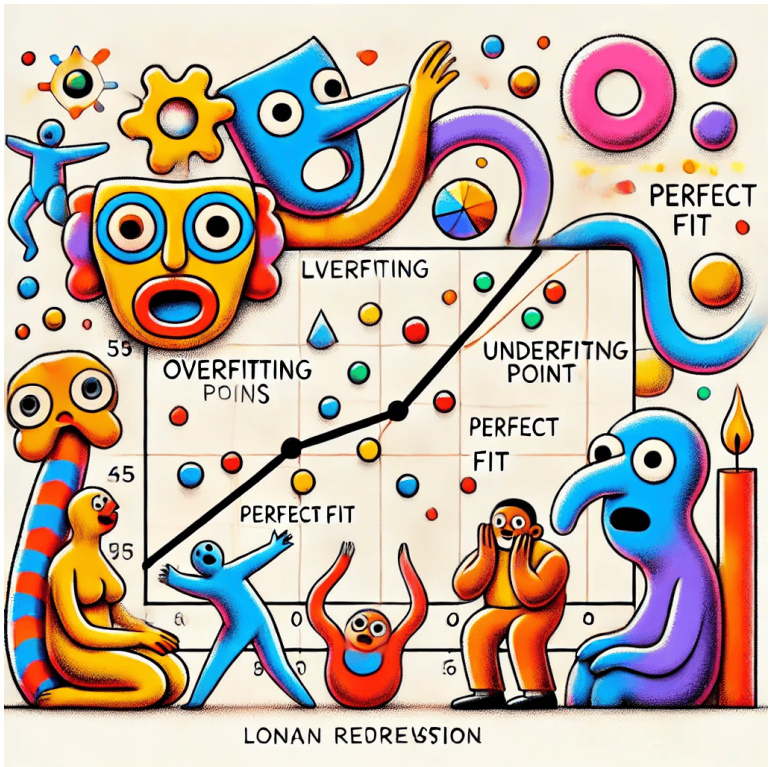
# Többszörös lineáris regresszió
# Egy lehetséges forgatókönyv

Matematikai statisztika
2024. október 28.



# 1. Adatfelvétel és Adattisztítás

Háziban csak az ott kérdezett dolgokat kell!!

## Adatok összegyűjtése

Felvesszük az összes releváns változót, beleértve a függő változót (célváltozó) és a független változókat (prediktorok).

## Adattisztítás

Ellenőrizzük az adatok minőségét és konzisztenciáját. Eltávolítjuk vagy pótoljuk a hiányzó adatokat, és korrigáljuk a hibás, irreleváns értékeket.

## Változók skálázása

A változókat egységes skálára hozzuk (pl. standardizálással), különösen ha a prediktorok eltérő mértékegységűek.

# 2. Exploratory Data Analysis (EDA)

## Deskriptív statisztika

Vizsgáljuk az adatok alapvető statisztikáit (átlag, medián, szórás, interkvartilis távolság) a kiugró értékek és mintázatok felismerésére.

## Korrelációs elemzés

Megvizsgáljuk a független változók közötti korrelációkat, hogy azonosítsuk a lehetséges multikollinearitást.

## Vizualizáció

Scatter plotokat, hisztogramokat és box plotokat készítünk az adatok eloszlásának és a függő változóhoz való kapcsolatuknak feltárására.

# 3. Outlierek és Hatásos Pontok Azonosítása

## Outlierek azonosítása

Az outlierek felismerésére használhatunk statisztikai teszteket, például box plotokat vagy Grubbs-tesztet, hogy azonosítsuk a kiugró értékeket.

## Hatásos pontok vizsgálata

A Cook-távolság és a hat mátrix elemei alapján megvizsgáljuk, hogy mely adatpontok gyakorolnak nagy hatást a modellre. Dönthetünk az outlierek eltávolításáról, transzformációjáról vagy súlyozásáról.

# 4. Változótranszformációk és Új Prediktorok

## Nemlineáris kapcsolatok transzformálása

Ha szükséges, logaritmus vagy más transzformációval linearizálhatjuk a változókat.

## Interakciós hatások és származtatott változók

- **Interakciós hatások**: Vizsgáljuk, hogy vannak-e olyan független változó interakciók, amelyek befolyásolják a célváltozót.
- **Származtatott változók**: Új változókat vezetünk be, például kvadratikus tagokat vagy főkomponenseket.

# 5. A Modell Specifikációja és Illesztése

## Modell kiválasztása

Meghatározzuk a többváltozós lineáris regressziós modellt és kiválasztjuk a legmegfelelőbb független változókat.

## Adatmegosztás (Train/Test Split)

A teljes adatot két vagy több részre osztjuk, például tanulási és tesztadatokra, hogy biztosítsuk a modell generalizálhatóságát.

## Modellillesztés

Illesztjük a modellt a tanulási adathalmazon, meghatározzuk az együtthatókat és ellenőrizzük a modell illeszkedését.

# 6. Modell Értékelése és Diagnosztika

## Modelldiagnosztikai statisztikák

Elemezzük a $R^2$-értéket, az RMSE-t és a szignifikanciaértékeket a modell magyarázóerejének értékelésére.

## Maradékok vizsgálata

A modellből származó reziduálisokat vizsgáljuk normalitás, homoszkedaszticitás és függetlenség szempontjából (histogram, Q-Q plot, reziduális plot).

## Multikollinearitás ellenőrzése

Kiszámítjuk a Variance Inflation Factor (VIF) értékeket, hogy azonosítsuk a multikollinearitás jelenlétét, amely torzíthatja az eredményeket.

# 7. Regularizáció (Szükség Esetén)

## Ridge és Lasso regresszió

Ha a modell túltanulásra hajlamos, Ridge vagy Lasso regularizációval csökkentjük az együtthatók nagyságát, különösen sok prediktor esetén.

## Keresztvalidáció alkalmazása

Keresztvalidációval teszteljük a modell teljesítményét különböző adathalmazokon, hogy optimalizáljuk a modell paramétereit.

# 8. Előrejelzés

## Előrejelzések készítése

A modell segítségével előrejelzéseket készítünk a célváltozóra, új vagy teszt adathalmazon.

## Előrejelzések kiértékelése

Az előrejelzések pontosságát a teszt adathalmazon értékeljük, és mérőszámokkal (MAE, MAPE, RMSE) ellenőrizzük az előrejelzési hibákat.

# 9. Modell Finomhangolása és Validálása

## Paraméterek finomhangolása

Szükség esetén optimalizáljuk a modellt az egyes prediktorok és a modell komplexitásának finomításával.

## Végső validáció

A modell teljesítményét véglegesen értékeljük a teszt adathalmazon, hogy megbizonyosodjunk az eredmények reprodukálhatóságáról.

# 10. Dokumentáció és Jelentéskészítés

## Eredmények interpretálása

A modellegyütthatókat és az eredményeket értelmezzük az eredeti probléma kontextusában.

## Jelentés készítése

Részletes jelentést készítünk, amely tartalmazza az adatfeldolgozás lépéseit, a modelldiagnosztikai elemzéseket, a modell teljesítményét és az előrejelzések pontosságát.

## Prezentáció

Az eredményeket vizuálisan bemutatjuk ábrák, grafikonok és interaktív dashboardok segítségével az érintettek számára.