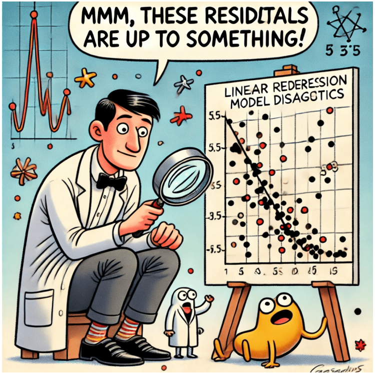
Regresszióanalízis: Egyszerű lineáris regresszió - illeszkedés- és modelldiagnosztika

Matematikai Statisztika
2024. október 21.

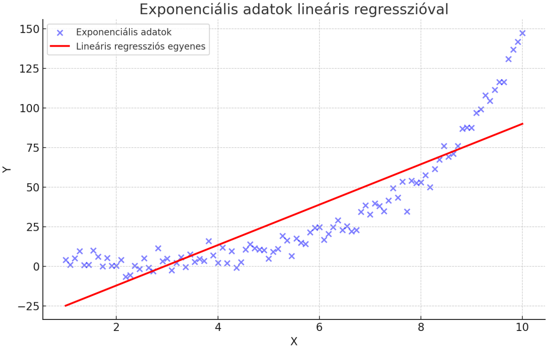


Miért van szükség illeszkedésvizsgálatra?

A lineáris regressziós modell univerzális, de nem mindenre jó...

A regresszióanalízis képletei minden adathalmazra alkalmazhatók, mivel a lineáris regressziós egyenletek általános formulákon alapulnak. Azonban:

- A regressziós modell akkor működik jól, ha a valós kapcsolat közel lineáris.
- Ha az adatok nem követnek lineáris mintázatot, a modell lehet, hogy pontosan kiszámítja az *a* és *b* értékeket, de ezek nem adnak valós vagy értelmes eredményeket.
- Az illeszkedésvizsgálat célja annak ellenőrzése, hogy a modell helyesen írja-e le az adatok kapcsolatát, és hogy az általunk kapott eredmények valóban megbízhatóak-e.



Miért van szükség modelldiagnosztikára?

A modelldiagnosztika fontossága

Még ha a regressziós modell látszólag illeszkedik is, előfordulhatnak olyan problémák, amelyek miatt az eredmények nem használhatók. Néhány fontos diagnosztikai szempont:

- **A hibatagok normalitása**: A modell alapfeltétele, hogy a hibatagok normális eloszlásúak. Ha ez nem teljesül, a becslések pontatlanok lehetnek.
- **Homoszkedaszticitás**: A hibatagok varianciájának állandónak kell lennie. Ha a variancia változik, az becslési problémákhoz vezet (heteroszkedaszticitás).
- **Függetlenség**: A megfigyelések hibatagjainak függetlenek kell lenniük. Ha korreláció van közöttük (pl. idősorokban), az torzítja a becsléseket.

Ezeket a problémákat diagnosztikai vizsgálatokkal azonosíthatjuk, és szükség esetén javíthatjuk a modellt.

Lineáris regresszió - SST, SSR, SSE definíciója

SST, SSR, SSE fogalmai

A regresszióanalízis során három fontos négyzetösszeget különböztetünk meg, amelyek az adatok szóródásának különböző összetevőit írják le:

- **Teljes négyzetösszeg (SST)**: Az összes megfigyelés szóródása az átlag körül:

  $$SST = \sum (Y_i - \overline{Y})^2$$

  Ez azt méri, hogy a függő változó (Y) összességében mennyire tér el az átlagától ($\overline{Y}$).
- **Regressziós négyzetösszeg (SSR)**: Az a szóródás, amit a regressziós modell magyaráz:

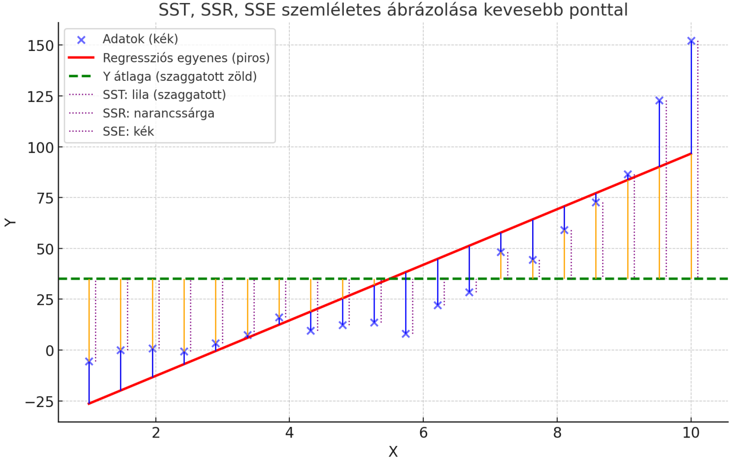
  $$SSR = \sum (\hat{Y}_i - \overline{Y})^2$$

  Ez azt mutatja meg, hogy a becsült értékek ($\hat{Y}_i$) mennyire térnek el az átlagtól, és ez az a rész, amit a magyarázó változók (pl. X) "elmagyaráznak".
- **Hibanégyzetösszeg (SSE)**: Az az eltérés, amit a modell nem magyaráz meg:

  $$SSE = \sum (Y_i - \hat{Y}_i)^2$$

  Ez az a rész, amit a modell nem tud magyarázni, azaz a valós Y_i értékek és a becsült $\hat{Y}_i$ értékek közötti különbség.

Grafikon



SST, SSR, SSE szemléletes ábrázolása kevesebb ponttal
Adatok (kék)
Regressziós egyenes (piros)
Y átlaga (szaggatott zöld)
SST: lila (szaggatott)
SSR: narancssárga
SSE: kék
150
125
100
75
50
25
0
−25
Y
2
4
6
8
10
X

Hibanégyzetösszeg magyarázó változó nélkül

Ha nincs magyarázó változó (tehát nincs X a modellben), akkor a regressziós modell egyszerűen az eredményváltozó átlagát ($\overline{Y}$) adja meg minden megfigyelésre. Ebben az esetben a becslés pontosságát a következő négyzetösszeg írja le:

$$SSE_{\text{nincs X}} = \sum (Y_i - \overline{Y})^2 = SST$$

Tehát, ha nincs magyarázó változó, a modell egyáltalán nem képes magyarázni az adatok varianciáját, így a teljes négyzetösszeg (SST) egyben a hibanégyzetösszeg (SSE) is. Ebben az esetben minden szóródás magyarázatlan marad.

A magyarázó változó bevonása

Amikor magyarázó változót (pl. X) vonunk be, akkor a modell képes a varianciából egy részt "elmagyarázni". Ez a rész a **regressziós négyzetösszeg** (SSR). A fennmaradó hibanégyzetösszeg (SSE) a következőképpen alakul:

$$SSE = SST - SSR$$

Az SSR az a rész, amit a modell magyaráz, és ennek bevonása csökkenti a hibanégyzetösszeget (SSE). Tehát minél nagyobb az SSR, annál jobban magyarázza a modell az adatok szóródását.

Lineáris regresszió - Determinációs együttható (R^2)

Mit mér a determinációs együttható?

Az R^2 mutató azt méri, hogy a regressziós modell milyen mértékben képes megmagyarázni az adatok variabilitását. A modell teljesítményének egy fontos indikátora, amely megmutatja, hogy a magyarázó változók mennyire vannak kapcsolatban a függő változóval.

Definíció:

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

ahol:

- SST (Teljes négyzetösszeg): Az eredményváltozó összes szóródása,
- SSR (Regressziós négyzetösszeg): Az a szóródás, amit a modell magyaráz,
- SSE (Hibanégyzetösszeg): Az a szóródás, amit a modell nem tud magyarázni.

Az R^2 interpretációja

Az R^2 értéke a következő tartományban helyezkedik el: $0 \leq R^2 \leq 1$. Ez az arány megmutatja, hogy az összes szóródás hány százalékát magyarázza a regressziós modell. Értelmezése:

- $R^2 = 0$: A modell egyáltalán nem magyarázza meg a varianciát (nincs kapcsolat a változók között),
- $R^2 = 1$: A modell teljes mértékben megmagyarázza a varianciát (tökéletes illeszkedés),
- **Köztes értékek**: Az R^2 értéke megmutatja, hogy az összes szóródás hány százalékát írhatjuk a modell számlájára.

Példa: Ha $R^2 = 0.85$, akkor a modell az adatok 85%-ának variabilitását magyarázza.

Lineáris regresszió - Együtthatók eloszlása

Miért fontos az együtthatók eloszlása?

A regressziós modell paramétereinek, a és b-nek a tesztelése kulcsfontosságú, mert segít eldönteni, hogy valóban van-e lineáris kapcsolat az X (magyarázó változó) és Y (függő változó) között. Ehhez ismernünk kell az együtthatók eloszlását.

Paraméterek eloszlása normális hibák esetén: Ha a modell hibatagjai (ε_i) normális eloszlásúak, akkor a becsült paraméterek eloszlása is normális lesz:

- a (intercept) eloszlása:

$$\hat{a} \sim N\left(a, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right)\right),$$

  ahol σ^2 a hibatagok varianciája, $\overline{X}$ az X-ek átlaga, és S_{XX} a X-ek szórása.

- b (meredekség) eloszlása:

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{XX}}\right),$$

  ahol $\hat{b}$ a regressziós meredekség becslése.

Mit jelent ez?

Mivel tudjuk, hogy a és b becsült értékei normális eloszlásúak, statisztikai teszteket végezhetünk a paraméterek szignifikanciájának meghatározására. Ezek az információk segítenek eldönteni, hogy a modell ténylegesen hasznos-e, vagy csupán véletlenszerű eredményeket produkál.

Hipotézisvizsgálat b-re (meredekségre)

A regressziós együtthatók szignifikanciájának teszteléséhez t-próbát alkalmazunk. A cél annak meghatározása, hogy a meredekség (b) szignifikánsan különbözik-e nullától, azaz van-e lineáris kapcsolat az X és Y változók között.

Nullhipotézis (H_0):

$$H_0 : b = 0 \quad \text{(nincs kapcsolat az } X \text{ és } Y \text{ között).}$$

Próbastatisztika:

$$t = \frac{\hat{b}}{SE(\hat{b})},$$

ahol a standard hiba $SE(\hat{b})$ a következőképpen számítható:

$$SE(\hat{b}) = \frac{\hat{\sigma}}{\sqrt{S_{XX}}},$$

ahol $\hat{\sigma}$ a becsült szórás, és S_{XX} a független változó szórása:

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Lineáris regresszió - Együtthatók tesztelése - b

Döntés a t-érték alapján

Az eloszlás alapján a próbastatisztika egy t-eloszlást követ $n-2$ szabadságfokkal. A nullhipotézist elvetjük, ha a t-érték nagyobb, mint a kritikus érték a választott szignifikanciaszinten (pl. 95

Döntés: Ha a t-érték t_{kritikus}, elvetjük a nullhipotézist, és azt mondjuk, hogy van szignifikáns lineáris kapcsolat az X és Y változók között.

Lineáris regresszió - Konstans (a) tesztelése

Hipotézisvizsgálat a-ra (interceptre)

A t-próbát alkalmazzuk az Y-tengelymetszet (a) teszteléséhez, hogy meghatározzuk, szignifikánsan különbözik-e nullától, azaz az Y-tengelyen való metszéspont létezik-e.

Nullhipotézis (H_0):

$$H_0 : a = 0 \quad \text{(nincs metszéspont az } Y\text{-tengellyel).}$$

Próbastatisztika:

$$t = \frac{\hat{a}}{SE(\hat{a})},$$

ahol a standard hiba $SE(\hat{a})$ a következőképpen számítható:

$$SE(\hat{a}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}},$$

ahol $\hat{\sigma}$ a becsült szórás, $\bar{x}$ az X átlagértéke, és S_{XX} a következőképpen számítható:

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Lineáris regresszió - Konstans (a) tesztelése

Döntés a t-érték alapján

Az eloszlás alapján a t-statisztika egy t-eloszlást követ $n-2$ szabadságfokkal. A nullhipotézist elvetjük, ha a t-érték nagyobb, mint a kritikus érték a választott szignifikanciaszinten (pl. 95%-os megbízhatóság esetén).

Döntés: Ha t-érték t_{kritikus}, elvetjük a nullhipotézist, és azt mondjuk, hogy az Y-tengelymetszet szignifikáns, tehát nem nulla.

Lineáris regresszió - Hibatagok függetlenségének vizsgálata

Miért fontos a hibatagok függetlensége?

A lineáris regresszió egyik alapfeltétele, hogy a hibatagok (ε_i) függetlenek egymástól. Ha a hibatagok korreláltak, az alábbi problémákat okozhat:

- **Torz becslések**: A korrelált hibatagok miatt a becsült paraméterek (a és b) pontatlanok lehetnek.
- **Megbízhatatlan konfidenciaintervallumok**: A hibatagok függetlenségének hiánya a konfidenciaintervallumok torzulásához vezethet.
- **Csökkent prediktív erő**: A modell prediktív teljesítménye csökken, ha a hibatagok nem függetlenek.

A függetlenség teszteléséhez használhatjuk a **Durbin-Watson statisztikát**, amely kimondottan az autokorreláció vizsgálatára szolgál.

Durbin-Watson statisztika képlete

A Durbin-Watson statisztika azt vizsgálja, hogy a hibatagok (ε_i) egymást követő értékei között van-e autokorreláció. A statisztika értéke a következőképpen számítható:

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

ahol:

- ε_i: Az i-edik hibatag,
- n: A megfigyelések száma.

Az DW statisztika értéke 0 és 4 között van. Az értelmezés:

- $DW \approx 2$: Nincs autokorreláció (függetlenség),
- $DW << 2$: Pozitív autokorreláció,
- $DW >> 2$: Negatív autokorreláció.

Lineáris regresszió - Durbin-Watson statisztika

Hipotézisvizsgálat a Durbin-Watson statisztikával

A Durbin-Watson statisztika alapján a következő hipotéziseket teszteljük:

- **Nullhipotézis (H_0)**: A hibatagok nem korreláltak (függetlenek),
- **Alternatív hipotézis (H_1)**: A hibatagok korreláltak.

Próbastatisztika eloszlása H_0 mellett: A Durbin-Watson statisztika nem követ hagyományos eloszlást (mint például a normális vagy t-eloszlás). A kritikus értékeket táblázatokból határozzuk meg, ahol az alsó (d_L) és felső (d_U) határok különböző szignifikanciaszintek mellett vannak megadva, a megfigyelések száma és a független változók száma alapján.

Döntés:

- Ha $DW < d_L$ vagy $DW > 4 - d_L$, elvetjük a nullhipotézist ,
- Ha $4 - d_U > DW > d_U$, nem vetjük el a nullhipotézist ,
- Ha $d_L \leq DW \leq d_U$ vagy $4 - d_L \geq DW \geq 4 - d_U$ a teszt nem dönthető el egyértelműen.

Homoszkedaszticitás fogalma

A **homoszkedaszticitás** azt jelenti, hogy a regressziós modell maradékainak (hibatagok, ε_i) varianciája állandó az X független változó különböző értékeinél. Ez a lineáris regresszió egyik fontos feltétele, mivel ha a hibatagok varianciája nem állandó, az torzíthatja a becsült paramétereket (a és b).

A homoszkedaszticitás megsértése esetén **heteroszkedaszticitás** lép fel, amely azt jelenti, hogy a maradékok szórása változik az X értékeivel, például az X nagyobb vagy kisebb értékeinél.

Breusch-Pagan próba

A **Breusch-Pagan próba** a leggyakrabban használt módszer a heteroszkedaszticitás tesztelésére lineáris regresszióban. A próba a hibatagok varianciájának változását vizsgálja az X változó értékeinek függvényében.

Nullhipotézis (H_0): A hibatagok varianciája állandó (**homoszkedaszticitás**).

Alternatív hipotézis (H_1): A hibatagok varianciája változó (**heteroszkedaszticitás**).

Lineáris regresszió - Hibatagok varianciáinak vizsgálata

Breusch-Pagan próba

A Breusch-Pagan próba egy segédregressziót alkalmaz, amely a hibatagok négyzetösszegét (ε_i^2) regresszálja az X független változó(k)ra. A statisztika az alábbi képlettel számítható:

$$BP = \frac{n \cdot R^2}{2},$$

ahol n a megfigyelések száma, R^2 pedig a segédregresszió determinációs együtthatójának értéke. A teszt statisztika χ^2-eloszlást követ a magyarázó változók számával megegyező szabadságfokkal.

Interpretáció és döntés

Ha a Breusch-Pagan próba alapján elvetjük a nullhipotézist, az azt jelenti, hogy a hibatagok varianciája nem állandó, vagyis heteroszkedaszticitás van jelen. Ekkor a következő lehetőségek állnak rendelkezésre a probléma kezelésére:

- **Súlyozott legkisebb négyzetek módszere (WLS)**: Súlyozzuk a megfigyeléseket, hogy korrigáljuk a változó varianciát.
- **Robusztus standard hibák**: Az együtthatók becsült értékei továbbra is használhatók, de a robusztus standard hibák jobban kezelik a heteroszkedaszticitást.
- **Transzformációk**: Transzformálhatjuk a függő változót (pl. logaritmus vagy négyzetgyök transzformáció), hogy stabilizáljuk a varianciát.

Lineáris regresszió - Hibatagok normalitásának ellenőrzése

Miért fontos a hibatagok normalitása?

A lineáris regresszió során az egyik feltétel az, hogy a maradékok (hibatagok, ε_i) normális eloszlásúak legyenek. Ez azért fontos, mert:

- **Konfidenciaintervallumok** és **hipotézisvizsgálatok**: A paraméterek becsléseire vonatkozó intervallumbecslések és hipotézisvizsgálatok a normalitásra támaszkodnak.
- **Torzításmentes becslések**: Ha a hibatagok nem normálisak, akkor a becslések torzak lehetnek, különösen kis mintaméret esetén.
- **Predikció pontossága**: A normalitás feltételezése befolyásolja az előrejelzési intervallumok és a prediktív modellek pontosságát.

Ha a hibatagok normalitása nem teljesül, akkor a modellből származó következtetések (például a t-próba vagy az F-próba) megbízhatatlanná válhatnak.

Lineáris regresszió - Normalitás tesztelése

Normalitás tesztelése - Grafikus módszerek

A hibatagok normalitásának ellenőrzésére gyakran alkalmaznak vizuális módszereket, amelyek gyorsan megmutatják, ha az eloszlás eltér a normálistól.

- **Q-Q plot (Kvantilis-kvantilis ábra)**: Ez az egyik leggyakrabban használt grafikus eszköz a hibatagok normalitásának ellenőrzésére. A Q-Q plot összehasonlítja a maradékok eloszlását egy elméleti normális eloszlással. Ha a hibatagok normális eloszlásúak, akkor a pontok nagyjából egy egyenesen helyezkednek el.
- **Histogram**: A maradékok hisztogramja egy gyors módja annak ellenőrzésére, hogy az eloszlás közelít-e a normálishoz. A normális eloszlású hibatagok haranggörbe alakot mutatnak.
- **Box plot**: A box plot a maradékok eloszlásának asszimetriáját és kiugró értékeit mutatja. Ha a maradékok normálisak, a box plot szimmetrikus lesz, kevés kiugró értékkel.

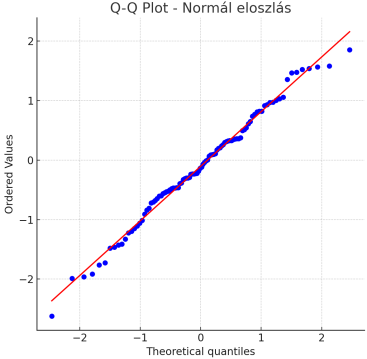
Ezek a vizuális módszerek gyors és egyszerű eszközt kínálnak a hibatagok eloszlásának ellenőrzésére.

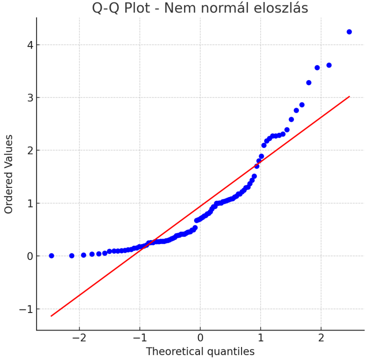
Q-Q Plot definíciója

Q-Q Plot definíció

A Q-Q plot (quantile-quantile plot) egy grafikus eszköz, amely két eloszlás kvantilisait veti össze egymással. Leggyakrabban azt használjuk, hogy egy mintabeli eloszlás kvantilisait hasonlítsuk össze egy referenciaeloszlás, például a normál eloszlás kvantiliseivel. Ha az adatok eloszlása megegyezik a referenciaeloszlással, a pontok egy egyenest alkotnak, amely az origón halad át, és meredeksége 1.

Használat: A normalitás tesztelésére gyakran használjuk: ha az adatok normál eloszlásúak, a Q-Q plot pontjai az egyenesen helyezkednek el.

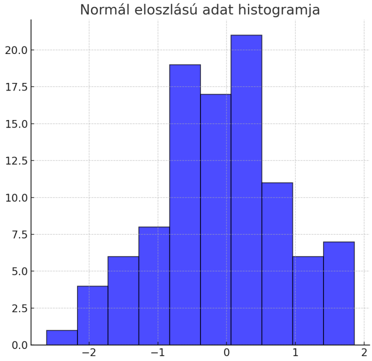


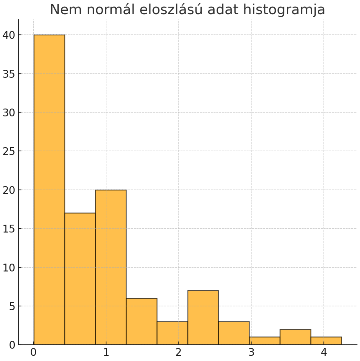


Histogram definíciója

Histogram definíció

A histogram egy olyan grafikus eszköz, amely egy adathalmaz eloszlását jeleníti meg. Az adatok tartományait oszlopdiagram formájában ábrázolja, ahol az oszlopok magassága az egyes tartományokba eső adatok gyakoriságát mutatja. A histogram gyakran használt eszköz az adatok eloszlásának vizsgálatára, például a normális eloszlás ellenőrzésére.





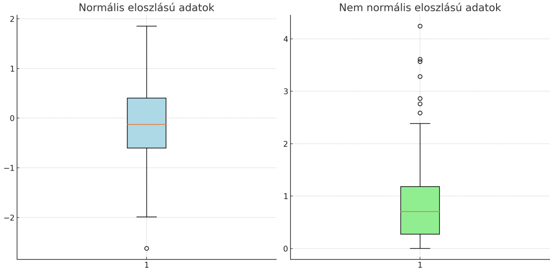
Box Plot definíciója

Box Plot definíció

A box plot (doboztáblázat) egy olyan grafikus eszköz, amely az adatok szóródását és eloszlását jeleníti meg. A box plot az adatok kvartiliseit mutatja:

- Az alsó és felső "doboz" szélei az első és harmadik kvartiliseket ($Q1$ és $Q3$) jelölik, azaz az adatok középső 50
- A doboz belsejében található vonal a mediánt jelöli.
- A dobozból kiinduló "bajuszok" az adatok szélső értékeit mutatják, miközben a kiugró értékek külön vannak jelölve.

A box plot segítségével vizsgálható, hogy az adatok normális eloszlásúak-e: szimmetrikus doboz és a medián középen normális eloszlásra utal.



Shapiro-Wilk teszt

A **Shapiro-Wilk teszt** az egyik legérzékenyebb módszer a normalitás tesztelésére, különösen akkor, ha a minta mérete kicsi. A teszt célja, hogy megvizsgálja, a hibatagok eloszlása mennyire tér el a normálistól.

Nullhipotézis (H_0): A hibatagok normális eloszlásúak.

Próbastatisztika: A Shapiro-Wilk teszt statisztikája:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \cdot \varepsilon_{(i)}\right)^2}{\sum_{i=1}^{n} \varepsilon_i^2},$$

ahol:

- $\varepsilon_{(i)}$: a rendezett hibatagok (azaz a hibatagok nagyság szerinti sorrendben),
- a_i: a normális eloszlás elméleti súlyai, amelyeket a mintabeli rendezett értékekre alapozva határoznak meg.

Shapiro-Wilk teszt folytatása

Shapiro-Wilk teszt magyarázat

Súlyok (a_i): Az a_i-k a normális eloszlásból származnak, és ezek az elméleti kvantilisek közötti távolságokat képviselik. Ezek a súlyok segítenek összehasonlítani a mintában szereplő rendezett értékeket a normál eloszlás rendezett kvantiliseivel.

Próbastatisztika eloszlása: A Shapiro-Wilk statisztika eloszlása empirikusan van meghatározva, ezért a kritikus értékeket táblázatokból vagy statisztikai szoftverekből kell lekérni.

Döntés: Ha a W-statisztika értéke elég alacsony a kritikus értékekhez képest, elvetjük a nullhipotézist, azaz azt mondjuk, hogy a minta nem normális eloszlású.

Előnyök:

- Nagyon érzékeny kis minták esetén is.
- Jó általános teljesítmény a normalitás ellenőrzésére.

Hátrányok:

- Nagy minták esetén kevésbé megbízható.
- Nehezebben értelmezhető a statisztika kiszámítása.

Kolmogorov-Szmirnov teszt

A **Kolmogorov-Szmirnov (K-S) teszt** egy általános teszt, amely azt vizsgálja, hogy a minta eloszlása mennyire tér el egy elméleti eloszlástól (jelen esetben normális eloszlás). A K-S tesztet gyakran használják a maradékok normalitásának ellenőrzésére is.
Nullhipotézis (H_0): A hibatagok eloszlása normális.

Próbastatisztika: A K-S statisztika a mintából és az elméleti eloszlásból származó kvantilisek legnagyobb eltérését méri:

$$D = \sup |F_n(\varepsilon_i) - F(\varepsilon_i)|,$$

ahol:

- $F_n(\varepsilon_i)$: az empirikus eloszlásfüggvény,
- $F(\varepsilon_i)$: a normális eloszlás elméleti eloszlásfüggvénye,
- sup: az eltérések maximuma.

Eloszlás: A K-S statisztika eloszlása aszimptotikusan **Kolmogorov-eloszlású**, amelyet szintén táblázatokból nyerünk.

Lineáris regresszió - Kolmogorov-Szmirnov teszt

Előnyök:

- Bármely elméleti eloszlással összehasonlítható (nem csak normális).
- Nagy mintákra is megbízható.

Hátrányok:

- Kicsi minták esetén kevésbé érzékeny.
- Csak a legnagyobb eltérést veszi figyelembe, nem érzékeny kisebb különbségekre.