (\mu_0 - \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}).$$

Látszik, hogy adott nullhipotézis, mintaelemszám és α egyértelműen meghatározza az elfogadási tartományt.

3.4.3 Várható érték hipotézis tesztelése normális esetben, amikor a szórás nem ismert

Ha a nullhipotézis az, hogy $\mu = \mu_0$, és $\sigma > 0$, akkor a nullhipotézisünk már egy félegyenes, nemcsak egy pont. A konfidencia intervallumok vizsgálatánál láttuk, hogy $\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n} = t_{n-1}$, vagyis ez egy potenciális tesztstatisztika, mivel ismerjük az eloszlását. Adott α-ra (azaz ekkora elsőfajú hibára) az

$$P(abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}) = t_{n-1,\alpha/2}) = \alpha.$$

egyenlet meghatározza $t_{n-1,\alpha/2}$-t.

Másképpen:

$$P\left[-t_{n-1,\alpha/2} \leq t_{n-1} = \sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n} \leq t_{n-1,\alpha/2}\right] = 1 - \alpha.$$

Az elfogadási kritérium ilyenkor a következő. Ha

$$abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}) \leq t_{n-1,\alpha/2},$$

akkor a $\mu = \mu_0$ hipotézist elfogadjuk, egyébként elvetjük.

Az ekvivalens elfogadási kritérium:

$$\overline{X}_n \in (\mu_0 - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}).$$

Látszik, hogy így megfogalmazva az elfogadási tartomány nem rögzített, a végpontjai változnak mintáról mintára.

Egyoldali próbák Eddig úgynevezett kétoldali tesztekkel foglalkoztunk. Egyoldali teszttel van dolgunk például, ha a nullhipotézis az, hogy $\mu = \mu_0$ és az alternatív hipotézis az, hogy $\mu > \mu_0$. Ilyenkor keressük azt a $t_{n-1,\alpha}$ értéket, amire

$$P(\sqrt{n}\frac{\overline{X}_n - \mu_0}{s_n}) > t_{n-1,\alpha}) = \alpha.$$

Ha

$$(\sqrt{n}\frac{\overline{X_n} - \mu_0}{s_n}) \leq t_{n-1,\alpha}$$

akkor a nullhipotézist elfogadjuk, egyébként elvetjük. Ugyanez ekvivalens módon: ha

$$\overline{X}_n \in (-\infty, \mu_0 + \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}),$$

akkor a hipozézist elfogadjuk, egyébként elvetjük.

Amennyiben a nullhipotézis az, hogy $\mu = \mu_0$ és az alternatív hipotézis $\mu < \mu_0$, akkor az egyoldalú elfogadási kritériumok:

$$\begin{aligned}(\sqrt{n}\frac{\overline{X_n} - \mu_0}{s_n}) &\geq -t_{n-1,\alpha} \\ \overline{X}_n &\in (\mu_0 - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}, +\infty).\end{aligned}$$

A féloldali próbák mögött az szokott meghúzódni, hogy a nullhipotézistől való különböző irányú eltéréseket esetleg különböző fontosságúnak tartjuk. Amikor "nem bánjuk", hogy esetleg $\mu > \mu_0$ is fennállhat, akkor ésszerű alternatívaként csak a $\mu < \mu_0$ "féloldali alternatívát" megfontolni. Nyilvánvaló, hogy a féloldali vagy kétoldali tesztelés melletti döntés mögött a konkrét problémához tartozó megfontolások kell, hogy álljanak.

Másodfajú hiba

Ismert szórás esete Tegyük fel, hogy az igazi várható érték

$$\mu = \mu_0 + \delta \ (\delta \neq 0).$$

Ha ismert szórással van dolgunk (z-próba), akkor a másodfajú hiba valószínűségét könnyen kiszámíthatjuk, mivel az alternatív hipotézis mellett az $\overline{X}_n$ eloszlása $N(\mu_0 + \delta, \frac{\sigma_0}{\sqrt{n}})$. A másodfajú hiba "bekövetkezik", ha $\overline{X}_n \in (\mu_0 - \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}})$, vagyis ha az átlag beleesik az elfogadási tartományba. Ennek valószínűsége, ha az alternatív ($\mu = \mu_0 + \delta$) hipotézis igaz:

$$F^{N(\mu_0+\delta, \frac{\sigma}{\sqrt{n}})}(\mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}) - F^{N(\mu_0+\delta, \frac{\sigma}{\sqrt{n}})}(\mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}).$$

Nyilvánvaló, hogy n növekedésével az elfogadási tartomány "összeszűkül" μ_0-ra. Továbbá a $N(\mu_0 + \delta, \frac{\sigma}{\sqrt{n}})$ eloszlás is egyre inkább koncentrálódik $\mu_0 + \delta$ körül. Vagyis a mintaelemszám növekedésével adott elsőfajú hiba valószínűség mellett a a másodfajú hiba valószínűsége 0-hoz, és a próba ereje 1-hez, tart. Ilyenkor azt mondjuk, hogy a próba konzisztens.

Ismeretlen szórás A δ eltéréshez tartozó másodfajú hiba mértékét akkor tudjuk megmondani, ha ismerjük az $\sqrt{n}\frac{\overline{X_n}-\mu_0}{s_n}$ valószínűségi változó eloszlását olyankor, amikor $\mu = \mu_0 + \delta$. Belátható, hogy ilyenkor $\sqrt{n}\frac{\overline{X_n}-\mu_0}{s_n}$ úgynevezett nem-centrális eloszlású t változó, $\frac{\delta\sqrt{n}}{\sigma}$ nem-centralitási paraméterrel. (Ez egy nem szimmetrikus eloszlás.) Ilyenkor a másodfajú hiba meghatározásához meg kell adnunk egy konkrét σ szórást is.) (Pontosabban egy $\frac{\delta}{\sigma}$ arányt.) A másodfajú hiba "bekövetkezik", ha $\overline{X_n} \in (\mu_0 - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}}, \mu_0 - \frac{s_n t_{n-1,\alpha/2}}{\sqrt{n}})$,tehát a

$$F^{t(n-1, \frac{\delta\sqrt{n}}{\sigma})}(t_{n-1,\alpha/2}) - F^{t(n-1, \frac{\delta\sqrt{n}}{\sigma})}(-t_{n-1,\alpha/2})$$

mennyiséget kell meghatározunk.

Általában δ/σ-t úgy szokták meghatározni, mint a még "elfogadható tévedést" a szórás arányában. Könnyen látható, hogy amennyiben az elfogadható tévedés nagyobb, akkor a másodfajú hiba valószínűsége csökken.

3.4.4 A p-érték megközelítés

A modern ökonometriában gyakran nem adnak meg szignifikancia szintet, és a kutatók az úgynevezett p-értéket közlik. A p-érték az a szignifikancia szint, amely mellett a nullhipotézist még éppen elfogadnánk. Például kétoldali t-próba esetén a p-érték:

$$p_v = 1 - (F_{t_{n-1}}(\sqrt{n}\frac{\overline{X_n} - \mu_0}{s_n}) - F_{t_{n-1}}(-\sqrt{n}\frac{\overline{X_n} - \mu_0}{s_n})).$$

Ha $\alpha = p_v$ lenne a szignifikancia szint, akkor még éppen elfogadnánk a nullhipotézist. Például ha egy adott próbánál

$$p_v = 0.28$$

jön ki eredménykémt, akkor minden "szokásos" szignifikancia szinten (0,1, 0,05, 0,01) elfogadnánk a nullhipotézist, de ha

$$p_v = 0.0028$$

akkor ugyanezeken a szinteken elutasítanánk. A p-érték közlésével mintegy a hipotézis elfogadás (elvetés) "erősségét" mérhetjük. Nagy (1-hez közeli) p-értéket általában úgy interpretálunk, hogy a próba határozottan elfogadja a nullhipotézist.

3.5 Előrejelzés

Szeretnénk "előrejelezni" egy x_{n+1} jövőbeli, vagy általában mintán kívüli, megfigyelés értékét, amelyről feltesszük, hogy független a meglevő mintánktól, de ugyanabból az eloszlásból származik. Tekintsük előrejelzésnek a mintaátlagot $\overline{X_n}$-t. Mit tudunk mondani ennek az előrejelzésnek a hibájáról? A várható négyzetes hiba:

$$\begin{aligned} E(x_{n+1} - \overline{X_n})^2 &= E((x_{n+1} - \mu) + (\mu - \overline{X_n}))^2 = \\ &= E(x_{n+1} - \mu)^2 + 2E((x_{n+1} - \mu)(\mu - \overline{X_n})) + E(\mu - \overline{X_n})^2 \\ &\quad E(x_{n+1} - \mu)^2 + E(\mu - \overline{X_n})^2 \\ &= \sigma^2 + \frac{\sigma^2}{n}. \end{aligned}$$

Itt kihasználtuk azt, hogy $E((x_{n+1} - \mu)(\mu - \overline{X_n})) = 0$, mivel x_{n+1} független $\overline{X_n}$-től.

Az előrejelzés négyzetes hibája tehát két részből áll: 1. a (redukálhatatlan) bizonytalanságból, amit σ^2-tel mérhetünk, és 2. a becslési hibából adódó varianciából, ami $\frac{\sigma^2}{n}$. Ez utóbbi csökkenthető a mintaelemszám növelésével. Egy harmadik hibaforrás lehetne, ha $\overline{X_n}$ torzított becslése lenne μ-nek. Azonban előfordulhat, hogy a torzítás nem növeli, hanem csökkenti a várható négyzetes hibát! Nem feltétlenül torzítatlan becslés vezet a legjobb előrejelzéshez.

3.6 Gyakorlatok R-ben

3.6.1 Véletlen számok és eloszlások

Generáljunk véletlen mintákat normális eloszlásokból, illetve határozzuk meg az eloszlás és sűrűségfüggvény értékeit!

(x=rnorm(30, 3,10)) # 30 elemű véletlen vektor, az N(3,10) eloszlásból

(x=rnorm(15)) # 15 elemű vektor az N(0,1) sztenderd normális eloszlásból

dnorm(0,3) # a standard normális sűrűségfüggvény értéke 23-nál
pnorm (15, 2,5) # az N(2,5) eloszlásfüggvény értéke 15-nél
qnorm(0.95) # a sztenderd normális eloszlásfüggvény inverzének értéke 0,95-nél

3.6.2 A Z (standard normális) eloszlás

Rajzoljuk ki a standard normális eloszlás sűrűségfüggvényét, és az ábrán jelöljünk be négy gyakran használt abszcissza értéket!

```
pln1 = function(x) {
dnorm(x)}
plot (pln1, main= "Z eloszlás", xlab= "Z 0.05 és 0.1 szintek",
xlim=c(-3,3))
par(new=T)
abline(v=qnorm(0.975),col=3,)
abline(v=qnorm(0.025),col=3)
abline(v=qnorm(0.95),col=2)
abline(v=qnorm(0.05),col=2)
```

A set.seed () utasítás megengedi, hogy rekonstruáljuk az eredményeket! Például:

```
set.seed(237)
rnorm (3)
set.seed(237)
rnorm(3)
# (237 helyett bármely másik pozitív egész szám megtette volna.)
```

3.6.3 A t eloszlás összevetve a standard normálissal

Próbálják ki a következőket!

```
qnorm(0.025)
qt(0.025,df=10)
qt(0.025,df=25)
qt(0.025,df=100)
# A standard normális eloszlás ábrája
n=1
pmean=0
pstd=1
alpha=0.025
pln1 = function(xmean) {
dnorm(xmean,pmean,pstd/sqrt(n))}
plot (pln1,pmean-3*pstd,pmean+3*pstd)
abline(v=qnorm(alpha))
abline(v=-qnorm(alpha))
# a t(99) eloszlás ábrája
n=100
```

```
pmean=0
pstd=1
alpha=0.025
df=n-1
pln1 = function(xmean) {
dt(xmean,df=df)}
plot (pln1,pmean-3*pstd,pmean+3*pstd)
abline(v=qt(alpha,df=df))
abline(v=-qt(alpha,df=df))

#Z-konfidencia intervallum meghatározása
alpha=0.05 # a konfidencia szint=1-alpha
zalpha=qnorm(1-alpha/2) # a z küszöbérték
pmean=15 # a generált minta átlaga
pstd=5 # a generált minta szórása
n=25 # minta elemszám
x1=rnorm(n,pmean,pstd) # minta generálás
mean(x1) # átlag
sd(x1) # variancia
(conl1=mean(x1)-(zalpha*pstd)/sqrt(n)) # a konfidencia intervallum alja
(conu1=mean(x1)+(zalpha*pstd)/sqrt(n)) # a konfidencia intervallum teteje
# Próbálják ki többször ugyanezt, illetve más n, alpha, pstd paraméterekkel!
```

3.6.4 t-konfidencia intervallum meghatározása

```
alpha=0.05
n=25
talpha=qt(1-alpha/2,df=n-1) # küszöbérték az n-1 szabadságfokú t eloszlásból
pmean=15
pstd=5
x1=rnorm(n,pmean,pstd)
mean(x1)
sd(x1)
(conl1=mean(x1)-(talpha*sd(x1))/sqrt(n)) # az intervallum a alső hatása
# itt most becsült szórással, nem pedig elméletivel számolunk
(conu1=mean(x1)+(talpha*sd(x1))/sqrt(n)) # az intervallum felső határa
# Próbálják ki többször ugyanezt más n, alpha, pstd paraméterekkel!
```

Kétoldali z-próba n=25 # mintaelemszám

```
alpha=0.05 # az elsőfajú hiba valószínűsége
(zalpha=qnorm(1-alpha/2))
pmean=15
pstd=10
x1=rnorm(n,pmean,pstd)
(mean(x1))
```

mu0=pmean # feltesszük, hogy a nullhipotézis az igazi várjató érték
(z= sqrt(n)* (mean(x1)-mu0)/ pstd)
abs(z-zalpha)
# A konkrét esetben elfogadjuk a nullhipotézist? Miért?
# Próbálják ki többször ugyanezt, illetve más n, alpha, pstd paraméterekkel!
# Egy másik módszer:
(zu=mu0+(zalpha*pstd)/sqrt(n)) # az elfogadási tartomány teteje
(zl=mu0-(lalpha*pstd)/sqrt(n)) # az elfogadási tartomány alja
Elfogadjuk a nullhipotézist? Miért?

Kétoldali t próba n=25
alpha=0.05
(talpha=qt(1-alpha/2, df=n-1))
pmean=15
pstd=10
x1=rnorm(n,pmean,pstd)
(mean(x1))
(sd(x1))
mu0=pmean # ismét az igazi várható érték a nullhipotézis
(t= sqrt(n)* (mean(x1)-mu0)/sd(x1))
abs(t-talpha)
# Elfogadja a teszt a nullhipotézist?
# Próbálják ki többször ugyanezt, illetve más n, alpha, pstd paraméterekkel!
# Egy másik módszer:
(tl=mu0-(talpha*sd(x1))/sqrt(n))
(tu=mu0+(talpha*sd(x1))/sqrt(n))
Mikor fogadjuk el a nullhipotézist?

Egyoldali t próba n=25
alpha=0.05
pmean=15
pstd=10
x1=rnorm(n,pmean,pstd)
mean(x1)
sd(x1)
(level=qt(1-alpha,df=n-1)) #figyelem alpha és nem alpha/2
(mu0=pmean)
(tu=mu0+(level*sd(x1))/sqrt(n))
Mi az alternatív hipotézis, ha ezt számoljuk ki? Elfogadjuk a nullhipotézist?

Másodfajú hiba: kétoldali z-teszt alpha=0.05
zalpha=qnorm(1-alpha/2)
delta=3 # ekkora az eltérés a nullhipotézistől
n=25

```
pstd=5
mu0=150
zu=mu0+zalpha*pstd/sqrt(n)
zl= mu0-zalpha*pstd/sqrt(n)
er2= pnorm(zu,mu0+delta,pstd)-pnorm(zl,mu0+delta,pstd) # a másodfajú
hiba valószínűsége
(power=1-er2) # a próba ereje
```

4 Bayes-i statisztika*

Ez a fejezet érdeklődő hallgatóknak szól. A szokásos bevezető ökonometria könyvek nem foglalkoznak a bayes-i megközelítéssel, de az ökonometriai gyakorlat egyre többször használ bayes-i módszereket. Ezért indokolt legalább a bayesianizmus alapjainak az ismerete, és annak megértése, hogyan és milyen indokokból különbözik a klasszikus megközelítéstől.

4.1 A klasszikus statisztika paradoxonjai

4.1.1 Becslési paradoxon

A Poisson eleoszlás írja le azt a valószínűséget, hogy esemény n alkalommal következik be egy adott időszakon belül, ha az egyes újabb bekövetkezések valószínűsége független a folyamat műltjától. Diszkrét eloszlásról lévén szó a valószínűségeloszlást végtelen 0 és 1 közti szám jellemzi, amelyek összege 1. Annak a valószínűsége, higy az esemény pontosan n-szer következik be:

$$p_n = \exp(-\lambda)\frac{\lambda^n}{n!}.$$

Itt az ismeretlen paraméter, amit egy statisztikus becsülni szeretne λ, ami az eloszlás elméleti várható értéke. Tegyük fel, hogy megfigyeljük a valóságban az eseményeket és azt találjuk, hogy pontosan n következett be. A klasszikus statisztika szellemében megpróbálhatunk találni egy becslőfüggvényt, amely minden n megigyeléshez hozzárendel egy számot, amit λ becslésének tekinthetünk. A becslőfüggvényt jelöljük $T(n)$-nel. Torzítatlan becslőfüggvényt keresünk, aminek várható értéke pontosan λ. Képletben kifejezve: re

$$\sum_{n=0}^{\infty} p_n T(n) = \sum_{n=0}^{\infty} \exp(-\lambda)\frac{\lambda^n}{n!}T(n) = \lambda.$$

Ebből

$$\sum_{n=0}^{\infty} \frac{\lambda^n}{n!}T(n) = \lambda \exp(\lambda).$$

A baloldal nem más, mint a Taylor-sora a jobboldalnak $\lambda = 0$ körül. Ezért

$$\begin{aligned} T(n) &= \frac{\partial^n(\exp(\lambda)\lambda)}{\partial n^n} \\ T(n) &= n. \end{aligned}$$

Vezessünk le most egy torzítatlan becslőfüggvényt λ^2-re (jelöljük $T_2(n)$-nel), ami egyébként a Poisson eloszlás varianciája. Használjuk újra a fenti módszert:

$$\sum_{n=0}^{\infty} \exp(-\lambda)\frac{\lambda^n}{n!}T_2(n) = \lambda^2$$

$$\begin{aligned} T_2(n) &= \frac{\partial^n(\exp(\lambda)\lambda^2)}{\partial n^n} \\ T_2(0) &= 0 \\ T_2(1) &= 0 \\ T_2(n) &= n(n-1). \end{aligned}$$

A két torzítatlan becslés nyivánvalóan ellentmond, és ez az ellentmondás $n = 1$-re elég "radikális".

4.1.2 Teszt paradoxon

Végrehajtunk egy kísérletet, és definiálunk egy tesztet 0.05 -ös elsőfajú hibával. Az elfogadási tartomány legyen mondjuk (-2,2). A tesztstatisztika 1.9, ezért a nullhipotézist elfogadjuk. Később azonban a statisztikus felfedezi, hogy a mérési eszköz korlátai miatt az eredmény nem lehetett volna nagyobb, mint 2, viszont a mérés során soha nem ütköztek bele a korlátba. Mégis kötelességtudóanm újraszámolja az elfogadási tartományt, és ennek alapján 1.9 már nincs benne, azaz a nullhipotézist elfogadják. (Kézenfekvő, hogy az új elfogadási tartomélny szűkebb mint a régi volt.) Ésszerű a hipotézis elfogadását olyan eseményektől függővé tenni, amik nem következtek be, habár bekövetkezhettek volna?

4.1.3 Konfidencia intervallum paradoxon

Legyen x egyenletes eloszlású a $(\theta - 1, \theta + 1)$ intervallumon. Most a becsülendő paraméter θ.

Van egy független mintánk 4 megfigyeléssel. Legyen $x_{\min} = \min(x_i)$,és $x_{\max} = \max(x_i)$. Ekkor

$$\begin{aligned} P(x_{\min} &> \theta) = 1/16 \\ P(\theta &> x_{\max}) = 1/16 \end{aligned}$$

$$P(x_{\min} < \theta < x_{\max}) = 7/8.$$

Tehát $(x_{\min}, x_{\max})$ egy konfidencia intervallum 7/8 konfidencia szinten.
Tegyük fel, hogy a mintában

$$\begin{aligned} x_{\min} &= 1.5 \\ x_{\max} &= 2.7. \end{aligned}$$

Ekkor viszont biztosan tudjuk, hogy θ $x_{\min}$ és $x_{\max}$ közé esik.

4.2 A likelihood elv

A klasszikus statisztikusok nagy jelentőséget tulajdonítottak az elégségs statisztika fogalmának. Intuitíve egy statisztika elégségs egy paraméter becsléséhez, ha mindent, amit megtudhatunk a paraméterről a mintából tartalmaz. Mit jelent ez pontosan?

Legyen a minta eloszlásfüggvénye

$$F(X \mid \theta),$$

ahol θ a kérdéses paraméter. Azt mondjuk, hogy $T(X)$ elégséges statisztikája θ-nak, ha a feltételes eloszlás

$$G(X \mid T(X))$$

nem függ θ-tól.

Könnyű belátni, hogy ha n független $0-1$ kísérletet hajtunk végre, akkor az 1-esek bekövetkezésének száma elégséges statisztikája az ismeretlen p valószínűségnek, ami a kísérlet kimenetelei által definiált binomiális eloszlás paramétere.

Formálisan egy statisztikai kísérletet úgy definiálhatunk, mint egy paraméter tér, Θ, egy mintatér, X, és egy likelihood függvény $f(x \mid \theta), x \in X$, $\theta \in \Theta$ hármasa.

$$E = \{\Theta, X, f(x \mid \theta), x \in X, \theta \in \Theta\}.$$

Ekkor evidenciának nevezünk bármely függvényt, amely értelmezési tartománya a kísérlet és a minta Descartes-szorzata.

Az elégségességgel kapcsolatos intuíciónk megfogalmazható az Elégségességi Elvben.

Elégségessgi Elv: Két ugyanabból a kísérletből származó azonos elégséges statisztika ugyanazt az evidenciát kell nyújtsa.

Egy másik racionálisnak tekinthető elv az, hogy, amennyiben két kísérletet, ahol a paraméter terek azonosak, randomizálunk $1/2$ valószínűséggel, akkor a kombinált kísérlet evidenciája ugyanaz kell legyen, mint a realizálódó kísérlet evidenciája (kondicionalitási elv).

Egy nevezetes tétel (Birnbaum Tétel) azt állítja, hogy ez a két elv együttesen ekvivalens az úgynevezett Likelihood Elvvel.

Likelihood Elv: Ha két kísérletnek ugyanaz a likelihood függvénye (pozitív konstans szorzótól eltekintve), akkor ugyanazt az evidenciát kell szolgáltatniuk.

Az előző paradoxonok megsértik a likelihood elvet. A becslési paradoxonnál nyilvánvaló, hogy ugyanaz a likelihood függvény és minta két egymással összeegyeztethetetlen evidenciát szolgáltat. A konfidencia intervallumos példában a likelihood függvény alapján a feltételezett minta valószínűsége 0.

Azt, hogy a klasszikus tesztelmélet megsérti a likelihoood elvet a következő példa illusztrálja legegyszerűbben. Tegyük fel, hogy független $0-1$ eseményeket figyelünk meg n-szer (1. kísérlet). Ekkor a minta eloszlását a szokásos binomiális eloszlás jellemzi:

$$P(k \mid p, n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

A 2. kísérletben azonban álljunk meg azután, hogy k alkalommal következett be az 1-es esemény. Itt az eloszlás negatív binomiális:

$$P(n \mid p, k) = \binom{n-1}{k-1} p^k (1-p)^{n-k}.$$

Ha az első kísérletben k 1-es megfigyelést teszünk, míg a másodikban n-t, akkor a két likelihood megegyzik. Ugyanakkor, mivel a valószínűségek mások, ezért az ezekből képzett tesztek vagy konfidencia intervallumok is mások lehetnek.

Tehát a klasszikus statisztikai procedúrák nem elégítik ki vagy az elégségességi, vagy a kondicionalitási elvet. Akik számára ezek a racionalitás szükséges feltételei más filozófiájú statisztikát kerestek. A bayes-i statisztika olyan, ami automatikusan kielégíti a Likelihood Elvet.

4.3 A bayes-i megközelítés

A bayes-i megköelítés egy lényeges kiinduló pontban más, mint a hagyományos. Feltételezi, hogy a paramétereknek van egy a vizsgálat kezdetén adott (*a priori* eloszlása), tehát a paraméterek maguk is véletlen változóknak tekinthetők.

Legyen ennek a véletlen eloszlásnak a sűrűségfüggvénye $p(\theta)$, és a likelihood vagy feltételes sűrűségfüggvény $f(x \mid \theta)$. Felhasználva a Bayes Tételt:

$$p(\theta \mid x) = \frac{f(x \mid \theta) p(\theta)}{\int\limits_{\Theta} f(x \mid \theta) p(\theta) d\theta'},$$

megkapjuk a paraméterek megfigyelés utáni (poszterior) eloszlását a minta bármely realizációára. Gyakran csak arra van szükségünk, hogy $f(x \mid \theta)p(\theta)$-t határozzuk meg, mivel

$$p(\theta \mid x) \propto f(x \mid \theta) p(\theta).$$

Ugyanis $\int\limits_{\Theta} f(x \mid \theta) p(\theta) d\theta'$ független a megfigyelt mintától.

A bayes-i frissítés egy szinte mechanikus procedúra:

$$p(\theta \mid x_1, x_2) \propto f(x_1, x_2 \mid \theta) p(\theta).$$

$$p(\theta \mid x_1, x_2) \propto f(x_2 \mid x_1, \theta) p(\theta \mid x_1) \propto f(x_2 \mid x_1, \theta) f(x \mid \theta) p(\theta).$$

Az újabb adatok beérkezésével az előző poszterior eloszlás prior eloszlás lesz, de a végeredmány független attól, hogy milyen sorrendben érkeztek az adatok.

A bayes-i megközelítés tehát azt mondja, hogy a megfigyelések funkciója az, hogy ezek fényében revideáljuk nézeteinket a paraméterek valószínűség eloszlásáról. Ez azonban nem olyan információ, amit a legtöbb felhasználó könnyen meg tud emészteni. Léteznek ezért a bayes-i elvek alapján pont és intervallum becslések, valamit tesztek is.

4.4 Bayes-i pontbecslés, intervallum becslés és tesztelés

4.4.1 Pontbecslés

Ehhez szükség van egy veszteségfüggvényra, ami azt méri milyen veszteség származik abból, ha $\widehat{\theta}$ a pontbecslésünk, miközben az "igazi" paraméter értéke θ. (Azaz hiszünk abban, hogy van igazi paraméterérték, például egy adott pénzdarabnál annak a valószínűsége, hogy a "fej" kerül felülre.) A veszteségfüggvény tehát

$$L(\theta, \widehat{\theta})$$

alakú, ahol általában feltesszük, hogy

$$L(\theta, \widehat{\theta}) = 0, \theta = \widehat{\theta}.$$

Ha adott egy $d(X)$ pontbecslés, akkor a várható veszteség: $R(\theta, d(X))) = E_\theta L(\theta, d(X))$, ahol a várható értéket θ poszterior eloszlása alapján számítjuk.

Az optimális *a priori* szabály:

$$d^B(X) = \arg\min_{d(X)} \int_\Theta R(\theta, d(X))p(\theta)d\theta.$$

Mivel

$$\int_\Theta (\int_X L(\theta, d(x))f(X \quad | \quad \theta)dx)p(\theta)d\theta =$$
$$\int_X (\int_\Theta L(\theta, d(x))p(\theta \quad | \quad x)d\theta)m(x)dx,$$

ahol $m(x)$ az X minta peremeloszlása, a minimalizációt úgy lehet elérni, hogy a poszterior várható veszteséget, $\int_\Theta L(\theta, d(X))p(\theta \mid X)d\theta)$, minden minta realizáció mellett minimalizáljuk. Ez a szabály tehát kielégíti a likelihood elvet.

Természetesen a bayes-i pontbecslés veszteségfüggvény függő. Néhány általános állítás azonban belátható. Például kvadratikus veszteségfüggvény esetén $d^B(x) = E(\theta \mid x)$. Ha a veszteséget a hiba abszolút értékében mérjük, akkor pedig $d^B(x)$ a mediánja a $p(\theta \mid x)$ poszterior eloszlásnak.

4.4.2 Intervallum becslés

Adjunk meg egy úgynevezett α hihetőségi szintet. Tekintsük a

$$P(\theta_A \leq \theta \leq \theta_B) = \alpha$$

összefüggéssel implicite definiált (θ_A, θ_B) intervallumokat. Ezek közül válasszuk ki a legrövidebbet.

4.4.3 Modellek tesztelése

Tegyük fel, hogy létezik két lehetséges hipotézisünk, azaz két különböző modellt fontolunk meg.

1. modell: $p_1(\theta_1), f_1((x \mid \theta_1)$
2. modell: $p_2(\theta_2), f_2((x \mid \theta_2)$.

A minta poszterior sűrűségfüggvényei:

$$f_i(x) = \int_\Theta f_i(x \mid \theta_i) p_i(\theta_i) d\theta_i', i = 1, 2.$$

A modellek összehasonlítása a Bayes faktoron

$$\frac{f_1(x)}{f_2(x)},$$

vagy a modellek poszterior valószínűségének arányán

$$\frac{p(M_1) f_1(x)}{p_2(M_2) f_2(x)}.$$

alapul. A Bayes faktor használata természetesen ekvivalens a $p_1(M_1) = p_2(M_2) = 1/2$ feltevéssel.

A modellválasztás az adatokon és *a priori* feltevéseken alapul, de nem függ olyan eseményektől, amelyek megtörténhettek volna, de nem történtek meg.

4.5 Nagy minta és bayes-i statisztika

Nagy n-re és független megfigyelésekre általában

(a) a prior eloszlás egyre kevésbé befolyásolja a poszteriort,

(b) a poszerior elfajult eloszláshoz konvergál,

(c) a poszterior aszimptotikusan normális, és várható értéke az igazi θ.

Vagyis nagy mintában nincs feltétlenül ellentmondás a klasszikus és a bayes-i megközelítésből származó eredmények között. A következő példa illusztrája ezt.

Tegyük fel, hogy van n megfigyelésünk egy $0-1$ változóról, és $P(x_i = 1) = \theta$.

Ekkor a likelihood függvény, mint tudjuk:

$$p(\mathbf{x} \mid \theta) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

Legyen a $p(\theta)$ prior eloszlás béta α és β paraméterekkel:

$$p(\theta) = \frac{\Gamma(\alpha) + \Gamma(\beta)}{\Gamma(\alpha + \beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}.$$

Ekkor a várható érték és a variancia:

$$\begin{aligned} E(\theta) &= \frac{\alpha}{\alpha + \beta} \\ var(\theta) &= \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}. \end{aligned}$$

Belátható, hogy a poszterior is béta eloszlású, és paraméterei:

$$\begin{aligned} \alpha_1 &= \alpha + \sum x_i, \\ \beta_1 &= \beta + n - \sum x_i. \end{aligned}$$

Ebből:

$$E(\theta \mid x) = \left(\frac{\alpha + \beta}{\alpha + \beta + n} \right) \frac{\alpha}{\alpha + \beta} + \left(\frac{n}{\alpha + \beta + n} \right) \frac{1}{n} \sum x_i$$

a bayes-i pontbecslés kvadratikus veszteségfüggvénnyel. A becslés tehát súlyozott átlaga a prior várható értéknek és a mintátlagnak. Láthatóan nagy n-re a mintaátlag súlya 1-hez közelít.

5 Többváltozós lineáris regresszió

5.1 Lineáris regresszió: Geometriai bevezető

Adott két pont a síkban x (koordináták: x_1, x_2) és y (koordináták y_1, y_2). (A szokással ellentétben itt a két tengely neve: 1. (vízszintes), és 2. (függőleges).) Keressük azt az $\widehat{y}$ pontot, amely az x által meghatározott egyenesen van, és az y ponttól való távolsága a legkisebb. A következő feladatot kell megoldani:

$$\min_{\beta} \left[(\beta x_1 - y_1)^2 + (\beta x_2 - y_2)^2 \right].$$

A célfüggvény deriváltját tegyük egyenlővé 0-val:

$$x_1(bx_1 - y_1) + x_2(bx_2 - y_2) = 0.$$

A megoldás:

$$b = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2},$$

ami átírható, mint

$$b = \cos(x, y) \frac{\|y\|}{\|x\|}.$$

Tehát

$$\widehat{y} = bx.$$

Az $\widehat{y}$ a merőleges vetítettje y-nak az x által meghatározott egyenesre, vagyis az $y - \widehat{y} = u$ vektor merőleges mind x-re, mind $\widehat{y}$-ra, és u a merőleges vetítettje y-nak, arra az egyenesre, ami merőleges az x által meghatározott egyenesre.

Az első vetítést a

$$\frac{1}{x_1^2 + x_2^2} \begin{bmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \end{bmatrix}$$

míg a másodikat a

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \frac{1}{x_1^2 + x_2^2} \begin{bmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

formula írja le.

5.2 Lineáris regresszió: lineáris algebra

A fenti síkgeometriai állításokat lehet általánosítani lineáris algebrai fogalmakkal az n-dimenziós valós vektortérre.

Adott egy $k < n$ elemű független vektorrendszer, amelyek alkossák egy X_{nxk} mátrix oszlopait. Legyen továbbá egy y_{nx1} vektor. Keressük meg azt a vektort az $\mathbf{X}$ oszlopai által kifeszített altérben, ami a legközelebb van y-hoz!

Mátrixjelölésekkel, keressük azt a $\mathbf{b}^{ols}$ k dimenziós vektort, amelyre

$$\mathbf{b}^{ols} = \arg\min_{b} [(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})]$$

ahol $\mathbf{b}$ $kx1$-es vektor. Jelöljük $\widehat{y}$-pal a legközelebbi pontot, azaz

$$\widehat{y} = \mathbf{X}\mathbf{b}^{ols}.$$

Tehát $\mathbf{b}^{ols}$ elemei a koordinátái $\mathbf{X}$ oszlopaira nézve az $\widehat{y}$-nak.
A célfüggvény kifejtve:

$$\mathbf{y}'\mathbf{y} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} - 2\mathbf{b}'\mathbf{X}'\mathbf{y}.$$

A deriváltat 0-val egyenlővé téve kapjuk a megoldást. A

$$2\mathbf{X}'\mathbf{X}\mathbf{b} - 2\mathbf{X}'\mathbf{y} = \mathbf{0}$$

normál egyenletekből:

$$\mathbf{b}^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y}.$$

(A formula láthatóan az általánosítása az előző alfejezetben b-re adott formulának, ami a $k = 1$ és $n = 2$ eset.)

Az $\widehat{\mathbf{y}}$ pont, ami

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y} = \mathbf{N}\mathbf{y}$$

alakban írható, egy lineáris projekció képe, ahol a projekció mátrixa $\mathbf{N}$. Egyszerú számolás igazolja ugyanis, hogy $\mathbf{N} = \mathbf{N}^2$ és $\mathbf{N}$ rangja k a feltételek alapján. A reziduális $\mathbf{u} = \widehat{\mathbf{y}} - \mathbf{y}$ vektor ortogonális az $\mathbf{X}$ által kifeszített altérre, és így X minden oszlopára és $\widehat{y}$-ra is:

$$\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'})\mathbf{y} = \mathbf{0}.$$

Az

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'} = \mathbf{I} - \mathbf{N}.$$

mátrix is projekció, azaz $\mathbf{M}^2 = \mathbf{M}$ és, belátható, hogy $\mathbf{M}$ rangja $n - k$. (Egy $\mathbf{P}$ projekciónak minden sajátértéke vagy 1 vagy 0, és a mátrix rangja megegyezik az 1-es sajátértékek számával.)

Továbbá definíció szerint:

$$\mathbf{u} = \mathbf{M}\mathbf{y},$$

vagyis $\mathbf{u}$ az $\mathbf{y}$ vetítettje az $\mathbf{X}$ által kifeszített altér ortogonális komplementerére.

5.3 Lineáris regresszió determinisztikus regresszorokkal: statisztika

A klasszikus statisztikai problémáknál mindig abból indulunk ki, hogy a megfigyelések valamilyen valószínűségi változó realizációi. A pontbecslés célja, hogy a megfigyelésekből következtessünk az ismeretlen eloszlás bizonyos tulajdonságaira.

5.3.1 A probléma megfogalmazása

Az Y_i valószínűségi változókat (célváltozó) az alábbi mechanizmus generálja:

$$Y_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + \varepsilon_i, i = 1, ...n$$

1. $\mathbf{X}_i$ nem-sztochasztikus és $k-1$ elemű vektor (regresszorok vagy kontrolváltozók)

2. ε_i iid $\sigma^2 > 0$ varianciával és $E(\varepsilon_i) = 0$,

Ebben a problémában ϵ_i-k "okozzák" az y_i-k véletlenszerűségét. Feltesszük, hogy ezeket nem tudjuk megfigyelni. Ezeket szokás "véletlen hibának" vagy "zajnak" nevezni. Szokásos interpretációjuk: az X mátrixhoz tartozó hatásokon kívüli összes hatás eredője. Később majd találkozunk egy másik interpretációval is: $\epsilon_i = y_i - E(y_i \mid X)$, az y_i változó és X feltétel melletti várható értékének az eltérése, mint valószínűségi változó.

Az ismeretlen paraméterek ebben a modellben $\beta_1, ...\beta_k$ és σ^2.

Mátrix alakban felírva a lineáris regressziós modellt:

$$\mathbf{Y}_{nx1} = \boldsymbol{\beta}_1 \mathbf{1}_{nx1} + \mathbf{X}'_{nx(k-1)} \boldsymbol{\beta}_{(k-1)x1} + \boldsymbol{\varepsilon}_{nx1}.$$

(Fontos: a "lineáris" jelző jelentése az, hogy a paraméterekben lineáris. Például X második oszlopa lehetne X első oszlopának a négyzete, amikor a leképezés nem-lineáris lenne az X-ben.))

A modell átfogalmazható, ha az $\mathbf{1}_{nx1}$ vektort "berakjuk" az $\mathbf{X}$ mátrixba. Így nem kell külön kiemelnünk a konstans (tengelymetszet, intercept) szerepét.

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

$$\begin{aligned} E(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') &= \sigma^2 \mathbf{I} \end{aligned}$$

A továbbiakban feltesszük azt is, hogy $det(\mathbf{X}'\mathbf{X}) \neq 0$.

5.3.2 A β becslése OLS-sel

Tegyük fel, hogy $\mathbf{y}$ az $\mathbf{Y}$ valószínűségi változókból vett n elemű iid minta.

Kézenfekvő ötlet a legkisebb négyzetek becslést az alábbi módon definiálni:

$$\mathbf{b}^{ols} = \arg\min_{\mathbf{b}} \left[(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \right].$$

Mint láttuk:

$$\mathbf{b}^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y}.$$

A legkisebb négyzetek becslőfüggvény tulajdonságai Nyilvánvalóan $\mathbf{b}^{ols}$ egy statisztika, vagyis a mintaelemek egy véltlen függvénye. Azok az általános statisztikai fogalmak, amelyeket megismertünk az előző fejezetben alkalmazhatóak.

Állítás: A $\mathbf{b}^{ols}$ torzítatlan becslése $\boldsymbol{\beta}$-nak, valamint BLUE (azaz legkisebb varianciájú torzítatlan, lineáris becslés). Továbbá ha $lim_{n\to\infty}((\mathbf{X}'\mathbf{X})/n) = Q$, akkor konzisztens is, Bizonyítás:

1. Torzítatlanság

$$\begin{aligned}
\mathbf{b}^{ols} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
\mathbf{b}^{ols} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\boldsymbol{\epsilon} \\
E(\mathbf{b}^{ols}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}E(\epsilon) = \boldsymbol{\beta}.
\end{aligned}$$

2.a BLUE tulajdonság

A $\mathbf{b}$ variancia és kovariancia mátrixa:

$$Var(\mathbf{b}^{ols}) = E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta})'$$

$$\begin{aligned}
&= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^{'}\mathbf{X})^{-1}.
\end{aligned}$$

Legyen $\mathbf{b}^c$ egy másik torzítatlan lineáris becslés:

$$\mathbf{b}^c = \mathbf{b}^{ols} + \mathbf{C}\mathbf{y}.$$

Kifejtve

$$\mathbf{b}^c = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}).$$

A formulából látszik, hogy ha $\mathbf{b}^c$ torzítatlan, akkor $\mathbf{C}\mathbf{X} = \mathbf{0}$ és kovarancia mátrixa

$$\begin{aligned}
Var(\mathbf{b}^c) &= E((\mathbf{b}^c - \boldsymbol{\beta})(\mathbf{b}^c - \boldsymbol{\beta})' \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\boldsymbol{\epsilon}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\boldsymbol{\epsilon})') = \\
&\quad \sigma^2((\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}\mathbf{C}').
\end{aligned}$$

Vagyis

$$Var(\mathbf{b}^c) = var(\mathbf{b}^{ols}) + \mathbf{Q},$$

ahol $\mathbf{Q}$ egy szimmetrikus pozitív (szemi)definit mátrix.

Ebből következik, hogy $\mathbf{b}^{ols}$ bármely nemnulla lineáris kombinációjának a varianciája kisebb, mint $\mathbf{b}^c$ ugyanazon lineáris kombinációjának a varianciája.

3. konzisztencia

Ha $\frac{1}{n}(\mathbf{X}'\mathbf{X})$ konvergál, akkor

$$\begin{aligned} \mathbf{b}^{ols} &= \boldsymbol{\beta} + n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\frac{\boldsymbol{\epsilon}}{n} \\ \operatorname*{Pr\,lim}_{n\to\infty} \mathbf{b}^{ols} &= \boldsymbol{\beta} + \mathbf{Q}^{-1} \operatorname*{Pr\,lim}_{n\to\infty} \frac{\mathbf{X}'\boldsymbol{\epsilon}}{n} = \boldsymbol{\beta}. \end{aligned}$$

(Az utolsó egyenlőség azért igaz, mert $\mathbf{X}' E(\frac{\boldsymbol{\epsilon}}{n}) = 0$, és $\lim_{n\to\infty} \mathbf{X}' var\frac{\boldsymbol{\epsilon}}{n} = 0$, a Nagy Számok Törvénye következménye.)

A σ^2 variancia pontbecslése A legkisebb négyzetek elv alapján csak a $\boldsymbol{\beta}$ vektorra kapunk becslést. Fontos azonban számunkra a σ^2 variancia is. Legyen

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b}^{ols}$$

az OLS reziduumok vektora. Kézenfekvőnek tűnhet σ^2-et $\frac{\mathbf{u}'\mathbf{u}}{n}$nel becsülni, ám így nem kapnánk torzítatlan becslést.

Legyen

$$s^2 = \frac{1}{n-k}\sum u_i^2 = \frac{\mathbf{u}'\mathbf{u}}{n-k}.$$

(Az s^2 pozitív s négyzetgyökét a regresszió standard hibájának nevezzük.)

Állítás: s^2 a σ^2 torzítatlan becslése, azaz

$$E(s^2) = \sigma^2.$$

Tehát a fentiek alapján: $s^2(\mathbf{X}'\mathbf{X})^{-1}$ torzítatlan becslése a $Var(\mathbf{b}^{ols})$ mátrixnak.

Bizonyítás:

$$\mathbf{u} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{M}\boldsymbol{\epsilon}$$

mivel

$$\mathbf{M}\mathbf{X} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{0}.$$

Tehát:

$$\mathbf{u}'\mathbf{u} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

Az idempotens mátrixok és a nyom tulajdonságai alapján

$$\mathbf{E}(\mathbf{u}'\mathbf{u}) = \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}) = \mathbf{E}(tr(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})) = \mathbf{E}(tr(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M})) = tr(\sigma^2\mathbf{IM}) = \sigma^2 tr(\mathbf{M}) = \boldsymbol{\sigma^2}(n-k).$$

(Itt kihasználjuk, hogy a tr lineáris operátor, valamint rendelkezik a következő "ciklikalitás" tulajdonsággal: $tr(ABC) = tr(CAB)$.)

Tehát:

$$E(\frac{\mathbf{u}'\mathbf{u}}{n-k}) = E(s^2) = \sigma^2.$$

Állítás: Ha minden ε_i normális eloszlású, akkor $(n-k)s^2/\sigma^2$ $n-k$ szabadságfokú χ^2 eloszlás Bizonyítás:

$$\frac{(n-k)s^2}{\sigma^2} = \frac{\mathbf{u}'\mathbf{u}}{\sigma^2} = \frac{\varepsilon'}{\sigma}\mathbf{M}\frac{\varepsilon}{\sigma} \sim \chi^2_{n-k}.$$

Állítás: Ekkor az $(s\sqrt{diag((X'X))^{-1}})^{-1}(\mathbf{b}^{ols}-\boldsymbol{\beta})$ vektor minden eleme t_{n-k} eloszlású. Bizonyítás:

A mivel $\mathbf{b}^{ols}$ normális változók lineáris kombinációja, maga is normális változók vektora. Az előzőekben láttuk, hogy a várható értéke β, és $(\sigma\sqrt{\mathbf{diag}((\mathbf{X}'\mathbf{X}))^{-1}})$ a $\mathbf{b}^{ols}$ vektor szórásainak vektora. Tehát $(\sigma\sqrt{\mathbf{diag}((\mathbf{X}'\mathbf{X}))^{-1}})^{-1}(\mathbf{b}^{ols}-\boldsymbol{\beta})$ vektor minden eleme standard normális. Azt kell még belátni, hogy $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ és $\mathbf{b}^{ols}-\boldsymbol{\beta}$ is függetlenek. Mivel

$$\begin{aligned}\mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}) &= \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) \\ &= \sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) = \mathbf{0},\end{aligned}$$

ez is teljesül.

5.3.3 Intervallum becslés

Ahogyan a populáció várható értékére nemcsak pontbecslést, hanem intervallum becslést is definiáltunk, most ezt megtesszük a $\boldsymbol{\beta}$ vektor komponenseire is. Ehhez fel kell tételeznünk a hibák normalitását.

Legyen $s\sqrt{diag_i((X'X))^{-1}} = SE_{b_i}$ a b_i^{ols} becslés sztenderd hibája. Ekkor a következő egyváltozós intervallum becslések adódnak annak alapján, hogy $(s\sqrt{diag((X'X))^{-1}})^{-1}(\mathbf{b}^{ols}-\boldsymbol{\beta})$ minden eleme t_{n-k} eloszlású:

$$b_i^{ols} - SE_{b_i} * t_{n-k,\alpha/2} \leq \beta_i \leq b_i^{ols} + SE_{b_i} * t_{n-k,\alpha/2}.$$

5.3.4 Egyparaméteres hipotézis vizsgálat

Megint csak átivihető a gondolatmenet egyes paraméterek hipotézis vizsgálatára is az előző fejezetből. Ha a nullhipotézis $\beta_i = \beta_{i0}$, akkor az egyparaméteres kétoldali t-tesztekben az elfogadási tartomány:

$$\beta_{i0} - SE_{b_i} * t_{n-k,\alpha/2} \leq b_i^{ols} \leq \beta_{i0} + SE_{b_i} * t_{n-k,\alpha/2}.$$

Hasonlítsuk össze a lineáris regressziót az egyváltozós becsléssel!

	Egyváltozós becslés	Lineáris regresszió
Első momentum	várható érték (μ)	regressziós paraméterek ($\boldsymbol{\beta}$)
Formula	$m = \frac{1}{n}\sum y_i$	$\mathbf{b} = (\mathbf{X\prime X})^{-1}\mathbf{X}'\mathbf{y}$
Tulajdonságok	torzítatlan,hatásos,konzisztens	torzítatlan,hatásos,konzisztens
Becslés varianciája	$var(m)=\frac{\sigma^2}{n}$	$var(\mathbf{b}) = \sigma^2(\mathbf{X\prime X})^{-1}$
Becslés szórása	$sd(m) = \frac{\sigma}{\sqrt{n}}$	$sd(\mathbf{b}) = \sigma\sqrt{diag((\mathbf{X\prime X})^{-1})}$
második momentum	variancia σ^2	hibatagok varianciája σ^2
σ^2torzítatlan becslése	$\frac{1}{n-1}\sum(y_i - m)^2$	$\frac{1}{n-k}\mathbf{u}'\mathbf{u}$
t-konfidencia interval.	$[m - sd(m)t_{\alpha,n-1}, m + std(m)t_{\alpha,n-1}$	$[b_j - sd(b_i)t_{\alpha,n-k}, b_j + sd(b_i)t_{\alpha,n-k}]$
t-próba elfog.tart.	$[\mu_0 - sd(m)t_{\alpha,n-1}, \mu_0 + std(m)t_{\alpha,n-1}$	$[b_{j0} - sd(b_i)t_{\alpha,n-k}, b_{j0} + sd(b_i)t_{\alpha,n-k}]$

A t-konfidencia intervallum és a t-próba érvényessége előfeltételezi a normalitást. Szignifikancia tesztről akkor beszélünk, ha a nullhipotézis $\mu_0 = 0$, vagy $b_{j0} = 0$.

5.3.5 Regresszió ortogonális regresszorokkal

Amennyiben az $\mathbf{X}'\mathbf{X}$ mátrix diagonális, akkor

$$b_j = b_{j_r},$$

ahol b_{i_r}, egy olyan egyváltozós regresszióból származó paraméter, amiben csak az $i-$edik regresszor szerepel.Tehát:

$$b_{j_r,} = \frac{\mathbf{X}_j'\mathbf{y}}{\mathbf{X}_j'\mathbf{X}_j}.$$

5.3.6 F-próbák

Szükség van arra, hogy a paramétereket ne csak külön-külön teszteljük, illetve ne csak külön-külön konfidencia intervallumokat határozzunk meg. Mostantól feltételezzük az ϵ vektor normalitását.

A hagyományos F próba Nullhipotézis: $\beta_2 = ...\beta_k = 0$ (azaz csak a konstans nem 0, vagyis itt van konstans a regresszióban, mely együtthatója β_1.) Ennek a hipotézisnek a tesztjét úgy interpretálhatjuk, higy meg akarunk bizonyosodni arról, hogy van-e egyáltalén értelme a regressziónak. Ha a nullhipotézist elfogadnánk, akkor a regressziós modell eldobható, legalábbis ezekkel a regresszorokkal.

Legyen

$$\begin{aligned} TSS &= (\mathbf{y}-\overline{\mathbf{y}}\mathbf{1})'(\mathbf{y}-\overline{\mathbf{y}}\mathbf{1}). \\ ESS &= \mathbf{u}'\mathbf{u} = (\mathbf{y}-\widehat{y})'(\mathbf{y}-\widehat{y}). \end{aligned}$$

TSS (teljes négyzetösszeg) nem más, mint az $\mathbf{y}$ empirikus varianciájának az n-szerese, ESS a hiba négyzetösszeg, pedig azt mutatja, hogy mennyire sikerült jó közelíteni az $\mathbf{y}$ megfigyeléseket a regresszióval. Ha csak a konstans szerepelne a regresszióban, akkor az OLS becslés $\overline{\mathbf{y}}$ lenne, vagyis TSS méri azt, hogy az $X-$ek nélkül milyen pontos közelítést tudnánk elérni.

Állítás: A nullhipotézis teljesülése esetén

$$\frac{TSS-ESS}{ESS}\frac{n-k}{k-1} \sim F_{k-1,n-k}.$$

Tehát a nullhipotézist elfogadjuk α szinten, ha

$$\frac{TSS-ESS}{ESS}\frac{n-k}{k-1} \leq F^{-1}_{k-1,n-k}(1-\alpha).$$

(A teszt szokás szerint féloldali, mivel az F változók nem-negatívak.)

Bizonyítás:

A nullhipotézis szerint TSS $n-1$ szabadságfokú χ^2 változó, és beláttuk, hogy ESS $n-k$ szabadságfokú χ^2 változó. Azt kell belátni, hogy $TSS-ESS = RSS$ $k-1$ szabadságfokú χ^2 változó, és független $ESS-$től. Ez az alábbi általános állításból következik.

Az eredmény jól interpetálható. Ha az X_i $(i = 2, ..k)$ változók elhagyása csak kicsit ront a közelítésen, akkor a nullhipotézist elfogadjuk, és ezeket fölöslegesnek tekintjük. Az F statisztika egy távolságmérték, ami pontosan definiálja azt, hogy mit tekintünk kicsinek.

Általános ANOVA Tétel lineáris regresszióra* Tekintsünk egy

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

regressziós modellt, ahol ϵ iid normális, és $\mathbf{X}$ rangja és rendje k. Legyen

$$\begin{aligned} \widehat{y} &= \mathbf{N}\mathbf{y} \\ \mathbf{u} &= (\mathbf{I}-\mathbf{N})\mathbf{Y} \end{aligned}$$

Tekintsünk most egy

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

regressziós modellt, ahol a $\mathbf{Z}$ oszlopai által kifeszített altér $l < k$ dimenziós altere az X oszlopai által kifeszített altérnek, és

$$\begin{aligned}\widehat{y}_R &= \mathbf{N}_R\mathbf{Y} \\ \mathbf{u}^R &= (\mathbf{I}-\mathbf{N}_R)\mathbf{Y}.\end{aligned}$$

Ekkor $\mathbf{u}'\mathbf{u}/((n-k)\sigma^2)$ $n-k$ szabadságfokú χ^2, $((\mathbf{N}_R-\mathbf{N})\mathbf{y})'((\mathbf{N}_R-\mathbf{N})\mathbf{y})$ $k-l$ szabadságfokú χ^2, $\mathbf{u}'\mathbf{u}/((n-k)/(k-l)\sigma^2)$ és $((\mathbf{N}_R-\mathbf{N})\mathbf{y})'((\mathbf{N}_R-\mathbf{N})\mathbf{y})$ függetlenek, valamint

$$((\mathbf{N}_R-\mathbf{N})\mathbf{y})'((\mathbf{N}_R-\mathbf{N})\mathbf{y}) = \mathbf{u}'\mathbf{u} - \mathbf{u}_R'\mathbf{u}_R.$$

Bizonyítás:

$$(\mathbf{N}-\mathbf{N}_R)^2 = \mathbf{N}+\mathbf{N}_R-\mathbf{2NN}_R.$$

Mivel

$$(\mathbf{I}-\mathbf{N})\mathbf{N}_R = 0$$

(az altér reláció miatt X_R oszlopai merőlegesek $u=(I-N)y$-ra) teljesül

$$\mathbf{NN}_R = \mathbf{N}_R.$$

Tehát

$$(\mathbf{N}-\mathbf{N}_R)^2 = \mathbf{N}-\mathbf{N}_R.$$

Az $\mathbf{N}-\mathbf{N}_R$ mátrixnak pontosan $k-l$ 1-es sajátértéke van. (Ismét az altér tulajdonság miatt.)

Továbbá

$$(\mathbf{N}-\mathbf{N}_R)\mathbf{N}_R = 0$$

vagyis a számlálóban és a nevezőben levő négyzetösszegek függetlenek egymástól.

Az F próba általánosítása Nullhipotézis: $\beta_{j+1} = ...\beta_k = 0$. (Vagyis: $k-j$ restrikció van.) Az "U" modell az, ahol nincsenek korlátozások a paraméterekre, és az "R" modell az, ahol a korlátozások érvényesek. Az előző tételből belátható a következő állítás.

Állítás:

$$\frac{ESS_R - ESS_U}{ESS_U}\frac{n-k}{k-j} \sim F_{k-j,n-k}.$$

Ez a teszt nyilvánvalóan a hagyományos F próba általánosítása. Ez a módszer mindig működik, ha az "R" modellt fel tudjuk írni, mint az eredeti "U" modell speciális esetét.

Általános variancia analízis Ha van egy n elemű minta és egy k rangú (független paraméterű) U modell, amelynek reziduális négyzetösszege ESS_U, és egy $j < k$ rangú R modell, amely "beágyazott" (speciális esete) az U modellnek ($m = k - j$ a korlátozások (restrikciók) száma), és amelynek reziduális négyzetösszege ESS_R, akkor

$$\frac{ESS_R - ESS_U}{ESS_U} \frac{n-k}{k-j} \sim F_{k-j,n-k}.$$

További általánosítás: paraméterek lineáris kombinációjának tesztjei Most legyen a nullhipotézis:

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

ahol $\mathbf{R}$ mxk ($m < k$) rendű mátrix. Ilyenkor is alkalmazhatjuk az általános ANOVA tételt, miután megbecsültük a korlátozott modell paramétereit. A következő feladatot kell megoldani a korlátozott modell becslésének meghatározásához:

$$\begin{aligned} & \max(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ \mathbf{Rb} &= \mathbf{r}. \end{aligned}$$

Viszont van egy "gyorsabb" módszer is. Ugyanis

$$\begin{aligned} Var(\mathbf{Rb}) &= E(\mathbf{Rb})(\mathbf{Rb})') = \mathbf{R}\mathbf{E}(\mathbf{bb}')\mathbf{R}' \\ &= \mathbf{R}\mathbf{V}ar(\mathbf{b})\mathbf{R}' \\ &= \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'. \end{aligned}$$

Állítás:

$$(\mathbf{Rb} - \mathbf{r})'(\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r}) \sim \chi^2_m.$$

Bizonyítás: $(\mathbf{Rb} - \mathbf{r})$ 0 várhatú értékű m elemű normális változó vektor, ha igaz a nullhipotézis. Ennek kovariancia mátrixa $\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$, tehát $(\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1/2}(\mathbf{Rb} - \mathbf{r})$ m elemű standard normális vektor.

Ebből következik (az általános ANOVA Tétel alapján), hogy

$$\frac{(\mathbf{Rb} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})}{ESS_U} \frac{n-k}{m} \sim F_{m,n-k}.$$

5.3.7 F statisztika: konfidencia tartomány

A konfidencia intervallumok és az egyváltozós tesztek meghatározásánál lényegében ugyanazokat a számításokat kellett elvégezni. Mindkét esetben egy intervallumot (a próbánál az elfogadási intervallumot) kell meghatározni, csak míg a konfidencia intervallumnál a középpont a becsült paraméter érték, addig az elfogadási

intervallumnál a nullhipotézisnek megfelelő paraméter érték. A fenti formulákból látszik, hogy az F próbánál az elfogadási tartomány egy ellipszoid. Nem meglepő, hogy készíthetünk konfidencia tartományt is, ahol az hasonlóképpen egy ellpiszoid lesz.

Megállapítottuk, hogy

$$var(\mathbf{b}^{ols}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Ezért $\left[\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})\right]^{\frac{1}{2}}(\mathbf{b}^{ols}-\boldsymbol{\beta}) \sim N(\mathbf{0},\mathbf{I})$, amiből

$$(\mathbf{b}^{ols}-\boldsymbol{\beta})'\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})(\mathbf{b}^{ols}-\boldsymbol{\beta}) \sim \chi^2_k.$$

Ha ezt osztjuk $s^2/\sigma^2 = \frac{\mathbf{u}'\mathbf{u}}{(n-k))\sigma^2}$-tel, akkor

$$(\mathbf{b}^{ols}-\boldsymbol{\beta})'\frac{1}{s^2}(\mathbf{X}'\mathbf{X})(\mathbf{b}^{ols}-\boldsymbol{\beta})\frac{1}{k} = (\mathbf{b}^{ols}-\boldsymbol{\beta})'\frac{1}{\mathbf{u}'\mathbf{u}}(\mathbf{X}'\mathbf{X})(\mathbf{b}^{ols}-\boldsymbol{\beta})\frac{n-k}{k} \sim F_{k,n-k}.$$

Ez meghatároz egy konfidencia tartományt (ellipszoid) a teljes paraméter vektorra tetszőleges $1-\alpha$ konfidencia szinten. Valamely b^* beletartozik ebbe a konfidencia tartományba, ha

$$(\mathbf{b}-\mathbf{b}^*)'\frac{1}{\mathbf{u}'\mathbf{u}}(\mathbf{X}'\mathbf{X})(\mathbf{b}-\mathbf{b}^*)\frac{n-k}{k} \leq F^{-1}_{k,n-k}(1-\alpha).$$

5.3.8 Maximum likelihood becslés normális esetben

A legkisebb négyzetek elve mellett alkalmazhatjuk a maximum likelihood elvet regresszióban is, ha konkrét feltevéssel élünk a hibák eloszlásáról. Most is feltesszük, hogy $\varepsilon_i \sim N(0,\sigma^2)$ $i = 1,...n$, és $E(\epsilon_i\epsilon_j) = 0, i \neq j$.

Ekkor a minta sűrűségfüggvénye:

$$\begin{aligned} L(\mathbf{y},\mathbf{X};\beta,\mathbf{s}^2) &= \sigma^{-n}(2\pi)^{-\frac{n}{2}}\exp(-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ &= \sigma^{-n}(2\pi)^{-\frac{n}{2}}\exp(-\frac{\sum_{i=1}^{n}(y_i-\mathbf{X}_i'\boldsymbol{\beta})^2}{2\sigma^2}). \end{aligned}$$

Rögzítjük a megfigyelések értékeit, ekkor megkapjuk a likelihood függvényt, majd annak a logaritmusát véve:

$$\ln L = -n\ln s - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_i(y_i-\mathbf{X}_i'\boldsymbol{\beta})^2.$$

Az ML elv szerint úgy választjuk meg a $(\mathbf{b}_{ML}, s^2_{ML})$ értékeket, hogy a (log)likelihood függvény a maximumát vegye fel az adott mintában.

A β ML becslése láthatóan ugyanaz, mint az OLS becslés, tehát normális változókat feltételezve az OLS elv ismét ugyanazt az eredményt adja, mint az ML elv. Az σ^2 szerinti elsőrendű feltétel:

$$-\frac{n}{s_{ML}} + \frac{1}{s_{ML}^3}\sum(y_i - X_i'\mathbf{b}^{ML})^2 = 0$$

megoldásaként

$$s_{ML}^2 = \frac{\sum(y_i - X_i'b^{ML})^2}{n} = \frac{ESS}{n}$$

adódik, ami, mint tudjuk, torzított becslése σ^2-nek.

5.3.9 Példák F és t-próbák több paraméterre vonatkozó korlátozása esetén

Ebben az alfejezetben megmutatjuk konkrét példákon keresztül, hogy hogyan lehet t-próbákat is alkalmazni több paraméterre vonatkozó hipotézis mellett is.

Példa: Induljunk ki a

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

regressziós modellből, és a nullhipotézis legyen:

$$\beta_2 - \beta_3 = 0.$$

1. megoldás: F próba a modell átparametrizálásával A nullhipotézis teljesülése esetén

$$\beta_3 = \beta_2$$

$$y = \beta_1 + \beta_2 x_2 + \beta_2 x_3 + \epsilon.$$

Tehát a modell átírható

$$y = \beta_1 + \beta_2 z + \epsilon$$

alakba, ahol $z = x_2 + x_3$ egy új változó. Ekkor ezt tekinthetjük a korlátozott (restricted) modellnek és teljesül

$$\frac{ESS_R - ESS_U}{ESS_U}\frac{n-3}{1} \sim F_{1,n-3}.$$

ami alapján tesztelhetjük a nullhipotézist.

2. megoldás: F próba átparametrizálás nélkül Ekkor

$$\begin{aligned} R &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \\ r &= \begin{bmatrix} 0 \end{bmatrix}. \end{aligned}$$

Az előzőek alapján:

$$\frac{(\mathbf{Rb}-\mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r})}{ESS_U}\frac{n-3}{1} \sim F_{1,n-3}.$$

3. megoldás: közvetett t próba Ekkor vezessünk be egy új változót: $\delta = \beta_2 - \beta_3$.

A modell és a nullhipotézis új alakja:

$$\begin{aligned} y &= \beta_1 + (\delta + \beta_3)x_2 + \beta_3 x_3 + \epsilon \\ &= \beta_1 + \delta x_2 + \beta_3 X + \epsilon. \end{aligned}$$

A nullhipotézis ebben a felírásban:

$$\delta = 0.$$

Ha d a δ paraméter OLS becslése, akkor

$$\frac{d}{se_d} \tilde{} t_{n-3},$$

és t próbát végezhetünk. (Itt se_d a d standard hibája, vagyis az $s^2(\mathbf{X}'\mathbf{X})^{-1}$ mátrix megfelelő diagonális elemének a négyzetgyöke.)

4. megoldás: közvetlen t próba A nullhipotézis teljesülése esetén

$$\frac{b_2 - b_3}{\sqrt{s_{b_2}^2 + s_{b_3}^2 - 2cov_s(b_2, b_3)}} \tilde{} t_{n-3},$$

ahol $s_{b_2}^2$és $s_{b_3}^2$ a b_2 és b_3 becsült varianciái, és $cov_s(b_2, b_3)$ a becsült kovarianciájuk. (Azaz az $s^2(\mathbf{X}'\mathbf{X})^{-1}$ mátrix megfelelő elemei.) Ez az összefüggés ismét egy t statisztikán alapuló tesztet tesz lehetővé, ahol a t statisztika pontosan megegyezik a közvetett próbában számolt t statisztikával.

Tanulságok Mivel

$$F_{1,n} = t_n^2$$

a p-értékek mind a négy esetben ugyanazok lesznek.

Ha egy restrikciónk van, akkor mindegy, hogy F vagy t próbát végzünk. A fenti példát nyilván általánosíthatnánk kettőnél több paraméterre adott, de

egyetlen, restrikcióra is. Több restrikció esetén van szükség F próbára mindenképpen, de itt is választhatunk több megoldás között.

Példa: Két restrikció esete:

$$\begin{aligned} 3\beta_2 - \beta_3 &= 0 \\ \beta_1 + \beta_2 + \beta_3 &= 0. \end{aligned}$$

1. megoldás: Bizonyos β-k kifejezése a többiek függvényeként, majd az így redukált dimenziójú modell becslése, és a korlátozott és általános modell variancia analízise.

Ekkor

$$\begin{aligned} \beta_3 &= 3\beta_2 \\ \beta_1 &= -\beta_2 - \beta_3. \\ y &= -\beta_2 - 3\beta_2 + \beta_2 x_2 + 3\beta_2 x_3 + \epsilon \\ y &= \beta_2(-4 + x_2 + 3x_3) + \epsilon. \end{aligned}$$

Tehát legyen

$$\begin{aligned} z_1 &= (-4 + x_2 + 3x_3) \\ y &= \beta_2 z_1 + \epsilon. \end{aligned}$$

2. megoldás Az m korlátozást tartalmazó modell átalakítható olyan alakba, hogy m zéró restrikciót kelljen tesztelni. Vezessünk be $\delta_1, , , \delta_m$ új változót és hagyjunk ki m darab β-t.

$$\begin{aligned} 3\beta_2 - \beta_3 &= 0 \\ \beta_1 + \beta_2 + \beta_3 &= 0. \end{aligned}$$

Legyen

$$\begin{aligned} \delta_1 &= 3\beta_2 - \beta_3 \\ \delta_2 &= \beta_1 + \beta_2 + \beta_3 \\ \beta_3 &= 3\beta_2 - \delta_1 \\ \beta_1 &= \delta_2 - \beta_2 - \beta_3. \end{aligned}$$

Ekkor a modell:

$$\begin{aligned} y &= (\delta_2 - \beta_2 - \beta_3) + \beta_2 x_2 + \beta_3 x_3 + \epsilon \\ y &= (\delta_2 - \beta_2 - 3\beta_2 + \delta_1) + \beta_2 x_2 + (3\beta_2 - \delta_1) x_3 + \epsilon \\ y &= \beta_2(-4 + x_2 + 3x_3) + \delta_1(1 - x_3) + \delta_2 + \epsilon \end{aligned}$$

Kell egy új változó:

$$z_2 = 1 - x_3.$$

Így felírva a korlátozatlan modell:

$$y = \beta_2 z_1 + \delta_1 z_2 + \delta_2 + \epsilon.$$

Ebben a modellben a $\delta_1 = \delta_2 = 0$ korlátozásokat teszteljük.

3. megoldás A korlátozott modell becslése nélkül a $\frac{(\mathbf{Rb}-\mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r})}{ESS_U}\frac{n-k}{m}$ tesztstatisztika használata.

Meg kell találnunk az $\mathbf{R}$ mátrixot és az $\mathbf{r}$ vektort. Ebben az esetben:

$$\begin{aligned} R &= \begin{bmatrix} 0 & 3 & -1 \\ 1 & 1 & 1 \end{bmatrix} \\ r &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

4. megoldás A korlátozott modell becslése korlátozott legkisebb négyzetekkel, majd a korlátozott és általános modell variancia analízise.

5.4 A regressziós paraméterek értelmezése

A β_j az y x_j változó szerinti deriváltja, tehát jelentése: egy egységnyi x_j növekedés hatása az y-ra várható értékben. Viszont a regressziós egyenlet csak a paraméterekben lineáris szükségképpen. Lehetséges például, hogy a modell

$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2 + \beta_5 x_1 x_2 + \epsilon$$

alakú. Ekkor például az x_1 szerinti derivált $= \beta_2 + 2\beta_3 x_1 + \beta_5 x_2$.

Nyilvánvalóan ésszerű követelmény, hogy a változók lineáris transzformációja lineárisan hasson a paraméterek értékére. Azaza, ha a magyarázó változó eleinte centiméterben volt megadva, de most áttérünk méterre, akkor az együttható a százszorosára növekedjen.

Legyen $\langle \mathbf{t} \rangle$ egy kxk-s diagonális mátrix. Az új mértékegyégben felírt $\mathbf{X}$ mátrix: $\mathbf{X}\langle \mathbf{t} \rangle$. (Pl $t_j = 0.01$ ha centiméterről méterre térünk át.)

Erre alkalmazva az OLS formulát:

$$\begin{aligned} \mathbf{b}^t &= ((\mathbf{X}\langle \mathbf{t} \rangle)'\mathbf{X}\langle \mathbf{t} \rangle)^{-1}(\mathbf{X}\langle \mathbf{t} \rangle)'\mathbf{y} \\ &= (\langle \mathbf{t} \rangle'\mathbf{X}'\mathbf{X}\langle \mathbf{t} \rangle)^{-1}\langle \mathbf{t} \rangle'\mathbf{X}'\mathbf{y} \\ &= (\langle \mathbf{t} \rangle')^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

Figyeljük meg, sehol sem használtuk ki, hogy $\langle \mathbf{t} \rangle$ diagonális. Tehát bármilyen $\mathbf{XT}$ változó transzformációra, ahol $\mathbf{T}$ nemszinguláris igaz, hogy $\mathbf{b}^T = (\mathbf{T}')^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Mi a hatása annak, ha egy j változót eltolunk, azaz $\mathbf{X}_j^T = \mathbf{X}_j + \mathbf{1}K_j$ a j-edik oszlopa az $\mathbf{X}$ mátrixnak. Tekintsük a minimalizálási problémát, ha van konstans a regresszióban:

$$\min_b \sum_i^n (y_i - b_1 - \mathbf{X}'_{2i}\mathbf{b}_2)^2.$$

Itt $\mathbf{X}'_{2i}$ és $\mathbf{b}_2$ a konstanst és együtthatóját nem tartalmazó vektorok. Ekkor a módosított feladat:

$$\min_b \sum_i (y_i - b_1 - b_j K_j - \mathbf{X}'_{2i}\mathbf{b}_2)^2.$$

A megoldás $\mathbf{b}_2$-re ugyanaz, de a módosított feladat megoldásának konstans paramétere:

$$b_1^T = b_1 + b_j K_j.$$

Vagyis, ha a változó eltolása értelmes művelet, akkor kell lennie konstansnak a regresszióban ahhoz, hogy a paraméter becslés ne változzon.

Egy speciális eset:

$$\mathbf{X}_j^T = \mathbf{X}_j - \mathbf{1}\overline{X_j}.$$

Ekkor $\mathbf{b}_2$ nem változik, és

$$b_1^T = b_1 - \sum_{j=2}^k b_j \overline{X_j}.$$

Mivel

$$\mathbf{1}'(\mathbf{y} - b_1 - \mathbf{X}_2\mathbf{b}_2) = 0$$

látható, hogy

$$b_1^T = \overline{y}.$$

Ha a

$$\mathbf{y}^T = \mathbf{y} - \mathbf{1}\overline{y}$$

transzformációt is végrehajtjuk, akkor

$$b_1^T = 0,$$

Vagyis ha minden változót (beleértve a célváltozót is) centralizálunk, akkor kihagyható a konstans a regresszióból, de ha "bennfelejtjük", akkor is 0 értéket fog kapni az OLS becslésben.

5.5 A regressziós modell minősége: modellválasztás

5.5.1 Az illeszkedés jósága

Tegyük fel, hogy van konstans a regresszióban. Legyen

$$\widehat{y_i}=b_1+\mathbf{X}_{i2}\mathbf{b}_2.$$

Ekkor

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}((\widehat{y_i}-\overline{y})+(y_i-\widehat{y_i}))^2\\ &= \sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2+\sum_{i=1}^{n}(y_i-\widehat{y_i})^2\end{aligned}$$

Bizonyítás:

Azt kell belátni,hogy $\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})(y_i-\widehat{y_i})=0$. Viszont ez igaz, mivel $(\widehat{\mathbf{y}}-\overline{y}\mathbf{1})'(\mathbf{y}-\widehat{\mathbf{y}})=(\widehat{\mathbf{y}}-\overline{y}\mathbf{1})'\mathbf{u}=0$.

A determinációs együttható definíciója:

$$R^2=\frac{TSS-ESS}{TSS}=1-\frac{ESS}{TSS},0\le R^2\le 1.$$

A fentiek alapján

$$\frac{R^2}{1-R^2}\frac{n-k}{k-1}$$

a hagyományos F teszt statisztika. Vagyis a nullhipotézist akkor fogadjuk el, ha az R^2 közel van $0-$hoz.

Gyakran kiszámolják a korrigált R^2 statisztikát is:

$$\overline{R}^2=1-\frac{(n-1)ESS}{TSS(n-k)}.$$

Itt a szabadságfokokkal korrigálunk. Úgy indokolhatjuk ezt a mutatót, hogy y (feltétel nélküli) varianciájának torzítatlan becslése $\frac{TSS}{n-1}$, és az ε varianciájának torzítatlan becslése $\frac{ESS}{n-k}$. A korrigált R^2 lehet negatív!

Figyelem: sem az R^2, sem korrigált változata nem abszolút modell minőségi kritérium!

5.5.2 Modell szelekciós kritériumok

Az F tesztekkel csak egymásba ágyazott modelleket tudunk összehasonlítani. A modell szelekciót információelméleti alapokon is el lehet végezni.

Entrópiának nevezzük valószínűségi eloszlások "bizonytalanságát" (megjósolhatatlanságát). Például a k dimenziós normális eloszlás entrópiája:

$$\frac{1}{2}\ln((2\pi e)^k\det(\Sigma)).$$

Láthatóan az eloszlás entrópiája kisebb (a modell informatívabb), ha a becsült paraméterek száma kisebb, és ha kovariancia mátrix közelebb van a 0-hoz. Ennek megfelelően egy becsült modell információtartalma nagyobb, ha

- kisebb az ESS (pontosabb), és ha
- kevesebb a becsült paraméter.

Az információs kritériumok ezt a két szempontot súlyozzák. A két talán legnépszerűbb az Akaike és a bayes-i kritérium.

Az Akaike (AIC) kritérium:

$$AIC = 2k - 2 \max \log L,$$

a Schwartz vagy bayes-i (BIC) kritérium:

$$BIC = \ln(n)k - 2 \max \log L.$$

A BIC jobban bünteti a sok paramétert, és annál inkább, minél több a megfigyelés. Két modell összehasonlításánál azt tekintjük jobbnak, ahol a kritériumok értéke kisebb.

5.6 Problémák a lineáris regressziós modellel

5.6.1 Multikollinearitás

Szigorú multikollinearitás áll fenn, ha $(\mathbf{X}'\mathbf{X})^{-1}$nem létezik. Ilyenkor az $\mathbf{X}$ oszlopai nem függetlenek. Nem szigorú multikollinearitásról beszélünk, ha $(\mathbf{X}'\mathbf{X})^{-1}$ létezik, de "nagyon nagy" abszolút értékben. (Mintha az $\mathbf{X}'\mathbf{X}$ közel lenne a nullmátrixhoz.) Ilyenkor a becslések varianciája nagy, tehát nagyon pontatlanok.

A multikollinearitás legnagyobb problémája az, hogy a t tesztek ereje csökken (nagy a másodfajú hiba valószínűsége), azaz gyakran fordul elő, hogy a nullhipotézist tévesen fogadjuk el.

A problémát jelzi a "variancia infláció tényező":

$$VIF = \frac{var(b_j)}{var(b_{j_r})},$$

ahol a $var(b_{jr})$ egy olyan regresszióból származik, amelyben csak a kérdéses (j-edik) változó szerepel. A VIF értéke legalább 1. Ha a VIF 5-nél nagyobb, akkor gyanítható, hogy a t tesztek ereje kicsi. A becslés varianciájának növekedése miatt az egyes paraméterek konfidencia intervallumai is tágak (semmitmondóak) lesznek.

5.6.2 Nem-normalitás

Mivel a t és F tesztek elvben csak normális hibatagok mellett igazak, fontos tudni, hogy a hibatagok normalitása feltevés teljesül-e. A nem-normalitást jelezheti a reziduum nem-normalitása. Számos teszt létezik annak a nullhipotézisnek a tesztelésére, hogy egy adott minta normális eloszlásból származik-e. Ilyen például a Jarque-Bera teszt:

$$
\begin{aligned}
JB &= \frac{n-k+1}{6}(S^2 + \frac{1}{4}(C-3)^2) \\
S &= \frac{m^3}{s^3} \\
C &= \frac{m_4}{s^4} \\
& JB \tilde{} asym\chi^2_2,
\end{aligned}
$$

ahol m_3 és m_4 a harmadik és negyedik centrális mintamomentum, S a minta ferdesége, és C a minta kurtosisa. Normális eloszlásnál a ferdeség 0, a kurtosis pedig 3.

Vizuálisan gyakran alkalmazzák a kvantilis-kvantilis ábrát, ahol az x tengelyen az elméleti normális kvantilis értéke, az y tengelyen pedig a mintakvantilis értéke van. Normalitás esetén az ábra közel 45 fokos egyenes lenne.

5.6.3 Kiugró értékek (outliers)

Ha vannak olyan megfigyelések, amelyek kiugró $\widehat{y_i} - y_i$ értéket adnak, akkor ez megnöveli a standard hibát és a paraméter becsléseket bizonytalanná teszi. Az outlier nagyon "messze" van a regressziós hipersíktól, nagy a hozzá tartozó reziduum. Megnöveli a becsült varianciát, ami hat a tesztekre (a szignifikancia ellen "hat"). Ha az outlier oka adathiba vagy kis valószínűségű véletlen, akkor ésszerű kihagyni a becslésből, de lehet figyelmeztető jele is annak, hogy a specifikáció rossz. A problémát észlelhetjük a "studentizált" reziduumok kiszámításával. Ilyenkor az egyes megfigyeléseket kihagyjuk a mintából, újra becslünk és a kihagyott megfigyelést osztjuk a megfelelő predikciós sztenderd hibával. Általában 3-nál nagyobb studentizált reziduumot szoktak outlier-nek tekinteni.

5.6.4 Nagyhatású magyarázó változó megfigyelések (high-leverage observations)

Az előző probléma akkor is fennállhat, ha valamely magyarázó változó egy bizonyos megfigyelése túlságosan nagy befolyással van az eredményre. Ennek megítéléséhez jól jön egy szintetikus mutató.

A kalapos értékek (hat-values) alatt a projekciós mátrix ($\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$) diagonálisát értjük. Ha valamely kalapos érték abszolút értéke több, mint kétszerese a többi kalapos érték abszolút értékének, akkor gyanús, hogy a megfelelő megfigyelés túl nagy hatással van a becslés értékére, és ha ráadásul a megfelelő y_i is outlier, akkor különösen. Tehát, ha kísérletet tervezünk arra is figyelni kell, hogy ne legyenek "kiugró" regresszor értékek, vagyis a regresszor vektorok legyenek "sűrűn" elhelyezve.

5.6.5 Egymásba ágyazott (nested) és nem egymásba ágyazott (non-nested) modellek

Amikor a nullhipotézis jellemezhető a fenntartott hipotézis altereként, akkor beágyazott (nested) hipotézisről beszélünk. Amikor nem, akkor úgynevezett non-nested hipotézis tesztelésről van szó. Ilyenkor a null és az alternatív hipotézis egyforma súlyú. (A célváltozó (eredményváltozó) mindig ugyanaz.) Az egyik legegyszerűbb tesztelési elv, ha a két hipotézis uniójaként adott modellt tekintjük fenntartott hipotézisnek, és ebben teszteljük a két hipotézist, mint nullhipotéziseket. Például ha az első modellben a regresszorok mátrixa $\mathbf{X}_1$ és a másodikban $\mathbf{X}_2$, akkor tekintsük azt a modellt, ahol a regresszorok mátrixa $[\mathbf{X}_1, \mathbf{X}_2]$, és teszteljük azt a nullhipotézist, hogy az X_1-hez tartozó paraméterek 0-k, majd azt, hogy az $\mathbf{X}_2$-höz tartozó paraméterek 0-k. Ez felfogható modellválasztási eljárásnak is: ha az egyiket elutasítjuk, míg a másikat nem, akkor a nem-elutasított modellt tekinthetjük jobbnak. Azonban a teszt eredménye nem feltétlenül vezet egyértelmű döntéshez.

5.7 Többváltozós lineáris regresszió sztochasztikus regresszorokkal

Regressziót olyankor használunk általában, amikor adott információkból (regresszorok értékei) következtetni szeretnénk a célváltozó várható értékére. A célváltozó és a "magyarázó változók" megkülönböztetés attól függ, hogy milyen kérdést tartunk érdekesnek. ’

5.7.1 Feltételes várható értékek

Legyen y egy valószínűségi változó, és $\mathbf{x}$ egy valószínűségi változó vektor. (Ebben a vektorban lehet elfajult valószínűségi változó is.)

Állítás: Az iterált várakozások tétele

$$E(y) = E(E(y \mid \mathbf{x})).$$

A feltétel nélküli várható érték a feltételes várható értékek várható értéke.

Állítás A feltételes várható értékre igaz, hogy létezik olyan ϵ, hogy

$$\begin{aligned} y &= E(y \mid \mathbf{x}) + \epsilon \\ E(\epsilon &\mid \mathbf{x}) = 0 \\ E(h(\mathbf{x})\epsilon) &= 0 \end{aligned}$$

minden $h(\mathbf{x})$ függvénnyel. Az utóbbiból következik, hogy ε minden x_i-vel is korrelálatlan.

Állítás

$$E(y \mid \mathbf{x}) = \arg\min_{m(\mathbf{x})}(E(y - m(\mathbf{x}))^2)$$

ahol $m(\mathbf{x})$ tetszőleges függvény.

Azaz a feltételes várhatóérték függvény közelíti y-t legjobban az $\mathbf{x}$ vektor függvényében a minimális várható négyzetes hiba (MMSE) tekintetében.

Példa: Legyen x és z független sztenderd normális változók, tehát $w = x^2$ elsőrendű χ^2, amelynek várható értéke 1, varianciája 2. Legyen $y = 3w + z$. Ekkor

$$E(y) = 3$$

$$\begin{aligned} E(y \mid x) &= 3x^2, \\ y - E(y \mid x) &= z \\ E(zx) &= 0 \\ E(y \mid z) &= 3 + z, \\ y - E(y \mid z) &= 3w - 3 \\ E((3w-3)z) &= 0. \end{aligned}$$

5.7.2 A lineáris projekció

A fentiek szerint a feltételes várható érték az y egyfajta legjobb közelítése $\mathbf{x}$ alapján. Feltehetjük azt a kérdést is, hogy van-e legjobb lineáris közelítése y-nak $\mathbf{x}$ függvényében, ahol a legjobb alatt ismét várható négyzetes eltérés minimalizálást értünk.

Tehát keressük azt a $\widehat{\boldsymbol{\beta}}$-ot, amely megoldja a

$$\min_{\boldsymbol{\beta}} E((y - \mathbf{x}'\boldsymbol{\beta})^2)$$

problémát. Az elsőrendű feltétel:

$$E(\mathbf{x}(y - \mathbf{x}'\widehat{\boldsymbol{\beta}})) = 0.$$

A megoldás:

$$\widehat{\boldsymbol{\beta}} = E((\mathbf{x}\mathbf{x}')^{-1})E(\mathbf{x}y).$$

Az $\mathbf{x}'\widehat{\boldsymbol{\beta}}$ valószínűségi változót y $\mathbf{x}$-re vonatkozó lineáris projekciójának nevezzük. Tehát, míg a feltételes várható érték függvény általában a legjobb MMSE prediktor, a lineáris projekció a legjobb lineáris MMSE prediktor.

Állítás Tekintsük az $\mathbf{x}$ egy x_j komponensét és azt a projekciót, ahol x_j-t x_{-j}-re (az $\mathbf{x}$ vektor kivéve x_j-t) vetítjük. Legyen $\widetilde{x_j}$ a "maradéktag" a többi projektor projekciójából a j-edik projektorra, azaz a többi projektor által "nem magyarázott" rész.

$$\widetilde{x_j} = x_j - \sum_{i \neq j} \widehat{\boldsymbol{\beta}}_{ji} x_i,$$

ahol $\widehat{\boldsymbol{\beta}}_{ji}$ -k a lineáris projekció paraméterei.

Állítás:

$$\widehat{\boldsymbol{\beta}}_j = \frac{cov(y, \widetilde{x_j})}{var(\widetilde{x_j})}$$

(Itt $\widehat{\boldsymbol{\beta}}_j$ az y projekciójának paraméterei.)
Bizonyítás:

$$y = \sum_{i=1}^{k} \widehat{\boldsymbol{\beta}}_i x_k + \varepsilon,$$

szorozzuk meg mindkét oldalt $\widetilde{x_j}$-vel, és vegyük a várható értéket. Mivel

$$\begin{aligned} cov(x_i, \widetilde{x_j}) &= 0, j \neq i \\ cov(x_j, \widetilde{x_j}) &= var(\widetilde{x_j}) \end{aligned}$$

$$cov(y, \widetilde{x_j}) = \widehat{\boldsymbol{\beta}}_j var(\widetilde{x_j}).$$

Igaz az is, hogy

$$\widehat{\boldsymbol{\beta}}_j = \frac{cov(\widetilde{y_{-j}}, \widetilde{x_j})}{var(\widetilde{x_j})}$$

ahol $\widetilde{y_{-j}}$ a "maradéktag" a többi ($\neq j$) regresszor projekciójából az y-ra.

$$\widetilde{y_{-j}} = y_j - \sum_{i \neq j} \widehat{\boldsymbol{\beta}}_{ji} x_i,$$

Másfelől

$$\widehat{\boldsymbol{\beta}}_j = \frac{cov(\widetilde{y_{-j}}, x_j)}{var(x_j)}$$

csak ha a j-edik projektor ortogonális a többire, vagyis ha a $\widetilde{x_j}$ maradéktag szórása ugyanaz, mint a j-edik változó szórása.

A példa folytatása: Tekintsük $y = 3x^2 + z$ projekcióját x-re és a konstansra,

$$\begin{aligned} E(x(3x^2 + z - a - bx)) &= 0, \\ E(3x^2 + z - a - bx) &= 0 \\ a &= 3, b = 0 \\ proj(y \quad &| \quad x, 1) = 3. \end{aligned}$$

Most pedig $y = 3x^2 + z$ projekcióját z-re és a konstansra,

$$\begin{aligned} E(z(3x^2 + z - a - bz)) &= 0 \\ E(3x^2 + z - a - bz) &= 0 \\ a &= 3, b = 1 \\ proj(y \quad &| \quad z, 1) = 3 + z. \end{aligned}$$

A példa mutatja, hogy a lineáris projekció egybeeshet ugyan a feltételes várható értékkel, de ez nem szükségszerű.

A projekció segítségével definiálható az egyszerű korreláció mellett a parciális korreláció::

$$r_{y,x_j} = \frac{cov(\widetilde{y_{-j}}, \widetilde{x_j})}{std(\widetilde{x_j})std(\widetilde{y_{-j}})}$$

ahol *std* jelöli a szórást. Míg a korreláció független egyéb változóktól, parciális korrelációt egy változó halmazon belül definiálunk, vagyis potenciálisan végtelenül sok parciális korrelációs mutató létezhet két változó között.

Állítás Ha a feltételes várható érték függvény lineáris, akkor a paraméterei ugyanazok, mint a lineáris projekció paraméterei.

Állítás

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta}(E(E((y \mid \mathbf{x}) - \mathbf{x}'\beta)^2))$$

azaz a lineáris projekció a legjobb lineáris közelítése a feltételes várható érték függvénynek.

Kiderül tehát, hogy a két probléma (lineáris projekció és a feltételes várható érték legjobb lineáris közelítése) megoldása egybeesik.

Együttes normalitás esetén biztosan teljesül, hogy a feltételes várható érték függvény lineáris, egyébként ez csak feltevés, vagy közelítés. Együttes normalitás esetén bármely változó felveheti az y szerepét.

5.7.3 A (paraméterekben) lineáris feltételes várható érték függvény paramétereinek becslése

Azt tételezzük fel, hogy egy alapsokaságban $(y, x_1, ...x_k)$ vektorral jellemezhető minden megfigyelt egyed, a mintavétel véletlenszerű, vagyis a megfigyeléseink egymástól függetlenek, és feltételezzük, hogy a célváltozó feltételes várható értéke lineáris a paraméterekben. Tehát

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i,$$

ahol minden ϵ_i 0 várható értékű, kölcsönösen független. Ekkor a feltételes várható érték függvény tulajdonságai miatt igaz az $E(\mathbf{x}_i\epsilon_i) = 0$. (A minta függetlensége miatt $E(\mathbf{x}_j\epsilon_i) = 0, j \neq i$ is igaz.) Nem szükségképpen teljesül azonban az, hogy

$$E(\epsilon_i^2 \mid x_i) = \sigma^2,$$

amit gyakran szintén felteszünk (homoszkedaszticitás).

A modellben lehetnek elfajult valószínűségi változók (konstans) is. A linearitás a paraméterekre vonatkozik. Elképzelhető például, hogy $x_3 = x_2^2$, ettől még a regressziós modell lineáris marad.

A momentumok módszerének elve A moentumok módszere a következő becslési elv: válasszuk meg a becsülendő paramétereket úgy, hogy az összes regresszor empirikus kovarianciája az eltérés vektorral legyen 0.

Legyenek a megfigyelések rendre $(y_i, \mathbf{x}_i)$ párok. Ekkor a momentumok módszeréből adódó egyenlet:

$$\sum_i (y_i - \sum_j x_{ij} b_j) x_{ik} = 0, \forall k$$

azaz mátrix alakban:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$$

Ebből

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

azaz a momentumok módszerével adott becslés ugyanaz, mint az OLS becslés.

Milyen eredmények érvényesek sztochasztikus regresszorokra? A torzítatlanság teljesül iid minta esetén, mivel ehhez az $E(\epsilon \mid \mathbf{X}) = \mathbf{0}$, feltételnek kell fennállnia, ami teljesül a feltételes várható érték függvényre.

Igaz marad a konzisztencia:

$$p \lim_{n\to\infty} \mathbf{b}_n^{ols} = \boldsymbol{\beta}.$$

Érvényes a Gauss-Markov Tétel**.** (Bármilyen X-re teljesül a hatásosság, tehát várható értékben is.)

A **t** és **F** tesztek **X** feltétel mellett érvényesek, illetve akkor feltétel nélkül is, ha az ϵ együttes eloszlása normális.

A $\mathbf{b}^{ols}$ aszimptotikus variancia és kovariancia mátrixa:

$$\sqrt{n}(\mathbf{b}_n^{ols}-\boldsymbol{\beta}) \sim asyN(\mathbf{0}, E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}\mathbf{x}'\varepsilon^2)E(\mathbf{x}\mathbf{x}')^{-1}),$$

vagyis a $\sqrt{n}(\mathbf{b}_n^{ols}-\boldsymbol{\beta})$ valószínűségi változók sorozata konvergál egy 0 várható értékű vektor normális eloszláshoz, amelynek a kovariancia mátrixa $\sigma^2 E(\mathbf{x}\mathbf{x}')^{-1})$-re egyszerűsödik, ha $var(\varepsilon \mid X) = \sigma^2$ minden **X**-re (homoszkedaszticitás) teljesül.

A kihagyott változó probléma Legyenek

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ x_2 &= \gamma x_1 + \epsilon_\gamma \end{aligned}$$

két projekció. Ekkor

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2(\gamma x_1 + \epsilon_\gamma) + \epsilon \\ &= (\beta_1 + \beta_2\gamma)x_1 + \beta_2\epsilon_\gamma + \epsilon. \end{aligned}$$

Mivel $E(x_1\epsilon_\gamma) = E(x_1\epsilon) = 0$ a projekció definíciója miatt, ezért $E(x_1(\gamma\epsilon_\gamma + \epsilon)) = 0$, és

$$y = \beta_{1s}x_1 + \epsilon_s$$

szintén projekció, ahol $\epsilon_s = \beta_2\epsilon_\gamma + \epsilon$, és $\beta_{1s} = \beta_1 + \beta_2\gamma$.

Ez a formula igaz a megfelelő becsült OLS együtthatókra is. Tekintsünk egy kétváltozós regressziót:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,$$

ahol feltesszük, hogy itt a feltételes várható érték függvény és a projekció gybeesik.

Ekkor

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix}.$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}.$$

$$b_1 = \frac{1}{\sum x_{1i}^2 - \sum x_{2i}^2}\left[\sum x_{2i}^2 \sum x_{1i}y_i - \sum x_{1i}x_{2i} \sum x_{2i}y_i\right]$$

$$b_2 = \frac{1}{\sum x_{1i}^2 - \sum x_{2i}^2}\left[\sum x_{1i}^2 \sum x_{2i}y_i - \sum x_{1i}x_{2i} \sum x_{1i}y_i\right]$$

Tekintsük a

$$y_i = \beta_{1s}x_{1i} + \epsilon_{si}$$

regressziót, ahol "kihagyjuk" x_2-t. Ekkor

$$b_{1s} = \frac{\sum x_{1i}y_i}{\sum x_{1i}^2}.$$

Tekintsük most a

$$x_{2i} = \gamma x_{1i} + \epsilon_{\gamma i}$$

regressziót (projekciót). Nyilván

$$\gamma = \frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2}.$$

A fentiekből kiszámolható, hogy

$$b_{1s} = b_1 + \gamma b_2.$$

Vagyis a teljes hatást (b_{1s}) megkaphatjuk a "saját hatás" (a b_1 paraméter) és a "közvetett hatás" γb_2 összegeként. Ezt az állítást általánosíthatjuk.

A regressziós modell partícionált formában

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

$$\mathbf{y} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\epsilon}$$

ahol $\mathbf{X}_1$ nxk_1 es és $\mathbf{X}_2$ $nx(k-k_1)$-es.
A normálegyenleteket is írjuk fel partícionált alakban:

$$\begin{matrix}\mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y}\end{matrix} = \begin{matrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{b}_1 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \mathbf{b}_2\end{matrix}.$$

Belátható, hogy

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).$$

Vezessük be a

$$\mathbf{b}_1^s = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

jelölést. Ekkor

$$\mathbf{b}_1 = \mathbf{b}_1^s - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2.$$

(Ezt szokás "kihagyott változó formulának" is nevezni.)

Vagyis az első változócsoporthoz tartozó paraméterek értékét megkapjuk, ha vesszük a paraméterek értékét egy "rövid regresszióban" (ahol csak az X_1 változók a regresszorok), és ezekből kivonjuk a "kihagyott változók" (az X_2 változók) hatását ($\mathbf{b}_2$) szorozva a bennhagyott változók hatásával a kihagyott változókra (ezek is regressziós paraméterek).

Látható, hogy amennyiben a két változó csoport korrelálatlan ($\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$), akkor a rövid és a teljes becslésből származó paraméterek megegyeznek

Legyen

$$\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$$

Ekkor az

$$\begin{aligned} \mathbf{X}_2^1 &= \mathbf{M}_1\mathbf{X}_2 \\ \mathbf{y}^1 &= \mathbf{M}_1\mathbf{y} \end{aligned}$$

vektorok az eltérés (regressziós reziduum) vektorok.

Belátható, hogy

$$\mathbf{b}_2 = (\mathbf{X}_2^{1\prime}\mathbf{X}_2^1)^{-1}\mathbf{X}_2^{1\prime}\mathbf{y}^1.$$

Szavakban: a 2-es változócsoport regressziós koefficienseit megkapjuk, ha

(1) Készítünk egy regressziós becslést az 1-es változókkal, mint regresszorokkal a 2-s változók mindegyikére, és ezeknek a regresszióknak meghatározzuk a reziduumait.

(2) Készítünk egy regressziós becslést az 1-es változókkal, mint regresszorokkal az y-ra, és ennek a regressziónak meghatározzuk a reziduumait.

(2) Ezután számolunk egy regressziót, ahol a baloldalon az y reziduumai, a jobboldalon pedig a 2-es változó csoport reziduumai vannak.

Igazolható, hogy

$$\mathbf{b}_2 = (\mathbf{X}_2^{1\prime}\mathbf{X}_2^1)^{-1}\mathbf{X}_2^{1\prime}\mathbf{y},$$

vagyis az $\mathbf{y}^1$ regressziós reziduum helyett használhatjuk az eredeti $\mathbf{y}$ vektort is.

Ebből az állításból következik, hogy amennyiben a 2-es csoportba csak valamely x_l tartozik, akkor

$$b_l = \frac{empcov(\widetilde{x_{-l}}, \widetilde{y_{-l}})}{empvar(\widetilde{x_{-l}})} = \frac{empcov(\widetilde{x_{-l}}, y)}{empvar(\widetilde{x_{-l}})}$$

ahol

$$\begin{aligned}\widetilde{x_{-l}} &= M_{-l}x_l \\ \widetilde{y_{-l}} &= M_{-l}y\end{aligned}$$

ahol M_{-l} az x_l elagyása utáni projekciós mátrix.

Ezeket a kovarianciákat és varianciákat empirikus parciális (ko)varianciának nevezzük, és a megfelelő korrelációs együtthatót empirikus parciális korrelációnak. (Vigyázat: a két változó közötti parciális kapcsolat függ attól, hogy milyen hatásokat "szűrünk ki".) Világos, hogy a parciális korreláció megegyezik a szokásos korrelációval, ha a szűrő változókkal nincs korreláció.

Példa:

$$\begin{aligned}\widehat{y} &= b_1x_1 + b_2x_2 \\ \widehat{x_2} &= b_{12}x_1 \\ \widehat{y_s} &= b_1^s x_1 \\ b_1^s &= b_1 + b_{12}b_2\end{aligned}$$

Torzítás kihagyott változó miatt Lásd a kihagyott változó formulát fent:

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).$$

Ebből

$$\mathbf{b}_1^s = \mathbf{b}_1 + \mathbf{B}_{12}\mathbf{b}_2$$

ahol

$$\mathbf{b}_1^s = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

$$\mathbf{B}_{12} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2.$$

Ha a változók kihagyása specifikációs hiba, vagyis nem ilyen alakú a feltételes várható érték függvény, akkor az OLS "torzít", mivel

$$\mathbf{E}(\mathbf{b}_1) \neq \mathbf{E}(\mathbf{b}_1^s),$$

kivéve, ha $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. Torzításról itt akkor beszélhetünk, ha a "hosszú" regresszió paramétereire vagyunk kíváncsiak. A $\mathbf{b}_1^s$ becslés torzított lesz ezekre nézve.

Specifikációs hiba Gyakran abban az értelemben használják a specifikációs hiba kifejezést, hogy valamilyen "érdekes" paraméter helyett egy "nem-érdekes" paraméter konzisztens becslését kapjuk OLS-sel. Például valaki a hosszú regresszió paramétereit szeretné megkapni, de a rövid regresszióból más paraméterre kap konzisztens becslést. Ilyenkor is gyakran "specifikációs hibáról" beszélnek. Specifikációs hibának tekintik gyakran azt is, amikor a modellben olyan regresszor van, amely paramétere a feltételes várható érték függvényben 0.

A specifikáció analízis egy alapvető eszköze a Ramsey-féle RESET teszt, ami egy olyan Wald F teszt, ahol az eredeti regressziót tekintjük korlátozott modellnek, és az általános modell tartalmazza a korlátozott modell becsült értékeinek (második, harmadik és negyedik) hatványait, mint regresszorokat. Vagyis a tesztstatisztika itt:

$$\frac{ESS_R - ESS_U}{ESS_U} \frac{n-(3+K)}{3},$$

mivel a regresszióba bevont új regresszorok száma 3. A teszt arra lehet alkalmas, hogy a nullhipotézis elvetése jelzi, hogy a specifikáció rossz, ám nem ad útmutatást arra, hogy mi a jó specifikáció.

Fölösleges változók Ismét tekintsük a kihagyott változó formulát. Ebből levezethető, hogy

$$\mathbf{Var}(\mathbf{b}_1^s) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$$

$$\mathbf{Var}(\mathbf{b}_1) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}.$$

Amiből

$$\mathbf{Var}(\mathbf{b}_1^s)^{-1} - \mathbf{Var}(\mathbf{b}_1)^{-1} = \sigma^2(\mathbf{X}_1'\mathbf{N}_2\mathbf{X}_1)$$

pozitív definit, ahol

$$\begin{aligned} \mathbf{N}_2 &= \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \\ \mathbf{M}_2 &= \mathbf{I} - \mathbf{N}_2. \end{aligned}$$

Vagyis a "rövid" regresszió kovariancia mátrixa mindig "kisebb", mint a "hosszú" regresszió kovariancia mátrixa. Tehát fölösleges változó bevétele azért probléma, mert rontja a becslés hatásosságát.

5.8 Három általános tesztelési elv

A lineáris regressziós modellben a normalitás feltétele mellett láttuk hogyan működnek az F és t tesztek. Létezik három általános tesztelési elv, amely alapján nagy mintás (aszimptotikus) teszteket készíthetünk paraméter restrikciókra általánosabb feltételek mellett is.

Ehhez vezessük be a Fisher-féle infomáció fogalmát.

$$\mathbf{F}(\mathbf{t}) = [-\mathbf{E}(\frac{\partial \log L}{\partial t_i \partial t_j})],$$

az úgynevezett Fisher-féle információs mátrix.
Például a normális lineáris regressziós modellben a likelihood függvény

$$L(\mathbf{y}, \mathbf{X}; b, \mathbf{s}^2) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_i'\boldsymbol{\beta})^2}{2\sigma^2}),$$

tehát a log-likelihood függvény:

$$\ln L = -n \ln \sigma - n \ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum (y_i - \mathbf{X}_i'\beta)^2.$$

Ha $\mathbf{b} = \mathbf{b}_{ML}$, akkor

$$\mathbf{F}(\mathbf{b}) = \mathbf{E}(\frac{X'X}{s^2}).$$

és

$$cov(asy(\mathbf{b}_n)) = \mathbf{F}(\mathbf{b})^{-1}.$$

Tekintsük most az általános (nem feltétlenül lineáris) paraméter restrikciós problémát, ahol a korlátozó feltétel:

$$\mathbf{R}(\boldsymbol{\beta}) = \mathbf{r}.$$

Itt $\mathbf{R}$ egy J komponensű függvény, tehát $\mathbf{r}$ egy J elemű vektor.
Legyen

$$\mathbf{R} = [\frac{\partial R_j}{\partial \beta_k}].$$

5.8.1 LR, mint tesztelési elv

A korlátozott és korlátozatlan becslés távolságát mérjük a loglikelihood függvény különbségének kétszeresében.

$$\xi_{LR} = -2(\log L(b_U) - \log L(b_R))$$

5.8.2 Wald, mint tesztelési elv

A korlátozott és korlátozatlan becslés távolságát mérjük a paramétervektorok távolságában.

$$\xi_W = (R(b_U) - r)' R_{b_u} F_{b_U} R'_{b_u} (R(b_U) - r),$$

5.8.3 LM, mint tesztelési elv

A korlátozott és korlátozatlan becslés távolságát mérjük a loglikelihood függvény deriváltjainak 0-tól vett távolságában. ("Score" tesztnek is szokás nevezni az LM tesztet.) A korlátozott becslésre igaz, hogy

$$\frac{\partial \log L}{\partial \mathbf{b}_R} - \mathbf{R}'_R \boldsymbol{\lambda} = 0$$

Az LM teszt statisztika

$$\xi_{LM} = \boldsymbol{\lambda}' R_R F_{b_R}^{-1} R'_R \boldsymbol{\lambda} = [\frac{\partial \log L}{\partial \mathbf{b}_R}]' F_{b_R}^{-1} [\frac{\partial \log L}{\partial \mathbf{b}_R}].$$

1. Ez a három tesztstatisztika aszimptotikusan ekvivalens és χ^2_J eloszlású.

Bizonyítás helyett érjük be egyfajta magyarázattal.

Legyen általában egy $F : R^n \to R$ differenciálható függvény, és $\boldsymbol{\Delta}\mathbf{F}_{x_0} = \mathbf{0}$. Ekkor

$$F(x_1) - F(x_o) \cong \frac{1}{2}(x_1 - x_0)' H_{F_{x_o}} (x_1 - x_0).$$

ahol H az F Hesse-mátrixa. Következésképpen

$$2(F(x_1) - F(x_o)) \cong (x_1 - x_0)' H_{F_{x_o}} (x_1 - x_0).$$

Ez "indokolja" az LR és W statisztikák aszimptotikus ekvivalenciáját, ha F-et a loglikelihood függvénnyel helyettesítjük, x_0 a (korlátozatlan) ML becslés, és x_1 a korlátozott ML becslés.

Másrészről

$$\begin{aligned}
\boldsymbol{\Delta}\mathbf{F}_{x_0} - \boldsymbol{\Delta}\mathbf{F}_{x_1} &\cong H_{F_{x_1}}(x_1 - x_0) \\
-H_{F_{x_1}}^{-\frac{1}{2}} \boldsymbol{\Delta}\mathbf{F}_{x_1} &\cong H_{F_{x_1}}^{\frac{1}{2}}(x_1 - x_0) \\
\boldsymbol{\Delta}\mathbf{F}'_{x_1} H_{F_{x_1}}^{-1} \boldsymbol{\Delta}\mathbf{F}_{x_1} &\cong (x_1 - x_0)' H_{F_{x_1}}(x_1 - x_0),
\end{aligned}$$

ami "indokolja" azt, hogy az LM asziptotikusan ekvivalens a másik kettővel.

2. A ξ_W-ből a szabadságfokokkal igazítva kapjuk a szokásos (Wald) F tesztet.

3. ξ_{LM} megkapható abból a segédregresszióból, ahol a célváltozó a korlátozott modell reziduum vektora, és a regresszorok az általános modell regresszorai. Ha R_a^2 a segédregresszióból számított determinációs együttható, akkor

$$\xi_{LM} = nR_a^2.$$

4. Belátható, hogy a többszörös lineáris regresszió speciális esetében, amikor a $\boldsymbol{\beta}$ paramétervektorra vonatkozó lineáris restrikciókat tesztelünk:

$$\begin{aligned}\xi_{LR} &= n\log(\frac{ESS_R}{ESS_U}) \\ \xi_W &= n\frac{ESS_R - ESS_U}{ESS_U} \\ \xi_{LM} &= n\frac{ESS_R - ESS_U}{ESS_R}.\end{aligned}$$

Ebből a következő nagyságrendi sorrend adódik:

$$\xi_{LM} \leq \xi_{LR} \leq \xi_W.$$

5.9 Gyakorlatok R-ben

5.9.1 Írjuk meg saját OLS becslésünket

#A következő programrész létrehoz egy 2000 elemű mesterséges mintát, ami négy változót tartalmaz.

```
n = 2000
x1 = rnorm(n,164,10)
x2= rnorm(n,0,10)
x3=rf(n,5,20)
y = x1+2*x2-10*x3+ rnorm(n,0,4)
adatok = data.frame(x1,x2,x3,y)
# a regresszorok mátrixa
X = cbind(1, adatok$x1,adatok$x2,adatok$x3)
# az X mátrixa transzponálása
XT=t(X)
# a transzponált és az eredeti mátrix szorzása (%*%= mátrix szorzás)
XTX=XT%*%X
# az OLS együttthatók előállítása (solve= invertálás)
(b=solve(XTX)%*%XT%*%y)
# a reziduum vektor előállítása
u=y-X%*%b
# a reziduumok négyzetösszegének kiszámolása
(ssr=sum(u*u))
# a regresszió standard hibája
(s=sqrt(ssr/(n-4)))
# az OLS együtthatók kovariancia mátrixának becslése
(cov=s*s* solve(XTX))
# a kovariancia matrix diagonálisa: a varianciák vektora
(std=sqrt(diag(cov)))
# a szokásos t-statisztikák
(t=b/std)
# A modell OLS becslése R utasítással
ols=lm(y ~adatok$x1+adatok$x2+adatok$x3)
```

summary(ols)
# az együttható becslések összevetése
b-ols$coeff
a kovariancia matrix becslések összevetése
cov-vcov(ols)
a standard hiba becslések összevetése
s-summary(ols)$sigma
a t-statisztikák összevetése
t-summary(ols)$coeff [,3]

5.9.2 OLS az R-ben

library (MASS)
attach (Boston)
# két változó együttes ábrája
plot(lstat, medv)
# egy egyszerű regresszió
ols0=lm(medv~lstat)
summary (ols0)
néhány többváltozós regresszió
(ols1=lm(medv~lstat+age))
(ols2 =lm(medv~.,data=Boston))
ols3=lm(medv~.-age,data=Boston)
summary (ols)
Értelmezzük ezt az outputot!
Call:
lm(formula = y ~adatok$x1 + adatok$x2 + adatok$x3)
Residuals:
Min 1Q Median 3Q Max
-13.673 -2.730 -0.025 2.837 16.009
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.067382 1.471864 0.725 0.468
adatok$x1 0.992008 0.008907 111.378 <2e-16 ***
adatok$x2 1.998966 0.008926 223.943 <2e-16 ***
adatok$x3 -9.780029 0.105554 -92.654 <2e-16 ***
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.997 on 1996 degrees of freedom
Multiple R-squared: 0.9728, Adjusted R-squared: 0.9728
F-statistic: 2.379e+04 on 3 and 1996 DF, p-value: < 2.2e-16

Konfidencia intervallum számítása
confint (ols3)
mi mit jelent az outputban?
2.5 % 97.5 %

(Intercept)	26.45557192	46.418281372
crim	-0.17251263	-0.043498576
zn	0.01958627	0.073081052
indus	-0.10014145	0.141265800
chas	1.00009928	4.377953122
nox	-24.94259876	-10.484480962
rm	3.01181752	4.616969603
dis	-1.85312219	-1.104100922
rad	0.17593586	0.435636023
tax	-0.01970656	-0.004950825
ptratio	-1.20821163	-0.696210713
black	0.00405936	0.014581947
lstat	-0.61742541	-0.430278274

5.9.3 Konfidencia régió, modellek összehasonlítása, tesztek több paraméterre

```
library (MASS)
    attach (Boston)
    ols1=lm(medv~dis+rad+tax)
    summary (ols1)
    # F konfidencia tartomány: azt kérdezzük, hogy a becsült paramétervektor
1,01-szerese benne van-e a 95 %-os konfidencia tartományban
    vcov(ols1)
    coef(ols1)
    bettest=1.01*coef(ols1)
    fconf=(t(coef(ols1))-t(bettest))%*%solve(vcov(ols1))%*%(coef(ols1)-bettest)*(1/ols1$rank)
    fconf-qf(0.95,ols1$rank, ols1$df)
    Feladat: Igaz-e, hogy a következő paramétervektor benne van a 10 %-os
konfidencia intervallumban?
    Konstans: ugyanaz, mint az OLS
    Dis együtthatója: 5%-nal nagyobb, mint az OLS együttható
    Rad: együtthatója 99 %-a az OLS együtthatónak
    Tax: együtthatója ugyanaz, mint az OLS együttható
    bettest1=c(coef(ols1)[1],coef(ols1)[2]*1.05, coef(ols1)[3]*0.99, coef(ols1)[4])

    # Egy ANOVA tábla két modell összevetésére
    (ols2 =lm(medv~.,data=Boston))
    ols3=lm(medv~.-age,data=Boston )
    anova(ols2,ols3)
    #Értelmezze az outputot!
    Analysis of Variance Table
    Model 1: medv ~crim + zn + indus + chas + nox + rm + age + dis + rad
+
    tax + ptratio + black + lstat
    Model 2: medv ~(crim + zn + indus + chas + nox + rm + age + dis + rad
+
```

tax + ptratio + black + lstat) - age
Res.Df RSS Df Sum of Sq F Pr(>F)
1 492 11079
2 493 11079 -1 0.061834 0.0027 0.9582

Elegáns F próba az Rb=r nullhipotézisre
Nullhipotézis:
25*b2-b3=0, b2-b4=0
r0=c(0,0)
r=as.matrix(r0)
R=rbind(x=c(0,25,-1,0),y=c(0,1,0,-1))
covmat=R%*%vcov(ols1)%*%t(R) # a korlátozott kovariancia mátrix kiszámolása
(ftest=t(R%*%coef(ols1)-r)%*%solve(covmat)%*%(R%*%coef(ols1)-r)*(1/length(r0))) # a tesztstatisztika kiszámolása
Miért nem jelenik meg a formulában az n?
ftest-qf(0.95, length(r0),ols1$df) # ha negatív, akkor a nullhipotézist elfogadjuk
Határozza meg a p-értéket!
(pval=1-pf(ftest, length(r0),ols1$df))
Tesztelje csak azt a hipotézist, hogy dis és tax együtthatója ugyanaz!
# Átparametrizált F teszt ugyanarra a nullhipotézisre
# a nullhipotézisből:
b2=b4
b3=25*b2
Az új egyenlet:
y=b1+b4*x2+25*b4*x3+b4*x4+e)
y=b1+b4(x2*25*x3+x4)+e
Új változó:
z=x2+25*x3+x4
z=dis+25*rad+tax
Az új tesztegyenlet:
ols2=lm(medv~z)
anova (ols2,ols1)
Analysis of Variance Table
Model 1: medv ~z
Model 2: medv ~dis + rad + tax
Res.Df RSS Df Sum of Sq F Pr(>F)
1 504 34855
2 502 32838 2 2017.4 15.42 3.167e-07 ***
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Hasonlítsa össze az eredményt az előző próbával!
Közvetett t próba
most csak a b2=b4 hipotézist teszteljük
Az új egyenlet:

y=b1+b4*x2+b3*x3+b4*x4+e
y=b1+b4*(x2+x4)*b3*x3+e)
Az új változó:
z1=x2+x4
y=b1+b4*z1+b3*x3+e
z1=dis+tax
Az új tesztegyenlet:
ols3=lm(medv~z1+rad) # a korlátozott modell becslése
summary(ols3) # elfogadjuk a nullhipotézist?
Call:
lm(formula = medv ~z1 + rad)
Residuals:
Min 1Q Median 3Q Max
-13.572 -4.884 -1.963 3.283 33.387
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.876740	1.376785	26.058	< 2e-16 ***
z1	-0.038765	0.005182	-7.480	3.33e-13 ***
rad	0.275235	0.099640	2.762	0.00595 **

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.08 on 503 degrees of freedom
Multiple R-squared: 0.2312, Adjusted R-squared: 0.2281
F-statistic: 75.62 on 2 and 503 DF, p-value: < 2.2e-16
# Az előző próba próba F-teszttel
ols4=lm(medv~z1+rad)
anova(ols1,ols4)
# Megegyezik az eredmény? Igaz-e, hogy t*t=F? Ugyanazok a p-értékek?
# Közvetlen t-próba
# most a b3=25*b2 hipotézist teszteljük
(tteszt=(coef(ols1)[[3]]-25*coef(ols1)[[2]])/sqrt(vcov(ols1)[[3,3]]+25*25*vcov(ols1)[[2,2]]-2*25*vcov(ols1)[[2,3]]))
# a t statisztika kiszámítása
abs(tteszt)-qt(0.975,df=ols1$df)
# ha ez negatív, akkor a nullhipotézist elfogadjuk 5 %-os szinten
Számítsa ki a p-értéket!
Számítsa ki a b2=b4 nullhipotézisre is a közvetlen t-tesztet!
(pvalue=2*(1-(pt(abs(tteszt),502))))
(pvalue=2*(pt(-abs(tteszt),502)))

# Ellenőrizzük, hogy a közvetett t próba ugyanezt az eredményt adná-e!
d= b3-25*b2
b3=d+25*b2
y=b1+b2*x2+(d+25*b2)*x3+b4*x4+e
y=b1+b2*(x2+25*x3)+d*x3+b4*x4+e
z2=x2+25*x3

y=b1+b2*z2+d*x3+b4*x4+e
z2=dis+25*rad
ols5=lm(medv~z2+rad+tax)
summary(ols5) # d a rad együtthatója
Ugyanaz, mint a közvetlen t-próba p-értéke?

5.9.4 Az OLS becslés diagnosztikái

# Studentizált reiziduumok és outlier-ek
require(MASS)
stres=studres(ols3)
plot (stres) #Mit tekinthetünk outlier-nek?

#High-leverage megfigyelések
lm.influence(ols1)
plot(hatvalues (ols1))
which.max (hatvalues (ols1)) # a legnagyobb hatású megfigyelés

#Normalitás tesztek
qqnorm(ols3$residuals) # a kvantilis - kvantilis ábra
require(tseries)
jarque.bera.test(ols3$residuals) # A nullhipotézis a hibatagok normalitása, mikor fogadjuk el?

Modellválasztás információs krltériummal
AIC(ols1,ols3) # Akaike
BIC(ols1,ols3) # bayes-i
Melyik a jobb modell?

#Specifikációs tesztek
require(lmtest)
resettest (ols4) # A Ramsey-féle RESET teszt. MIkor fogadjuk el a specifikációt, ami a nullhipotézis?
ols3=lm(medv~.-rad,data=Boston)
encomptest (ol4,ol6) # Melyik modellt fogadjuk el?

Előrejelzési és konfidencia intervallumok
predict(ols,interval="predict") # a mintaadatok előrejelzési intervallumai
predict(ols,interval="confidence") # a mintaadatok konfidencia intervallumai
# Mi a különbség a két intervallum között?
uj=data.frame(Catholic=0.5,Education=0.2, Agriculture=0.1) # egy új előrejelzéshez "prediktorok"
predict(ols,interval="predict",newdata=uj) # új adatokhoz predikciós intervallum
predict(ols,interval="con.dence",newdata=uj) #új adatokhoz konfidencia (átlagos előrejelzési) intervallum

5.9.5 Kihagyott változó formula

Lássuk be tapasztalatilag, hogy például "Examination " együtthatója a "hosszú" regresszióban megkapható a kihagyott változó formulából ("Education" a kihagyott változó)!

attach(swiss)
ols=lm(Fertility~Education+Examination+Agriculture)
coef(ols)[3] # Examination együtthatója a hosszú regresszióban
coef(ols)[2] # Education regressziója a hosszú regresszióban
olsmined=lm(Fertility~Examination+Agriculture) # a rövid regresszió
coef(olsmined)[2] # Examination együtthatója a rövid regresszióban
olsed=lm(Education~Examination+Agriculture) # a kihagyott változót "magyarázó" regresszió
coef(olsed)[2] # Examination együtthatója a kihagyott változót "magyarázó" regresszióban
coef(olsmined)[2]- coef(ols)[3]- coef(ols)[2]* coef(olsed)[2] # 0-nak kell lennie
Számolja ki ugyanezt "Agriculture" együtthatójával is!

LM, LR és Wald-tesztek attach(swiss)

#A Lagrange-multiplikátor vagy score teszt
olh=lm(Fertility~Examination+Education+Catholic+Agriculture)
olr=lm(Fertility~Examination+Education)
an=anova(olr,olh) # az anova output megadja a szükséges reziduális négyzetösszegeket
sc=as.matrix(an$RSS) # ezeket mátrixként definiáljuk
(scoretest=47*(sc[1,1]-sc[2,1])/sc[1,1]) # kiszámítjuk az LM statisztikát
# LM másodszor
olaux=lm(olr$residuals~Examination + Education + Catholic + Agriculure) # kisegítő regresszió
(lmutest=47*summary(olaux)$r.squared) # a kisegítő regresszió R-négyzete értékéből számoljuk az LM statisztikát.
Van különbség?

#LR teszt
(lratio=47*(log(sc[1,1])-log(sc[2,1]))) # az LR érték
lrtest(olh,olr) # ugyanez a "lmtest" package utasításával

#Wald teszt
(wchi=47*(sc[1,1]-sc[2,1])/sc[2,1]) # Wald khi-négyzet teszt
# Mit tapasztal a három tesztstatisztika egymáshoz viszonízott értékeiről?
# a khi2 érték
qchisq(0.95,2)
waldtest(olr,olh) # Wald-teszt az "lmtest"-ben
anova(olr,olh) # mit tapasztal?

6 Heteroszkedaszticitás

A homoszkedaszticitási feltevés:

$$E(\epsilon_i^2 \mid X) = \sigma^2,$$

minden i-re nem következik a feltételes várható érték létezéséből, hanem "extra" feltevés. Az OLS becslés milyen tulajdonságai maradnak igazak, ha nincs homoszkedaszticitás?

Állítás: A torzítatlanság és konzisztencia megmarad, viszont az OLS paraméter becslés nem hatásos. Bizonyítás: Az OLS becslés torzítatlansága és konzisztenciája bizonyításánál nem használjuk ki a homoszkedaszticitási feltételt.

A hatásosság hiányát megérthetjük, ha a normalitási feltevés mellett tekintjük a regressziós modell ML becslését nem homoszkedasztikus hibákkal. .

Ekkor a loglikelihood függvény:

$$\ln L = -n \ln \det(diag \langle \sigma_i \rangle) - n \ln(\sqrt{2\pi}) - \frac{1}{2} \sum_{i=}^{n} (\frac{y_i - \mathbf{X}_i' \beta}{\sigma_i^2})^2,$$

illetve az ebből levezetett normálegyenletek:

$$\begin{aligned} \sum_{i=1}^{n} x_{ij} (\frac{y_i - \sum_{l=1}^{k} \beta_l x_{il}}{\sigma_i^2}) &= 0, \forall j, j = 1, ...k, \\ \sum_{i=1}^{n} \frac{x_{ij}}{\sigma_i} (\frac{y_i}{\sigma_i} - \sum_{l=1}^{k} \beta_l \frac{x_{il}}{\sigma_i}) &= 0, \forall j, j = 1, ...k. \end{aligned}$$

Mátrix alakban felírva

$$\mathbf{X}'((\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \langle \sigma_1, ...\sigma_n \rangle^{-1}) = 0.$$

Ha $\sigma_i = \sigma$ (homoszkedaszticitás), akkor a megoldás nyilván független σ-tól. Egyébként pedig úgy kapjuk meg, mintha az eredeti megfigyeléseket a megfelelő σ_i-vel osztva "új" megfigyeléseket tennénk, és ezekre számolnánk ki a normálegyenletek megoldását.

A heteroszkedaszticitás nem tűnik nagyon súlyos problémának az OLS becslés minősége szempontjából. Viszont komolyabb probléma, hogy az OLS becslésből számított tesztek inkonzisztensek lesznek. Ugyanis:

Állítás: Torzított és inkonzisztens az OLS becslőfüggvény kovariancia mátrixának szokásos becslése, $s^2\mathbf{X}'\mathbf{X})^{-1}$. Az OLS becslés kovariancia mátrixa:

$$\mathbf{var}(\mathbf{b}^{OLS}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})(\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})).$$

Az $E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}) = \sigma^2 E(\mathbf{X}'\mathbf{X})$ egyszerűsítés nem igaz heteroszkedaszticitás esetén. A variancia-kovariancia mátrix becsléséhez meg kell becsülni a $E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X})$ negyedik momentum mátrixot. (Az iid esetben az $\boldsymbol{\epsilon}\boldsymbol{\epsilon}'$ diagonális.)

Legyen

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b}^{OLS},$$

ekkor a negyedik momentum mátrix becslése:

$$\mathbf{S} = \frac{\sum_i \mathbf{x}_i \mathbf{x}_i^{'} u_i^2}{n}.$$

Így a heteroszkedaszticitással korrigált kovariancia mátrix becslése:

$$var(\mathbf{b}^{OLS}) = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{S}(\mathbf{X}^{'}\mathbf{X})^{-1}.$$

Ennek diagonális elemeit lehet használni a t statisztikák mostmár korrekt meghatározásához.

Tehát az OLS becslést lehet használni teszteléshez is, csak a hatásosság marad megoldatlan probléma.

6.1 Heteroszkedaszticitási tesztek

6.1.1 F teszt

Legyen két különböző varianciájú, ám azonos várható értékű független normális sokaság (A és B).

Ekkor

$$\begin{aligned} \frac{s_A^2}{\sigma_A^2} &\sim \chi^2_{n_A-1}, \\ \frac{s_B^2}{\sigma_B^2} &\sim \chi^2_{n_B-1}. \end{aligned}$$

Az $\sigma_A^2 = \sigma_B^2$ nullhipotézis fennállása esetén:

$$\frac{s_A^2}{s_B^2}\frac{n_B-1}{n_A-1} \sim F_{n_A-1,n_B-1}.$$

A számláló és a nevező természetesen megfordítható. Regressziós esetben ennek általánosítása a Goldfeld-Quandt teszt.

6.1.2 LM tesztek

A nullhipotézis a homoszkedaszticitás. A nullhipotézist akkor utasítjuk el, ha találunk olyan változókat, amelyek szignifikánsan "magyarázzák" a maradéktagot. Kell tehát egy specifikus feltevés a varianciáról.

Lépések:

1. OLS becslés a paraméterekre.
2. Reziduumok előállítás.

$$u = y - Xb^{OLS}.$$

3. Varianciát "magyarázó" segédregresszió.

Például a

$$\sigma_i^2 = \alpha + \beta_{1h} Z_{1i} + ... + \beta_{Ph} Z_{Pi}$$

specifikus feltevésnek megfelelő

$$u_i^2 = \alpha + \beta_{1h} Z_{1i} + ... + \beta_{Ph} Z_{Pi} + \eta_i$$

segédregressziót becsülhetjük.

4. A segédregresszió nR_s^2-ének kiszámolása, ami aszimptotikusan χ_P^2, ahol $P+1$ a segédregresszióban becsült paraméterek száma.

Itt is különböző tesztstatisztikák számíthatók ugyanabból a regresszióból! Meg kell különböztetni a tesztstatisztikák eloszlását (pl. t, F, vagy χ^2 próba) a tesztelési elvektől (pl. Wald, LM, LR). A három utóbbi elv használata aszimptotikusan biztosan χ^2 tesztstatisztikához vezet, és nem függ a "kisminta" normalitásától.

6.2 Becslési módszerek heteroszkedaszticitás esetén

6.2.1 Az általánosított legkisebb négyzetek módszere (GLS)

A fenntartott regressziós feltevések:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}.$$

A homoszkedaszticitási feltevés

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}) = \sigma^2 \mathbf{I},$$

ahol I nxn-es mátrix, most nem teljesül. Heteroszkedaszticitás esetén:

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}) = \mathbf{diag}\left\langle \sigma_1^2, ..\sigma_i^2, ...\sigma_n^2 \right\rangle,$$

ahol σ_i^2 függhet az $\mathbf{X}$-től.

Tegyük fel, hogy

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}) = \boldsymbol{\Omega},$$

ami általánosabb, mint a heteroszkedaszticitás esete, ahol

$$\boldsymbol{\Omega} = \mathbf{diag}\left\langle \sigma_1^2, ..\sigma_i^2, ...\sigma_n^2 \right\rangle.$$

Ekkor legyen a transzformált hiba vektor

$$\boldsymbol{\epsilon}^* = \boldsymbol{\Omega}^{-\frac{1}{2}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

(Ez mindig létezik, mivel létezik a $\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\Omega}^{\frac{1}{2}} = \boldsymbol{\Omega}$ felbontás pozitív definit mátrixoknál.)

Nyilvánvalóan

$$E(\boldsymbol{\epsilon}^*\boldsymbol{\epsilon}^{*\prime}) = (\boldsymbol{\Omega}^{-\frac{1}{2}}\epsilon)(\epsilon'\boldsymbol{\Omega}^{-\frac{1}{2}}) = \mathbf{I}.$$

Tekintsük a transzformált

$$\begin{array}{ccc} \boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{y} & = \boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\epsilon} & \\ \mathbf{y}^* & = & \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \end{array}$$

modellt. Ez homoszkedasztikus és OLS becslése:

$$\mathbf{b}^{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y})$$

torzítatlan, konzisztens, és hatásos becslése $\boldsymbol{\beta}$-nak.

Továbbá igaz, hogy

$$var(\mathbf{b}^{GLS}) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}.$$

Tiszta heteroszkedaszticitás esetén Ω diagonális mátrix, és ennek megfelelően az inverze is. A GLS becslés úgy interpretálható, hogy minden megfigyelést a szórásának reciprokával súlyozunk, majd az így transzformált változókra írjuk fel az OLS-t. (Súlyozott legkisebb négyzetek.)

6.2.2 Megvalósítható GLS

Mivel Ω nem ismert, szükség van egy konzisztens $\widehat{\Omega}$ becslésére. Ha ilyennel rendelkezünk, akkor a megvalósítható GLS becslés:

$$\mathbf{b}^{FGLS} = (\mathbf{X}'\widehat{\boldsymbol{\Omega}^{-1}}\mathbf{X})^{-1}(\mathbf{X}'\widehat{\boldsymbol{\Omega}^{-1}}\mathbf{y}).$$

Mint azt a tesztelésnél is tettük, feltehetjük Ω diagonális elemeire, hogy :

$$\sigma_i^2 = f(Z_i).$$

ahol Z_i valamilyen "varianclát magyarázó" vektora, amely változók nem feltétlenül jelennek meg a regresszió magyarázó változói között. Az ismeretlen σ_i^2-ekre vonatkozó becsléseket ugyanúgy végezzük el, mint a tesztelésnél, ahol az OLS reziduumok négyzetét (u_i^2), tekintjük a varianciák "megfigyelésének". Ezután a megvalósítható súlyozott LS-t alkalmazhatjuk az $\widehat{u_i^2}$ közelítő értékek négyzetgyökének reciprokával, mint súlyokkal.

6.3 Gyakorlatok R-ben

6.3.1 Heteroszkedaszticitás: észlelés és tesztelés

```
require (lmtest)
    # Generáljunk változókat
    x1=rnorm(100,1,1)
    x2=1.01*x1+rnorm(100,0,1)
    ex1=exp(x1)
    ex2=exp(x2)
    par(mfrow=c(2,2))
    plot (ex1,ex2)
    # a következő regresszió rosszul specifikált
    ole=lm(ex2~ex1)
    plot (ole$residuals)
    plot (x1,x2)
    # következő regresszió viszint jól specifikált
    oll=lm(x2~x1)
    plot (oll$residuals)
    # A Breusch-Pagan teszt
    bptest(ole)
    bptest(oll)
    #Melyik specifikációt fogadjuk el?
    # A White-teszt
    bptest(ole,~ex1+I(ex1^2))
    bptest(oll,~x1+I(x1^2))
    #Melyik specifikációt fogadjuk el?
```

6.3.2 Heteroszkedaszticitás konzisztens kovariancia mátrix

```
require (sandwich)
    vcov(ole) # a szokásos kovariancia mátrix
    vcovHC(ole) # a heteroszkedaszticitás-konzisztens kovariancia mátrix
    vcov(oll)
    vcovHC(oll)
    #Hol lát különbséget?
```

Heteroszkedaszticitás és becslés require (MASS)

```
# 1. feltevés: a varianciák az ex1 négyzetével arányosak
ex12=ex1^2
Weight=diag(sqrt(ex12))
olehs=lm.gls(ex2~ex1,W=Weight,inverse=T) # 1. becslés
olew=lm(ex2~ex1,weights=1/sqrt(ex12)) # 2. becslés
coef(olehs) #hasonlítsuk össze a koefficienseket
coef(olew)
coef(ole)
summary(olew)
summary(ole)
#Melyik a jobb modell?
AIC(ole,olew)
# Figyelem: a "jobb" modell sem jól specifikált!
# feltevés: a White-tesztnek megfelelő modell írja le a varianciákat a jól specifikált regresszióban
olw=lm(resid(oll)~x1+I(x1^2))
weightw=olw$.tted.values
ollw=lm(x2~x1,weights=1/sqrt(weightw))
summary(oll)
summary(ollw)
#Hasonlítsuk össze a két becslést!
```

7 Instrumentális változók

7.1 Az IV (instrumentális változók) módszer általában

Tegyük fel, hogy fennáll egy

$$y = \boldsymbol{\beta}\mathbf{x} + \varepsilon$$

összefüggés, ahol $E(\mathbf{x}\epsilon) \neq 0$. Szeretnénk $\boldsymbol{\beta}$-t megbecsülni, azaz $\boldsymbol{\beta}$ az **érdekes paraméter**. A feltevésből nyilvánvaló, hogy itt nem az $E(y \mid \mathbf{x})$ feltételes várható érték függvényéről van szó, hanem valami olyasmiről, amit gyakran strukturális összefüggésnek neveznek (pl. egy keresleti függvény).

Ekkor

$$\boldsymbol{\beta} \neq \mathbf{E}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{E}(\mathbf{y}\mathbf{x}').$$

Viszont tegyük fel, hogy létezik egy $\mathbf{z}$ valószínűségi vektor, amely ugyanannyi elemű, mint $\mathbf{x}$, $\mathbf{E}(\mathbf{z}\boldsymbol{\epsilon}) = \mathbf{0}$,és $\mathbf{E}(\mathbf{x}\mathbf{z}')$ nemszinguláris. Tehát

$$\mathbf{E}(\mathbf{y}\mathbf{z}') = \boldsymbol{\beta}\mathbf{E}(\mathbf{x}\mathbf{z}'),$$

vagyis

$$\boldsymbol{\beta} = \mathbf{E}(\mathbf{x}\mathbf{z}')^{-1}\mathbf{E}(\mathbf{y}\mathbf{z}').$$

Ez az elméleti összefüggés. Az ennek megfelelő becslőfüggvény:

$$\mathbf{b}^{iv} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

és belátható, hogy $\mathbf{b}^{iv}$ konzisztens becslése $\boldsymbol{\beta}$-nak általános feltételek ($plim\frac{\mathbf{Z}'\mathbf{X}}{n}$ nemszinguláris, $plim\frac{\mathbf{Z}'Z}{n}$ pozitív definit, $plim\frac{\mathbf{Z}'\varepsilon}{n} = 0$.) mellett.

A $\mathbf{z}$ vektor elemeit instrumentumoknak nevezzük, és a becslést instrumentális változó (IV) becslésnek. Az IV becslés problémája, hogy sok instrumentum létezhet, amik határértékben ugyanazt a becslést adhatják, de vannak köztük hatékonyak és kevésbé hatékonyak.

7.1.1 Indirekt legkisebb négyzetek (IOLS)

Tekintsük az $\mathbf{x}$ vektor regresszióját $\mathbf{z}$-re és az $\mathbf{y}$ regresszióját $\mathbf{z}$-re.

Ekkor

$$\mathbf{B}_{x,z} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

és

$$\mathbf{b}_{y,z} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

Látszik, hogy

$$\mathbf{b}^{iv} = \mathbf{B}_{x,z}^{-1}\mathbf{b}_{y,z},$$

vagyis az IV becslést megkaphatjuk indirekt módon a két OLS becslésből is.

7.1.2 A kétfokozatú legkisebb négyzetek (2SLS)

Legyen most $\mathbf{z}$ nagyobb elemszámú, mint $\mathbf{x}$. Tekintsük $\mathbf{x}$ vetítését $\mathbf{z}$-re:

$$\widehat{X} = \mathbf{B}_{x,z}\mathbf{Z}.$$

Ekkor $\widehat{\mathbf{x}}$ ugyanúgy tekinthető instrumentumnak, mint $\mathbf{z}$, és készíthetünk vele, mint instrumentummal, egy IV becslést. A

$$\mathbf{b}^{2sls} = (\widehat{X}'\mathbf{X})^{-1}\widehat{X}'\mathbf{y}$$

konzisztens becslőfüggvény.
Mivel

$$\begin{aligned}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}) &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \\ \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} &= (\widehat{\mathbf{X}}'\mathbf{X})\end{aligned}$$

ezért

$$\mathbf{b}^{2sls} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\mathbf{y}.$$

Ezért szokás ezt a becslést kétfokozatú OLS-nek nevezni. Az első fokozatban elkészítjük a $\widehat{X} = B_{x,z}X$ OLS becslést, majd a második fokozatban az $\widehat{\mathbf{X}}$ regresszorokkal készítünk OLS becslést y-ra, mint célváltozóra.

7.2 Az IV módszer alkalmazásai

7.2.1 Hiba a változókban modell

Tegyük fel, hogy

$$E(y_i \mid x_i) = \beta x_i$$

de x_i-t csak "zajosan" tudjuk megfigyelni, vagyis csak

$$x_i^* = x_i + u_i.$$

megfigyeléseink vannak, ahol

$$\begin{aligned}E(u_i \mid x_i) &= 0 \\ E(u_i^2) &= \sigma^2 \\ E(y_i u_i) &= 0.\end{aligned}$$

Ekkor

$$\begin{aligned} \beta^z &= \frac{E(x^*y)}{E(x^{*^2})} = \frac{E(xy)}{E(x^2)+\sigma^2} \\ abs(\frac{E(xy)}{E(x^2)+\sigma^2}) &< abs(\frac{E(xy)}{E(x^2)}) = abs(\beta). \end{aligned}$$

Az OLS becslés az elméleti kovarianciákat az empirikus kovarianciákkal helyettesíti.

$$b = \frac{\mathbf{x}^{*\prime}\mathbf{y}}{\mathbf{x}^{*\prime}\mathbf{x}^*} = \frac{\mathbf{x}^{*\prime}\mathbf{y}}{n} : \frac{\mathbf{x}^{*\prime}\mathbf{x}^*}{n}.$$

Tehát a b OLS becslés (0 felé) torzított, sőt inkonzisztens, becslése β-nak, az "érdekes" paraméternek.

Ugyanakkor (például, ha minden fenti változó normális), akkor

$$\begin{aligned} E(y \quad | \quad x^*) &= \beta' x^*, \\ \beta' &= \frac{E(xy)}{E(x^2)+\sigma^2} \end{aligned}$$

és

$$\beta' = E(b).$$

Tehát "jó" becslést kapunk a feltételes várható érték függvényre a lineáris projekció becsléséből. A probléma az, hogy nem β', hanem β az érdekes paraméter. Mit tehetünk, ha β-ra vagyunk kíváncsiak?

Megoldás, ha találunk olyan z megfigyelhető változót (instrumentumot), amelyre

$$\begin{aligned} E(zx) &\neq 0 \\ E(zu) &= 0 \\ E(z\epsilon) &= 0. \end{aligned}$$

Ekkor:

$$\begin{aligned} E(zy) &= \beta E(zx^*) \\ \beta &= \frac{E(zy)}{E(zx^*)}. \end{aligned}$$

Azaz mintából becsülve β-t:

$$b_z = \frac{\mathbf{z}'\mathbf{y}}{\mathbf{z}'\mathbf{x}}.$$

A fentiek alapján b_z konzisztens becslése β-nak.

$$p\lim b_z = \beta.$$

7.2.2 Béregyenlet: klasszikus mikrökonometriai probléma

Legyen y_i az i-edik egyed bére és x_i az oktatási "inputja", például az oktatási rendszerben töltött ideje. Tegyük fel, hogy érvényes a következő reláció:

$$\begin{aligned} y_i &= \beta x_i + \epsilon_i, \\ E(x\epsilon) &\neq 0. \end{aligned}$$

A feltevés az, hogy a bér függ nem-megfigyelhető pszichológiai és egyéb tényezőktől, amelyeket nem tudunk megfigyelni, és amik "benne vannak" ϵ-ban. Instrumentumnak (z) kell valami, ami korrelál az oktatási inputtal, de nem korrelál a fenti pszichológiai tényezőkkel, vagy bármi mással, ami befolyásolhatja a bért, és "benne van" ϵ-ban.

Lehetséges instrumentum például a lakhely távolsága az iskolától vagy a születési dátum az éven belül. Ugyanis létezik szignifikánsan $0-$tól különböző korreláció ezen ismérvek és az iskolában töltött évek száma között, ugyanakkor ezek vélhetően olyasmik, amiknek nem lenne helye a béregyenletben (az együtthatójuk 0 lenne), és nem is korrelálnak olyan változókkal, amiknek helyük van ott. Tehát feltehetjük, hogy

$$\begin{aligned} E(zx) &\neq 0, \\ E(z\epsilon) &= 0. \end{aligned}$$

Ebből következik, hogy

$$\beta = \frac{E(zy)}{E(zx)}.$$

A fenti elméleti összefüggésnek megfelelő becslés:

$$\begin{aligned} b_z &= \frac{z'\mathbf{y}}{z'x}, \\ p\lim b_z &= \beta. \end{aligned}$$

Egy másik (indirekt) út is ugyanehhez az eredményhez vezet. Legyen

$$\widehat{x} = \frac{E(zx)}{E(z^2)} z = \gamma z.$$

vagyis az $proj(x/z)$ lineáris projekció eredménye. Látszik, hogy a $proj(y/\widehat{x})$ projekció együtthatója:

$$\frac{E(y\widehat{x})}{E(\widehat{x}^2)} = \frac{\gamma E(zy)}{\gamma^2 E(z^2)} = \frac{E(zy)}{E(zx)} = \beta.$$

Vagyis becslésként ez nem más, mint a kétfokozatú LS:

$$\begin{aligned}\widehat{x} &= \frac{\mathbf{z'x}}{\mathbf{z'z}}x \\ \frac{\widehat{x}'y}{\widehat{x}'\widehat{x}} &= \frac{\mathbf{z'y}}{\mathbf{z'z}} : \frac{\mathbf{z'x}}{\mathbf{z'z}} = b_z.\end{aligned}$$

A fenti modell általánosítása, ha mindkét lehetséges érvényes instrumentumot (pl. lakhely távolsága (z_1) és születési dátum (z_2)) figyelembe veszünk a becslésnél. Ekkor is létezik a $proj(x/(z_1, z_2))$ projekció, és $\widehat{x}$ most ennek a képe. A $proj(y/\widehat{x})$ projekció konzisztens becslése β-nak. Ez hatékonyabb, mint z_1 és z_2 külön-külön.

Amikor $E(x\epsilon) \neq 0$ endogén regresszorról szoktunk beszélni. Hogyan kezeljük az egyéb regresszorokat, amelyekre $E(x_j\epsilon) = 0$? (Az exogén regresszorok.) Ezek tekinthetők önmaguk instrumentumának. A béregyenletekben általában sok más ilyen regresszort is találunk. Az első fázisbeli becslésnél a projekciót az összes regresszorra és összes instrumentumra kell elvégezni.

7.2.3 Kereslet-kínálati modell: a szimultán strukturális modell alapesete

Legyen a kínálati és az inverz keresleti függvény:

$$\begin{aligned}q &= \alpha_s p + \beta c + u \\ p &= \alpha_d q + \gamma y + v,\end{aligned}$$

ahol q a mennyiség, p az ár, c a határköltség, és y a jövedelem.

Oldjuk meg ezt a két egyenletet q-ra és p-re. Ekkor megkapjuk a kereslet-kínálati modell redukált formáját:

$$\begin{aligned}q &= \frac{1}{1-\alpha_d\alpha_s}((\alpha_s\gamma y + \beta c) + u + \alpha_s\nu) = \pi_{cq}c + \pi_{yq}y + \varepsilon_q \\ p &= \frac{1}{1-\alpha_d\alpha_s}((\alpha_d\beta c + \gamma y) + v + \alpha_d u = \pi_{cp}c + \pi_{yp}y + \varepsilon_p.\end{aligned}$$

A paraméterekre igaz, hogy

$$\begin{aligned}\pi_{cq} &= \frac{\beta}{1-\alpha_d\alpha_s} \\ \pi_{yq} &= \frac{\alpha_s\gamma}{1-\alpha_d\alpha_s} \\ \pi_{yp} &= \frac{\gamma}{1-\alpha_d\alpha_s} \\ \pi_{cp} &= \frac{\alpha_d\beta}{1-\alpha_d\alpha_s}.\end{aligned}$$

A regresszorok és hibatagok korreláltsága miatt az OLS inkonzisztens becslést as mindkét egyenletre. A fenti egyenletek paramétereinek konzisztens becslése viszont megvalósítható instrumentális becsléssel. Mivel $cov(p, y) \neq 0$ és $cov(u, y) = 0$, ezért y alkalmas instrumentuma p-nek a kínálati egyenletben, és

$$\begin{aligned} cov(q.y) &= \alpha_s cov(p, y) + \beta cov(c, y) \\ cov(q, c) &= \alpha_s cov(c, p) + \beta var(c), \end{aligned}$$

amiből meghatározhatók a strukturális paraméterek a kínálati egyenletben.

Ugyanígy belátható, hogy c alkalmas instrumentuma q-nak a keresleti egyenletben.

7.2.4 Strukturális ökonometria modell

A fenti kereslet-kínálati modell általánosítható, mint a

$$\mathbf{Ay} = \mathbf{Bz} + \mathbf{u}$$

strukturális forma.

A kereslet-kínálat modell speciális esetében:

$$A = \begin{matrix} 1 & -\alpha_s \\ -\alpha_d & 1 \end{matrix} .B = \begin{matrix} \beta & 0 \\ 0 & \gamma \end{matrix} , y = \begin{matrix} q \\ p \end{matrix} , z = \begin{matrix} c \\ y \end{matrix} ,$$

ahol $\mathbf{A}$ és $\mathbf{B}$ mátrixok elemeinek egy részét ismertnek tételeztük fel.

A modell redukált formája:

$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{Bz} + \mathbf{A}^{-1}\mathbf{u} = \mathbf{\Pi z} + \boldsymbol{\varepsilon}.$$

Az OLS konzisztens lenne, ha $E(\boldsymbol{\varepsilon}\mathbf{z}) = \mathbf{0}$ lenne, de ez az ökönometrikusok számára érdekes esetekben általában nem teljesül. Látszik, hogy a strukturális paraméterek mindig meghatározzák a redukált forma paramétereket. Megfordítva viszont három lehetőség van.

Éppen identifikált eset Amennyiben figyelembe véve az *a priori* restrikciókat

$$\mathbf{A}^{-1}\mathbf{B} = \mathbf{\Pi}$$

teljesül, akkor ennek az egyenletrendszernek a megoldása egyértelmű adott $\mathbf{\Pi}$-re, és a redukált forma együtthatóinak becslése megadja az $\mathbf{A}$ és $\mathbf{B}$ "hiányzó" paramétereinek konzisztens becslését is "számolással".

Ez teljesül a fenti példában. (A π együtthatók száma ugyanannyi, mint a nem előre meghatározott a és b együttthatók száma.)

Alulidentifikált (nemidentifikált) eset Figyelembe véve az *a priori* restrikciókat $\mathbf{\Pi}$ elemei meghatározzák $\mathbf{A}$ és $\mathbf{B}$ elemeit, de nem egyértelműen. Ekkor a strukturális paramétereket nem tudjuk identifikálni. Ilyenkor a π együtthatók száma kevesebb, mint a nem előre meghatározott a és b együtthatók száma. Ezek a fogalmak egyes egyenletekre, illetve egyes paraméterekre is értelmezhetőek.

Például ha adott $\mathbf{\Pi}$-re nincs egyértelmű megoldás, akkor összességében alulidentifikált, de előfordulhat, hogy egyes egyenletek identifikáltak, mások nem. Például a:

$$\begin{aligned} q_i &= \alpha_s p_i + \beta c_i + u_i \\ p_i &= \alpha_d q_i + v_i \end{aligned}$$

modellben α_d identifikált, míg α_s és β nem. A redukált forma egyenletei most:

$$\begin{aligned} q_i &= \frac{1}{1-\alpha_s\alpha_d}(\beta c_i + u_i + \alpha_s v_i) \\ p_i &= \frac{\alpha_d}{1-\alpha_s\alpha_d}(\beta c_i + u_i + \alpha_s v_i + v_i) \\ \pi_q &= \frac{1}{1-\alpha_s\alpha_d}\beta \\ \pi_c &= \frac{\alpha_d}{1-\alpha_s\alpha_d}\beta. \end{aligned}$$

A második egyenlet identifikált, mivel

$$\frac{\pi_c}{\pi_q} = \alpha_d,$$

az első egyenlet paraméterei viszont nem identifikáltak, hiszen a

$$\begin{aligned} (1-\alpha_s\alpha_d)\pi_c - \alpha_d\beta &= 0 \\ (1-\alpha_s\alpha_d)\pi_q - \beta &= 0 \end{aligned}$$

rendszernek végtelen sok megoldása van, mivel az első egyenlet α_d-szerese a másodiknak.

Túlidentifikáció Figyelembe véve az *a priori* restrikciókat $\mathbf{\Pi}$ elemei nem képesek meghatározni $\mathbf{A}$ és $\mathbf{B}$ elemeit. (A π együtthatók száma nagyobb, mint a nem előre meghatározott a és b együtthatók száma.)

Mivel több instrumentumot is használhatunk, igazából ez a szokásos eset. Mi a megoldás a becslésre? Természetesen a kétfokozatú becslés.

Első lépésben becsüljük a redukált formát, és meghatározzuk az ez alapján "előrejelzett" változókat.

$$\begin{aligned} \widehat{q} &= \pi_{cq}c + \pi_{yq}y \\ \widehat{p} &= \pi_{cp}c + \pi_{yp}y. \end{aligned}$$

Második lépésben elkészítjük a becslést az előrejelzett értékekkel, mint regresszorokkal.

$$\begin{aligned} q &= \alpha_s\widehat{p} + \beta c + u' \\ p &= \alpha_d\widehat{q} + \gamma y + v'. \end{aligned}$$

Mivel a jobboldali változók korrelálatlanok a hibataggal a regressziós paraméterek konzisztensen becsülhetők.

Feltehetjük a kérdést, hogy mi az összefüggés a kétfokozatú becslés és az instrumentális becslés között?

Például a q egyenletéhez tartozó első (elméleti) normálegyenlet:

$$cov(q, \widehat{p}) = \alpha_s var(\widehat{p}) + \beta cov(\widehat{p}, c).$$

Ha $\widehat{p}$ -t instrumentumnak tekintjük, és az egyenletet instrumentálisan becsüljük, akkor

$$cov(q, \widehat{p}) = \alpha_s cov(p, \widehat{p}) + \beta cov(\widehat{p}, c).$$

viszont

$$var(\widehat{p}) = cov(p, \widehat{p})$$

mivel

$$\begin{aligned} p &= \widehat{p} + u_p \\ E(u_p &\mid \widehat{p}) = 0. \end{aligned}$$

Vagyis ezzel az instrumentum választással igaz, hogy

$$Z'X = Z'Z$$

Tehát ugyanazt a becslést kapjuk, ha az egyenletet két fokozatban becsüljük, vagy pedig instrumentálisan a $\widehat{p}_i$ instrumentummal. Belátható, hogy éppen identifikált esetben az indirekt LS módszer is ugyanezt az eredményt adja. Viszont egyéb instrumentális becslések is létezhetnek, amelyek szintén konzisztensek, de aktuálisan más becsléseket adnak.

7.3 Gyakorlatok R-ben

```
require (AER)
    data("SwissLabor")
    attach(SwissLabor)
    # egy instrumentum mátrix létrehozása
    un=rep(1,length(age))
    Z=cbind(un, foreign,education)
    # a regresszorok mátrixa
    X=cbind(un,age,education)
    # az instrumentális becslés ("foreign" az "age" instrumentuma)
    (bz=solve(t(Z)%*%X)%*%t(Z)%*%income)
    # a "kalapos" instrumentum mátrix előállítása (első fázisú LS)
    instr1=fitted(lm(age~foreign+education))
    # az LS második fázisa: a "kalapos" regresszorokkal ls "education"-nel, ami
önmaga instrumentuma
    (b12=coef(lm(income~instr1+education)))
    # instrumentális becslés a "kalapos" instrumentum vektorral
    Zhat1=cbind(instr1,education,un)
    (b1z=solve(t(Zhat1)%*%X)%*%t(Zhat1)%*%income)
    # Instrumentális becslés két instrumentummal (foreign és oldkids)
    # a "kalap" instrumentum mátrix előállítása (első fázisú LS) két instrumen-
tummal (foreign és oldkids)
    instr2=fitted(lm(age~foreign+oldkids+education))
    # az LS második fázisa: a "kalap", mint regresszor vektorral
    (b22= coef(lm(income~instr2+education)))
    # instrumentális becslés a "kalap" instrumentum vektorral
    Zhat2=cbind(instr2,education,un)
    (b2z=solve(t(Zhat2)%*%X)%*%t(Zhat2)%*%income)
    # Ugyanezek a becslések az AER package-ból
    ivreg1 <- ivreg(income ~age+education | foreign+education)
    coef(ivreg1)
    ivreg2 <- ivreg(income ~age+education | foreign+oldkids+education)
    coef(ivreg2)
    summary(ivreg1)
    summary(ivreg2)
    #Tesztek
    coeftest(ivreg1, vcov=vcovHC)
    coeftest(ivreg2, vcov=vcovHC)
```

8 Kvalitatív változók

8.1 Kvalitatív magyarázó változók

Legyen A egy lehetséges tulajdonsága a megfigyeléseknek. Ha egy i $(i = 1, ...n)$ egyed rendelkezik ezzel a tulajdonsággal, akkor azt mondjuk, hogy $i \in A$.

Az A tulajdonság dummy változója egy D_A n elemű vektor, amelyre

$$\begin{aligned} D_{Ai} &= 1, ha\ i \in A, \\ D_{Ai} &= 0, ha\ i \notin A. \end{aligned}$$

Legyen két lehetséges tulajdonság (A és B). Tegyük fel, hogy

$$y_i = \alpha_k + \epsilon_i, k \in \{A, B\},$$

vagyis a várható érték tulajdonság (csoport) függő, míg a variancia nem.

Ekkor felírhatjuk a következő regressziót:

$$y = \alpha + \alpha'_B D_B + \epsilon.$$

(Itt A a "kontrollcsoport", és B a "kezelt" csoport.)

Világos, hogy

$$\begin{aligned} \alpha_A &= \alpha \\ \alpha_B &= \alpha + \alpha'_B \end{aligned}$$

Itt α OLS becslése nyilván a kontrollcsoport átlaga, α'_B pedig a másik ("kezelt") csoport átlagának és a kontrollcsoport átlagának a különbsége.

Egy ekvivalens felírás:

$$y = \alpha'_A D_A + \alpha'_B D_B + \epsilon,$$

ahol

$$\begin{aligned} \alpha_A &= \alpha'_A \\ \alpha_B &= \alpha'_B \end{aligned}$$

Itt az OLS becslés az egyes csoportátlagok.

Megjegyzések:

1. A dummy változóknál használt számértékek esetlegesek, 0 és1 helyett bármely két különböző érték megtenné.

2. A konstans és a két dummy együtt szinguláris $\mathbf{X}'\mathbf{X}$ mátrixot eredményezne, és nem létezne egyértelmű megoldása az OLS problémának.

3. A hipotézisek értelemszerűen felírásfüggőek. Például az a nullhipotézis, hogy a két várható érték egyenlő, az első felírásban azonos azzal, hogy D_B együtthatója 0, míg a második felírásban azzal, hogy D_A és D_B együtthatói megegyeznek. Több lehetséges módon is lehet tesztelni bármelyik hipotézist.

8.1.1 Általánosítások

Több kategória Tegyük fel, hogy három tulajdonságunk (kategóriánk) van. Az első típusú regresszió most:

$$y = \alpha + \alpha'_B D_B + \alpha'_C D_C + \epsilon,$$

a második típusú:

$$y = \alpha'_A D_A + \alpha'_B D_B + \alpha'_C D_C + \epsilon.$$

Több ismérv Tegyük fel, hogy több ismérv van. Az egyik szerint A és B, a másik szerint a, b és c tulajdonságok vannak.

Potenciálisan 6 különböző várható értékünk van:

	a	b	c
A	α_{Aa}	α_{Ab}	α_{Ac}
B	α_{Ba}	α_{Bb}	α_{Bc}

Hogyan fogalmazzuk meg a problémát regresszióként?

Definiáljunk minden csoportra egy külön dummyt, pl. D_{Bc} stb. Ha ezekkel futtatjuk a regressziót, akkor nincsenek megkötéseink a hat különböző várható értékre.

Egy alternatív megfogalmazása ugyanannak, ha a regresszióban van egy konstans, három "főhatás dummy" (D_b, D_c, D_B), valamint két úgynevezett interakciós dummy $(D_{Bb} = D_B D_b, D_{Bc} = D_B D_c)$. (Itt az Aa halmaz a kontrollcsoport.) Ez összesen hat paraméter, és ugyanúgy nem vagyunk korlátozva, mint az előző esetben.

Gyakran azonban a következő leegyszerűsítést tesszük. A regresszióban csak a konstanst és a három (D_b, D_c, D_B) "főhatás dummy-t" szerepeltetjük. Ez négy paraméter, amiből kiszámolható mind a hat várható érték. Itt azonban vannak restrikciók. Például mivel

$$\alpha_{Bb} = \alpha + \alpha_B + \alpha_b$$

és

$$\begin{aligned} \alpha_{Aa} &= \alpha \\ \alpha_{Ba} &= \alpha + \alpha_B \\ \alpha_{Ab} &= \alpha + \alpha_b \end{aligned}$$

ezért

$$\alpha_{Bb} = \alpha_{Ba} + \alpha_{Ab} - \alpha_{Aa}.$$

Tehát az így egyszerűsített modell két szabadságfokot veszít (6 helyett csak 4 szabad paraméter van). Felfogható a "teljes" modell egy korlátozott változatának, tehát a szokásos F-próbák alkalmazhazóak a rstrikció tesztelésére. Az egyszerűsítés indoka, hogy sok ismérv és sok tulajdonság esetén a "teljes" modell nagyon sok paramétert tartalmaz.

Példák Tegyük fel, hogy feltevésünk szerint a férfiak jövedelme leírható, mint

$$y_{iF} = \mu_F + \epsilon_i,$$

ahol ϵ_i 0 várható értékű, az i-edik "egyed" sajátos jellemzőit megtestesítő valószínűségi változó. A nők jövedelme pedig

$$y_{iN} = \mu_N + \epsilon_i$$

alakban írható. (Az ϵ_i-kre a szokásos feltevések érvényesek.) Ez a modell egyetlen egyenletben is felírható:

$$y = \mu_N D_N + \mu_F D_F + \epsilon,$$

ahol D_N egy nő-dummy, vagyis egy olyan változó, amely 1-es értéket vesz fel, amikor az i-edik megfigyelt egyén nő, és 0-t, amikor férfi. A D_F pedig egy férfi-dummy az értelemszerű módosításokkal. Mivel D_N és D_F korrelálatlanok (belső szorzatuk = 0), tudjuk, hogy a paramétereik OLS becslése független, vagyis ugyanazt kapjuk eredményként az összesített regresszióból, mint amit a két külön regresszióból kapnánk. (Például μ_N megegyezik a mintában szereplő nők jövedelmének átlagával.)

Vannak azonban további ekvivalens alternatívák is. Például:

$$y = \mu + \mu'_F D_F + \epsilon,$$

ahol most az első változó a szokásos konstans (1-es érték minden megfigyelésre.) Tudjuk ismét, hogy lényegi változás nincs, most μ becslése meg fog egyezni μ_N becslésével, míg μ'_F becslése μ_F és μ_N becslésének különbsége lesz. Ami azt jelenti, hogy $\widehat{y_i}$ (a becsült érték) változatlan marad.

Tegyük fel, hogy most a budapesti, városi és községi jövedelmek várható értékére vagyunk kíváncsiak. Természetes általánosítás, hogy most egy

$$y = \mu_B D_B + \mu_V D_V + \mu_K D_K + \epsilon$$

egyenletben gondolkodjunk, ahol D_B egy Budapest-dummy, és így tovább. Itt is kiválaszthatnánk a három kategória közül bármelyiket referenciaként, és az egyenletet "konstans+2 dummy" alakban is felírhatnánk ekvivalens módon.

Lépjünk tovább, és tegyük fel azt a kérdést, hogy van-e különbség a jövedelmekben nemek és településformák szerint. Első látásra természetesnek tűnik a következő modell:

$$y = \mu_N D_N + \mu_F D_F + \mu_B D_B + \mu_V D_V + \mu_K D_K + \epsilon,$$

így 5 paraméterünk lesz.

Azonban alaposabban megfontolva itt egy "restrikciót" is bevezettünk. A két nem és a három település típus alapján összesen 6 kategóriát tudunk megkülönböztetni (FB,NB,FV,NV,FK,NK), tehát potenciálisan 6 különböző átlagjövedelmünk

van. A fenti egyenletben az a "restrikció", hogy feltételezi minden településtípusban ugyanaz a különbség a férfiak és a nők várható jövedelme között. Hogyan lehet a korlátozatlan egyenletet felírni?

$$\begin{aligned} y &= \mu_{NB}(D_N * D_B) + \mu_{FB}(D_F * D_B) + \mu_{NV}(D_N * D_V) + \\ &\quad \mu_{FV}(D_F * D_V) + \mu_{NK}(D_N * D_K) + \mu_{FK}(D_F * D_K) + \epsilon, \end{aligned}$$

ahol a * a vektorok komponensenkénti szorzását jelenti. Ez az általános modell és ebben már tesztelhető a fenti restrikció is.

Tegyük fel, hogy azt a kérdést vizsgáljuk, hogy hogyan hat a befejezett iskolaévekben (S) kifejezett tanultság a jövedelmekre. Most, eltekintve a településformáktól, természetesnek tűnik a

$$y = \mu_N D_N + \mu_F D_F + \theta S + \epsilon,$$

regresszió becslése. Azonban itt megint van egy restrikció: ez az egyenlet felteszi, hogy az iskolákban eltöltött idő "hatása" a jövedelemre nemtől független. Általánosítsunk újra:

$$y = \mu_N D_N + \mu_F D_F + \theta_N(D_N * S) + \theta_F(D_F * S) + \epsilon.$$

8.2 Kvalitatív függő változók (klasszifikációs probléma)

Gyakran a célváltozó csak két lehetséges értéket vesz fel. Például egy adott hitelfelvevő visszafizeti a hitelt vagy nem. A gyakorlatban fontos probléma, hogy megpóbáljuk megjósolni, hogy egy adott tulajdonságú hitelfelvevő milyen valószínűséggel kerül egyik vagy másik állapotba. A modellt többféleképpen is megfogalmazhatjuk valószínűségi feltevésekkel.

8.2.1 Lineáris valószínűségi modell

Tegyük fel, hogy ez egy korrekt modell:

$$\begin{aligned} y_i &= \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i \\ E(\epsilon_i &\mid X) = 0 \\ y_i &\in \{0, 1\}. \end{aligned}$$

Látható, hogy ϵ_i minden $\mathbf{x}_i$-re $\mathbf{x}_i$-től függő bináris valószínűségi változó, és nem lehet homoszkedasztikus. Emiatt a lineáris valószínűségi modell nem népszerű feltevés, habár elvben az OKS becslés aszimptotikusan konzisztens.

8.2.2 Probit modell

Itt feltesszük, hogy létezik bizonyos magyarázó változók és a véletlen függvényeként egy nem megfigyelhető (folytonos) u_i index. Ha ez az index elér egy küszöbértéket,

amit az általánosság megsértése nélkül vehetünk 0-nak, akkor az adós visszafizeti a hitelt, egyébként pedig nem. A probit modell jellemzője az, hogy a véletlen hatás normális eloszlású.

$$\begin{aligned} y_i &= 0, u_i = \beta x_i + \epsilon_i \leq 0 \\ y_i &= 1, u_i = \beta x_i + \epsilon_i > 0 \end{aligned}$$

$E(\epsilon_i \mid X) = 0)$ és ϵ_i iid standard normális. Levezethető az egyes megfigyelséek valószínűsége:

$$P(y_i = 0) = P(\epsilon_i \leq -\beta' x_i) = F(\frac{-\beta' x_i}{\sigma}),$$

$$P(y_i = 1) = P(\epsilon_i > -\beta' x_i) = 1 - F(\frac{-\beta' x_i}{\sigma}),$$

ahol F a standard normális eloszlásfüggvény.

Tehát

$$P(y_i = 0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\beta' x_i}{\sigma}} \exp(-\frac{s^2}{2}) ds),$$

$$P(y_i = 1) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\beta' x_i}{\sigma}} \exp(-\frac{s^2}{2}) ds).$$

Feltéve, hogy a mintaelemek függetlenek, ezen valószínűsggek szorzataként kapjuk a likelihood függvényt, amelynek maximalizálásával becslést kaphatunk az ismeretlen paraméterekre.

Ekkor

$$E(y_i \mid x_i) = P(y_i = 1)$$

vagyis a feltételes várható érték megegyezik $y_i = 1$ valószínűségével. Levezethető x_{ij} marginális hatása a feltételes várható értékre:

$$\frac{\partial E(y_i \mid x_i)}{\partial x_{ij}} = \beta_j \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\frac{\beta' x_i}{\sigma})^2}{2}) ds).$$

8.2.3 Logit modell

A logit modell a probithoz képest más feltevéssel él az exogén változók és a bekövetkezési valószínűségek relációjáról.

$$\log(\frac{P_i}{1-P_i}) = \beta' x_i,$$
$$P_i = \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)},$$

ahol P_i annak a valószínűsége, hogy $y_i = 1$. A likelihood függvény ismét ezen valószínűségek szorzata, ha a mintaelemek függetlenségét is feltételezzük.

Ekkor a feltételes várható érték:

$$E(y_i \mid x_i) = P_i,$$

és a marginális hatás $\beta_j(P_i - P_i^2) = \beta_j P_i(1-P_i)$, azaz

$$\frac{\partial E(y_i \mid x_i)}{\partial x_{ij}} = \beta_j \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)} - \beta_j \frac{(\exp(\beta' x_i))^2}{(1+\exp(\beta' x_i))^2}.$$

A klasszifikációs modell jósága Klasszifikációs modellek jóságának összeméréséhez gyakran használják az úgynevezett konfúziós mátrixot. Ennek előállításához szükséges egy döntési szabály, amely úgy szól, hogy "mindig jósold azt az állapotot, aminek nagyobb a valószínűsége". Vagyis, ha egy adott megfigyelésre a modell 0,5-nél nagyobb valószínűséget jósol annak, hogy a hitel vissza lesz fizetve, akkor tekintsd ezt a modell jóslatának (klasszifikációjának). Így a mintaelemeket 4 csoportra oszthatjuk: 1. azok, amelyeknél 0 értéket jósolunk, és a megfigyelés is 0, 2. azok, amelyeknél 1 értéket jósolunk, és megfigyelés is 1, 3. azok, ahol a megfigyelés 0 és a jóslat 1, és 4. azok, ahol a megfigyelés 1 és a jóslat 0. Az 1. és 2. csoport aránya az összesen belül jelzi a klasszifikáció pontosságát. Ez mégsem igazán szerencsés mérőszám, amennyiben nagyon eltérő mértékben vannak 0 és 1 megfigyeléseink a mintában. Egy többet mondó mérce az úgynevezett ROC görbe. Ez a következőképpen konstruálható.

Tekintsünk valamilyen klasszifikációs kritériumot és számítsuk ki az igaz 1-es arányt (TPR, az 1-esek hány százalékát klasszifikáljuk valóban 1-esnek), és a hamis 1-es arányt (FPR, a 0-sok hány százalékét klasszifikáljuk tévesen 1-esnek). Például a kritérium legyen olyan, hogy ha valamely 1-es megfigyelés becsült valószínűsége nagyobb, mint p, akkor azt 1-es típusúnak klasszifikáljuk.Ha $p = 1$, akkor semmit sem klasszifikálunk 1-esnek, tehát TPR=FPR=0. Ha viszont $p = 0$, akkor TPR=FPR=1. Csökkentve p-t 1-től 0-ig mind a két arány általában nő, és képük egy nem véletlenszerű modellben a 45 fokos egyenes fölött helyezkedik el. Ez a ROC görbe, amely alatti terület nagyságát gyakran tekintik a modell jóságát összefoglaló mutatónak. (A maximális terület akkor állna elő, ha bármely $p < 1$-re a modell eltalálná a helyes klasszifikációt, mivel a becsült valószínűség akkor és csakis akkor lenne 1, ha a megfigyelés valóban 1-es.)

8.3 Gyakorlatok R-ben

8.3.1 Kvalitatív magyarázó változók

require (ISLR)

```
data(Carseats) # itt vannak a kvalitatív változók
attach(Carseats)
require(caret) # ez egy dummy változükat létrehozó package
dumo=data.frame(Urban, ShelveLoc) # a két kvalitatív változóból egy
adatstruktúrát hozunk létre
durmi<- dummyVars(" ~.", data = dumo) # a dummy változók létrehozása
dummies <- data.frame(predict(durmi, newdata = dumo))
print(dummies)
(ol=lm(Sales~Urban.No+ShelveLoc.Good+ShelveLoc.Bad,data=dummies))
# becslés az előállított dummy változókkal
(crUnSb=dummies$Urban.No*dummies$ShelveLoc.Bad) # szorzat dummy
készítése
(ol1=lm(Sales~crUnSb+ShelveLoc.Good, data=dummies)) # becslés fel-
használva szorzat dummy-t is
```

8.3.2 Kvalitatív függő változók

```
require (AER)
    data("SwissLabor")
    attach(SwissLabor)
    plot(participation ~age)
    plot(participation ~income)
    #Probit becslés
    swpr <- glm(participation ~age+income+education, family = binomial(link
= "probit"))
    summary(swpr)
    par(mfrow=c(2,1))
    plot(fitted(swpr)) # a az egyed becsült hasznossága
    plot(predict(swpr)) # az y=1 valószínűsége
    #Logit becslés
    swlg <- glm(participation ~age+income+education, family = binomial(link
= "logit"))
    summary(swlg)
    par(mfrow=c(2,1))
    plot(predict(swlg))
    plot(fitted(swlg))
    # Mekkora a különbség a kétfajta modell becslése között?
    plot (predict(swlg), predict(swpr))
    plot (fitted(swlg),fitted(swpr))
    # Marginális hatások a két modellben
    (avd <- mean(dnorm(predict(swpr)))) # y várható értéke
    (efpr = avd * coef(swpr))
    (eff=mean(fitted(swlg) * (1 - fitted(swlg)))) # likelihood arányok átlaga
    (eg=eff*coef(swlg))
    # Találunk különbséget?
    require (ROCR)
```

```
# A probit modell minősége
table(true = SwissLabor$participation, pred = round(fitted(swpr))) #siker-
mutató tábla
predpr <- prediction(fitted(swpr),SwissLabor$participation)
plot(performance(predpr, "acc")) # előrejelzés pontossága
plot(performance(predpr, "tpr", "fpr")) # kétfajta hiba relációja
abline(0, 1)
# A logit modell minősége
table(true = SwissLabor$participation, pred = round(fitted(swlg))) # sik-
ermutató tábla
predlg <- prediction(fitted(swlg),SwissLabor$participation)
plot(performance(predlg, "acc"))
plot(performance(predlg, "tpr", "fpr"))
abline(0, 1)
# szignifikancia tesztek
coeftest(swpr, vcov = sandwich)
coeftest(swlg, vcov = sandwich)
# tesztelhetünk több, mint egy restrikciót is
swpr1 <- glm(participation ~age, family = binomial(link = "probit"))
waldtest(swpr1,swpr)
swlg1 <- glm(participation ~age, family = binomial(link = "logit"))
waldtest(swlg1,swlg)
```

9 Statisztikai tanulás*

Ez a fejezet rövid bevezetést nyújt a statisztikai tanulás filozófiájába, néhány példán keresztül illusztrálva is azt.

Az általános probléma a következőképpen írható le. Legyen X az inputok (magyarázó változók) tere. Tegyük fel, hogy létezik egy $F : X \to Y$ függvény, ahol $Y = R$ a célváltozó, amit "igazi" kapcsolatnak nevezhetünk. Ez a kapcsolat azonban csak zajosan figyelhető meg. Szeretnénk egy olyan $f : X \to Y$ becslőfüggvényhez eljutni, amely valamilyen értelemben optimálisan közelíti az igazi kapcsolatot.

A statisztikus kiindul egy F_β függvényhalmazból, ahol $f_\beta(X) \in F_\beta$. Például lineáris regresszió esetén β a paraméter vektor. Az optimalitás értelmezéséhez szükség van egy veszteségfüggvényre is, $L(Y, f_\beta(X))$. A statisztikus megpróbálja megtalálni f^*_β-t, ami a várható veszteséget minimalizálja. Azonban kissé különösen jár el, az eredeti céfüggvény helyett egy perturbált veszteségfüggvényt, $E_X L(Y, f_\beta(X)) + h(\beta, \lambda)$, minimalizál, ahol λ (egy "hiperparaméter") bünteti a becslőfüggvény komplexitását.

Példa: ridge regresszió A perturbált célfüggvény:

$$SSR + \lambda \boldsymbol{\beta}'\boldsymbol{\beta},$$

ahol SSR a szokásos négyzetes eltérés, és λ a büntető paraméter.
A megoldás:

$$\mathbf{b}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

Nyilvánvalóan a büntetés úgy hat, hogy nagy λ-ra $\mathbf{b}^{ridge}$ egyre inkább a "semmitmondó" 0 felé közelít. Mi indokolhatja ezt?

9.1 A torzítás-variancia átváltás

A feltételezett additív zaj mellett a modell:

$$\begin{aligned} y &= f(x) + \epsilon \\ E(\epsilon) &= 0 \\ var(\epsilon) &= \sigma^2. \end{aligned}$$

Ha a veszteségfüggény kvadratikus, $(y - f_\beta(x))^2$, akkor egy a mintában nem szereplő x_0 megfigyelésnél a várható veszteség:

$$\begin{aligned} err(x_0) &= E(y - f_\beta(x_0))^2 = \\ &\sigma^2 + (Ef_\beta(x_0) - f(x_0))^2 + E\left[f_\beta(x_0) - Ef_\beta(x_0)\right]^2. \end{aligned}$$

A veszteségnek három komponense van: 1. a redukálhatatlan hibából fakadó tévedés (σ^2), 2. a négyzetes torzítás ($(Ef_\beta(x_0) - f(x_0))^2)$), és 3. a becslőfüggvény varianciája ($E\left[f_\beta(x_0) - Ef_\beta(x_0)\right]^2$). Minél komplexebb egy becslőfüggvény, annál nagyobb a varianciája. Például lineáris regresszióban arányosan nő a becsült paraméterek számával. A tapasztalat azt mutatja, hogy némi torzítást érdemes lehet elcserélni a variancia csökkentésére.

Hogyan lehet ezt a gyakorlatban megvalósítani? Az első fázisban a hiperparaméterek különböző értékeire becsüljük meg a legjobb modellt. Majd ezeket a legjobb modelleket validáljuk, választjuk ki közülük a legjobbat, lényegében a legjobban teljesítő hiperparamétert. A validálásra legegyszerűbb módszer, ha a megfigyelések egy részét nem használjuk fel a becslési fázisban, majd ezeken a félretett adatokon kipróbáljuk az egyes hiperparaméterekhez tartozó legjobb modeleket, és kiszámoljuk mindegyikhez mekkora veszteség tartozik. A legkisebb ex post veszteséggel rendelkező modellt tekintjük a tanulási folyamat végtermékének.

Amennyiben nem akarunk adatokat félretenni, alkalmazhatjuk a K-szoros keresztvalidálás módszerét.

Legyen $\kappa : \{1, ...N\} \rightarrow \{1, , , K\}$ egy indikátorfüggvény, amely az adatokat K azonos elemszámú részhalmazba osztja. Legyen továbbá $f_\beta^{-k}(x, \lambda)$ a λ hiperparaméterrel indexált optimális függvény, amelyet úgy kapunk, hogy a becslésnél csak azokat az adatokat használjuk, amelyek nem tartoznak a k-adik halmazba. Számoljuk ki a veszteséget minden egyes megfigyelésre, ezt jelöljük $L(y_i, f_\beta^{-\kappa(i)}(x_i, \lambda))$-vel. Ezután ezeknek vegyük az átlagát minden hiperparaméterre:

$$CV(\lambda) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\beta^{-\kappa(i)}(x_i, \lambda)).$$

A kiválasztott modell azzal a λ-val készült becslés, amelyik $CV(\lambda)$-t minimalizálja. Ezt az eljárást a következő speciális algoritmus példáján illusztráljuk.

9.2 A CART (klasszifikációs és regressziós fa)

A CART az egyik legnépszerűbb tanuló algoritmus, lehet kvalitatív és folytonos célváltozóval is használni. Itt a több (P) kvalitatív célváltozós esetet mutatjuk be.

9.2.1 Fa építés

Induljunk ki a teljes adathalmazból, ahol összesen n megfigyelés van. A fa kialakítása lényegében azt jelenti, hogy a teljes n elemű halmazt minden lépésben 1-gyel növelt számú diszjunkt részhalmazra bontjuk, mégpedig az input tér partícionálásával. Maga a fa egy olyan gráf, ahol az egyes csúcspontok (amelyeknek megfelel a megfigyelések egy részhalmaza) utód-szülő kapcsolatban vannak egymással. Minden felmenőnek pontosan két utóda van, egészen a

végpontokig, amelyeknek nincs utódjuk, és amelyek a végső partíciót reprezentálják.

Induljunk ki a gyökérből, ami a teljes megfigyelés halmaz. Az algoritmus lényegében egy fát épít a megfigyelésekből, amely minden csomópontja a megfigyelések egy halmazát reprezentálja. A gyökér (R) az összes megfigyelést tartalmazza. Ekkor

$$p_{hR} = \frac{n_i}{n}, h = 1, ..P$$

az h-adik osztály *a priori* valószínűsége (relatív gyakorisága) a mintában. Természetes klasszifikációs szabály:

$$\tau(R) = \arg\max_h p_{hR}.$$

Tegyük fel, hogy a teljes input halmazt valahogyan két részhalmazra bontjuk, és mindkét részhalmazra alkalmazzuk ugyanezt a szabályt. Általában, ha a h-adik típus relatív gyakorisága egy A halmazban:

$$p_{hA} = \frac{n_{hA}}{n_A},$$

akkor a klasszifikációs szabály:

$$\tau(R) = \arg\max_h p_{hA}.$$

A két részhalmazra bontás nyilván nagyon sokféleképpen megtörténhet, a CART (egyik) lényeges tulajdonsága az az elv, ami alapján elvégezzük ezt a felbontást. A cél az, hogy minél inkább csökkentsük a fa teljes entrópiáját. Egy csomópont entrópiája:

$$I(A) = -\sum_{h=1}^{P} p_{hA} \log p_{hA}.$$

Egy teljes leszámlálása a lehetőségeknek és az optimális felbontás (split) választása csak elvben kivitelezhető, és nem is célszerű. A CART a következőképpen jár el: veszi az első magyarázó változót és annak összes lehetséges bináris megbontását (ha a változó rendezett, akkor csak a rendezett felbontásokat). Minden egyes lehetséges felbontásra kiszámolja az entrópia csökkenését, és kiválasztja azt, amelyik a legnagyobb javulást éri el. Ugyanezt megteszi a második, harmadik, k-adik változóval is. Majd azt a változót és azt a felbontást választja, amelyre a legnagyobb az entrópia csökkenése. Ez a felbontás (vágás) egy három csúcspontú fát eredményez, ahol most két végpont van. A következő vágásnál már mind a két végpontra el kell végezni az egyes változók összes lehetséges felosztását, de az újabb vágás most is csak egy végpontot érint, vagyis a fa mérete újra kettővel, a végpontok száma pedig 1-gyel nő. A következő lépésben most már a három végpont valamelyikén haladunk tovább a legnagyobb entrópia csökkenés elvének figyelembevételével. Mivel véges számú adatunk van a faépítés egyszer meg fog állni, de a gyakorlatban az algoritmusok már akkor

is leállnak, amikor a végpontokhoz tartozó megfigyelések száma az összes megfigyelés számához képest nagyon kicsivé válik. Minden végpontnak megfelel az input tér egy diszjunkt részhalmaza, és ezek uniója kiadja az input teret. Ez a módszer az input tér homogenizálásának fogható fel, mivel ugyanahhoz a végponthoz hasonló elemek tartoznak abban az értelemben, hogy a hozzájuk tartozó célváltozó értékek közel lesznek egymáshoz. A lényeg persze az, hogy azonosítjuk az input tér azon részhalmazait, ahol ez a homogenitás érvényesül.

9.2.2 A fa metszése

Az így épített fa egy nem-paraméteres becslésnek tekinthető, ahol az input tér diszjunkt részhalmazaira a fenti klasszifikációs szabályt alkalmazzuk. Ez a becslés „túlilleszt", vagyis túlságosan pontosan adja vissza az empirikus adatokat, aminek következtében rossz az általánosítási képessége, azaz a mintán kívül pontatlan lesz az előrejelzés. A metszési művelet a nagy és nagyon komplex fát, egyre kevésbé komplex al-fákra „metszi" vissza, amelyek adott komplexitás (végpontok száma) feltétel mellett optimálisak, vagyis a legkisebb entrópiával rendelkeznek. Belátható, hogy ez a metszési műveletsor ekvivalens avval, hogy definiálunk egy új célfüggvényt, ami tartalmazza nemcsak az entrópiát, hanem egy komplexitási büntetőfüggvényt is, és egy adott komplexitási büntető paraméter mellett ezt a módosított célfüggvényt minimalizáló részfát választjuk. Definiáljuk a perturbált veszteségfüggvényt bármely T_d fára a következőképpen:

$$\sum_{h=1}^{n} L(\tau(x_h), \tau(T_d(x_h)) + \lambda \left|T_d\right|,$$

ahol $T_d(x_h)$ az a végpont, amelyhez x_h tartozik, és $|T_d|$ a végpontok számossága T_d-ben.

Ha $\lambda = 0$ akkor nyilvánvalóan T az optimális fa a perturbált veszteségfüggvény szerint. Azonban növelve λ-t egyre rövidebb részfák válnak optimálissá. Bizonyíthatóan a folyamat egymásba ágyazott részfák sorozatát indukálja. Ha λ nagyon nagy, akkor már csak a gyökér lesz optimális. Tehát $0 = \lambda_0 < \lambda_1 < \lambda_2 < ...\lambda_M = \infty$, és minden $[\lambda_i, \lambda_{i+1}]$ intervallumban van egy optimális fa. Így egy rész-fa sorozatot kapunk, amelynek egyik végén áll a maximális fa (ahol a komplexitás büntető paramétere 0), a másik végén pedig az osztatlan fa (ahol a büntető paraméter végtelen nagy).

9.2.3 Validáció: a legjobb rész-fa kiválasztása

A legjobb rész-fát (ami ekvivalens az optimális komplexitással) a CART keresztvalidációval határozza meg. Határozzuk meg a β-kat a következőképpen:

$$\begin{aligned}\beta_1 &= 0\\ \beta_2 &= \sqrt{\lambda_1\lambda_2}\\ &\ldots\\ \beta_M &= \infty.\end{aligned}$$

Csináljunk K-szoros kereszt-validációt, ahol minden $n-\frac{n}{K}$almintára M modelt becsülünk, egyet minden β-ra. Számoljuk ki az entrópiákat a kihagyott megfigyelésekre, és átlagoljuk a K részmintára. Minden β_j-re lesz egy entrópiánk (veszteségünk), és azt a β_j-t választjuk, amelyre ez a legkisebb.

9.3 Gyakorlat R-ben

```
#Klasszifikációs fát készítünk logit vagy probit modell helyett
  library (tree)
  tree.partic =tree(participation ~age+income+education)
  # Csak default paramétereket használunk
  plot(tree.partic)
  text(tree.partic,pretty=0)
  summary(tree.partic)
  tree.partic
  #Milyen információt kapunk az outputból?
```

10 Felhasznált irodalom

Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2008.

Casella, George, and Roger L. Berger. "Statistical inference. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks." Cole Advanced Books & Software, Pacific Grove, CA, 1990.

Greene, William H. Econometric analysis. Pearson Education India, 2003.

Hastie, T. - Tinshirani, R. - Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer-Verlag, New York. 2009

R. Ramanathan: Bevezetés az ökonometriába alkalmazásokkal, PANEM, Budapest, 2003.