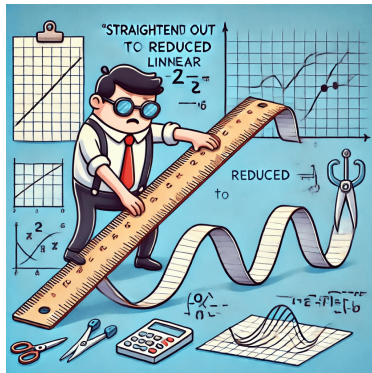


Regresszióanalízis: lineárisra visszavezethető regressziók

Matematikai Statisztika
2024. október 21.



Általános definíció

A lineárisra visszavezethető regresszió olyan modell, amely alapvetően a lineáris regresszióhoz hasonlóan működik, de a változók között nem feltétlenül lineáris kapcsolat áll fenn. Az ilyen modelleknél különböző transzformációk alkalmazásával a nemlineáris kapcsolatokat is lineárisá tehetjük, így azokat a klasszikus lineáris regresszió módszereivel vizsgálhatjuk.

Lineáris modell és a transzformációk

A lineáris regresszió általános formája:

$$Y = \alpha + \beta X + \varepsilon$$

ahol Y a függő változó, X a magyarázó változó, α és β a paraméterek, és ε a hibatag.

A lineárisra visszavezethető modellekben különböző transzformációkat alkalmazunk a magyarázó vagy függő változókra, hogy lineáris kapcsolatot nyerjünk.

Logaritmikus regresszió

Ha a kapcsolat a magyarázó és a függő változó között exponenciális jellegű, logaritmikus transzformációt alkalmazhatunk. Ekkor a modell:

$$\log(Y) = \alpha + \beta X + \varepsilon$$

Ez akkor alkalmazható, ha az eredményváltozó szórása növekszik a magyarázó változóval.

Polinomiális regresszió

Amikor a változók közötti kapcsolat nem lineáris, de jól közelíthető polinomiális függvényvel, akkor polinomiális regressziót alkalmazunk. Például egy másodfokú modell:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Ez a modell lineáris az α , β_1 , és β_2 paraméterekre nézve, így a klasszikus lineáris regresszió módszerével becsülhető.

Log-lineáris modell

A log-lineáris modell olyan esetekben hasznos, amikor mind a magyarázó, mind a függő változó logaritmikus transzformációval lineárisra tehető:

$$\log(Y) = \alpha + \beta \log(X) + \varepsilon$$

Ez a modell jól alkalmazható például gazdasági adatok esetén, ahol a változók között multiplikatív kapcsolatok lehetnek.

Inverz regresszió

Az inverz regressziót akkor alkalmazzuk, ha az X változó növekedésével a Y csökken. Ekkor a modell így néz ki:

$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$

Ez a modell is lineárisra visszavezethető, mivel a $\frac{1}{X}$ lineáris alakban szerepel.

Lineárisra visszavezethető regressziók

A lineárisra visszavezethető regressziók olyan esetekben hasznosak, amikor a változók között nemlineáris kapcsolat van, de az megfelelő transzformációkkal lineárisrá tehető. A leggyakrabban alkalmazott transzformációk a logaritmikus, polinomiális és inverz transzformációk, amelyekkel a klasszikus lineáris regressziós módszereket használhatjuk.

Alkalmazások

Ezek a modellek számos területen használhatók, például gazdasági modellezésben, biológiai adatok elemzésében vagy akár fizikai folyamatok leírásában, ahol a változók közötti kapcsolat nemlineáris, de transzformálható.

Miért hasznos a logaritmikus transzformáció?

Amikor a függő változó (Y) gyors növekedést mutat a magyarázó változó (X) értékével, gyakran érdemes logaritmikus transzformációt alkalmazni, hogy a modell lineárisabb legyen.

Például: ha Y értékei kezdetben gyorsan nőnek, de később a növekedés üteme lassul, egy logaritmikus transzformáció javíthatja a modell előrejelző képességét.

Alapmodell

A logaritmikus regresszió formája:

$$\log(Y) = \alpha + \beta X + \varepsilon$$

ahol α az y -tengely metszéspontja, β pedig a regressziós meredekség.

Példa: Értékesítés és reklámköltség kapcsolata

Vizsgáljuk egy vállalat reklámköltsége (X) és az értékesítés (Y) közötti kapcsolatot. A reklámra fordított költségek növelésével az értékesítés gyorsan növekedhet, de a növekedés lassuló tendenciát mutathat.

Adatok

Reklámköltség (X)	Értékesítés (Y)
1	10
2	15
3	19
4	22
5	23

A fenti adatok logaritmikus transzformáció után várhatóan egy lineárisabb kapcsolatot mutatnak.

Logaritmikus transzformáció alkalmazása

A Y változó logaritmusát vesszük, hogy a növekedést lineárisabb formában vizsgáljuk.

$$\log(Y) = \alpha + \beta X + \varepsilon$$

Az átalakított értékek így alakulnak:

Reklámköltség (X)	Értékesítés (Y)	$\log(Y)$
1	10	2.30
2	15	2.71
3	19	2.94
4	22	3.09
5	23	3.14

Becslés a transzformált adatokra

Az adatok logaritmikus transzformációja után lineáris regressziót alkalmazunk a $\log(Y)$ függvényre.

$$\log(\hat{Y}) = \alpha + \beta X$$

Az eredeti Y érték visszanyerése

A becsült $\log(Y)$ -ból vissza tudjuk nyerni az eredeti Y értékeket az exponenciális transzformációval:

$$\hat{Y} = e^{\alpha + \beta X}$$

Ez a modell lehetővé teszi az eredeti skálán való előrejelzést.

Végső előrejelzési formula

Az értékesítés becslése az alábbi képlettel történik:

$$\hat{Y} = e^{2.05 + 0.22X}$$

Miért használunk polinomiális regressziót?

Amikor a függő változó (Y) és a magyarázó változó (X) közötti kapcsolat nem lineáris, de görbe jellegű, polinomiális regressziót alkalmazunk. Ilyen helyzetekben magasabb fokú polinomot használunk a kapcsolat leírására.

Alapmodell

A másodfokú polinomiális regresszió általános formája:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

ahol α az y -tengely metszéspontja, β_1 a lineáris hatás, β_2 pedig a kvadratikus hatás.

Példa: Sebesség és fékút kapcsolata

Vizsgáljuk egy autó sebessége (X) és a fékút (Y) közötti kapcsolatot. Az autó sebességének növekedésével a fékút is nő, de nem lineáris módon, hanem négyzetes összefüggés alapján.

Adatok

Sebesség (km/h)	Fékút (m)
30	14
40	20
50	28
60	40
70	54

A fenti adatok alapján a másodfokú polinomiális regresszió modellje várhatóan pontosabban írja le a kapcsolatot.

Modell illesztése

A másodfokú polinomiális modell illesztéséhez az alábbi egyenletet használjuk:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Az adatpontok alapján számítjuk ki az α , β_1 és β_2 paramétereket.

Transzformált adatok

A modellben az X^2 kifejezés kvadratikus hatást vezet be:

Sebesség (X)	Fékút (Y)	X^2
30	14	900
40	20	1600
50	28	2500
60	40	3600
70	54	4900

Az előrejelzés modellje

A becsült paraméterek alapján a polinomiális modell végső formája:

$$\hat{Y} = 5 + 0.5X + 0.02X^2$$

Ez az egyenlet használható az autó sebessége és a fékút közötti kapcsolat előrejelzésére.

Előrejelzési példa

Egy 80 km/h sebességgel haladó autó esetében a fékút becslése:

$$\hat{Y} = 5 + 0.5(80) + 0.02(80^2) = 5 + 40 + 128 = 173 \text{ méter}$$

A polinomiális modell tehát jelentősen eltérhet a lineáristól, különösen nagyobb X értékeknél.

Log-lineáris modell

A log-lineáris modell alkalmazásával mind a magyarázó, mind a függő változót logaritmussal transzformáljuk. Ez akkor hasznos, ha a változók között multiplikatív kapcsolat van.

Példa: A cég éves bevétele (Y) és az alkalmazottak száma (X) közötti kapcsolat. Feltételezzük, hogy a bevétel az alkalmazottak számával arányosan növekszik, de a növekedés üteme csökken.

$$\log(Y) = \alpha + \beta \log(X) + \varepsilon$$

Adatok:

Alkalmazottak száma (X)	Bevétel (Y millió)
10	1
50	10
100	20
200	35
300	40

Ez a transzformáció segít lineáris kapcsolatot feltárni a nemlineáris adatok között.

Miért alkalmazunk log-lineáris modellt?

A log-lineáris modell alkalmazása akkor indokolt, ha a magyarázó (X) és függő (Y) változó közötti kapcsolat multiplikatív jellegű, és nem lineáris. Ilyen esetekben mindkét változót logaritmussal transzformáljuk, hogy a kapcsolat lineárisra váljon.

Alkalmazási példa: Amikor egy cég éves bevétele az alkalmazottak számával arányosan nő, de a növekedés üteme csökken, log-lineáris transzformációval egyszerűsíthetjük a kapcsolat modellezését.

Alapmodell

A log-lineáris modell általános formája:

$$\log(Y) = \alpha + \beta \log(X) + \varepsilon$$

ahol α az y -tengely metszéspontja, β a regressziós együttható, és ε a hibatag.

Példa: Alkalmazottak száma és bevétel kapcsolata

Egy vállalat éves bevétele (Y) és az alkalmazottak száma (X) közötti kapcsolatot vizsgáljuk. Feltételezzük, hogy a bevétel növekszik az alkalmazottak számával, de a növekedés üteme egyre kisebb lesz.

Adatok

Az alábbi adatok állnak rendelkezésünkre:

Alkalmazottak száma (X)	Bevétel (Y millió)
10	1
50	10
100	20
200	35
300	40

Ezek az adatok mutatják a nemlineáris kapcsolatot a magyarázó változó (X) és a függő változó (Y) között.

A logaritmikus transzformáció

Mindkét változóra (alkalmazottak száma és bevétel) logaritmikus transzformációt alkalmazunk, hogy a kapcsolat lineárisabbá váljon:

$$\log(Y) = \alpha + \beta \log(X) + \varepsilon$$

A transzformáció eredménye az alábbi táblázatban látható:

Alkalmazottak száma (X)	Bevétel (Y millió)	$\log(X)$	$\log(Y)$
10	1	2.30	0
50	10	3.91	2.30
100	20	4.61	3.00
200	35	5.30	3.56
300	40	5.70	3.69

Lineáris kapcsolat feltárása

A logaritmikus transzformáció után a kapcsolat közel lineáris lesz, ami lehetővé teszi a lineáris regresszió alkalmazását a transzformált adatokra.

A log-lineáris modell paramétereinek becslése

A logaritmikus transzformált adatokat használva illesztjük a modellt:

$$\log(Y) = \alpha + \beta \log(X)$$

Az illesztett paraméterek becslése:

$$\hat{\alpha} = -3.28 \quad \text{és} \quad \hat{\beta} = 1.15$$

Ez azt jelenti, hogy a becsült egyenlet így alakul:

$$\log(\hat{Y}) = -3.28 + 1.15 \log(X)$$

Az eredeti Y visszanyerése

A transzformált $\log(Y)$ -ból vissza tudjuk nyerni az eredeti Y értékeket:

$$\hat{Y} = e^{-3.28 + 1.15 \log(X)}$$

Az előrejelzés végső formája

Az illesztett log-lineáris modell alapján az alkalmazottak számának függvényében a bevétel becslése:

$$\hat{Y} = e^{-3.28} \cdot X^{1.15}$$

Ez a képlet lehetővé teszi a bevétel előrejelzését az alkalmazottak számának ismeretében.

Előrejelzési példa

Egy 500 alkalmazottal rendelkező vállalat becsült éves bevétele:

$$\hat{Y} = e^{-3.28} \cdot 500^{1.15} = 0.038 \cdot 1213.09 \approx 46.99 \text{ millió}$$

Ez a becslés közel 47 millió forintos bevételt jelent.

Inverz regresszió

Az inverz regressziót akkor alkalmazzuk, ha a függő változó csökken a magyarázó változó növekedésével. Ilyenkor az X reciprokát használjuk magyarázó változóként.

Példa: A munkavégzés hatékonysága (Y) és a dolgozók szüneteinek száma (X) közötti kapcsolat vizsgálata. Több szünet esetén a munka hatékonysága csökken.

$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$

Adatok:

Szünetek száma (X)	Hatékonyság (Y)
1	80
2	60
3	50
4	45
5	40

Ez az inverz modell jól leírja a csökkenő hatékonyság és a szünetek közötti kapcsolatot.

Mikor alkalmazzuk az inverz regressziót?

Az inverz regressziót akkor alkalmazzuk, ha a függő változó (Y) értéke csökken a magyarázó változó (X) növekedésével. Ilyen esetekben a magyarázó változó reciprokát használjuk, hogy lineárisabb kapcsolatot érjünk el.

Alapmodell

Az inverz regresszió általános formája:

$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$

ahol α az y -tengely metszéspontja, β pedig a reciprok változó hatásának nagyságát mutatja.

Példa: Hatékonyság és szünetek kapcsolata

Vizsgáljuk egy vállalat munkavégzésének hatékonyságát (Y) és a dolgozók által tartott szünetek számát (X). Több szünet esetén a munka hatékonysága csökken, és ezt az összefüggést az inverz regresszió segítségével modellezzük.

Adatok

Az alábbi adatok állnak rendelkezésünkre:

Szünetek száma (X)	Hatékonyság (Y)
1	80
2	60
3	50
4	45
5	40

Ez a minta azt mutatja, hogy a szünetek számának növekedésével a hatékonyság csökken.

Reciprok változó alkalmazása

Az inverz regresszióban a magyarázó változó reciprokát ($\frac{1}{X}$) használjuk, hogy lineáris kapcsolatot kapjunk. Az alábbi táblázat tartalmazza az eredeti adatokat és az átalakított reciprok változókat:

Szünetek száma (X)	Hatékonyság (Y)	Reciprok ($\frac{1}{X}$)
1	80	1.00
2	60	0.50
3	50	0.33
4	45	0.25
5	40	0.20

Lineáris kapcsolat feltárása

A reciprok transzformáció után a magyarázó változó ($\frac{1}{X}$) lineárisabb kapcsolatot mutat a függő változóval (Y), ami lehetővé teszi a lineáris regresszió alkalmazását a transzformált adatokra.

Az inverz regresszió modellje

Az inverz transzformált adatokat használva illesztjük a modellt:

$$Y = \alpha + \frac{\beta}{X}$$

Az illesztett paraméterek becslése:

$$\hat{\alpha} = 35 \quad \text{és} \quad \hat{\beta} = 45$$

Ez azt jelenti, hogy a becsült egyenlet így alakul:

$$\hat{Y} = 35 + \frac{45}{X}$$

Előrejelzés a modell alapján

Az illesztett modell segítségével bármilyen szünetek száma (X) esetén megjósolható a hatékonyság (Y). Például ha 6 szünetet tartanak:

$$\hat{Y} = 35 + \frac{45}{6} \approx 42.5$$

Ez azt jelenti, hogy 6 szünet esetén a hatékonyság körülbelül 42.5 lesz.

Az előrejelzés végső formája

Az illesztett inverz modell alapján a szünetek számának függvényében a hatékonyság becslése:

$$\hat{Y} = 35 + \frac{45}{X}$$

Ez a képlet pontosan megmutatja, hogy hogyan csökken a hatékonyság a szünetek számának növekedésével.

Alkalmazhatóság és korlátok

Ez a modell jól alkalmazható olyan helyzetekben, ahol egy változó (pl. hatékonyság) csökken a másik változó (pl. szünetek száma) növekedésével. Azonban fontos figyelembe venni, hogy a reciprok változók esetében a túl nagy értékeknél a hatás mértéke gyorsan csökken, ami egyes esetekben korlátozhatja a modell érvényességét.

Összegzés

A lineárisra visszavezethető regressziók olyan esetekben alkalmazhatók, amikor a változók közötti kapcsolat nemlineáris, de megfelelő transzformációval lineáris formába hozható. Példák a logaritmikus, polinomiális, log-lineáris és inverz regressziók. Ezekkel a módszerekkel bonyolult, nemlineáris kapcsolatokat is sikeresen modellezhetünk.