Econometrics with Python

Olivér Nagy

John von Neumann University - MNB Institute
Central Bank of Hungary

2023. november 23.

Általános információk

- Megajánlott jegy a heti workshop-okon elkészített feladatok alapján:
  - Elméleti blokk (30%)
  - Gyakorlati blokk (70%)
- Év végi számonkérés:
  - Zárthelyi december 4.-én 3 óra (2 sáv) hosszan
  - Pót zárthelyi december 8.-án 3 óra (2 sáv) hosszan

Ponthatárok	Érdemjegy
90 - 100	5
80 - 89	4
66 - 79	3
50 - 65	2
0 - 49	1

Általános információk (2)

- Ajánlott irodalom:
  - Hayashi, F. (2000). Econometrics. Princeton University Press
  - Wasserman, L. (2010). All of Statistics: A Concise Corse in Statistical Inference. Springer
  - Efron, B. and Hastie T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press
  - Hansen, B. (2022). Econometrics. Princeton University Press
  - Taleb, N. N. (2023). Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications. STEM Academic Press
- Email: nagy.oliver@nje.hu

Elméleti szegmens felépítése

1. Statisztikai alapok
2. Nagy mintás / aszimptotikus tulajdonságok
3. Hagyományos legkisebb négyzetek módszere
4. Hipotézis vizsgálat
5. Modell diagnosztika
6. Általánosított legkisebb négyzetek módszere
7. Maximum Likelihood elvű becslés
8. Bootstrap

Várható érték

Várható érték: $\mathbb{E}(X) = \mathbb{E}\,X = \int x\,dF(x) = \int xf(x)\,dx = \mu = \mu_X$

A lusta statisztikus szabály (LOTUS)

Ha $Y = g(X)$, akkor:

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \int g(x)f(x)\,dx$$

Várható érték tulajdonságok

1. Ha $X_1, ..., X_n$ véletlen változók és $a_1, ..., a_n$ konstans értékek, akkor:

$$\mathbb{E}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i\,\mathbb{E}(X_i)$$

2. Ha $X_1, ..., X_n$ **függetlenek** akkor:

$$\mathbb{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbb{E}(X_i)$$

Variancia

Variancia: $\mathbb{V}(X) = \mathbb{V}\,X = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 f(x)\,dx = \sigma_X^2$

Szórás: $sd(X) = \sqrt{\mathbb{V}(X)} = \sigma_X$

Variancia tulajdonságok

Feltéve, hogy a variancia jól definiált, akkor a következő tulajdonságokkal rendelkezik:

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$
2. Ha **a** és **b** konstans értékekek akkor:
   $\mathbb{V}(\mathbf{a}X + \mathbf{b}) = \mathbf{a}^2\,\mathbb{V}(X)$
3. Ha $X_1, ..., X_n$ **függetlenek** és $a_1, ..., a_n$ konstans értékek, akkor:

$$\mathbb{V}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2\,\mathbb{V}(X_i)$$

Kovariancia és korreláció

Ha X és Y véletlen változók, akkor a **kovariancia** és a **korreláció** az X és Y között fennálló **lineáris** kapcsolat erősségét méri.

Kovariancia: $Cov(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y))$

Korreláció: $\rho = \rho_{X,Y} = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, ahol $-1 \leq \rho(X,Y) \leq 1$

Nem független véletlen változók varianciája

- Ha $X_1, ..., X_n$ véletlen változók és $a_1, ..., a_n$ konstans értékek, akkor:

$$\mathbb{V}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \, \mathbb{V}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j Cov(X_i, X_j)$$

Minta tulajdonságok

Ha $X_1, ..., X_n$ véletlen változók akkor a **minta átlag ($\overline{X}_n$)**:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1} {}_n X_i$$

és a **minta variancia**(S_n^2):

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Theorem

*Ha $X_1, ..., X_n$ független, azonos eloszlású (**i.i.d.**) véletlen változók és $\mu = \mathbb{E}(X_i)$, valamint $\sigma^2 = \mathbb{V}(X_i)$ akkor:*

$$\mathbb{E}(\overline{X}_n) = \mu,\ \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n},\ \mathbb{E}(S_n^2) = \sigma^2$$

Momentumok

Centrális momentum: X véletlen változó k.-ik centrális momentuma:

$$\mu_k \equiv \mathbb{E}([X - \mu]^k)$$

Studentization / Standardization: Ha X véletlen változó μ várható értékkel és σ^2 varianciával, akkor X **standardizált** transzformáltja (Z):

$$Z = \frac{x - \mu}{\sigma}$$

Ferdeség / Skewness: Ha X véletlen változó, akkor a **ferdesége**

$$\mathbb{E}(Z^3) = \frac{\mathbb{E}([X - \mathbb{E}(X)]^3)}{\mathbb{E}([X - \mathbb{E}(X)]^2)^{\frac{3}{2}}} = \frac{\mu_3}{(\sigma^2)^{\frac{3}{2}}}$$

Csúcsosság / Kurtosis: Ha X véletlen változó, akkor a **csúcsossága**

$$\mathbb{E}(Z^4) = \frac{\mathbb{E}([X - \mathbb{E}(X)]^4)}{\mathbb{E}([X - \mathbb{E}(X)]^2)^2} = \frac{\mu_4}{(\sigma^2)^2}$$

Limit

Határérték

Legyen x_n valós számokból álló nem sztochasztikus sorozat. Ha bármely ϵ pozitív valós számhoz tartozik olyan N természetes szám (*index*), hogy minden $n > N$ esetén $|x_n - x| < \epsilon$, akkor x értéket az x_n sorozat **határértékének** nevezzük. Jelölése: $x_n \to x$.

A határérték tehát egy olyan pont amelyet a sorozat (x_m) megközelít, és idővel, mindig a közelőben is marad. Lehet hogy a sorozat, soha se éri el a határértékét, de ha a sorozat elemszáma kellően nagy (n $>$ N), onnantól fogva mindig ϵ távolságon belül marad x határtértékhez képest.

A véletlen számok határértékét / határait, több formában is tudjuk értelmezni, ezeket nézzük meg a következők során.

Konvergencia - Eloszlás 1.

Legyen X_n véletlen számokból álló sorozat, és X egy véletlen változó. F_n jelölje az X_n-hez tartozó kumulált eloszláls függvényt F pedig X eloszlás függvényét.

Convergence in Distribution

X_n véletlen változók sorozata **eloszlásában konvergál** X-hez, ha

$$\lim_{n\to\infty} F_n(t) = F(t)$$

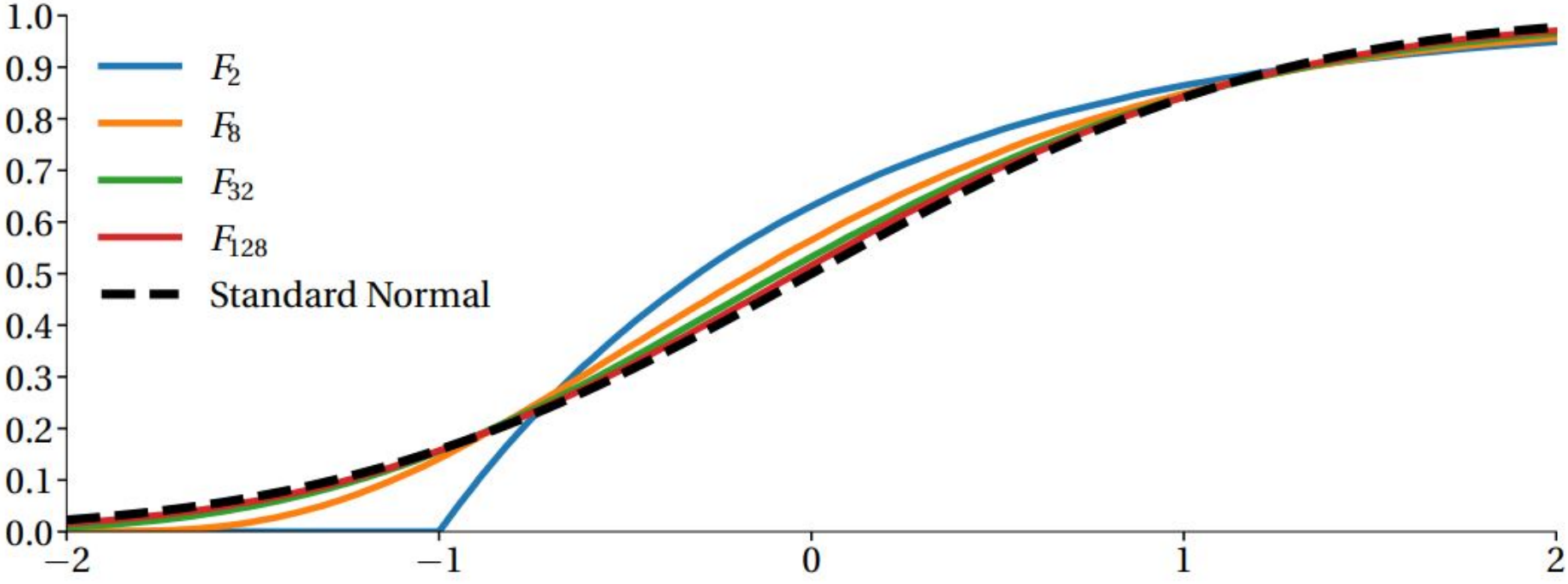
Jelölése: $X_n \xrightarrow{d} X$

A eloszlás konvergencia azt jelenti, hogy a sorozat határoló eloszlása megegyezik egy $(convergent)$ véletlen változó eloszlásával.

Continuous Mapping Theorem

Ha $X_n \xrightarrow{d} X$ és $g(x)$ függvény egy nulla valószínűségű halmazon kívül folytonos, akkor $g(X_n) \xrightarrow{d} g(X)$.

Konvergencia - Eloszlás 2.



Az 1. ábrán az Fi kumulatív eloszlás függvények sorozata látható, amint konvergálnak a standard normális kumulatív eloszláshoz, az elemszám növekedése során.

Konvergencia - Valószínűség, Kvadratikus átlag

Convergence in Probability

X_n véletlen változók sorozata **valószínűségében konvergál** X-hez akkor és csak akkor, ha:

$$\lim_{n\to\infty} Pr(|X_n - X| < \epsilon) = 1 \, \forall \epsilon > 0$$

Jelölése: $X_n \xrightarrow{p} X$, illetve $\operatorname{plim} X_n = X$

Convergence in Mean Square / Quadratic Mean / L_2

X_n véletlen változók sorozata **kvadratikus átlagban konvergál** X-hez akkor és csak akkor, ha:

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0$$

Jelölése: $X_n \xrightarrow{m.s.} X$, illetve $X_n \xrightarrow{qm} X$

A kvadratikus átlagban értelmezett konvergencia elég erős ahhoz, igaz legyen: $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ és $\lim_{n\to\infty} \mathbb{V}[X_n] = \mathbb{V}[X]$

Konvergencia - Szinte biztos

Almost sure convergence

X_n véletlen változók sorozata **szinte biztos konvergál** X-hez akkor és csak akkor, ha:

$$\lim_{n\to\infty} Pr(X_n - X = 0) = 1$$

Jelölése: $X_n \xrightarrow{a.s.} X$

Konvergenciák közti implikációk

- $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{m.s.} X$
- $X_n \xrightarrow{m.s.} X \implies X_n \xrightarrow{p} X$
- $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$
- Akkor és csak akkor, ha $Pr(X = c) = 1$: $X_n \xrightarrow{d} X \implies X_n \xrightarrow{p} X$

Konvergencia - Kapcsolatok

Konvergencia implikációk

Legyenek X_n, Y_n véletlen változók sorozatai, legyenek X, Y véletlen változók, c skalár, és legyen g folytonos függvény.

1. $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n + Y_n \xrightarrow{p} X + Y$
2. $X_n \xrightarrow{m.s.} X, Y_n \xrightarrow{m.s.} Y \implies X_n + Y_n \xrightarrow{m.s.} X + Y$
3. $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} c \implies X_n + Y_n \xrightarrow{d} X + c$
4. $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n Y_n \xrightarrow{p} XY$
5. $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} c \implies X_n Y_n \xrightarrow{d} cX$
6. $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$
7. $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$

A **3.** és **5.** részeket **Slutsky tételnek** nevezük.

Konzisztencia és torzítatlanság

Legyen $\hat{\theta}_n$ egy θ-ra vonatkozó becsült paraméterek sorozata n pedig a becsült minta nagysága. Ebben az esetben $\hat{\theta}_n$ véletlen változók sorozataként értelmezhető.

Consistency

$\hat{\theta}_n$ **konzisztens** becslője θ-nak akkor ha:

$$\hat{\theta}_n \xrightarrow{p} \theta$$

Bias

Torzításának nevezzük a becsült paraméter várható értéke és a tényleges (nem megfigyelhető) paraméter közti különbséget:

$$B[\hat{\theta}_n] = \mathbb{E}[\hat{\theta}_n] - \theta$$

Ha $B[\hat{\theta}_n] = 0$, akkor a becslő torzítatlan.

Nagy számok törvényei

Chebychev's Weak Law of Large Numbers (WLLN)

Ha $X_1, ..., X_n$ független, azonos eloszlású (**i.i.d.**) véletlen változók és feltételezzük, hogy

$$\mathbb{E}[X_n] < \infty, \forall n \in \mathbb{N}$$

akkor:

$$\overline{X}_n \xrightarrow{\text{p}} \mu$$

Kolmogorov's Strong Law of Large Numbers (SLLN)

Ha $X_1, ..., X_n$ független, azonos eloszlású (**i.i.d.**) véletlen változók és feltételezzük, hogy

$$\mathbb{E}[X_n] < \infty, \forall n \in \mathbb{N}$$

akkor:

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu$$

Iterált logaritmus

Law of the iterated logarithm (LIL)

Legyenek $X_1, ..., X_n$ független, azonos eloszlású (**i.i.d.**) véletlen változók, 0 várható értékkel, és egységnyi varianciával. Legyen $S_n = X_1 + ... + X_n$. Ekkor

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2n \ln \ln n}} = 1 \quad a.s.$$

Az iterált logaritmus tétel a nagy számok törvényei és a centrális határeloszlás tételek között helyezkedik el. Az előbbiek "pontszerűen", az utóbbiak eloszlás szintjén tudnak információval szolgálni véletlen változók sorozatának konvergenciájáról. E kettő között, a LIL egy sávot jelöl ki, amelyen belül fog tartózkodni a sorozatunk, ahogy növeljük az elemszámot.

Centrális határeloszlás tétel

Lindeberg-Levy Central Limit Theorem (CLT)

Legyenek $X_1, ..., X_n$ független, azonos eloszlású (**i.i.d.**) véletlen változók, és feltételezzük, hogy:

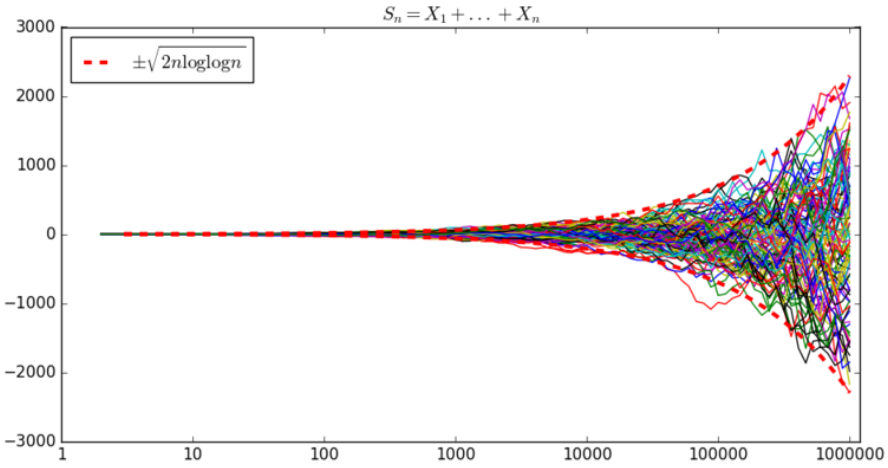
$$\mu \equiv \mathbb{E}[X_i],\ \sigma^2 \equiv V[X_i] < \infty, \text{ és } \sigma^2 > 0$$

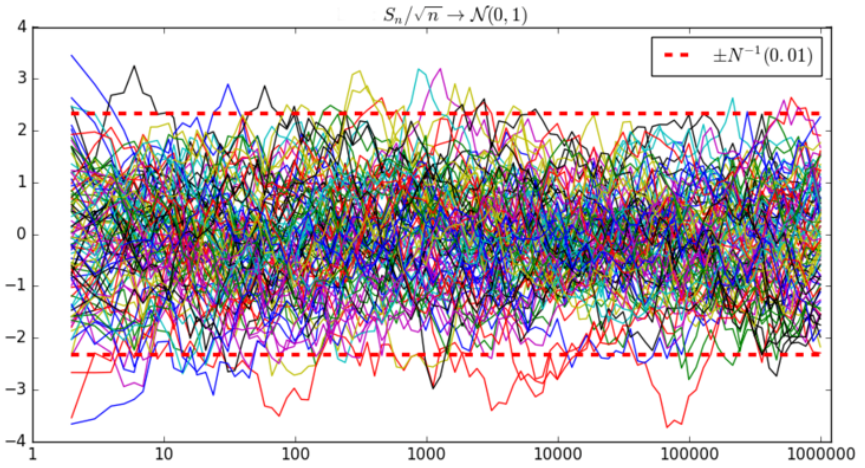
akkor,

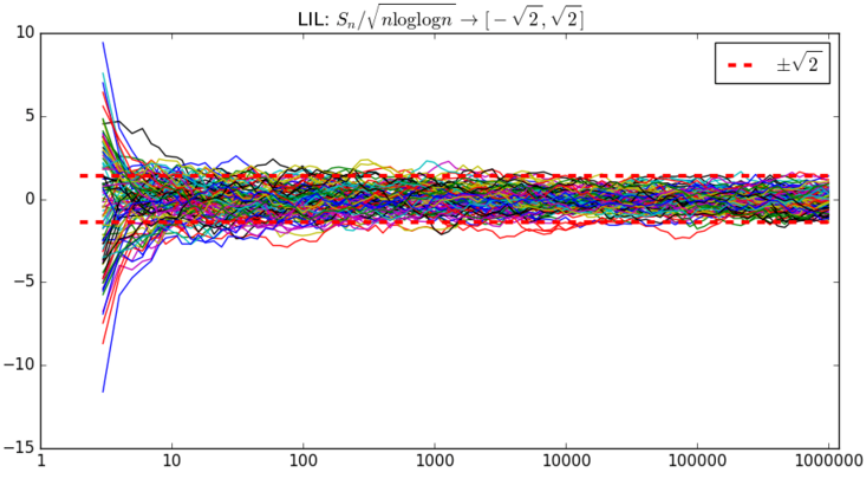
$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

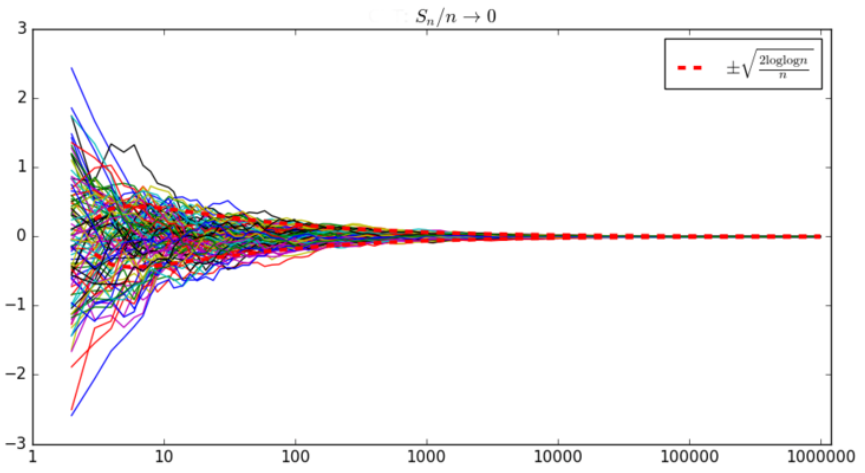
A CLT tehát azt mondja, hogy a valószínűségi változók átlagának studentizált / sztenderdizált értéke eloszlálásában a standard normális eloszláshoz tart.

Átlagok konvergenciája









Véletlen változók konvergenciája a főbb konvergenicia törvények / tételek mentén

Delta módszer

Ha X_n véletlen változók sorozata eloszlásában Normális eloszláshoz tart, akkor a **delta módszer** lehetővé teszi, hogy meghatározzuk $g(X_n)$ eloszlás konvergenciáját.

Delta Method

Tegyük fel, hogy:

$$\frac{\sqrt{n}(X_n-\mu)}{\sigma} \xrightarrow{d} N(0,1)$$

és g differenciálható függvény oly módon, hogy $g'(\mu) \neq 0$, akkor

$$\frac{\sqrt{n}(g(X_n)-g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} N(0,1)$$

Alternatív megfogalmazásban:

$$X_n \xrightarrow{d} N(\mu, \frac{\sigma^2}{2}) \implies g(X_n) \xrightarrow{d} N(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n})$$

Multivariate Delta Method

Legyen X_n véletlen vektorok sorozata, és tegyük fel, hogy:

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \Sigma)$$

és ∇_μ jelöli ∇g μ pontban kiértékelt nemnulla elemeit, akkor:

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, \nabla_\mu^T \Sigma \nabla_\mu)$$

Hatékonyság

A hatékonyság abban segít nekünk, hogy a **konzisztens, aszimptotikus normális** (CAN), azonos kovergencia sebességel rendelkező becslők között rangsort tudjunk felállítani.

Relative efficiency

Legyen $\tilde{\theta}_n$ és $\hat{\theta}_n$ két $\sqrt{n}$-konzisztens aszimptotikusan normális becslője θ_0-nak. Ha $\tilde{\theta_n}$ **aszimptotikus varianciája** $(avar)$ kisebb mint $\hat{\theta}_n$ aszimptotikus varianciája, tehát

$$avar(\tilde{\theta}_n) < avar(\hat{\theta}_n)$$

akkor $\tilde{\theta}_n$ **relatív hatékonyabb**, mint $\hat{\theta}_n$.

Asymptotically Efficient Estimator

Legyen $\tilde{\theta}_n$ és $\hat{\theta}_n$ két $\sqrt{n}$-konzisztens aszimptotikusan normális becslője θ_0-nak. Ha

$$avar(\tilde{\theta}_n) < avar(\hat{\theta}_n)$$

minden $\hat{\theta}_n$ esetén, akkor $\tilde{\theta}_n$ **hatékony becslője** θ_0-nak.

Review: alapszakon / kvantitatív alapokon elsajátított ismeret

- Többváltozós lineáris regressziós modell (MLR) definíciója:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i \tag{1}$$

  ahol $i = 1, ..., n$

- A model y változó alakulását próbálja leírni $x_1, ..., x_k$ függvényében
- y_i függő változó (hivatkozunk rá LHS, vagy outcome változóként)
- $x_1, ..., x_k$ független változók (RHS, prediktor változók)
- ϵ_i a hibatag

Review: OLS elvű becslés

- A hagyományos legkisebb négyzetek módszere (OLS) egy becslési elv, ami lehetővé teszi lineáris regressziós modellek modell paramétereinek ($\hat{\beta}$) becslését, a **négyzetes távolságok** minimalizálásán keresztül:

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_k x_{ki})^2$$

- Legyen $Y = [y_1, ..., y_n]^T, \beta = [\beta_0, \beta_1, ..., \beta_n]^T, \epsilon = [\epsilon_1, ..., \epsilon_n]^T$, és

$$X = \begin{bmatrix} 1 & x_{11} & ... & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & ... & x_n k \end{bmatrix}.$$

- Ebben az esetben (1) egyenlet mátrixos formában átírható a következő módon:

$$Y = X\beta + \epsilon \tag{2}$$

Review: Lineáris regresszió OLS elvű becslőjének levezetése

- Az OLS elvű becslő a maradéktagok ($residuals$) négyzetének az összegét (SSR/RSS) minimalizálja, amit mátrixos formában a következő módon tudunk felírni:

$$SSR = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta).$$

- SSR konkáv β függvényében, ezért az elsőrendű feltétel (FOC) elégséges az egyenlet minimalizálásához.
- Alkalmazva a következő azonosságot $(Y^T X\beta)^T = \beta^T X^T Y$ azt kapjuk, hogy

$$\epsilon^T \epsilon = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta.$$

- Az FOC alapján

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} = -2X^T Y + 2X^T X\hat{\beta} = 0$$

$$\implies \hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

Review: Projection matrix, Annihilator matrix I.

- A projekciós mátrix (projection matrix), egy szimmetrikus, idempotens mátrix, amely egy változónak az X által felölelt térre való vetítését adja, jelölése P_X:

$$P_X = X(X^T X)^{-1} X^T$$

- Az annihilátor mátrix (annihilator matrix), egy szimmetrikus idempotens mátrix, amely egy változó vetületét adja X nullterére, jelölése M_X:

$$M_X = I_n - X(X^T X)^{-1} X^T$$

P_X és M_X mátrixokat előnyös tulajdonságaik miatt használjuk. Mind a függő változó becsült értéke, mind a hiba tag kifejezhető velük a függő változó függvényében:

$$\hat{Y} = P_X Y \epsilon = M_X Y$$

Review: Projection matrix, Annihilator matrix II.

- A mátrixok idempotensek és ortogonálisak:

$$P_X P_X = P_X$$
$$M_X M_X = M_X$$
$$P_X M_X = 0$$

- A projekciós mátrix tehát visszadja Y azon részét, amely az X által kifeszített térben található, míg az annihilátor mátrix Y azon részét, amely X nullterében van. Ebből következik, hogy Y felbontható a következő módon

$$Y = P_X Y + M_X Y$$

- A függő változó négyzete szintén dekomponálható ezekkel a mátrixokkal:

$$Y^T Y = Y^T P_X Y + Y^T M_X Y$$

Gauss-Markov feltevések

Ahhoz, hogy az OLS elven becsült lineáris regressziónak fennálljanak bizonyos előnyös tulajdonságai, meghatározott feltevéseknek teljesülniük kell:

1. **Modell linearitás**: Feltesszük, hogy az Y és az X közötti kapcsolat lineárisan leírható
2. Az adatok a sokaság **véletlen mintái**. A legtöbbször használt i.i.d. ennél szigorúbb, tehát teljesíti a véletlen mintás feltevést.
3. X mátrix **teljes oszlop rangú**. Ezen feltevés alpján tehát nincs egzakt multikolinearitás. Ha ezt nem teszük fel, akkor $(X^T X)$ nem lenne invertálható tetszőleges X esetén.
4. $\mathbb{E}[\epsilon|X] = 0$. Ez a feltevés (0 feltételes átlag) alapján a hibatagok átlaga 0. Ebből következik, hogy $\mathbb{E}(Y) = X\beta$.
5. $\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I$. Ezen feltevés szerint a hibatagok homoszkedasztikusak és autokorrelálatlanok.

OLS torzítatlanság

A Gauss-Markov feltevések (1.-4.) alapján $\hat{\beta}_{OLS}$ torzítatlan becslője β-nak. $\mathbb{E}[\hat{\beta}_{OLS}] - \beta = 0$

Bizonyítás

$$\mathbb{E}[\hat{\beta}_{OLS}] = \mathbb{E}[(X^T X)^{-1} X^T Y] = \mathbb{E}[(X^T X)^{-1} X^T (X\beta + \epsilon)]$$

$$= \mathbb{E}[(X^T X)^{-1} X^T X\beta] + \mathbb{E}[(X^T X)^{-1} X^T \epsilon]$$

- Ismerjük fel, hogy $(X^T X)^{-1} X^T X = I$, és $\mathbb{E}[\beta] = \beta$ így

$$= \beta + \mathbb{E}[(X^T X)^{-1} X^T \epsilon]$$

- A 4. Gauss-Markov feltevés ($\mathbb{E}(\epsilon|X) = 0$) alapján

$$= \beta$$

$$\implies \mathbb{E}[\hat{\beta}_{OLS}] - \beta = 0$$

OLS kovariancia mátrix

A Gauss-Markov feltevések (1.-5.) alapján $\mathbb{V}[\hat{\beta}_{OLS}] = \sigma^2(X^TX)^{-1}$

Bizonyítás

$$\mathbb{V}[\hat{\beta}_{OLS}] = \mathbb{E}[(\hat{\beta}_{OLS} - \mathbb{E}[\hat{\beta}_{OLS}])(\hat{\beta}_{OLS} - \mathbb{E}[\hat{\beta}_{OLS}])^T]$$

$$= \mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)^T]$$

$$= \mathbb{E}[[(X^TX)^{-1}X^T\epsilon] * [(X^TX)^{-1}X^T\epsilon)^T]]$$

$$= \mathbb{E}[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}]$$

$$= (\mathrm{X}^TX)^{-1}X^T\,\mathbb{E}[\epsilon\epsilon^T]X(X^TX)^{-1}$$

- Az 5. Gauss-Markov feltevés ($\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I$) alapján

$$= (\mathrm{X}^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}$$

$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

OLS becsült koefficiens eloszlás

Feltéve, hogy a Gauss-Markov (1.-5.) feltevések teljesülnek, és a hibatagok eloszlása normális, akkor

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X^T X)^{-1}).$$

A $\hat{\beta}$ normális eloszlású valószínüségi változók lineáris kombinációja, amiből következik a normalitása.

OLS konzisztencia

A Gauss-Markov feltevések (1.-3.) alapján az OLS-becslő konzisztens: $\hat{\beta}_{OLS} \xrightarrow{p} \beta$

Konvergencia (pongyola áttekintés)

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y = (\tfrac{1}{N} X^T X)^{-1} (\tfrac{1}{N} X^T Y)$$

Felismerve, hogy $X^T X = \sum_{i=1}^{N} \underline{x_i} * \underline{x_i}^T$, a WLLN segítségével belátható, hogy $\frac{1}{N} \sum_{i=1}^{N} \underline{x_i} * \underline{x_i}^T \xrightarrow{p} \mathbb{E}[X^T X]$ (analóg módon $X^T Y$ esetén is belátható). A konvergencia implikációkat felhasználva:

$$\hat{\beta}_{OLS} \xrightarrow{p} \mathbb{E}[(X^T X)^{-1} X^T Y]$$

Mivel tudjuk, hogy,

$$\mathbb{E}[\beta] = \mathbb{E}[(X^T X)^{-1} X^T Y]$$

ezért

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta$$

Gauss-Markov tétel

Gauss-Markov Theorem

Ha teljesül mindegyik feltétel (1.-5.), akkor a **lineáris, torzítatlan becslők körében** az OLS-becslő **minimális varianciájú** (azaz hatásos).

Más szavakkal elmondva, az OLS-becslő **B**est, **L**inear, **U**nbiased, **E**fficient becslő, tehát **BLUE** estimator.

Z-érték

A hipotézis, amit a becsült modell paraméterek egyikére (k) határozunk meg:

$$H_0 : \hat{\beta}_k = b_k$$

Ebben az esetben b_k előre ismert érték, amihez mért távolságot fogjuk mérni a **null hipotézis** keretében. A null hipotézist az **alternatív hipotéssiel** szemben vizsgáljuk ($H_1 : \hat{\beta}_k \neq b_k$), α szignifikancia szint mellett.

Z-value

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$(\hat{\beta}_k - b_k)|X \sim N(0, \sigma^2((X^T X)^{-1})_{kk}),$$

ahol $((X^T X)^{-1})_{kk})$ a $(X^T X)^{-1}$ mátrix főátlójának k-adik eleme.
Ekkor:

$$z_k \equiv \frac{\hat{\beta}_k - b_k}{\sqrt{\sigma^2((X^T X)^{-1})_{kk}}} \sim N(0, 1).$$

Z-érték tulajdonságai

A Z értéknek nagyon jó statisztikai tulajdonságai vannak, **feltéve ha** ismerjük a varianciát. Ekkor z_k

1. értéke a mintából kiszámolható (értsd, a nevező kiszámolható becslés nélkül),
2. eloszlása nem függ az X eloszlásától,
3. eloszlása előzetesen ismert, tehát nem függ ismeretlen paraméterektől.

Nuisance parameters: Azok az ismeretlen paraméterek, amelyktől a test statisztika eloszlása függ.

Mivel a gyakorlatban a σ^2 értékét nem ismerjük, ezért a kézenfekvő megoldás, hogy az ismeretlen variancia helyére az OLS becslésből számított minta varianciát (s^2) illesztjük.

$$s^2 = \frac{\epsilon^T \epsilon}{n-K},$$

ahol n a minta elemszáma, K pedig a becsült paraméterek száma.

T-érték

A minta variancia behelyettesítése után a teszt statisztika értékét t-értéknek nevezzük. Ebben az esetben a nevezőbe a β_k paraméter OLS elvű becsült értékének standard hiábja kerül, amit $SE(\hat{\beta}_k)$-val jelölünk.

$$SE(\hat{\beta}_k) \equiv \sqrt{s^2((X^T X)^{-1})_{kk})}$$

Mivel s^2 egy valószínűségi változó (hiszen a minta függvényében változik az értéke), ez a behelyettesítés megváltoztatja a teszt statisztika eloszlását. Szerencsére az új eloszlás szintén nem függ függ ismeretlen paraméterektől, sem X-től.

T-value

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$t_k \equiv \frac{\hat{\beta}_k - b_k}{SE(\hat{\beta}_k)} \equiv \frac{\hat{\beta}_k - b_k}{\sqrt{s^2((X^T X)^{-1})_{kk})}} \sim t(n-K). \tag{3}$$

T-test

A t-érték alapú null hipozézis tesztelést t-teszt-nek nevezzük. T-tesztet a következő módon végzünk:

1. Számold ki a t-értéket a (3) egyenlet alapján.
2. Számold ki a kritikus értéket: $(n-K)$ szabadságfokú t eloszlás inverz kumulatív eloszlás függvényébe (CDF^{-1} v. PPF) helyettesítsd be az $\alpha/2$ értékét (regressziós paramétereket jellemzően 2 oldalon tesztelünk). A kapott értéket $t_{\alpha/2}$-vel jelöljük.
3. Amennyiben az 1. lépésben számolt t-érték a 2. lépésben számolt $[-t_{\alpha/2}; t_{\alpha/2}]$ intervallumba esik, akkor a H_0 null hipotézist nem tudjuk elvetni. Ha az intervallumon kívül, akkor 1-α konfidencia szint mellett H_0 null hipotézist elvetjük.

Amennyiben a becsült paraméterre szeretnénk **konfidencia intervallumot** (CI) meghatározni, akkor a t-érték definíciójából kiindulva, és azt átrendezve megkapjuk az $[\hat{\beta}_k - SE(\hat{\beta}_k)t_{\alpha/2}; \hat{\beta}_k + SE(\hat{\beta}_k)t_{\alpha/2}]$ intervallumot, amely már $\hat{\beta}_k$ mértékegységében értelmezhető.

P-érték alapú döntéshozatal t-teszt esetén

A t-érték alapú null hipozézis tesztelést t-teszt-nek nevezzük. T-tesztet a következő módon végzünk:

1. Számold ki a t-értéket a (3) egyenlet alapján.
2. $(n - K)$ szabadságfokú t eloszlás kumulatív eloszlás függvényébe (CDF), helyettesítsd be az előző lépésben kapott t-értéket.
3. A 2. lépésben kapott értéket vond ki 1-ből, a kapott értéket szorozd meg 2-vel.

$$\text{p-value} = min[CDF_t(t); 1 - CDF_t(t)] * 2$$

4. Ha p $> \alpha$, akkor H_0-t nem tudjuk elvetni, ellenkező esetben, H_0-t elvetjük.

Lineáris hipotézisek

A null hipotézis nem feltetélenül 1 paraméterre fogalmazhatók meg, számos esetben a model több paraméterét egyszerre, azok **lineáris kombinációját** kívánjuk vizsgálni. Ebben az esetben egyenletrendszer formájában tesztelünk a következő módon:

$$H_0 : R\hat{\beta} = r, \tag{4}$$

ahol, R restrikciós mátrix $(m \times K)$ és r pedig a restrikciós egyenletrendszer skalárjainak vektora $(m \times 1)$.

Példa

Szeretnénk megvizsgálni egy olyan regressziót, amelyben 4 jobboldali változónk van (az első változó a konstans). Amennyiben azt vizsgáljuk, hogy $\beta_2 = \beta_3$ és $\beta_4 = 0$, akkor az a (4) egyenlet formájában a következő módon írható fel:

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

Wald statisztika

A wald statisztika az $R\hat{\beta}$ vektor és az r vektor közötti távolságot vizsgálja. Amennyiben a null hipotézis igaz, akkor az $R\hat{\beta} - r \approx 0$.

Wald statistics

Kis mintás környezetbe $R\hat{\beta} - r$ eloszlása:

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X^T X)^{-1} R^T)$$

Ez alapján a nullhipotézis H_0: $R\hat{\beta} - r = 0$, és az alternatív H_1: $R\hat{\beta} - r \neq 0$, a teszt statisztika értéke pedig:

$$W_{Infeasible} = \frac{(R\hat{\beta} - r)^T [R(X^T X)^{-1} R^T]^{-1} (R\hat{\beta} - r)}{\sigma^2} \sim \chi^2_m$$

Ez a statisztika közvetlen formában nem használható, mert a sokasági variancia ismeretlen, ezért ismét behelyettesítéssel kell élnünk.

Kis mintás Wald teszt

Ahhoz, hogy a Wald statisztika értéke kiszámolható legyen, be kell helyettesítenünk a minta varinaciát (s^2) a sokasági variancia helyére.

Finite-sample Wald test

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$W = \frac{(R\hat{\beta} - r)^T [R(X^T X)^{-1} R^T]^{-1} (R\hat{\beta} - r)/m}{s^2} \sim F_{m,n-K}.$$

A variancia behelyettesítés hatására megváltozott a Wald statisztika eloszlása, amelyre a hipotézisvizsgálat során oda kell figyelnünk. A hipotézis vizsgálat, a t-teszttel analóg módon végezhető el.

Illeszkedés

A becsült paraméterekre vonatkozó hipotézisvizsgálatok mellett azt is meg kell állapítanunk, hogy a modell milyen jól illeszkedik az adatokra. Erre lineáris regresszió során többek között a (Centrált) R^2 és a korrigált R^2 mutatók szolgálnak. Ezen mutatók azt regadják meg, hogy a **a minta variancia mekkora részét magyarázza a modell**.

Centered R^2 és adjusted R^2

$$R^2 = 1 - \frac{\epsilon^T \epsilon}{\tilde{Y}^T \tilde{Y}}$$

ahol $\tilde{Y} = Y - \overline{Y}$, tehát a célváltozó átlagtól szűrt értékei. Mivel az R^2 értéke minden hozzáadott változóval nő, ezért ezt az értéket korrigálni érdemes, hogy a modell bővítés mellett megmaradjon az összehasonlíthatóság.

$$Adj.R^2 = 1 - \frac{n-1}{n-K-1}(1-R^2),$$

R^2 mátrix műveletekkel

A centrált R^2 mutató felírható tisztán mátrix műveletek segítségével is:

$$
\begin{aligned}
R^2 &= 1 - \frac{\epsilon^T \epsilon}{\tilde{Y}^T \tilde{Y}} \\
&= 1 - \frac{Y^T (I - P_X)^T (I - P_X) Y}{Y^T (I - J)^T (I - J) Y} \\
&= 1 - \frac{Y^T (I - P_X) Y}{Y^T (I - J) Y} \\
&= 1 - \frac{Y^T (M_X) Y}{Y^T (I - J) Y},
\end{aligned}
$$

ahol I, P, J, M_X mátrixok idempotensek, és rendre az identity mátrixot, projekciós mátrixot, egyesek mátrixát és az annihilációs mátrixot jelölik.

Általánosított regressziós modell

Az 5. Gauss-Markov feltevés ($\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I$) szerint a hibatagok homoszkedasztikusak és autokorrelálatlanok. Ez a feltevés gyakran túlzottan szigorú, ezért most némelyest lazítunk rajta.

Ha a $\mathbb{E}[\epsilon\epsilon^T|X]$ mátrix főátlójának elemei nem azonosak, akkor a hibatagok nem homoszkedasztikusak, amennyiben a főátlón kívüli elemek különböznek nullától, akkor a hibatagok között korreláció áll fent.

Bármely tetszőleges pozitív skalár σ^2 mellett definiálható $V(X) \equiv \mathbb{E}[\epsilon\epsilon^T|X]/\sigma^2$ és feltesszük, hogy $V(X)$ nem szinguláris, és előre ismert. Ekkor:

$$\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 V(X),$$

ahol $V(X)$ n nem szinguláris mátrix.

Amennyiben az 5. Gauss-Markov feltevést felváltjuk az előző egyenlent eredményére ($\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 V(X)$) – amely annyit feltételez, hogy a feltételes második momentum nem szinguláris – akkor az úgynevezett **általánosított regressziós modellt** kapjuk.

Következmények

Azok a következtetések, amelyek az 5. feltevés meglétét alkalmazták nem lesznek validak az általánosított regressziós modell esetén. Tételesen:

1. Az OLS elvű becslés esetén a Gauss-Markov tétel nem érvényes, a

$$\hat{\beta} \equiv (X^T X)^{-1} X^T Y$$

   nem a legkisebb varianciájú a lineáris, torzítatlan becslők körében (**nem BLUE**)

2. A **t-érték nem követ t-eloszlást**, tehát a t-teszt nem valid
3. A **kis mintás Wald statisztika nem követ F-eloszlást**, tehát a F-teszt nem valid
4. Az **OLS becslő továbbra is torzítatlan**, mert a levezetéshez erre a feltevésre nem volt szükség (lásd itt)

G-M feltevések az általánosított regresszió esetén

Mivel $V(X)$ – továbbiakban V – szimmetrikus és pozitív szemidefinit, ezért létezik olyan n×n C mátix, amelyre igaz, hogy:

$$V^{-1} = C^T C.$$

Ez dekompozíció nem egyedi (*not unique*), tehát több C mátrixra is igaz. Ez alapján felírható egy új regressziós model a következő transzformációval:

$$\tilde{Y} \equiv CY, \tilde{X} \equiv CX, \tilde{\epsilon} \equiv C\epsilon.$$

1. **Modell linearitás**: $\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$
2. Az adatok a sokaság **véletlen mintái**.
3. $rank(\tilde{X}) = rank(X)$, mivel C nem szinguláris. $\tilde{X}$ **teljes oszlop rangú**.
4. $\mathbb{E}[\tilde{\epsilon}|\tilde{X}] = C\,\mathbb{E}[\epsilon|X] = 0$. A hibatagok átlaga transzformációt követően is 0.
5. $\mathbb{E}[\tilde{\epsilon}\tilde{\epsilon}^T|\tilde{X}] = \sigma^2 CVC^T = \sigma^2 I$.

GLS becslés

Mivel a transzformált modellre teljesülnek a Gauss-Markov feltevések, ezért teljesül rá a Gauss-Markov tétel, tehát az általánosított regressziós modellre alkalmazott OLS elvű becslő minimális varianciájú lesz a lineáris, torzítatlan becslők körében. Ezt a becslőt nevezzük **általánosított legkisebb négyzetek módszerének, vagy GLS elvű becslőnek**.

GLS coefficient estimation

$$\begin{aligned}\hat{\beta}_{GLS} &= (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y} \\ &= [(CX)^T(CX)]^{-1}(CX)^TCy \\ &= (X^TC^TCX)^{-1}(X^TC^TCy) \\ &= (X^TV^{-1}X)^{-1}(X^TV^{-1}y)\end{aligned}$$

GLS conditional varaince

$$\mathbb{V}(\hat{\beta}_{GLS}|X) = \sigma^2(X^TV^{-1}X)^{-1}$$

GLS tulajdonságok

Kismintás tulajdonságok:

1. A GLS elvű becslő torzítatlan ($\mathbb{E}(\hat{\beta}_{GLS}|X) = \beta$)
2. A feltételes variancia arányos V-vel:$\mathbb{V}(\hat{\beta}_{GLS}|X) = \sigma^2(X^T V^{-1} X)^{-1}$
3. A GLS elvű becslő hatékony, tehát minden lineáris, torzítatlan becslőnek legalább akkora a feltételes varianciája, mint a GLS feltételes varianciája.

Limitációk:

1. Amennyiben a 4. Gauss-Markov feltevés sérül ($\mathbb{E}(\tilde{\epsilon}|\tilde{X}) \neq 0$), úgy a GLS nem lesz torzítatlan. Ez a tuladonság az OLS esetén is fennáll, de utóbbi nagymintás tulajdonságai ezt a hátrányt orvosolják. A GLS esetében nincsenek ilyen pozitív nagymintás tulajdonságok.
2. A gyakorlatban nem ismerjük V mátrixot, és ezt becsülnünk kell. Ebben az esetben **Megvalósítható GLS/Feasible GLS** becslőről beszélünk.

FGLS becslés lépései

A Feasible GLS elvű becslés lépései:

1. Becsüld meg $\hat{\beta}$ koefficienseket OLS elvű becslővel.
2. A kapott koefficiens vektor segítségével számítsd ki a modell hibáit (ϵ).
3. A négyzetes hibákra becsülj egy kiegészítő regressziót ($\hat{\omega}$).
4. A becsült koefficiensek ($\hat{\omega}$) alapján elítsd elő $\hat{V}$ variancia mátrixot.
5. $\hat{V}$-t felhasználva transzformáld X és Y változókat ($\tilde{X}$, $\tilde{Y}$).
6. A transzformált változókon becsüld meg $\hat{\beta}_{GLS}$ koefficienseket OLS elvű becslővel.
7. $\hat{\beta}_{GLS}$ segítségével a transzformált változókból számold ki a hibatagokat

Likelihood függvény

Likelihood function

Ha $X_1, ..., X_n$ i.i.d. véletlen változók, és $f(x|\theta)$ eloszlás függvény (PDF). Ekkor a **likelihood függvény**t úgy definiáljuk, hogy

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i|\theta).$$

A **log-likelihood függvény**t pedig a következő módon írjuk fel:
$\ell(\theta) = log\mathcal{L}(\theta)$

Maximum Likelihood Estimator

A **maximum likelihood** becslő (MLE), jelölése $\hat{\theta}_n$, nem más, mint θ azon értéke, amely mellett a likelihood függvény $(\mathcal{L}(\theta))$ a legnagyobb.

Likelihood függvény: megjegyzés

Mivel a logaritmus egy szigorúan monoton növekvő transzformáció, ezért a **log-likelihood és a likelihood függvények ugyanazon θ értékek mellett veszik fel a szélső értékeiket**. Ebből következik, hogy a log-likelihood függvény maximumának megkeresése ugyanarra a válaszra vezet. A gyakorlatban gyakrabban alkalmazzuk a log-likelihoodot, mert könnyebben kiszámítható.

Ha $\mathcal{L}(\theta)$-t megszorozzuk bármely olyan tetszőleg pozitív c számmal, amely nem függ θ-tól, akkor a MLE nem változik. Emiatt a likelihood függvényben lévő konstans értékeket gyakran el is hagyjuk.

MLE tulajdonságok

Bizonyos regularitási feltételek teljesülése esetén az MLE elvű becslő olyan jó tulajdonságokkal rendelkezik, amelyek népszerű becslési módszerré teszik.

Tulajdonságok

1. Invariancia: Ha $\hat{\theta}_{MLE}$ a legjobb becslésünk θ-ra, akkor, $g(\hat{\theta}_{MLE})$ a a legjobb becslés $g(\theta)$-ra
2. Konzisztencia: $\hat{\theta} \xrightarrow{p} \theta$
3. Aszimptotikus normalitás: $\frac{\hat{\theta}-\theta}{se} \xrightarrow{d} N(0,1)$
4. Aszimptotikus hatékonyság

Regresszió

A hipotézis vizsgálat során, annak érdekében, hogy következtetni tudjunk a becsült paraméterekről, feltevéseket tettünk a hibatagok eloszlására vonatkozóan. A Maximum Likelihood (MLE) elvű becslés során ezt a feltevést a becslési folyamat során is felhasználjuk.

Ha feltesszük, hogy a hibatagok:

1. normális eloszlásúak
2. homoszkedasztikusak
3. feltételesen korrelálatlanok

akkor a likelihood függvény:

$$\mathcal{L}(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp -\frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2}.$$

A log-likelihood függvény:

$$\ell(\beta, \sigma^2) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2}$$

Elsőrendű feltételek

A log-likelihood függvény maximuma az ismeretlen paraméterek függvényében úgy számolható ki, hogy vesszük azok parciális deriváltjait:

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = \frac{X^T(Y - X\hat{\beta})}{\sigma^2} = 0$$

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}{2\hat{\sigma}^4} = 0.$$

Ezen egyenleteket felhasználva arra a megoldásra jutunk, hogy:

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2_{MLE} = n^{-1}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = n^{-1}\epsilon^T\epsilon$$

Ahogyan látható, a MLE elven becsült regressziós koefficiensek **megegyeznek** az OLS elven becsült koeffíciensekkel, azonban a varianciák különböznek. Az MLE a korrigálatlan minta varianciát adja vissza. (Emiatt lesz csak aszimptotikusan torzítatlan a becslés)

Hatékonyság

Best Unbiased Estimator

Amennyiben **feltesszük**, hogy a Gauss-Markov feltevések teljesültek, és a hibatag normális eloszlású, akkor

$$\hat{\beta} = \hat{\beta}_{MLE}$$

a legkisebb varianciájú β torzítatlan becslői körében.

A Gauss-Markov tételhez képest ezen tétel kis különbsége annál lényegesebb. A MLE elvű becslés esetén - a feltételek teljesülése mellett - már nem csak a lineáris, hanem a nem-lineáris torzítatlan becslők körében sem találunk kisebb varianciájut.

Bootstrap elmélet

Számos olyan eset lehet, hogy egy minta statisztikának (pl. átlag, medián) valamilyen módon szeretnénk számszerűsíteni a becslési bizonytalanságát. A bizonytalanságot megpróbálhatjuk leírni a standard hibával, vagy a konfidencia intervallummal. Az eddig tanult módszerek során, jellemzően előzetes feltevést tettünk a hibatagok eloszlására, amelyek következtében analítikus megoldásra jutottunk. A **bootstrap** lehetőséget biztosít, hogy kevésbé szigorú feltevések mellett számszerűsítsünk bizonytalanságot, amelynek az ára, hogy **szimuláció**t kell futtatnunk.

Bootstrap elv

Legyen $T_n = g(X_1, X_2, ..., X_n)$ egy **statisztika**, tehát T_n az adat függvénye. Tegyük fel, hogy szeretnénk megismertni $\mathbb{V}_F(T_n)$-t, ami T_n varianciája, amely érték függ F ismeretlen eloszlástól. A bootstrap elméleti szempontból 2 lépésből áll:

1. Becsüld meg $\mathbb{V}_F(T_n)$-t $\mathbb{V}_{\hat{F}}(T_n)$ segítségével
2. Adj közelítést $\mathbb{V}_{\hat{F}}(T_n)$ értékére szimulációval

Szimuláció

Emlékeztető

Tegyük fel, hogy H eloszlásból húzunk $X_1, X_2, ..., X_J$ i.i.d. mintát. A nagyszámok gyenge törvénye alapján, amennyiben g függvény átlaga véges:

$$\frac{1}{J}\sum_{i=1}^{J} g(X_i) \xrightarrow{p} \int g(x)dH(x) = \mathbb{E}(g(X)),$$

ahogy $J \to \infty$.

Tehát, ha egy tetszőleges eloszlásból mintát veszünk, akkor kellően nagy minta esetén a minta átlaggal közelíteni tudjuk a várható értéket. Szimuláció során J-t (a minta méretét) olyan nagyra tudjuk állítani, amekkorára szeretnénk, h helyére pedig tetszőleges (*véges átlagú*) függvényt helyettesíthetünk be (pl.: variancia, csúcsosság, konfidencia intervallum).

Variancia becslés

A kérdés tehát az, hogy hogyan tudunk egy ismeretlen eloszlásból származó statisztikának kiszámolni a varianciáját? És hogyan tudjuk T_n eloszlását szimulálni, ha az adatok eloszlása $\hat{F}_n$ empirikus eloszlás?

A válasz, hogy $X_1^*, X_2^*, ..., X_n^*$ megfigyeléseket szimuláljuk $\hat{F}_n$-ből, és kiszámítjuk T_n^* értékét. Ez megfelel 1 húzásnak T_n eloszlásából. Az ötlet a következő:

$$\text{Ismeretlen DGP } F \implies X_1, X_2, ..., X_n \implies T_n = g(X_1, X_2, ..., X_n)$$
$$\text{Bootstrap } \hat{F}_n \implies X_1^*, X_2^*, ..., X_n^* \implies T_n^* = g(X_1^*, X_2^*, ..., X_n^*)$$

Az utolsó kérdés, hogy hogyan szimuláljuk $X_1^*, X_2^*, ..., X_n^*$ megfigyeléseket $\hat{F}_n$-ből? **Amennyiben $\hat{F}_n$ empirikus eloszlásból húzunk egy elemet, az megfelel annak ha az eredeti adatgeneráló folyamatból kapnánk a megfigyelést.** Tehát, ha $X_1^*, X_2^*, ..., X_n^*$ szeretnénk szimulálni, akkor elégséges, ha **visszatevéses mintavétel segítségével n darabos mintát** húzunk a számunkra rendelkezésre álló $X_1, X_2, ..., X_n$ adathalmazból. Ha ezt elvégezzük J alkalommal, akkor J darab n elemű **(bootstrap) minta** áll a rendelkezésünkre.

Variancia becslés lépései

Empirikus bootstrap steps

1. Húzz $X_1^*, X_2^*, ..., X_n^*$ megfigyelést $\hat{F}_n$ empirikus eloszlásból.
2. Számold ki $T_n^* = g(X_1^*, X_2^*, ..., X_n^*)$ értékét.
3. Ismételd meg az 1. és 2. lépést J alkalommal, hogy megkapd $T_{n,1}^*, T_{n,2}^*, ..., T_{n,J}^*$ statisztikákat.
4. Számold ki

$$\mathbb{V}_{boot} = \frac{1}{J} \sum_{i=1}^{J} (T_{n,i}^* - \overline{T_n^*})^2$$

Sematikusan ábrázolva a következő 2 lépcsős közelítéssel élünk:

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{\hat{F}}(T_n) \approx \mathbb{V}_{boot}(T_n^*)$$

Konfidencia intervallum becslés

A konfidencia intervallum becslésére 3 módszert vizsgálunk meg

1. Normális intervallum: A legegyszerűbb módszer, ahol felteszük, hogy T_n normális eloszlást közelít (gyakorlatban t eloszlás is használható)

$$CI = [T_n \pm z_{\alpha/2} sd_{boot}]$$

2. Pivot intervallum: Ez a módszer a valódi és a minta statisztika érték közötti különbséget a minta és a bootstrap minta különbségével közelíti

$$CI = [2T_n - T^*_{n,1-\alpha/2}, 2T_n - T^*_{n,\alpha/2}],$$

ahol a $T^*_{n,1-\alpha/2}, T^*_{n,\alpha/2}$ a bootrappelt minta T statisztikáinak **sorba rendezett elemei** közül az $1-\alpha/2, \alpha/2$ helyen megtalálható értékei.

3. Percentilis intervallum: Ezt **csak akkor** add megfelelő közelítést, ha a minta eloszlása közelíti a sokasági eloszlást

$$CI = [T^*_{n,\alpha/2}, T^*_{n,1-\alpha/2}]$$

Paired Bootstrap

Lineáris regresszió esetén is több lehetőségünk van a becsült paraméterek standard hibáinak / konfidencia intervallumainak a becslésére.

Párosított bootstrap (paired bootstrap): Az empirikus bootstrap során alkalmazott logikát használjuk, olyan módon, hogy $X_1^*, X_2^*, ..., X_n^*$ megfigylések helyett $(X_1^*, Y_1^*), (X_2^*, Y_2^*), ..., (X_n^*, Y_n^*)$ "párokat" húzunk, ahol X_i **lehet vektor is**. J bootstrap minta húzása után a következő struktúra áll rendelkezésünkre:

$$\begin{array}{cccc} (X_{1,1}^*, Y_{1,1}^*), & (X_{2,1}^*, Y_{2,1}^*), & \dots & (X_{n,1}^*, Y_{n,1}^*) \\ (X_{1,2}^*, Y_{1,2}^*), & (X_{2,2}^*, Y_{2,2}^*), & \dots & (X_{n,2}^*, Y_{n,2}^*) \\ \vdots & \vdots & & \vdots \\ (X_{1,J}^*, Y_{1,J}^*), & (X_{2,J}^*, Y_{2,J}^*), & \dots & (X_{n,J}^*, Y_{n,J}^*) \end{array}$$

A J darab n méretű bootstrap mintára futtatunk J számú regressziót. Ebből kapunk J számú $\hat{\beta}^*$ koefficiens vektort, amelyből már kiszámítatjuk a varianciát, az MSE-t és jellemzően normális intervallum módszerrel a konfidencia intervallumot. **Ez a módszer kifejezetten érzékeny az outlier-ekre**.

Residual Bootstrap

Residual bootstrap: A residual bootstrap során először megbecsüljük az eredeti adatokon a regressziót, majd kiszámoljuk a hibatagokat:

$$\epsilon_i = Y_i - X_i\hat{\beta}.$$

Ezt követően ismételen $(X_1^*, Y_1^*), (X_2^*, Y_2^*), ..., (X_n^*, Y_n^*)$ "párokat" húzunk, de Y_i^*-t nem közvetlenül vesszük, hanem a ϵ hibatagokat mintavétellezzük visszetevéssel. A mintapárokat a következő módon állítjuk elő:

$$X_i^* = X_i$$
$$Y_i^* = X_i\hat{\beta} + \epsilon_i^*$$

ahol ϵ_i^* értéket függetlenül mintavétellezük X_i-től. Ezt megismételjük J darab bootstrap mintára, majd futtatunk J darab regressziót, amelyekből kapott paraméter becslések alapján kiszámítjuk a számunkra érdekes bizonytalansági mutatót. **Ez a megoldás addig működik stabilan, ameddig a hibatagok homoszkedasztikusak.**

Wild Bootstrap

Heteroszkedaszticitás jelenléte esetén a residual bootstrapet átalakítjuk. Mivel a **hibatagok értéke függ X értékétől**, ezért nem mintavételezhetjük külön őket.

Y_i^* értékét továbbra is mi állítjuk elő, de a wild bootstrap során $V_1, V_2, ..., V_n \sim N(0,1)$ i.i.d. mintát generálunk, amelyet felhasználunk olyan módon:

$$X_i^* = X_i$$

$$Y_i^* = X_i\hat{\beta} + V_i * e_i,$$

ahol e_i az i-edik célváltozóból számolt hibatag X_i jobboldali változók segítségével. *Érdemes megjegyezni, hogy V_i eloszlása különbözhet a normálistól, itt a könnyebb érthetőség miatt alkalmaztuk $N(0,1)$ eloszlást.*

Block Bootstrap

Tegyük fel, hogy $X_1, X_2, ..., X_n$ nem i.i.d. minta, hanem idősor. Ebben az esetben X értékeinek **sorrendje nem elhanyagolható**, hiszen az egymáshoz közel szereplő értékek (index szerint közel) között **magas korreláció** is lehet. Ekkor a meglévő idősorunkat feldaraboljuk k hosszúságú átfedő idősor szeletre a következő struktúrában:

$$B_k = \begin{matrix} X_1, & X_2 & ... & X_k \\ X_2, & X_3 & ... & X_{k+1} \\ \vdots & \vdots & & \vdots \\ X_{n-k+1}, & X_{n-k+2} & ... & X_n \end{matrix}$$

Lehetőleg k értékét úgy válasszuk meg, hogy tetszőleges X_i és X_j ($|j-i| > m$) esetén a korreláció elhanyagolható legyen.

Ekkor az új bootstrap mintát úgy állítjuk elő, hogy B_k-ból húzunk n/k darab sort visszatevéses mintavétellel, amelyeket egymás után elhelyezünk és létrehozunk egy új $X_1^*, X_2^*, ..., X_n^*$ mintát. Ezt követően az empirikus bootstrapnél tanultak szerint J darab mintát hozunk létre, majd kiszámítjuk a számunkra releváns mutatókat.

Parametric Bootstrap

Az eddig áttekintett módszerek mindegyike az úgynevezett **nem-parametrikus bootstrap** kategóriába esett, mivel minden alkalommal $\hat{F}$ empirikus eloszlásból indultunk ki. **Parametrikus Bootstap** esetén feltevéssel élünk a sokaság eloszlását illetően.

Ebben az esetben a minta adatokat felhasználhatjuk, hogy a **sokasági eloszlás paramétereit megbecsüljük**, majd a becsült paramétereket (pl.: location $= \overline{X}$, scale $= \hat{\sigma}$, shape $= \hat{\alpha}$) behelyettesítjük a feltételezett H eloszlásba. Ezt követően mintát generálunk $X_1^*, X_2^*, ..., X_n^* \sim H(\overline{X}, \hat{\sigma}, \hat{\alpha})$, amely egy új bootstrap minta lesz. Ezen mintagenerálást belyehettesítjük az empirikus bootstrap során leírtak 1. lépésének helyébe, és analóg módon folytatjuk azt.

Köszönöm, és sok sikert a tantárgyhoz!