Regressziós vizsgálatok

KAPCSOLATVIZSGÁLATI MÓDSZEREK

A változók közötti összefüggések lehetnek:

1. *Függvényszerű (teljes meghatározottságú összefüggés): a*z egyik ismérv változása minden esetben a másik ismérv változását idézi elő.

   Az egyikből változóból egyértelműen (egyenlettel) lehet következtetni a másikra.

2. Függetlenség

A változók között nincs összefüggés
az egyikből nem következtethetünk a másikra.

3. Sztochasztikus összefüggés

Az egyik ismérv változásából csak tendenciaszerűen (valószínűségi jelleggel) következtethetünk a másik ismérv változására. Pl. akinek több pénze van, általában többet költ autókra.

A sztochasztikus kapcsolatvizsgálat kérdései:

Fő szempont: milyen típusú kapcsolatot vizsgálunk:

milyen változók, ismérvek között?

1. Asszociációs kapcsolat

Minőségi ismérv, minősítéses jellemzők között: pl. iskolai végzettség-beosztás

2. Vegyes kapcsolat

minőségi és mennyiségi ismérvek közötti sztochasztikus kapcsolat

(pl. nem - kereset; beosztás – életkor)

3. Korrelációs kapcsolat

Mindkét vizsgált jellemző mennyiségi ismérv: életkor – szívferekvencia

Két változó oksági kapcsolatának feltételei:

a) Az ok időben megelőzi az okozatot.

b) A kettő között empirikus együttjárás van.

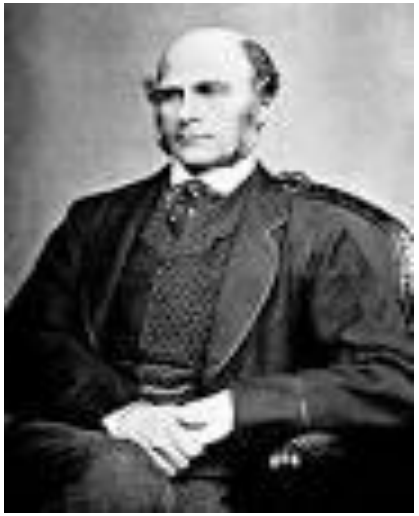
c) A kapcsolat nem egy harmadik változó okozza.

Regressziószámítás

Regresszió:

a változók közötti kapcsolat elemzésének elterjedt eszköze.

Vizsgálja: egy kitüntetett, a vizsgálat tárgyát képező változó, amelyet *eredményváltozónak* (vagy függő változónak, response) nevezünk, hogyan függ egy vagy több ún. *magyarázó* (vagy független, prediktor) *változótól*.



Born: February 16, 1822, Birmingham, United Kingdom
Died: January 17, 1911,

Ha a magyarázó változók száma (k) több (k>1), akkor sok(több)változós lineáris modellről beszélünk:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon$$

Feltételezzük, hogy valamennyi változóra *n* számú megfigyelésünk van, amelyeket célszerűen vektorokba, illetve mátrixba rendezhetünk:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1\, x_{11}\; x_{21} \cdots x_{k1} \\ 1\, x_{12}\; x_{22} \cdots x_{k2} \\ \vdots\;\vdots\;\;\vdots\;\;\ddots\;\vdots \\ 1\, x_{1n}\; x_{2n} \cdots x_{kn} \end{bmatrix}, \quad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \quad \text{és} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

A modell feltételek vizsgálata

- A **multikollinearitás** úgy is megfogalmazható, hogy a magyarázó változók között korreláció van.
- Multikollineáris esetben mind a becslés, mind a paraméterek értelmezése megnehezedik, hiszen a magyarázó változók hatásait nem lehet egyértelműen szétválasztani.
- Minden változó hatása minden más változóban is megjelenik, a becslések bizonytalanná válnak.

Multikollinearitás

- Mintabeli tulajdonság – mintán kívül nem alkalmazható.
- **Ellenőrzése:**
  - Többszörös determinációs együtthatóval,
  - |R|=0 multikollinearitás; |R|=1 a vált. független
  - VIF-mutató,
  - Tolerancia mutató.

- Ez a mutató azt mutatja, hogy a *j*-edik változó becsült együtthatójának tényleges varianciája hányszorosa annak, ami a multikollinearitás teljes hiányának esete lenne.
- Ezért ezt a mutatószámot a *j*-edik változóhoz tartozó variancia inflációs tényezőnek (Variance Inflator Factor) VIF_j mutatónak nevezzük:

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

VIF-mutató

- $1 < VIF \leq \infty$

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

- VIF=1 ha $R_j^2=0$ (amikor a j. magyarázó változó nem korrelál a többi magyarázó változóval)
- VIF$\Rightarrow\infty$ $R_j^2=1$ (a j. magyarázó változó pontosan kifejezhető a többi lineáris kombinációjaként)

- $1 < VIF \leq 2$ - gyenge multikollinearitás
  $2 < VIF \leq 5$ - erős zavaró multikollinearitás
  $5 < VIF$ - nagyon erős, káros multikollinearitás

- A *VIFj -mutató reciprokát tolerancia**mutatónak** nevezzük.*

$$Tolerancia = \frac{1}{VIF_j}$$

- ***Értéke:*** $0 \leq \text{Tolerancia} \leq 1$.
- Minél nagyobb a multikollinearitás mértéke annál közelebb van a mutató értéke a nullához.

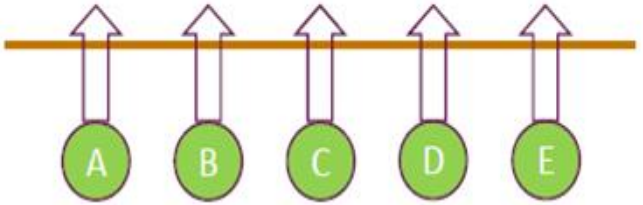
Káros multikollinearitás esetén…

- megkeressük azokat a magyarázó változókat, amelyek a zavart okozzák, és elhagyjuk őket a modellből;
- az egymással nagyon szoros kapcsolatban álló magyarázó változókat egy új változóban összevonjuk(főkomponensek), amely másabb lesz, mint az eredeti, de hordozza azok információtartalmát.

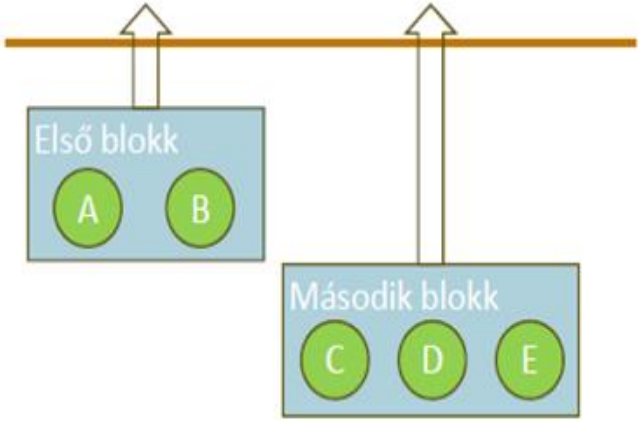
- Ridge regresszió (torzított, de kisebb varianciájú becslőfüggvényt ad, mint a legkisebb négyzetek becslőfüggvénye)

Modell építési lehetőségek

- **Módok:**
  - **Enter (forced entry)**
    - A modellbe egyszerre lépnek be a változók

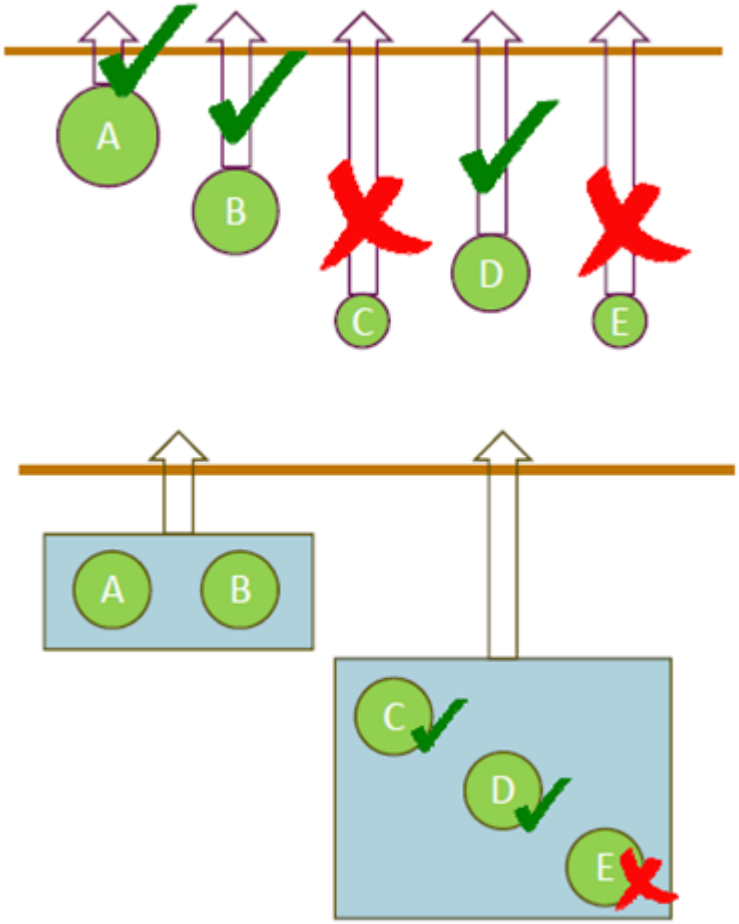


- **Hierarchikus (blokkonként – blockwise model)**
  - A változók sorrendjét a kutató adja meg (előzetes tudása, sejtése alapján)
  - Először azok a változók kerülnek a modellbe,
    - melyek a hipotézisek szempontjából fontosabban
    - melynek már ismerjük a hatását
    - melynek várhatóan legnagyobb a hatása
  - Egy blokkba
    - több változó is tartozhat,
    - és ezeknek külön megadhatjuk a belépési módját



Modell építési lehetőségek

- **Módok:**
  - **Stepwise módszerek**
    - A modellbe való kerülést matematikai kritériumok szabják meg
    - Ha egy prediktor nem tud szignfikáns mértékben hozzáadni a modellhez, nem kerül bele
    - Több fajta is létezik:
      - Forward, Backward, Stepwise
  - **Összetett modellek**
    - A beléptetés módjait lehet kombinálni
    - Pl. egyik blokkban Enter módszerrel lépnek be a változók, másik blokkban stepwise-zal



Milyen eljárást használjunk: Blockwise, Enter vagy Stepwise?

- **Ha előzetesen van elképzelésünk a megoldásról, akkor Blockwise (hierarchikus) módszer ajánlott:**
  - a számunkra fontos változókat hangsúlyozottan használhatjuk
- **Exploratív vizsgálatnál Enter módszer ajánlott:**
  - a változó modellbe kerülését nem befolyásolja az előzetes tudás vagy elképzelés,
  - pontos képet kapunk a függő és független változók kapcsolatáról, illetve a független változók kapcsolatrendszeréről.
- **Takarékos vagy gazdaságos (parsimonious) modell keresésénél Stepwise:**
  - a lehető legkevesebb prediktorral a lehető legjobb becslést tenni,
  - hátránya: a változókról önmagukban kevés információt kapunk,
  - overfitting veszély!
  - támadják: véletlen és matematikai döntéseken múlik, a kutatónak „nincs beleszólása” a modellbe.

Többszörös regresszió használatához a feltételek

- **Linearitás a függő változóval:** ha ez nincs, akkor alábecsüljük az y-t, pontatlan a modell.
- **Mintaszám:** kis elemszám növeli a β-hibát. Ökölszabály betartása többváltozós vizsgálatoknál.
- **Nincsenek több dimenziós extrém (outlier) értékek:** az együtthatók torzítását eredményezik.
- **Nincs multikollinearitás:** az összefüggő változók nem értelmezhetők. Ki kell hagyni a kevésbé fontos változót.
- **Minden változónak van varianciája:** nincs konstans változónk.
- **Nincs kovariáns** (külső befolyásoló változó).
- **Független hibák:** a belső korrelációk a CI-t, a szignifikancia értékeket torzítják.
- **Hiba normális eloszlása:** a normalitás sérülése, más feltételek sérüléseként keletkezik.
- **Változók típusai:** dummy-változó is engedett (pl. nem).
- Fennáll a homoszkedaszticitás (varianciák homogenitása vagy szóráshomogenitás): a heteroszkedaszticitás rontja a konfidencia-intervallumokat, torzítja a szignifikancia értékeket.

SAS Code : Squared Partial and Semi-Partial Correlation

- In PROC REG, the **PCORR2** option tells SAS to produce squared-partial correlation and **SCORR2** option tells SAS to produce squared semi-partial correlation. The **STB** option is used to generate standardized estimate and **TOL** is used to calculate tolerance.
- Proc Reg data= Readin;
  Model Overall = VAR1 - VAR5 / SCORR2
  PCORR2 STB TOL ;
  Run;

Parciális és semi parciális korreláció

Variable	Parameter Estimate	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II	Squared Partial Corr Type II
Intercept	-1.19483	0.0649	0	.	.
VAR1	0.76324	<.0001	0.66197	0.18325	0.42836
VAR2	0.13198	0.4219	0.10608	0.00365	0.01472
VAR3	0.48898	0.0008	0.32506	0.07129	0.2257
VAR4	-0.18431	0.2715	-0.10547	0.00689	0.02742
VAR5	0.00052549	0.1848	0.12424	0.01009	0.03961

The **squared semi-partial correlation between Overall and VAR1** tells us model R-square is added by 0.18325 if VAR1 is included in the model.

The **squared partial correlation between Overall and VAR1** tells us the proportion of variance in Overall that is not explained by the other independent variables, 43% is explained by VAR1.

Shrinkage módszerek

- **Ridge regresszió**
- **Lasso**
- **Legkisebb szög regresszió (2004)**
- **Bilineáris regresszió**