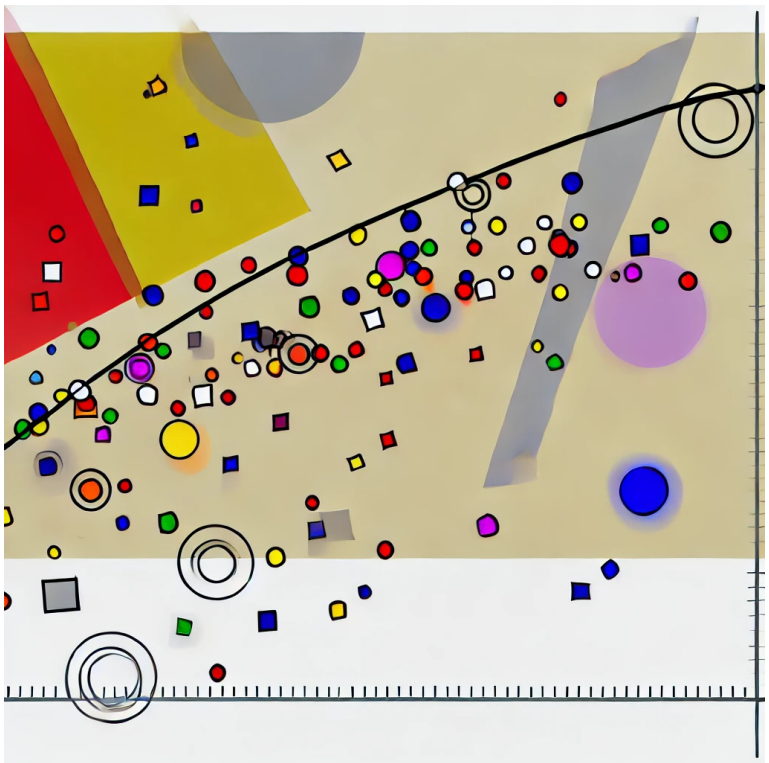
Regresszióanalízis: Bevezető

Matematikai statisztika
2024. október 21.



Bevezetés + Hol tartottunk?

Korábbi témák

Az előző előadásokon a **hipotézisvizsgálat** különböző módszereit vettük sorra, ideértve a változók függetlenségének tesztelését. Néhány fontos módszer:

- χ^2-**függetlenségvizsgálat**: Statisztikailag szignifikáns kapcsolat megállapítása két változó között.
- **Varianciaanalízis** (ANOVA): Többváltozós elemzések a csoportok közötti különbségek feltárására.
- **Nem-paraméteres tesztek**: A Friedman- és Kruskal-Wallis-próbák elemzése.

Ezek a módszerek segítenek eldönteni, hogy érdemes-e továbblépni az összefüggések leírása felé.

A mai óra céljai

Ma arra fókuszálunk, hogy **milyen kapcsolat van a változók között**, ha függetlenségüket elutasítjuk. A cél az, hogy megvizsgáljuk a változók közötti kapcsolatot és annak leírását.

Következő lépés

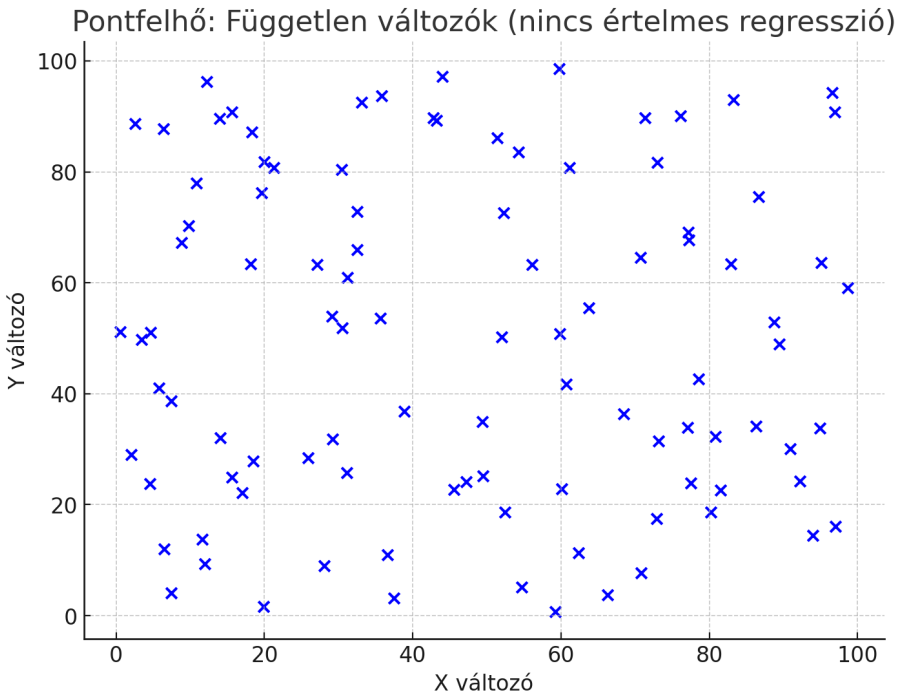
Kapcsolat leírása

A függetlenség elutasítását követően a változók közötti kapcsolat leírásának egyik módja:

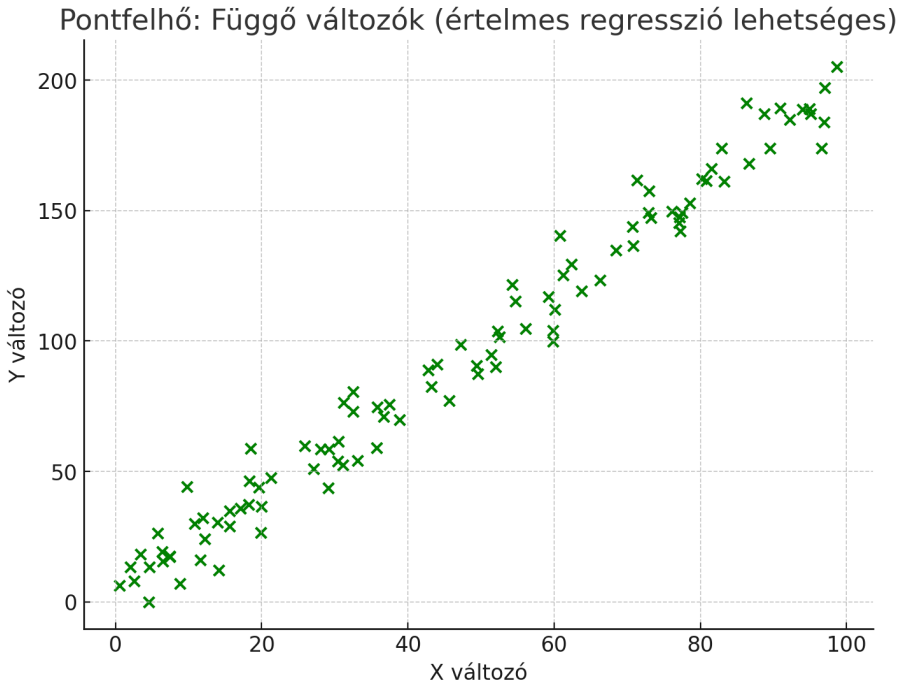
- **Korreláció kiszámítása**: Ez az első lépés, amikor egy egyszerű statisztikai mutatószámot használunk a változók közötti kapcsolat erősségének meghatározására.
- **Regresszióanalízis**: Ha a kapcsolat szignifikáns, regresszióanalízissel modellezzük a változók közötti viszonyt, hogy függvényszerűen jellemezhessük őket.

Mindkét módszer segít abban, hogy a változók közötti kapcsolatot világosabbá tegyük, és értelmezhetőbb modellt alkossunk.

Következő lépés



nincs regresszió



regresszió lehetséges

Regresszióanalízis fogalma

A regresszióanalízis egy statisztikai módszer, amely azt vizsgálja, hogyan függ egy függő változó (eredmény változó) egy vagy több független változótól (magyarázó változók). Célja egy olyan modell felállítása, amely megbecsüli a függő változó értékeit a magyarázó változók alapján.

A regresszióanalízis pár típusa:

- **Egyszerű lineáris regresszió**: A függő változó és a magyarázó változó közötti kapcsolat egy egyenes vonallal írható le.
- **Többszörös (lineáris) regresszió**: Több független változót használunk a függő változó becsléséhez.
- **Nemlineáris regresszió**: A kapcsolat nem egy egyenes vonallal írható le, hanem valamilyen bonyolultabb függvénnyel, például exponenciális vagy logaritmikus alakban.

A megfelelő modell kiválasztása kulcsfontosságú a pontos előrejelzésekhez.

Mikor használjuk a regresszióanalízist?

Alkalmazási területek

A regresszióanalízis széles körben használható a következő területeken:

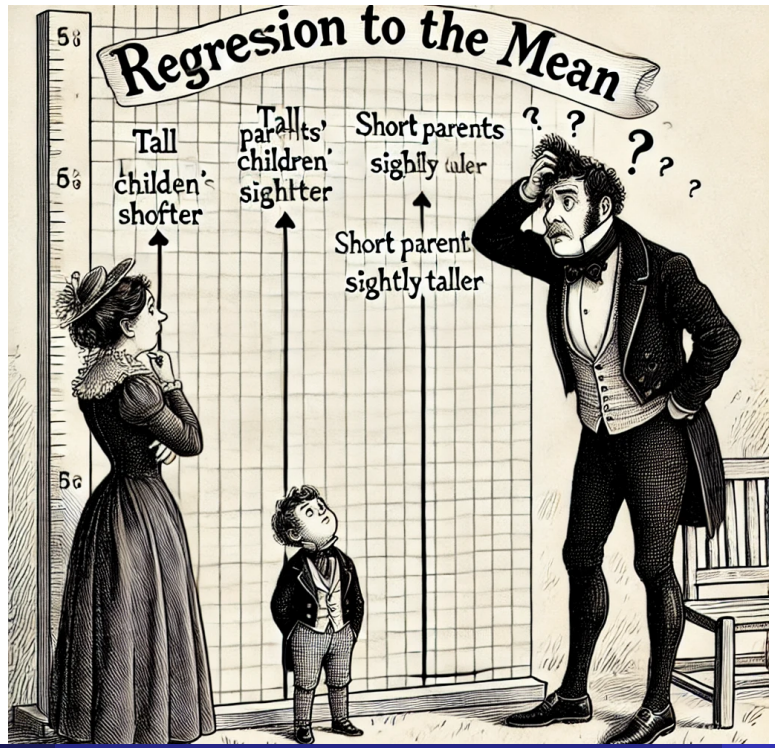
- **Gazdasági előrejelzések**: Hogyan változik a piaci ár egy adott változó hatására, például a kereslet vagy kínálat függvényében?
- **Orvosi kutatások**: Mennyiben befolyásolja egy gyógyszer adagolása a betegségek kimenetelét, és milyen mértékben csökkenti a tüneteket?
- **Marketing**: Milyen tényezők, például az ár, a reklámköltségek vagy az ügyfél-elégedettség befolyásolják egy termék eladásait?
- **Pszichológia**: Hogyan függ a vizsgázók eredménye a tanulási időtől, és milyen egyéb tényezők, például a stressz szintje vagy a tanulási módszer befolyásolják a teljesítményt?
- **Környezetvédelem**: Hogyan hat az üvegházhatású gázok kibocsátása a globális felmelegedés mértékére?
- **Sportanalízis**: Mely tényezők befolyásolják egy sportoló teljesítményét, például a fizikai edzés időtartama vagy a táplálkozás?

A regresszióanalízis tehát sokféle területen alkalmazható, és segítségével fontos döntéshozási folyamatok támogathatók.

Galton és az átlaghoz való visszatérés - Az első megfigyelések

Történeti áttekintés

A regresszióanalízis története Sir Francis Galton nevéhez fűződik, aki a 19. század végén dolgozott a módszeren. Galton az öröklődés hatásait vizsgálta, és azt találta, hogy a szülők magassága és a gyermekeik magassága közötti kapcsolatban visszafelé közelítés (regresszió) figyelhető meg. Innen ered a módszer elnevezése is: "regresszió" a középérték felé.



Galton és az átlaghoz való visszatérés - Az első megfigyelések

Galton korai munkássága

Sir Francis Galton, brit tudós a 19. század végén vizsgálni kezdte a szülők és gyermekeik közötti öröklődési kapcsolatot. Kutatásai során megfigyelte, hogy a szülők és gyermekeik magassága között van összefüggés, de nem olyan erős, mint azt korábban gondolták.
Galton észrevette, hogy bár a magas szülők gyerekei általában magasak, ők mégsem voltak olyan magasak, mint a szüleik. Ugyanakkor az alacsony szülők gyerekei általában magasabbak voltak, mint a szüleik, de nem annyira, hogy elérjék az átlagon felüli magasságot.

Átlaghoz való visszatérés

Galton ezt a jelenséget az "átlaghoz való visszatérés" (regression to the mean) elvének nevezte el. Azt tapasztalta, hogy a szélsőséges tulajdonságok (például nagyon magas vagy nagyon alacsony szülők) esetén a gyerekek hajlamosak közelebb kerülni az átlaghoz.

Az első matematikai modellek

Galton matematikai elemzés során először az apa és a fia testmagassága közötti kapcsolatot vizsgálta. Ezt egy egyszerű lineáris modell segítségével írta le:

$$Y = m_2 + r(X - m_1),$$

ahol:

- Y: a fiú testmagassága,
- X: az apa testmagassága,
- m_1: az apák átlagos magassága,
- m_2: a fiúk átlagos magassága,
- r: a korrelációs együttható.

Ezt a modellt használva Galton megmutatta, hogy a fiúk magassága általában kevésbé tér el az átlagtól, mint az apáké.

Pearson és a korrelációs együttható

Korrelációs együttható bevezetése

Karl Pearson, Galton tanítványa, bevezette a **korrelációs együtthatót**, amely leírja két változó közötti **lineáris kapcsolat erősségét**. Ez az egyik legfontosabb statisztikai mérőszám a modern adatelemzésben.

Formulája:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

ahol $\text{cov}(X, Y)$ a kovariancia, σ_X és σ_Y pedig a szórásokat jelenti.

Pearson munkája lehetővé tette a változók közötti lineáris kapcsolat kvantitatív mérését, ami ma is az adatelemzés egyik alapja.

A lineáris regresszió formalizálása

Matematikai formalizálás

Pearson továbbfejlesztette a **lineáris regresszió** fogalmát, amely egy modellt állít fel a függő (Y) és független (X) változók közötti kapcsolat leírására.
Modell:

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

ahol a a **konstans**, b a **meredekség**, ϵ pedig a **hibatag**.
Ez a modell lehetőséget adott arra, hogy pontosan becsüljük a két változó közötti lineáris összefüggést, amely alapja a modern statisztikai elemzéseknek.