

Regressziószámítás

A **regresszió** egyenes egyenlete:

$$y = b_0 + b_1 \cdot x$$

$$\text{Ahol } b_1 = \frac{\sum dx \cdot dy}{\sum d^2x} \text{ és } b_0 = \bar{y} - b_1 \cdot \bar{x}$$

A regressziós egyenes egyenletében szereplő regressziós paraméterek közül b_1 az egyenes meredeksége. A b_0 érték kevésbé jelentős, ez azt adja meg, hogy a magyarázó változó nulla értékéhez milyen y érték tartozik.

[Megnézem a kapcsolódó epizódot](#)

$$\text{A regressziós egyenes egyenlete } \hat{y} = \hat{b}_0 + \hat{b}_1 \cdot x$$

Ez egy **lineáris függvény**, ami mindegyik x -hez hozzárendel valamilyen y -t. Ezek általában eltérnek a valódi y -októl. Ezeket az eltéréseket reziduumoknak nevezzük.

[Megnézem a kapcsolódó epizódot](#)

A reziduumokból képzett mutató az úgynevezett SSE, jelentése sum of squares of the errors vagyis eltérés-négyzetösszeg.

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Ha a **regresszió** tökéletesen illeszkedik, akkor az $e_i = y_i - \hat{y}_i$ különbségek mindegyike 0, így SSE=0. Ha az illeszkedés nem tökéletes, akkor SSE egy pozitív érték, ami az illeszkedés pontatlanságát méri.

[Megnézem a kapcsolódó epizódot](#)

Ha az SSE értékeit elosztjuk a megfigyelt pontok számával és a kapott eredménynek vesszük a gyökét, akkor kapjuk a reziduális szórást:

$$s_e^* = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{\sum e_i^2}{n}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

[Megnézem a kapcsolódó epizódot](#)

Az illeszkedés egy mérőszáma a lineáris **korrelációs együttható**:

$$r = \frac{\sum dx \cdot dy}{\sqrt{\sum d^2x \cdot \sum d^2y}}$$

A lineáris **korrelációs együttható** azt méri, hogy x és y között milyen szoros lineáris kapcsolat van. Értéke mindig $-1 \leq r \leq 1$.

[Megnézem a kapcsolódó epizódot](#)

A magyarázóerőt méri az úgynevezett determinációs együttható, melynek jele R^2 . Ez a kétváltozós lineáris modell esetében megegyezik r^2 -tel.

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

Itt SSE az eltérés-négyzetösszeg, míg SSR az úgynevezett regressziós, vagy magyarázó négyzetösszeg, SST pedig a teljes négyzetösszeg, a köztük lévő kapcsolat pedig...

$$SST = \sum d^2 y \quad SSR = \sum (\hat{y}_i - \bar{\hat{y}})^2 = b_1^2 \sum d^2 x \quad SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

[Megnézem a kapcsolódó epizódot](#)

A [regresszió](#) egyenes egyenlete:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

Amból

$$\lg \hat{y} = \lg \hat{b}_0 + \hat{b}_1 \cdot \lg x$$

$$\text{Ahol } \hat{b}_1 = \frac{\sum d \lg x \cdot d \lg y}{\sum d^2 \lg x} \text{ és } \lg \hat{b}_0 = \overline{\lg y} - \overline{\lg x} \cdot \hat{b}_1$$

[Megnézem a kapcsolódó epizódot](#)

A [regresszió](#) egyenes egyenlete:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

Amból

$$\lg \hat{y} = \lg \hat{b}_0 + x \cdot \hat{b}_1$$

$$\text{Ahol } \lg \hat{b}_1 = \frac{\sum dx \cdot d \lg y}{\sum d^2 x} \text{ és } \lg \hat{b}_0 = \overline{\lg y} - \bar{x} \cdot \lg \hat{b}_1$$

[Megnézem a kapcsolódó epizódot](#)

Az elaszticitás két összefüggő jelenség közti kapcsolat.

$$\text{Lineáris modellben: } El(\hat{y}, x) = \frac{\hat{b}_1 x}{\hat{y}} = \frac{\hat{b}_1 x}{\hat{b}_0 + \hat{b}_1 x}$$

$$\text{Hatványkitevős modellben: } El(\hat{y}, x) = \hat{b}_1$$

$$\text{Exponenciális modellben: } El(\hat{y}, x) = x \cdot \ln \hat{b}_1$$

[Megnézem a kapcsolódó epizódot](#)

- I. A magyarázó változók nem valószínűségi változók.
- II. A magyarázó változók lineárisan független rendszert alkotnak.
- III. Az eredményváltozó közel lineáris függvénye a magyarázó változónak.
- IV. Az ϵ hibatag feltételes eloszlása normális, várható értéke nulla.
- V. Az ϵ hibatag különböző x -ekhez tartozó értékei korrelálatlanok.

[Megnézem a kapcsolódó epizódot](#)

A paraméterek becslése:

$$\hat{b}_i \pm t_{1-\frac{\alpha}{2}} \cdot (n - k - 1) \cdot s_{\hat{b}_i}$$

A [regresszió](#) becslése:

$$\hat{y}_* \pm t_{1-\frac{\alpha}{2}} \cdot (n - k - 1) \cdot s_{\hat{y}_*}$$

[Megnézem a kapcsolódó epizódot](#)

A többváltozós regressziós modelleket olyankor alkalmazzuk, amikor az eredményváltozó alakulását több magyarázó változó tükrében vizsgáljuk.

A többváltozós [lineáris regresszió](#) egyenlete:

$$y = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_k x_k + \epsilon$$

Az y eredményváltozó itt k darab magyarázó változótól és a hibatagtól függ.

A képletben a \hat{b}_0 paraméter a tengelymetszet, a többi \hat{b}_i paraméter pedig azt jelenti, hogy az i -edik magyarázó változó egy egységgel történő változása, mennyivel változtatja az \hat{y} értéket, ha a többi magyarázó változót rögzítjük.

[Megnézem a kapcsolódó epizódot](#)

A kétváltozós esethez hasonlóan a [korreláció](#) itt is a változók közti kapcsolat szorosságát írja le, csak hogy itt egy fokkal rosszabb a helyzet, ugyanis most bármely két változó korrelációját vizsgálhatjuk. Ezt tartalmazza a korrelációmátrix.

$$R = \begin{pmatrix} 1 & r_{y1} & r_{y2} & \dots & r_{yk} \\ r_{1y} & 1 & r_{12} & \dots & r_{1k} \\ r_{2y} & r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{ky} & r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}$$

Itt r_{ij} az x_i és az x_j magyarázó változók közti korrelációt írja le, tehát például r_{12} az x_1 és az x_2 közti korrelációt jelenti.

r_{iy} pedig az x_i magyarázó változó és az y eredményváltozó közti kapcsolatot jelenti.

Mivel $r_{ij} = r_{ji}$ a [korreláció-mátrix](#) szimmetrikus. Az áttekinthetőbb felírás kedvéért a felső háromszöget el is szokták hagyni.

[Megnézem a kapcsolódó epizódot](#)

A [lineáris regresszió](#) egyenlete: $\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \dots + \hat{b}_kx_k$

A tesztelés úgy zajlik, hogy nullhipotézisnek tekintjük a $H_0 : b_i = 0$ feltevést, ellenhipotézisnek pedig azt, hogy $H_1 : b_i \neq 0$.

A nullhipotézis azt állítja, hogy a modellben a b_i paraméter szignifikánsan nulla, vagyis az i -edik magyarázó változó felesleges, annak hatása az eredményváltozóra nulla. Az ellenhipotézis ezzel szemben az, hogy $b_i \neq 0$ vagyis az i -edik magyarázó változónak a regresszióban nem nulla hatása van.

[Megnézem a kapcsolódó epizódot](#)

| Szóródás oka | Négyzetösszeg | Szabadságfok | Átlagos négyzetösszeg | F |
|----------------------------|---------------|--------------|-------------------------------|-----------------------|
| Regresszió | SSR | k | $MSR = \frac{SSR}{k}$ | $F = \frac{MSR}{MSE}$ |
| Hiba | SSE | $n - k - 1$ | $MSE = \frac{SSE}{n - k - 1}$ | |
| Teljes | SST | $n - 1$ | | |

[Megnézem a kapcsolódó epizódot](#)

A multikollinearitás röviden összefoglalva azt jelenti, hogy két vagy több magyarázó változó között túl szoros [korrelációs kapcsolat](#) van, és ez zavarja a becslést.

A multikollinearitás mérésére az úgynevezett VIF (variance inflator factor) variancia növelő faktor van forgalomban.

$$VIF_j = \frac{1}{1 - R_j^2}$$

A képletben szereplő R_j^2 a j -edik magyarázó változó és az összes többi magyarázó változó közti determinációs együttható.

[Megnézem a kapcsolódó epizódot](#)

Az auto**korreláció** a **regresszió** maradéktagjának a saját későbbi értékeivel való korrelációját jelenti, vagyis egyfajta szabályszerűséget a maradékváltozóban. Ideális esetben a maradéktagnak véletlenszerűnek kell lennie, bármiféle szabályszerűségért a magyarázó változók felelnek a regresszióban.

Az auto**korreláció** tesztelésére a Durbin-Watson-tesztet használjuk.

[Megnézem a kapcsolódó epizódot](#)

A Durbin-Watson-teszt lényegében egy hipotézisvizsgálat, aminek részletezésére nem térünk ki, mindössze a használatát nézzük meg.

Maga a próbafüggvény

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

A **szignifikanciaszint** α , a próba elvégzése pedig az alábbi módon történik:

d_L és d_U értékeket kikeressük a táblázatból,

n = a megfigyelések száma,

k = a magyarázó változók száma,

végül megnézzük, hogy a próbafüggvény melyik tartományba esik.

[Megnézem a kapcsolódó epizódot](#)
