Többszörös lineáris regresszió

Matematikai statisztika
2024. október 28.

A regresszióanalízis feladata, célja

Feladat

Keressünk egy olyan **eredményváltozó** (Y) és **magyarázó változók** ($X_1, \ldots, X_p$) közötti kapcsolatot, amely előrejelzést adhat Y-re. Ezt a kapcsolatot egy f függvény segítségével modellezzük, amely meghatározza Y várható értékét adott $X_1, \ldots, X_p$ értékei mellett.

Cél

A cél az, hogy egy olyan $f(X_1, \ldots, X_p)$ függvényt találjunk, amely minimalizálja a négyzetes hibát, azaz:

$$\mathbb{E}((Y - f(X_1, \ldots, X_p))^2).$$

A négyzetes hibaminimalizálás révén a modell a legjobban illeszkedik az adatokhoz.

Függvénycsalád

A függvénycsalád kiválasztása

A f függvény olyan függvénycsaládból kerül ki, amely illeszkedik a probléma természetéhez. Tipikusan a következő lehetőségek közül választunk:

- **Lineáris függvénycsalád**: $f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, amely a lineáris regresszió alapja.
- **Polinomiális függvények**: $f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_1 X_1^2 + \cdots + \gamma_p X_p^2$, ahol magasabb rendű tagokat is tartalmazunk a nemlineáris kapcsolatok modellezésére.
- **Nemlineáris függvénycsaládok**: Ide tartozhatnak exponenciális, logaritmikus és más, bonyolultabb függvényformák, például $f(X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j \ln(X_j)$ vagy $f(X_1, \ldots, X_p) = \beta_0 + e^{\sum_{j=1}^{p} \beta_j X_j}$, ahol a kapcsolat nemlineáris.
- **Következtetési modellek**: Ilyenek például a döntési fák vagy neuronhálók, amelyek komplex kapcsolatok megtalálására alkalmasak a változók között.

A megfelelő függvénycsalád kiválasztása az adatok természetéhez és a vizsgálni kívánt kapcsolat komplexitásához igazodik. Mi a lineáris függvénycsaládd al foglalkozunk.

Egyszerű lineáris regresszió - modellfeltevések

Modell

Az egyszerű lineáris regresszió esetén:

$$Y = a + bX + \varepsilon,$$

ahol ε a hibatag.

Modellfeltevések

- $\mathbb{E}(\varepsilon) = 0$ (várható érték nulla),
- $\mathrm{Var}(\varepsilon) = \sigma^2$ (konstans szórás),
- ε-ek függetlenek és normális eloszlásúak.

Egyszerű lineáris regresszió - számítások

Számítások

Az egyszerű lineáris regresszió során a következőket számoltuk ki:

Részletek

- **Együtthatók pontbecslése**: a és b becslése a legkisebb négyzetek módszerével.
- **Hiba becslése**: $\hat{\sigma}$, a hibatag szórásának becslése.
- **Együtthatók eloszlása**: a és b becsléseinek normális eloszlása.
- **Intervallumbecslés**: Konfidenciaintervallum a és b számára.
- **Konfidenciaintervallum az előrejelzésre**: A becslés pontosságát kifejező intervallum.
- **Előrejelzési intervallum**: Az egyes új megfigyelések várható értéke.

ELR -Illeszkedésdiagnosztika

Determinációs együttható (R^2)

Az R^2 megmutatja, hogy a modellünk mennyire képes magyarázni az Y változó varianciáját.

Formulák

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

ahol SSE a maradéknégyzetösszeg, SST pedig az összes variancia.

ELR - Modelldiagnosztika

Együtthatók tesztelése

Nullhipotézisek megfogalmazása az egyes együtthatók (a, b) szignifikanciájának ellenőrzésére.

Modellfeltevések tesztelése

Ellenőrizzük az alábbi feltételeket:

- **Normalitás**: A hibák normális eloszlásúak-e?
- **Homogenitás**: A hibák varianciája állandó-e?
- **Függetlenség**: A megfigyelések között nincs-e autokorreláció?

Tesztelési eszközök

- χ^2-teszt
- Kolmogorov-Smirnov-próba
- egyéb tesztek
- Vizuális ellenőrzések (pl. normális Q-Q plot)

Többszörös lineáris regresszió - Elméleti modell (1. rész)

Feladat

Adottak az Y és $X_1, \ldots, X_k$ valószínűségi változók, amelyek közös eloszlásfüggvénnyel rendelkeznek. A cél az, hogy meghatározzuk a $\beta_0, \beta_1, \ldots, \beta_k$ paramétereket úgy, hogy az alábbi kifejezés minimális legyen:

$$\mathbb{E}(\varepsilon^2),$$

ahol $\varepsilon = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ az előrejelzési hiba. Ez a kifejezés a négyzetes hiba várható értékét adja meg, amelyet minimalizálni szeretnénk.

Értelmezés

Az ε a **hiba-tag**, amely azt mutatja meg, hogy mennyire tér el az Y megfigyelt érték az $X_1, \ldots, X_k$ változókkal előrejelzett értéktől. Az együtthatók $(\beta_0, \beta_1, \ldots, \beta_k)$ kiválasztásával minimalizálni kívánjuk a hiba négyzetes várható értékét.

Többszörös lineáris regresszió - Elméleti modell (2. rész)

Elnevezések és célok

- **Művelet célja**: Az optimális $\beta_0, \beta_1, \ldots, \beta_k$ paraméterek kiszámítása, amelyek minimalizálják a teljes négyzetes hibát.
- Y: **magyarázott változó** vagy eredményváltozó, amelyet előrejelezni szeretnénk.
- X_i: **magyarázó változók**, amelyek segítségével Y értékét próbáljuk modellezni.
- ε: **hiba**, amely Y és az előrejelzett érték különbsége.

Megoldási lépések

1. **Modell megalkotása**: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, ahol ε a hibatag.
2. **Hiba várható értékének minimalizálása**: Minimalizáljuk $\mathbb{E}(\varepsilon^2)$-t a β paraméterekre vonatkozóan.
3. **Parciális deriválás**: Kiszámítjuk a parciális deriváltakat a $\beta_0, \beta_1, \ldots, \beta_k$ változókra, majd egyenlővé tesszük nullával.
4. **Egyenletrendszer megoldása**: Megoldjuk az így kapott egyenletrendszert a paraméterekre.

Többszörös lineáris regresszió - Statisztikai modell

Statisztikai minta

Tekintsük a következő statisztikai mintát, amely n darab megfigyelésből áll:

$$(Y_1, X_{1,1}, X_{1,2}, \ldots, X_{1,k}), (Y_2, X_{2,1}, X_{2,2}, \ldots, X_{2,k}), \ldots, (Y_n, X_{n,1}, X_{n,2}, \ldots, X_{n,k}),$$

ahol minden Y_i az Y eloszlásból, $X_{i,1}$ pedig az X_1 eloszlásból származik. Itt Y_i az i-edik megfigyeléshez tartozó **magyarázott változó** értéke, míg $X_{i,j}$ az i-edik megfigyelés j-edik **magyarázó változója**.

Modell felírása

A statisztikai mintára a többszörös lineáris regressziós modell egyenletei így írhatók fel:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \ldots + b_k X_{i,k} + \varepsilon_i,$$

ahol $i = 1, \ldots, n$. Itt $b_0, b_1, \ldots, b_k$ az ismeretlen együtthatók, és ε_i az i-edik megfigyeléshez tartozó hiba, amely az előrejelzési modell és a tényleges Y_i érték közötti eltérést mutatja meg.

Gyakorlati példa

Tegyük fel, hogy egy lakás árát (Y) becsüljük különböző tényezők ($X_1, X_2, \ldots, X_k$) alapján, mint például:

- X_1: a lakás területe (négyzetméterben),
- X_2: a szobák száma,
- X_3: az emelet száma, stb.

A célunk egy olyan modell, amely ezeknek a tényezőknek a lineáris kombinációja segítségével becsli Y értékét. Ezzel a modellel közelítjük a lakás értékét, azaz:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \varepsilon,$$

ahol ε a hibatag, amely a tényleges lakásár és a modell által becsült érték közötti eltérést jelenti.

Többszörös lineáris regresszió - Modellfeltevések

Modellfeltevések

A többszörös lineáris regresszió során bizonyos statisztikai feltevéseket teszünk, amelyek szükségesek ahhoz, hogy a modell eredményei megbízhatóak legyenek. Ezek a feltevések biztosítják, hogy a becsült együtthatók értelmezhetők, és hogy a statisztikai tesztek érvényesek legyenek.

Feltevések 1-3

- **1. Lineáris modell**: Feltételezzük, hogy a magyarázott változó, Y, lineáris kapcsolatban áll a magyarázó változókkal. A modell formája:

$$Y_i = b_0 + b_1 X_{i,1} + \ldots + b_k X_{i,k} + \varepsilon_i,$$

  ahol $b_0, b_1, \ldots, b_k$ az ismeretlen együtthatók, és ε_i az i-edik megfigyeléshez tartozó hibatag. A lineáris kapcsolat feltételezése az alapja a lineáris regressziós modellnek.
- **2. Nincs egzakt multikollinearitás**: A magyarázó változók $(X_1, \ldots, X_k)$ nem lehetnek teljesen lineárisan függők, vagyis a $\mathbf{X}$ mátrix rangja $k+1$ kell legyen. Ez a feltétel biztosítja, hogy minden egyes magyarázó változó hozzáadott információt nyújtson a modell számára, és ne legyen köztük redundancia.
- **3. Erős exogenitás**: A hibák várható értéke adott magyarázó változók esetén nulla, azaz $\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0$. Ez azt jelenti, hogy a hibák nem függnek a magyarázó változóktól, így a magyarázó változók értékei nem befolyásolják a hibákat. Ez a feltétel garantálja, hogy az együtthatóink torzítatlan becslést adnak.

Többszörös lineáris regresszió - További feltevések

Feltevések 4-6

- **4. Homoszkedaszticitás**: A hibák varianciája állandó, függetlenül a magyarázó változók értékétől, azaz $\sigma^2(\varepsilon_i|\mathbf{X}) = \sigma_0^2$. Ezt az egyenletes szóródást nevezik homoszkedaszticitásnak, és biztosítja, hogy a becslések egyenletes pontossággal működjenek a különböző megfigyelési értékeknél.
- **5. Nincs autokorreláció**: Két különböző megfigyelés hibái függetlenek egymástól, azaz $\mathrm{Cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}) = 0$ minden $i \neq j$ esetén. Ez a feltétel különösen fontos idősoros adatoknál, ahol a hibák közötti kapcsolat torzíthatja az eredményeket. Ha nincs autokorreláció, akkor minden egyes megfigyelés hibája egyedileg járul hozzá a modellhez.
- **6. Normális eloszlású hibák**: A hibák normális eloszlásúak, azaz $\varepsilon_i \sim N(0, \sigma_0)$. Ez a feltétel elsősorban a statisztikai tesztek szempontjából fontos, mivel biztosítja, hogy a becslések eloszlása normális legyen, ami lehetővé teszi a pontos konfidenciaintervallumok és hipotézisvizsgálatok elvégzését.

Többszörös lineáris regresszió - A minta és az egyenletek felírása

Adatmodell

A regressziós elemzés során rendelkezésünkre áll n darab megfigyelés a magyarázott változóra és a magyarázó változókra vonatkozóan. Jelölje $Y = (y_1, y_2, \ldots, y_n)$ a magyarázott változó vektorát, azaz a megfigyelt értékeket, amelyeket előre szeretnénk jelezni. A magyarázó változók $X_1, X_2, \ldots, X_k$, amelyek a modell paramétereinek becsléséhez szükségesek.
A célunk az, hogy megtaláljuk az optimális $b_0, b_1, \ldots, b_k$ becsléseket úgy, hogy a modell a lehető legjobban illeszkedjen a megfigyelésekhez. Ezt az illeszkedést úgy érjük el, hogy minimalizáljuk a **reziduális négyzetösszeget**, amely az alábbi formában írható fel:

$$\text{Reziduális négyzetösszeg} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

ahol $\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_k x_{i,k}$ az i-edik megfigyelés előrejelzett értéke.

Többszörös lineáris regresszió - A minta és az egyenletek felírása

A modell adatos alakja

A többszörös lineáris regressziós modell minden egyes megfigyelés esetén a következő formát ölti:

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_k x_{i,k} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Itt:

- y_i: Az i-edik megfigyelés magyarázott változójának tényleges értéke.
- $x_{i,1}, x_{i,2}, \ldots, x_{i,k}$: Az i-edik megfigyelés magyarázó változóinak értékei, amelyekkel a magyarázott változót előrejelzünk.
- $b_0, b_1, \ldots, b_k$: Az ismeretlen paraméterek (együtthatók), amelyeket a modell illesztésével becslünk meg.
- ε_i: Az i-edik megfigyelés hibája vagy reziduálisa, amely a tényleges és a becsült érték közötti különbséget jelenti.

A hibatag, ε_i, az alábbi feltételeknek felel meg a lineáris regressziós modell feltevései alapján:

- A hibák várható értéke nulla, azaz $\mathbb{E}(\varepsilon_i) = 0$.
- A hibák varianciája konstans, vagyis $\mathrm{Var}(\varepsilon_i) = \sigma^2$, és függetlenek a magyarázó változók értékeitől (homoszkedaszticitás).
- A hibák függetlenek egymástól, azaz $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ minden $i \neq j$ esetén.
- A hibák normális eloszlásúak: $\varepsilon_i \sim N(0, \sigma^2)$, ami lehetővé teszi a klasszikus statisztikai teszteket.

Többszörös lineáris regresszió - Modell mátrix alakban (1. rész)

Mátrix forma

A többszörös lineáris regressziós modell kényelmesen ábrázolható mátrix formában, ami leegyszerűsíti a paraméterek számítását és az összefüggések átláthatóságát.

Jelölések mátrix formához

A következő jelöléseket használjuk a modell mátrix alakú felírásához:

- $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ az **eredményváltozók vektora**, amely az összes megfigyelt y_i értéket tartalmazza.
- $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$ a **paraméterek vektora**, amelyek a modell együtthatóit jelentik. Ezek az értékek adják meg, hogyan kapcsolódnak a magyarázó változók az eredményváltozóhoz.

Többszörös lineáris regresszió - Modell mátrix alakban (2. rész)

További jelölések

- $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ a **hibák vektora**, amely a megfigyelt és a modell által becsült értékek közötti eltéréseket tartalmazza.
- $\mathbf{X}$ egy $n \times (k+1)$-es mátrix, amely az **magyarázó változókat** tartalmazza, és az első oszlopa minden esetben 1-esekből áll, hogy a konstans tagot (az b_0 paramétert) is figyelembe vegyük:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{bmatrix}.$$

A modell mátrix alakja

A fenti jelölésekkel a többszörös lineáris regressziós modell egyszerűen felírható mátrix formában:

$$\mathbf{y} = \mathbf{Xb} + \varepsilon.$$

Ez az egyenlet azt fejezi ki, hogy az eredményváltozók vektorát a magyarázó változók mátrixa és a paraméterek vektora lineáris kombinációja, plusz a hibatagok vektora adja meg.

Többszörös lineáris regresszió - Megoldás 1. lépés: Reziduális hiba

Reziduális hiba (reziduum), $\varepsilon(\mathbf{b})$

Tegyük fel, hogy $\mathbf{b}$ egy tetszőleges becslővektor, amely a modell paramétereit tartalmazza. A reziduális hiba (vagy reziduum) kifejezése az alábbi:

$$\varepsilon(\mathbf{b}) = \mathbf{y} - \mathbf{X}\mathbf{b}.$$

Ez a kifejezés a tényleges megfigyelések ($\mathbf{y}$) és a modell által adott előrejelzések ($\mathbf{X}\mathbf{b}$) különbségét adja. A reziduális hiba mutatja meg, hogy az egyes megfigyelt értékek mennyire térnek el a becsült értékektől.

Hibanégyzetösszeg minimalizálása

A legkisebb négyzetek módszerének célja a reziduális hibák négyzetösszegének minimalizálása, amelyet az alábbi függvénnyel jelölünk:

$$V(\mathbf{b}) = \varepsilon(\mathbf{b})^T \varepsilon(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}).$$

A minimalizálás során $V(\mathbf{b})$ függvényt optimalizáljuk úgy, hogy a deriváltat nullára állítjuk, és így találjuk meg az optimális $\mathbf{b}$ becslést.

Többszörös lineáris regresszió - Megoldás 2. lépés: Első és második deriváltak

Első derivált

A $V(\mathbf{b})$ függvény minimalizálásához először vesszük az első deriváltat $\mathbf{b}$ szerint, majd nullára állítjuk, hogy megtaláljuk a minimumot:

$$\frac{\partial V(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = 0.$$

Az egyenlet bal oldalán található két kifejezés a megfigyelési és a becslési értékek közötti kapcsolatot tükrözi.

Második derivált (Hess-mátrix)

A második derivált, más néven a Hess-mátrix, az alábbi alakot ölti:

$$\frac{\partial^2 V(\mathbf{b})}{\partial \mathbf{b}^2} = 2\mathbf{X}^T\mathbf{X}.$$

Ez a mátrix pozitív szemidefinit (≥ 0), ami garantálja, hogy a minimalizálás valóban minimumot ad. Ezáltal biztosítva van, hogy a megoldásunk stabil és érvényes.

Többszörös lineáris regresszió - Legkisebb négyzetes megoldás

Becslés megoldása

Az első deriált nullára állítása után az egyenletet rendezve kapjuk a legkisebb négyzetes megoldást, amely a paraméterek optimális becslését adja:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Ez a képlet akkor alkalmazható, ha $\mathbf{X}^T\mathbf{X}$ invertálható, vagyis ha $\mathbf{X}$ teljes rangú. Ez biztosítja, hogy minden magyarázó változó hozzáadott információt nyújt a modell számára, és nincsenek redundáns magyarázó változók.

Megjegyzés

Az így kapott $\hat{\mathbf{b}}$ becslővektor tartalmazza azokat az együtthatókat, amelyek minimalizálják a reziduális hibák négyzetösszegét, így optimális illeszkedést biztosítanak a megfigyelési adatokhoz.

Hat mátrix és a becsült értékek meghatározása

A többszörös lineáris regressziós modellben a megfigyelési vektor ($\mathbf{y}$) becsült értékei $\hat{\mathbf{y}}$-nal jelölhetők. Ezek az értékek a becsült paramétervektor ($\hat{\mathbf{b}}$) segítségével határozhatók meg:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Ezzel a kifejezéssel minden megfigyeléshez hozzárendeljük a modell által becsült y értékeket, amelyek a magyarázó változók lineáris kombinációjából származnak.

Projekciós mátrix, P

A $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ kifejezést **hat mátrixnak** nevezzük, és jelöljük P-vel:

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Ez a mátrix projekciós mátrix, amelynek fontos tulajdonsága, hogy $P^2 = P$, ami azt jelenti, hogy P **idempotens**. A projekciós mátrix alkalmazása $\mathbf{y}$-ra az $\mathbf{X}$ által generált oszlopteret adja meg, így $\mathbf{y}$-t az $\mathbf{X}$ oszloptere által generált altérre vetíti.

Többszörös lineáris regresszió - Együtthatók jelentése

Együtthatók értelmezése a regressziós modellben

A többszörös lineáris regressziós modell becsült együtthatói ($\hat{\mathbf{b}}$) fontos információkat nyújtanak arról, hogyan befolyásolják a magyarázó változók az eredményváltozót (Y).

Magyarázó változók együtthatói ($\hat{b}_i$, $i \geq 1$)

Az egyes magyarázó változókhoz rendelt együtthatók ($\hat{b}_i$) értelmezése: Ha a X_i magyarázó változó értéke egy egységgel nő, miközben a többi változó értéke változatlan, akkor az eredményváltozó (Y) várhatóan $\hat{b}_i$ értékkel változik. Ez az együttható tehát megmutatja az X_i változó **marginalis hatását** az Y változóra.

Konstans ($\hat{b}_0$) értelmezése

A $\hat{b}_0$ együttható (intercept) az Y változó becsült értékét jelenti, amikor minden magyarázó változó értéke nulla. Más szóval, $\hat{b}_0$ az Y várható értékét mutatja meg az $X_1, X_2, \ldots, X_k$ változók nulla értéke esetén, így ez az érték egy **alapállapotot** képvisel a modellben.

A skálaproblematika és a standardizálás szükségessége

Különböző skálájú változók esetén az együtthatók összehasonlítása nehézséget okoz. Például, egy kis együtthatóval rendelkező változó jelentős hatással lehet a magyarázott változóra, ha az X_i változó nagyságrendekkel nagyobb értékeket vesz fel, mint más magyarázó változók. Ezt a **skálaproblematikát** oldja meg a standardizálás, hiszen így az együtthatók hatása közvetlenül összevethető.

Bevezetés: Standardizált együtthatók (Béta együtthatók)

A standardizált együtthatókat, vagy más néven **béta együtthatókat** azért vezetjük be, hogy a különböző skálájú változók hatásait összehasonlíthatóvá tegyük a modellben. Az itt szereplő β-k konkrét számértékeket jelölnek, ezért jelölésben néha zavart okozhatnak, de ezek mindig standardizált értékeket képviselnek.

TLR - Standardizált együtthatók 2.

Definíció

A standardizált együttható $\hat{\beta}_i$ kiszámítása az alábbi módon történik:

$$\hat{\beta}_i = \frac{s_i}{s_Y} \hat{b}_i,$$

ahol:

- s_i az X_i magyarázó változó empirikus szórása,
- s_Y az Y eredményváltozó empirikus szórása.

Jelentés és interpretáció

A standardizált együtthatók segítenek megérteni, hogy egy adott változó (skálától függetlenül) mennyire befolyásolja a célváltozót. Minél nagyobb az adott $\hat{\beta}_i$ értéke, annál jelentősebb hatása van az X_i változónak Y-ra.

OLS becslés tulajdonságai 1. rész - Torzítatlanság

Torzítatlanság

Az OLS (Ordinary Least Squares) becslés egyik legfontosabb tulajdonsága, hogy **torzítatlan**. Ez azt jelenti, hogy a becslés várható értéke megegyezik a valódi paraméterértékekkel, azaz:

$$\mathbb{E}(\hat{\mathbf{b}}) = \mathbf{b}.$$

Ez a tulajdonság azt biztosítja, hogy hosszú távon, számos minta alapján az OLS becslések átlaga pontosan a valódi paraméterértékeket adja vissza. Más szavakkal, az OLS becslés **nincs szisztematikusan eltolva** a valódi értékekhez képest.

Torzítatlanság feltétele

A torzítatlanság feltétele, hogy a hibatagok várható értéke nulla legyen és független legyen a magyarázó változóktól, azaz $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$. Ha ez a feltétel teljesül, akkor az OLS becslő várható értéke a valódi paraméterekre áll be.

OLS becslés tulajdonságai 2. rész - Konzisztencia

Konzisztencia

Az OLS becslés **konzisztens**, ami azt jelenti, hogy ahogy a minta elemszáma, n, növekszik, a becsült paraméterek ($\hat{\mathbf{b}}$) egyre pontosabban közelítik a valódi $\mathbf{b}$ paramétereket. Formálisan:

$$\lim_{n \to \infty} \hat{\mathbf{b}} = \mathbf{b}.$$

Ez a tulajdonság garantálja, hogy elegendően nagy minta esetén az OLS becslés megbízható eredményeket ad.

Konzisztencia feltételei

A konzisztencia feltételei közé tartozik, hogy a hibatagok várható értéke nulla legyen, és a magyarázó változók eloszlásának legyen egy megfelelő, stabil struktúrája nagy minta esetén. Ezenkívül a magyarázó változók függetlensége a hibatagoktól is fontos a konzisztencia biztosításához.

OLS becslés tulajdonságai 3. rész - Hatásosság

Hatásosság

A **Gauss-Markov tétel** szerint, amennyiben a hibatagok várható értéke nulla, varianciájuk állandó (homoszkedaszticitás), és nincsenek korrelálva egymással, akkor az OLS becslés **hatásos** az összes torzítatlan lineáris becslés között. Ez azt jelenti, hogy:

$$\text{Var}(\hat{\mathbf{b}}) \leq \text{Var}(\tilde{\mathbf{b}})$$

minden más, torzítatlan becslővektor $\tilde{\mathbf{b}}$ esetén.

Hatásosság következménye

A hatásosság azt biztosítja, hogy az OLS becslésnek a legkisebb varianciája van a torzítatlan becslések között, vagyis **leghatékonyabb**. Ez különösen fontos a becslések megbízhatósága szempontjából, hiszen kisebb variancia nagyobb pontosságot jelent.

OLS becslés tulajdonságai 4. rész - Normalitás

Normalitás

Ha a hibatagok normális eloszlásúak, akkor az OLS becslések is normális eloszlást követnek. Ekkor a becsült paramétervektor ($\hat{\mathbf{b}}$) eloszlása:

$$\hat{\mathbf{b}} \sim N\left(\mathbf{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

Ez az eloszlás lehetővé teszi, hogy konfidenciaintervallumokat és hipotézisvizsgálatokat végezzünk a becsült paraméterekre.

Normalitás nagy mintaszám esetén

Ha a hibatagok eloszlása nem normális, nagy mintaszám esetén a **centrális határeloszlás-tétel** alapján az OLS becslések közel normális eloszlást követnek. Ez nagy mintaszám esetén szintén biztosítja a statisztikai tesztek érvényességét és a konfidenciaintervallumok használatát.

Hiba varianciájának becslése, $\hat{\sigma}^2$

A modell hibájának varianciáját, azaz a reziduális varianciát σ^2 becslésére az alábbi formulát használjuk:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

ahol:

- n: az összes megfigyelés száma,
- k: a magyarázó változók száma,
- $(y_i - \hat{y}_i)^2$: az i-edik megfigyelés reziduális négyzete, amely a valódi és a becsült érték eltérését méri.

Az $n - k - 1$ a szabadságfok, ami a paraméterek miatt történő korrekciót veszi figyelembe.

A hiba becslésének szerepe

A hiba becslés fontos szerepet játszik a konfidenciaintervallumok és az előrejelzési intervallumok meghatározásában. Meghatározza az intervallumok szélességét és a modell illesztésének pontosságát, ami kulcsfontosságú az eredmények megbízhatósága szempontjából.

Többszörös lineáris regresszió - Együtthatók eloszlása és intervallumbecslés

Együtthatók eloszlása

Ha a hibatagok normális eloszlásúak, akkor az OLS-becslésből származó együtthatók, $\hat{\mathbf{b}}$, is normális eloszlást követnek:

$$\hat{\mathbf{b}} \sim N\left(\mathbf{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

Ez azt jelenti, hogy $\hat{\mathbf{b}}$ várható értéke a valódi paramétervektor, $\mathbf{b}$, varianciája pedig $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Ezen eloszlás ismerete lehetővé teszi, hogy konfidenciaintervallumokat és hipotézisvizsgálatokat végezzünk az együtthatókra.

Konfidenciaintervallum az együtthatókra

Az egyes együtthatók ($\hat{b}_j$) $100(1-\alpha)\%$-os konfidenciaintervalluma az alábbiak szerint adható meg:

$$\hat{b}_j \pm t_{\alpha/2,n-k-1} \cdot \sqrt{\hat{\sigma}^2 \cdot \left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}},$$

ahol:

- $\hat{\sigma}^2$: a modell hiba varianciájának becslése,
- $\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}$: a kovarianciamátrix főátlójának j-edik eleme, amely $\hat{b}_j$ varianciáját adja.

Többszörös lineáris regresszió - Konfidenciaintervallum és előrejelzési intervallum

Konfidenciaintervallum az előrejelzésre

Egy adott $X = \mathbf{x}_0$ értéknél az Y várható értékére vonatkozó $100(1-\alpha)\%$-os konfidenciaintervallum:

$$\hat{y}_0 \pm t_{\alpha/2,n-k-1} \cdot \hat{\sigma} \cdot \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0},$$

ahol $\hat{y}_0 = \mathbf{x}_0^T \hat{\mathbf{b}}$ a becsült érték. Ez az intervallum az Y várható értékének bizonytalanságát méri a kiválasztott $\mathbf{x}_0$ pontban.

Előrejelzési intervallum egy új megfigyelésre

Az előrejelzési intervallum szélesebb, mivel a modell hibavarianciáját is tartalmazza. Egy új y_0 megfigyelés előrejelzési intervalluma:

$$\hat{y}_0 \pm t_{\alpha/2,n-k-1} \cdot \hat{\sigma} \cdot \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}.$$

Az előrejelzési intervallum figyelembe veszi az új megfigyelés bizonytalanságát és a modell illeszkedési hibáját is, így nagyobb lefedettséget biztosít.

Lineáris regresszió R-ben - Nem standardizált együtthatók

Adatok generálása és a modell futtatása

Először generáljunk adatokat a modellhez és illesszünk egy lineáris regressziós modellt.

```
# Adatok generálása
set.seed(123)
n <- 100
X1 <- rnorm(n, mean = 10, sd = 2)
X2 <- rnorm(n, mean = 20, sd = 5)
Y <- 4.32 + 0.2845 * X1 + 0.6953 * X2 + rnorm(n, mean = 0, sd = 1.5)

# Adatok adatkeretben
data <- data.frame(Y = Y, X1 = X1, X2 = X2)

# Lineáris regressziós modell illesztése
model <- lm(Y ~ X1 + X2, data = data)
summary(model)
```

R output - Nem standardizált együtthatók

A `summary(model)` parancs eredménye:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.3206     1.2974    3.33   0.0013 **
X1            0.2845     0.0556    5.12   <2e-16 ***
X2            0.6953     0.0234   29.73   <2e-16 ***
```

Lineáris regresszió R-ben - Standardizált együtthatók számítása

Standardizált változók és együtthatók

A standardizált együtthatók számításához először a változókat standardizáljuk.

```
# Standardizált változók létrehozása
data_std <- as.data.frame(scale(data))

# Standardizált modell illesztése
model_std <- lm(Y ~ X1 + X2, data = data_std)
summary(model_std)
```

R output - Standardizált együtthatók

A `summary(model_std)` parancs eredménye:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0          0        NA        NA
X1             0.3087     0.0584    5.28     <2e-16 ***
X2             0.7051     0.0223   31.63     <2e-16 ***
```

Értelmezés

A standardizált együtthatók lehetővé teszik a magyarázó változók relatív hatásának összehasonlítását az eredményváltozóra. Az X_2 változó hatása erősebb az Y-ra, mint X_1-é.

Hiba varianciájának becslése

A hiba varianciájának becslését az alábbi módon kaphatjuk meg.

```
# Hiba varianci j nak becsl se
sigma_squared <- sum(residuals(model)^2) / (n - length(model$coefficients))
sigma_squared
```

R output - Hiba varianciája

```
[1] 2.1
```

Hiba varianciájának becslése és konfidenciaintervallumok R-ben

Konfidenciaintervallumok az együtthatókra

Az együtthatók 95%-os konfidenciaintervallumai R-ben:

```
confint(model, level = 0.95)
```

R output - Konfidenciaintervallumok

```
                2.5 %    97.5 %
(Intercept)    1.7508    6.8904
X1             0.1742    0.3948
X2             0.6487    0.7419
```

Új megfigyelés előkészítése

Készítsünk egy új megfigyelést, amelyhez előrejelzési és konfidenciaintervallumot számítunk:

```
# Uj megfigyeles
new_data <- data.frame(X1 = 12, X2 = 25)

# Elorejelzesi intervallum
predict(model, newdata = new_data, interval = "prediction", level = 0.95)
# Konfidenciaintervallum
predict(model, newdata = new_data, interval = "confidence", level = 0.95)
```

R output - Előrejelzési és konfidenciaintervallum

```
# Elorejelzesi intervallum:
       fit       lwr       upr
1 20.695   18.344    23.046

# Konfidenciaintervallum:
       fit       lwr       upr
1 20.695   19.253    22.137
```

Előrejelzési és konfidenciaintervallum egy új megfigyelésre R-ben

Értelmezés

Az előrejelzési intervallum szélesebb, mivel figyelembe veszi az új megfigyelés bizonytalanságát. A konfidenciaintervallum az Y várható értékének pontosságát mutatja adott $X_1 = 12$, $X_2 = 25$ mellett.