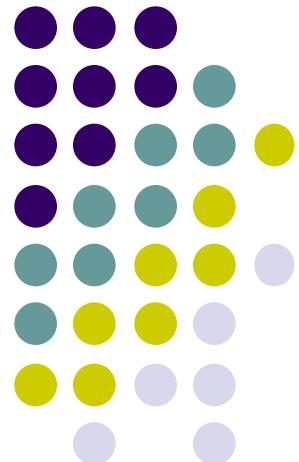




TÖBBVÁLTOZÓS ADATELEMZÉS

Lineáris regresszió

2020.10.05.



© Dr. Vékás Péter, e-mail: peter.vekas@uni-corvinus.hu

BCE Matematikai és Statisztikai Modellezés Intézet



Lineáris regresszió

- Numerikus változót becsül más változók lineáris függvényeként.
- Az egyik legnépszerűbb statisztikai módszer, használható például lakásárak, kamatok, stb., szinte akármilyen mennyiség becslésére.
- Legkisebb négyzetek elve (*OLS* = *ordinary least squares*): a becsült és tényleges értékek különbségeinek négyzetösszegét minimalizáljuk.
- Az OLS-t a 18 éves Gauss találta ki 1795-ben.



Alkalmazások

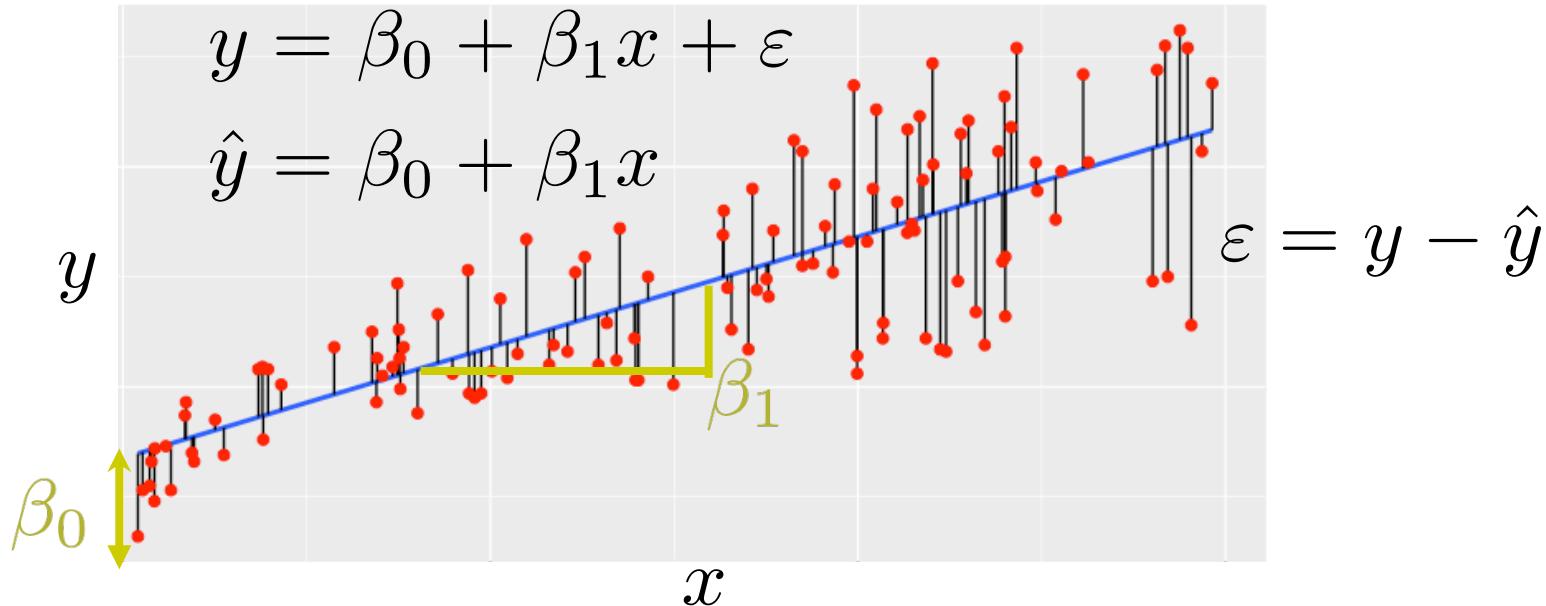
- Mitől függnek az autóárak?
- Melyik lakásokat lehet érdemes befektetési céllal megvenni?
- Megéri-e továbbtanulni?
- Van-e nemek szerinti diszkrimináció a fizetésekben?
- stb. stb.

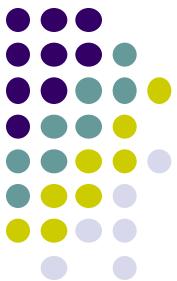


Lineáris regresszió két dimenzióban

Interaktív demó:

<https://www.geogebra.org/m/AuRrgqNV>





Terminológia

- y :
kimenet
eredményváltozó
függő változó
- x :
prediktor
magyarázó változó
független változó

Lineáris regresszió több dimenzióban

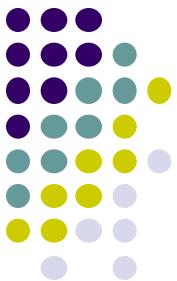


- A becsült kimenet a prediktorok lineáris függvénye:

$$\hat{y} = \beta_0 + \sum_{j=1}^p \beta_j x_j$$

- A valóság nem tökéletes, ezért kell hibatag:

$$y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \varepsilon$$



Vektorok

Kimenet:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Becsült kimenet:

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

Együtthatók:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$



Prediktorok mátrixa

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

A β_0 konstans oszlopa csupa 1, mert minden megfigyelés egyenletében 1-szer szerepel.

A többi prediktor értékei egy-egy oszlopban szerepelnek.

Minden megfigyeléshez egy-egy sor tartozik



Legkisebb négyzetek (OLS)

- Minimalizáljuk a hibatagok négyzetösszegét:

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min .$$

- Az optimális együtthatók mátrixműveletekkel előállíthatók az adatokból:

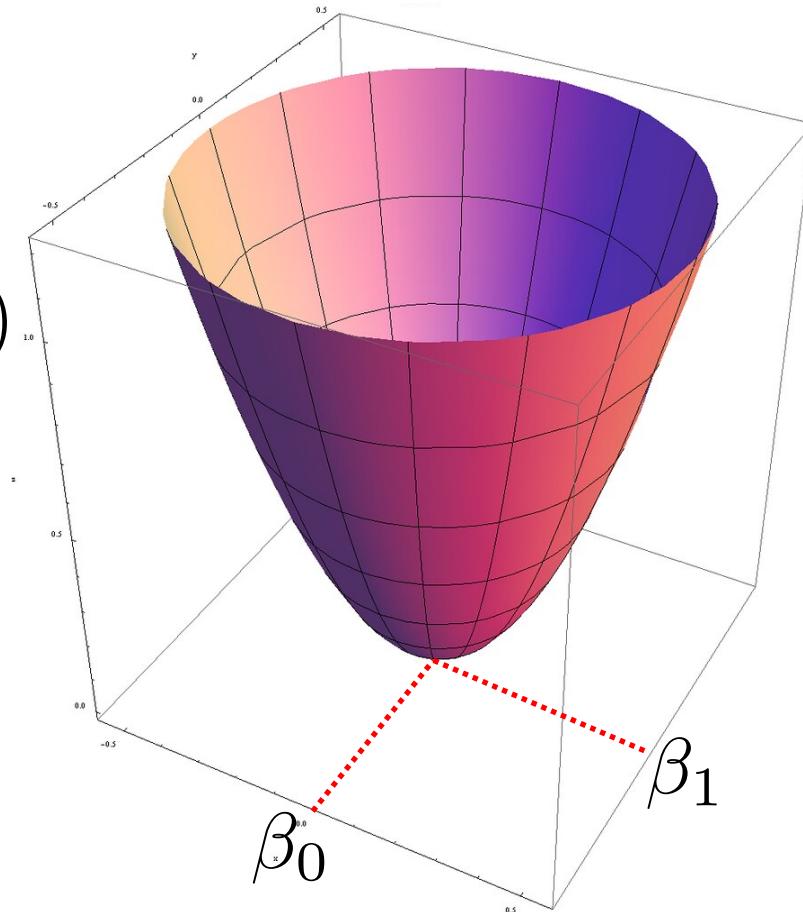
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

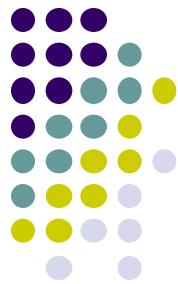
- Létezik a hibatagok abszolút értékeinek összegét minimalizáló módszer is, de nem terjedt el (*LAD = Least Absolute Deviations*).



Legkisebb négyzetek (OLS)

$$SSE(\beta_0, \beta_1)$$





Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés

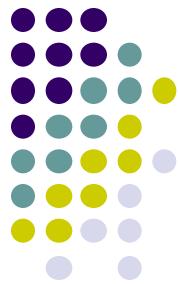


2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>



Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés



2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>



Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés

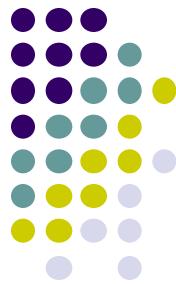


2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>



Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés



2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>



Adatelőkészítés

- Adatok beszerzése, tisztítása, transzformációja, pótlása, stb.
- Pénzben kifejezett, *pozitív* (!) változókat (például ár, árfolyam, jövedelem, munkabér, vagyon stb.) logaritmikusan szokás transzformálni (így relatív, százalékos változásokat értelmezünk).
- Kategorikus prediktorokból dummy változókat kell képezni (R-ben automatikus).



Dummy változók

- Csak numerikus változók szerepelhetnek a lineáris regresszióban. Kategorikus prediktorokból *dummy* változókat kell képezni.
- R-ben minden kategória automatikusan kap dummy változót (értéke 1 az adott kategóriában, 0 egyébként), *kivéve* az ábécében az elsőt.
- Ez a *referenciakategória*: a többi kategória hatását ehhez viszonyítjuk.



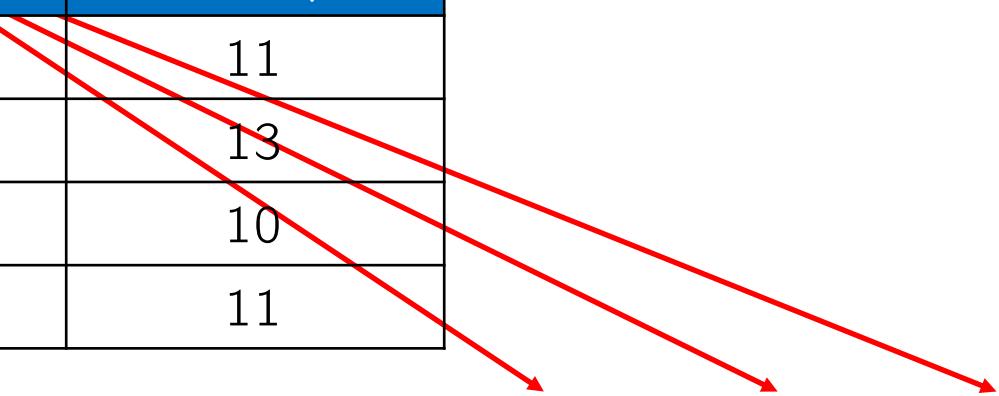
Kategorikus prediktor

Név	Egyetem	Óraszám/hét
Anna	BCE	11
Béla	BME	13
Csilla	BCE	10
Dániel	ELTE	11



Dummy változók

Név	Egyetem	Óraszám/hét
Anna	BCE	11
Béla	BME	13
Csilla	BCE	10
Dániel	ELTE	11



Név	Egyetem	Óraszám/hét	BCE	BME	ELTE
Anna	BCE	11	1	0	0
Béla	BME	13	0	1	0
Csilla	ELTE	10	1	0	0
Dániel	BCE	11	0	0	1



Dummy változók

Név	Egyetem	Óraszám/hét
Anna	BCE	11
Béla	BME	13
Csilla	BCE	10
Dániel	ELTE	11

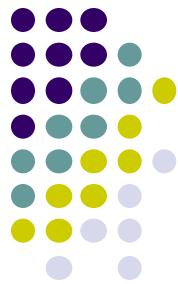
Referencia-kategória:
az ábécében
az első

Név	Egyetem	Óraszám/hét	BCE	BME	ELTE
Anna	BCE	11	1	0	0
Béla	BME	13	0	1	0
Csilla	ELTE	10	1	0	0
Dániel	BCE	11	0	0	1



Mintaméret

- Statisztikai hüvelykujj-szabály: legyen legalább 5-ször, de inkább 10-szer annyi megfigyelés, mint becsült paraméter.
- Különben a becslések nagyon bizonytalanok lesznek.
- Például ne építsünk 10 prediktorral modellt 27 EU-tagra!



Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés



2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>

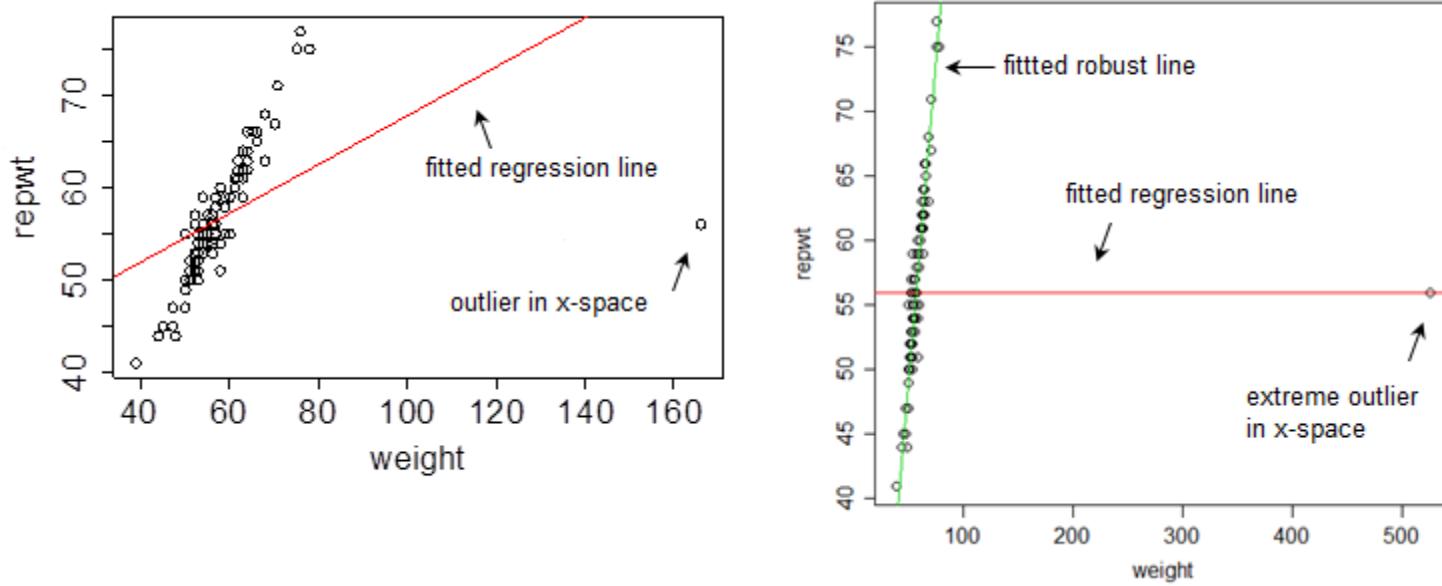


Problémák a lineáris regresszióban

- a. Kilógó értékek
- b. Multikollinearitás
- c. Hibatagok nemnormalitása
- d. Heteroszkedaszticitás
- e. Nemlinearitás
- f. Felesleges prediktorok



a. Kilógó értékek: durva és még durvább



Forrás: <https://stats.stackexchange.com>

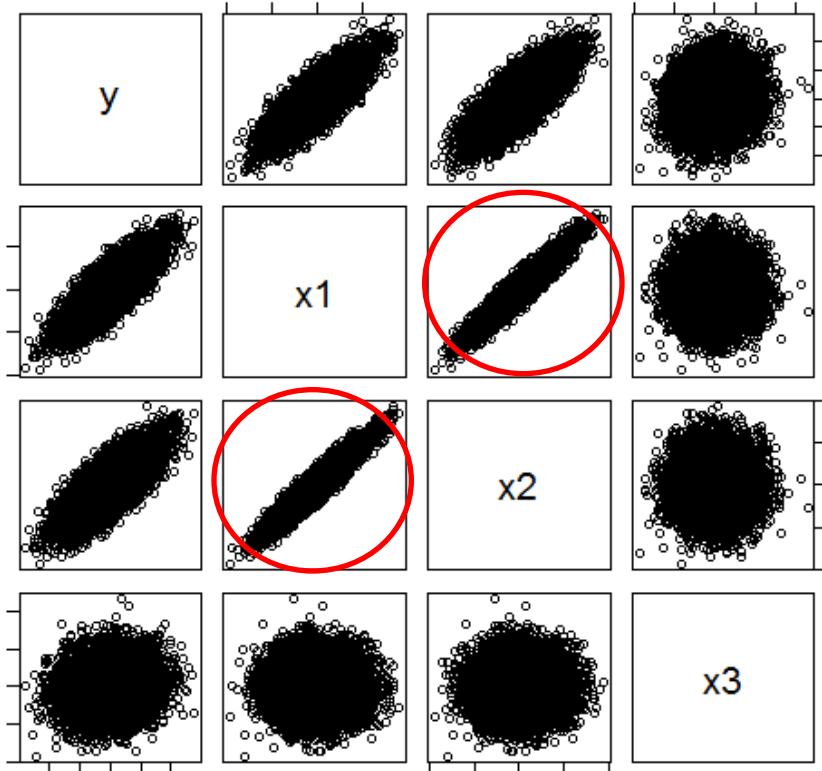


a. Kilógó értékek

- A különc megfigyelések maguk felé téríthatik el a becsült függvényt, elrontva az illeszkedést.
- Az OLS módszer négyzetre emeli a hibákat, így a távoli pontok hatása óriási lehet!
- *Studentizált hibatag*: -3 és 3 között OK, egyébként kilógó érték.
- Az utóbbiakat töröljük, esetleg külön elemezzük, mint érdekes eseteket.
- Adathibákat is találhatunk így (pl. $BMI = 300$).



b. Multikollinearitás: itt problémás lehet



Forrás: <https://stats.stackexchange.com/>



b. Multikollinearitás

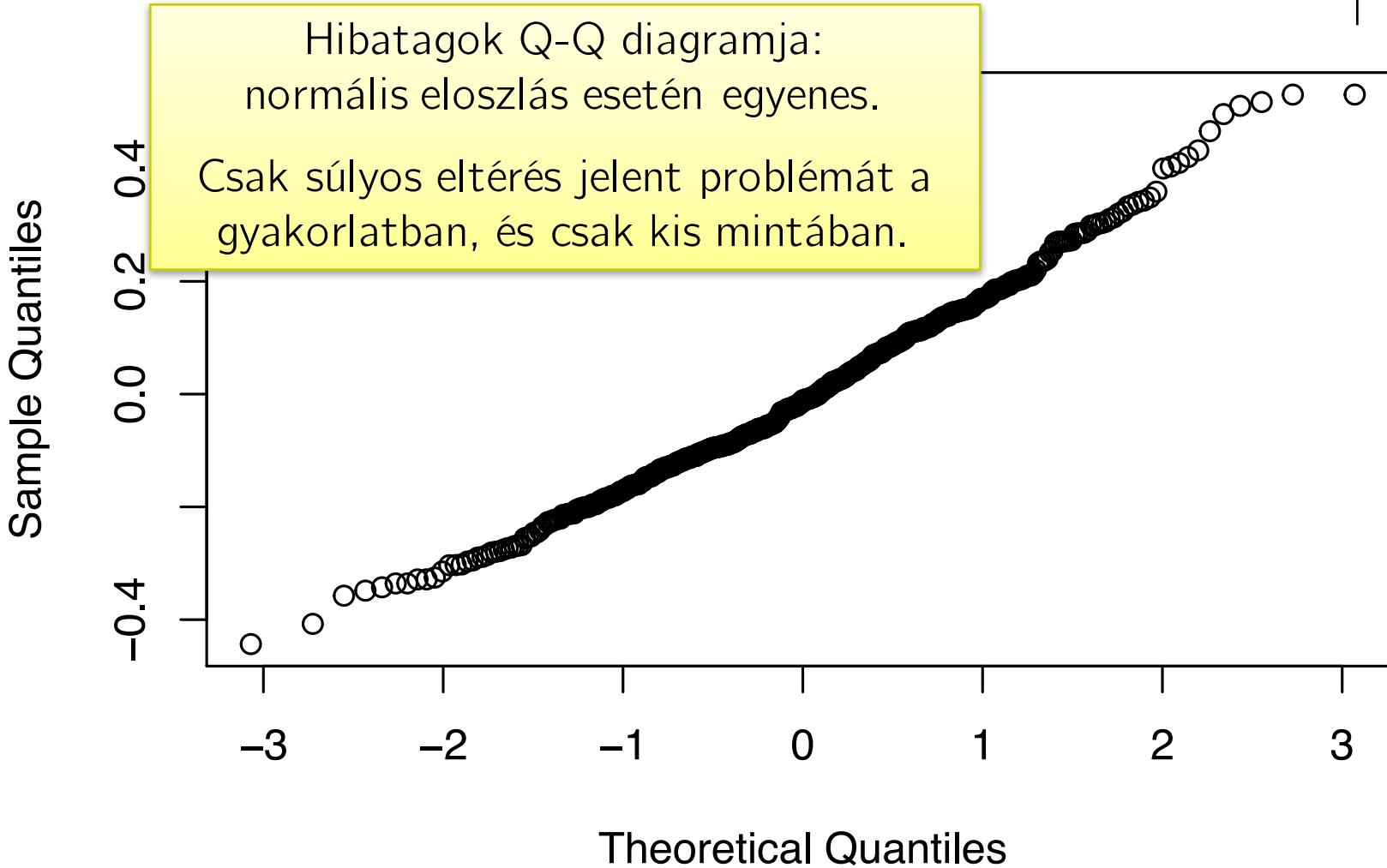
- *Multikollinearitás*: a prediktorok szorosan korrelálnak egymással.
- Bizonytalanná teszi a becsült együtthatókat, megnehezíti a modell értelmezését.
- Extrém esetben a becslés el sem végezhető!
- Változónkénti VIF: 5 felett zavaró, 10 felett pláne. Az egyik ilyen prediktor elhagyható.
- Kivétel: nem baj, ha egy prediktor korrelál a saját transzformáltjával (például négyzetével).



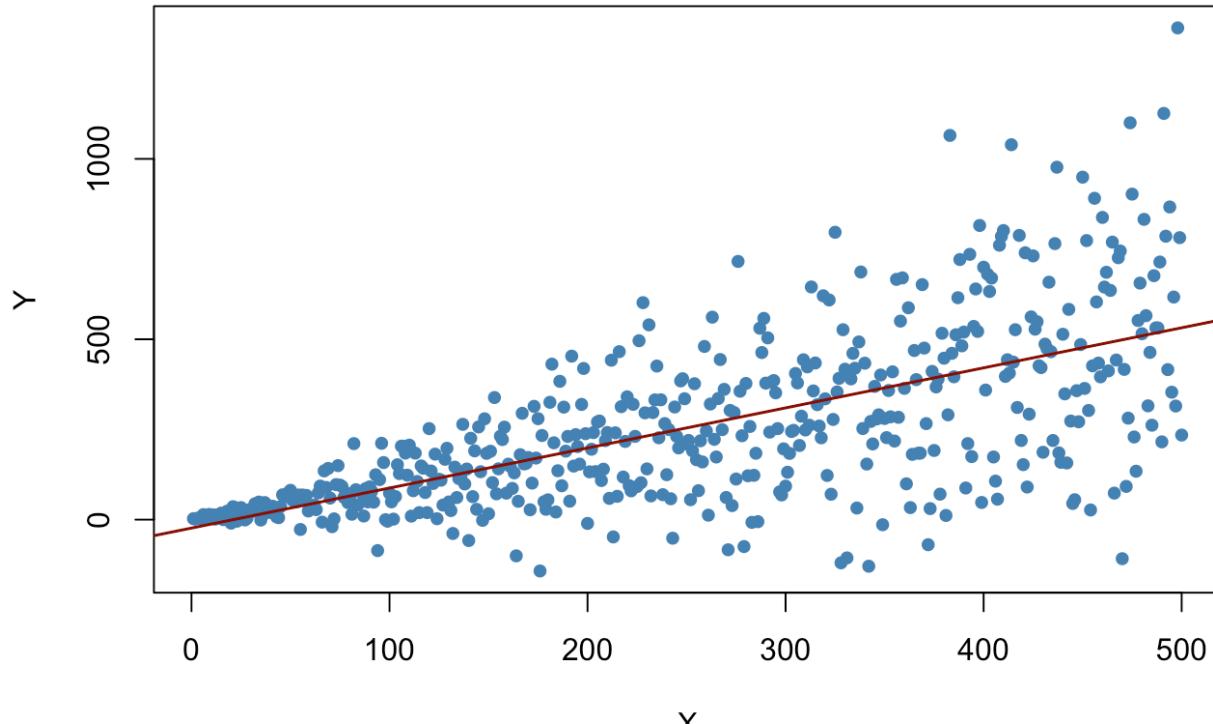
c. Hibatagok nemnormalitása

- A későbbi tesztek feltételezik a hibatagok normális eloszlását, egyébként tévedhetnek.
- FONTOS: csak a hibatagok normális eloszlása feltétel, a változóké nem!
- Nagy mintában (szimmetrikus hibatagok: $n > 100$, erősen ferde eloszlás: $n > 500$) nem jelent problémát.
- Balra ferde hibatagoknál segíthet y helyett $\ln y$ használata. Egyébként robustus regresszió használható (itt nem).

c. Hibatagok nemnormalitása: itt minden OK



d. Heteroszkedaszticitás: itt elég erős



Forrás: <https://www.econometrics-with-r.org/>

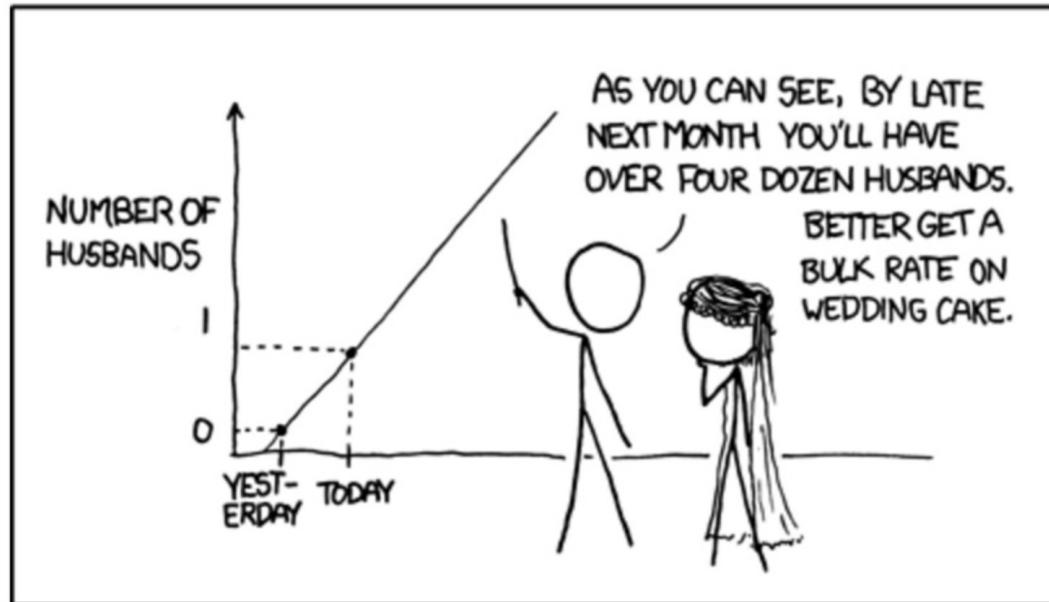


d. Heteroszkedaszticitás

- A későbbi tesztek feltételezik, hogy a hibatagok szórása állandó, azaz nem függ a prediktoroktól (*homoszkedaszticitás*).
- Egyébként téves eredményt adhatnak.
- Breusch-Pagan teszt:
 H_0 : a hibatagok homoszkedasztikusak, elvetése esetén a robustus teszteket kell elvégezni a későbbieken.



e. Nemlinearitás



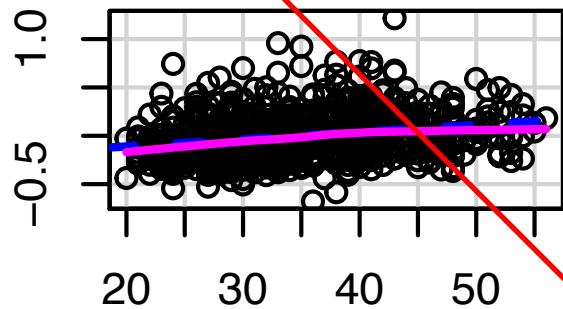
Forrás: [https://www.explainxkcd.com/
wiki/index.php/605:_Extrapolating](https://www.explainxkcd.com/wiki/index.php/605:_Extrapolating)



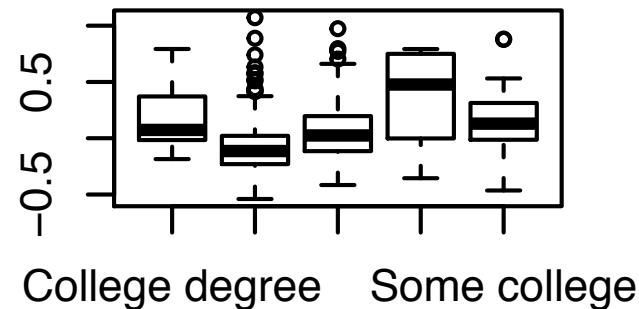
e. Nemlinearitás

- Ha a prediktorok hatása nemlineáris, a lineáris modell torzított, hamis összefüggéseket mutat.
- *RESET teszt*:
 H_0 : a magyarázó változók hatása lineáris, elvetése esetén vizsgálhatók a *CR diagramok*.
- Ezek egy-egy prediktor függvényében ábrázolják a kimenet többi prediktor által meg nem magyarázott részét, LOESS simítással.

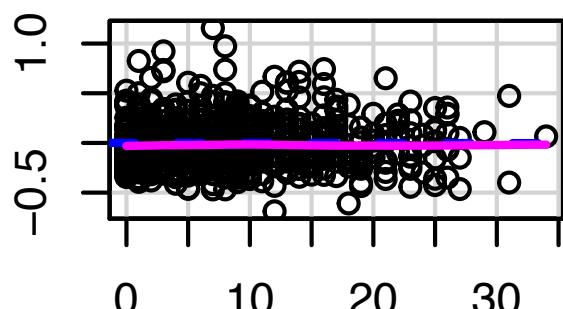
e. Nemlinearitás: a *debtinc* problémásnak tűnik



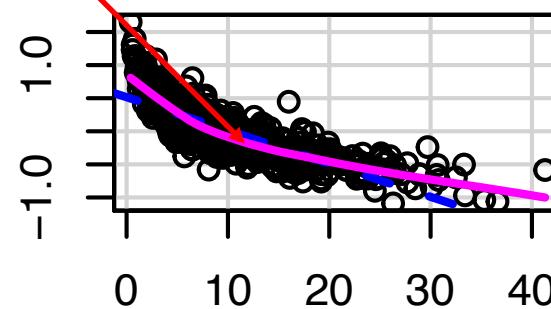
age



ed

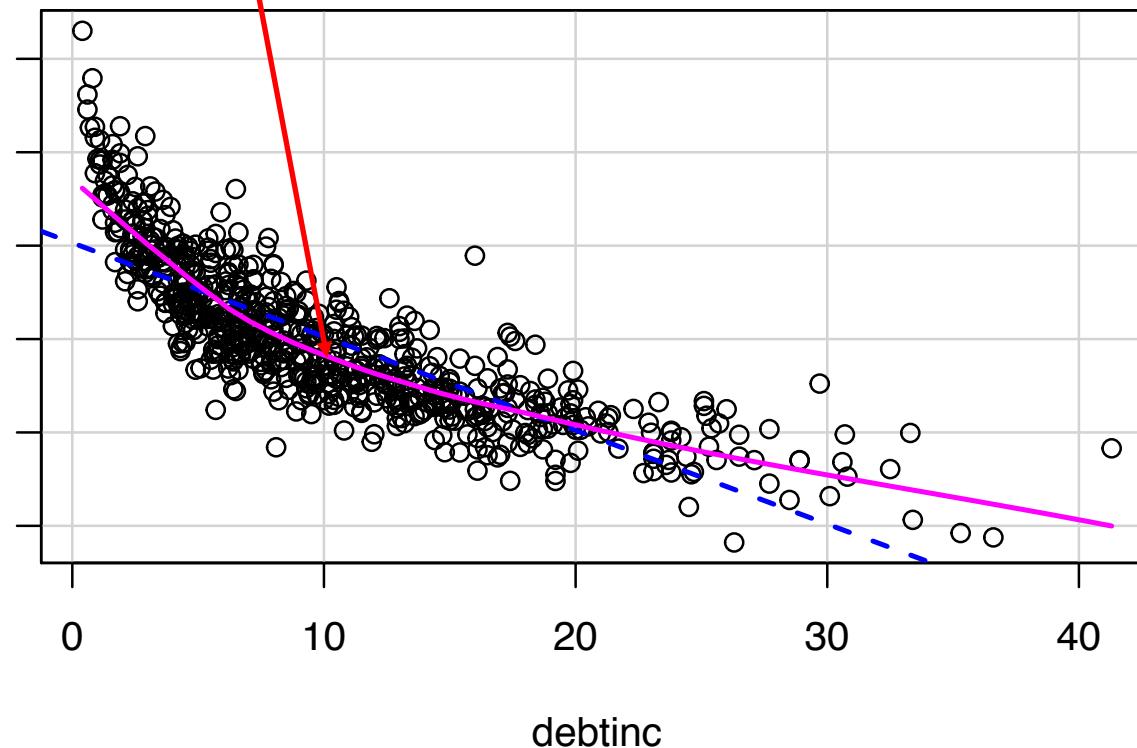
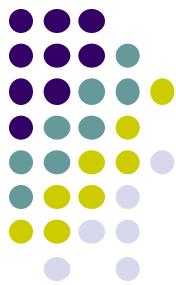


address



debtinc

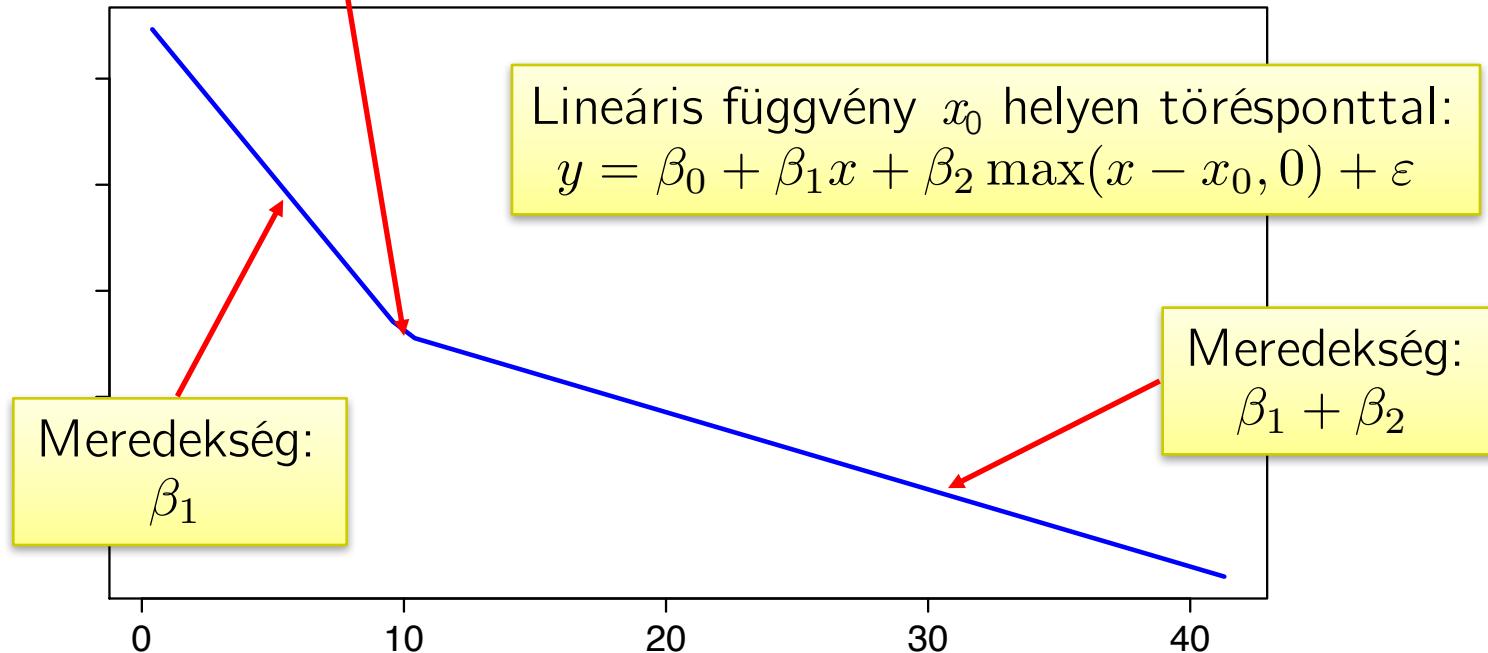
e. Nemlinearitás: $debtinc = 10$ körül töréspont

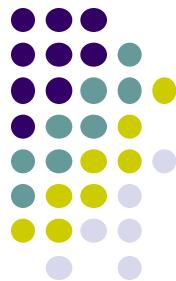




e. Nemlinearitás:

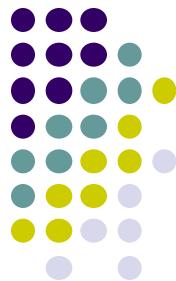
debtinc = 10 körül töréspont





e. Nemlinearitás: lehetséges megoldások

- A prediktorok transzformálása segíthet:
 - Lineáris függvény törésponttal (itt csak ez)
 - Másodfokú függvény: x és x^2 is prediktor
 - Egyéb függvény (például logaritmus)
 - Prediktorok kategorikussá alakítása (például dummy: $D = 1$, ha $debtinc > 10$)
 - Keresztszorzatok: prediktorok szorzata



f. Felesleges prediktorok

- *Parszimónia elve (Ockham borotvája):*
A legegyszerűbb működő magyarázatot fogadjuk el egy jelenségre.
- A statisztikában azt jelenti, hogy az azonos teljesítményű modellek közül a legkevesebb paraméterrel rendelkezőt preferáljuk.
- Különben nehezen értelmezhető, feleslegesen bonyolult a modell.



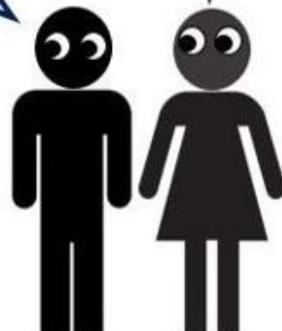
Ockham's Razor

Why did the tree fall down?



"I agree."

"It was the wind. It is the simpler explanation."



Two Explanations

1. The wind knocked down the tree.
2. Two meteorites. One hit the tree and knocked it down. Then it hit the other meteorite, thus obliterating evidence of its existence.

***When there are two explanations,
choose the simpler one***

Forrás: <https://marketbusinessnews.com/>



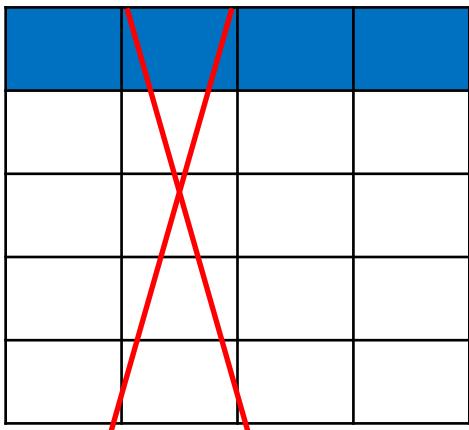
f. Felesleges prediktorok tesztjei

- *t-teszt:* $H_0 : \beta_j = 0$,
elfogadása esetén törölhető a j -edik prediktor
(a β_0 konstanst nem szokás törölni).
- *Globális F-teszt:* $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$,
elfogadása esetén az összes prediktor törölhető,
a modell egy konstansból áll (üres).
- *Parciális F-teszt:* $H_0 : \text{bizonyos } j\text{-kre } \beta_j = 0$,
elfogadása esetén törölhető a prediktorok adott
csoportja.

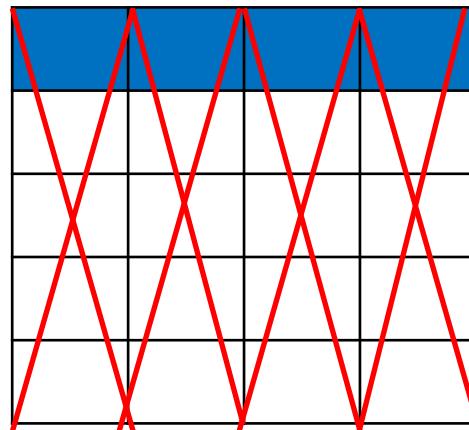


f. Felesleges prediktorok tesztjei

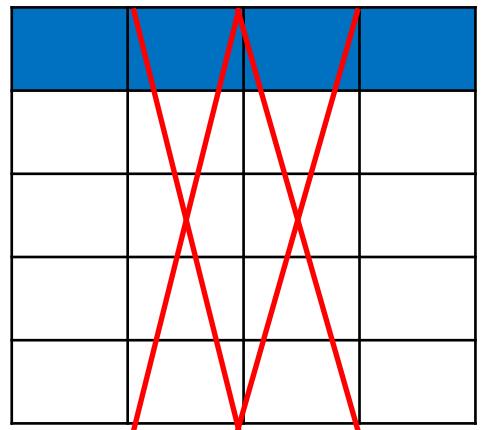
t-teszt

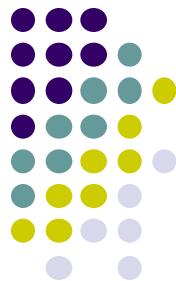


Globális *F*-teszt:



Parciális *F*-teszt:





f. Felesleges prediktorok kiválogatása

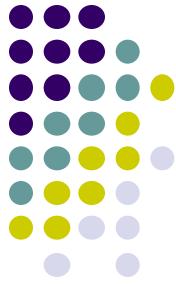
- A teljes modellből parciális F -teszttel halászhatjuk ki a felesleges prediktorokat.
- Lehetőleg szignifikánsak legyenek a bennmaradó prediktorok.
- Vannak automatikus válogató módszerek is (később a kurzuson!).



Diagnosztika (összefoglaló)

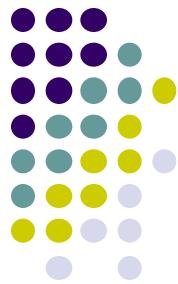
Jelenség	Miért baj?	Diagnózis	Megoldás
a. Kilógó értékek	Torzított modell	Stud. hibatagok	Megfigyelések elhagyása
b. Multikollinearitás	Bizonytalan együtthatók	VIF	Prediktorok elhagyása
c. Hibatagok nemnormalitása	Hamis tesztek	Q-Q ábra, hisztogram	Balra ferde hibatagok: $\ln y$ használata
d. Heteroszkedaszticitás	Hamis tesztek	Breusch-Pagan teszt	Robusztus tesztek
e. Nemlinearitás	Torzított modell	RESET teszt, CR ábrák	Prediktorok transzformálása
f. Felesleges prediktorok	Nehéz értelmezés	t és F tesztek	Prediktorok elhagyása

+ Autokorreláció: keresztmetszetben nem (csak idősort, panel, földrajzi).



Diagnosztika (összefoglaló)

- Ez egy javasolt sorrend, de a legjobb addig pofozgatni a modellt, amíg minden probléma (többé-kevésbé) meg nem oldódik.
- Van rengeteg egyéb teszt, diagram és technika a problémák felismerésére és javítására!



Lineáris regressziós modellezés lépései



5. Értelmezés



4. Diagnosztika



3. Adatelőkészítés

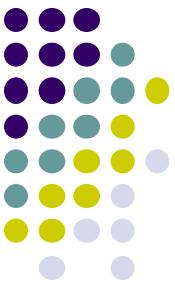


2. Adatgyűjtés



1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: <https://flaticon.com>



Értelmezés

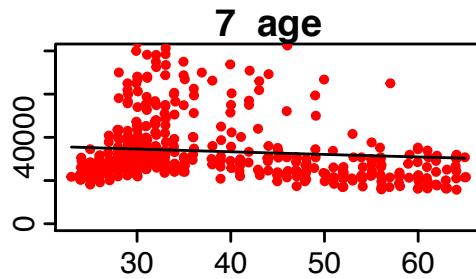
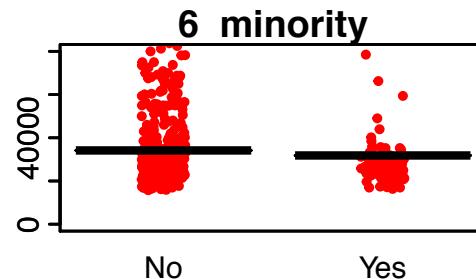
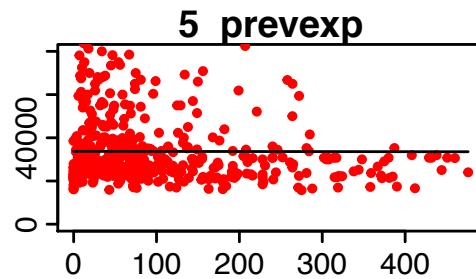
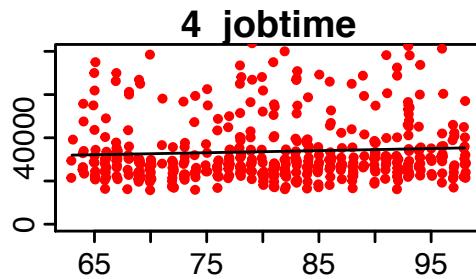
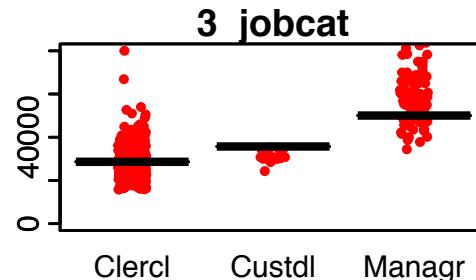
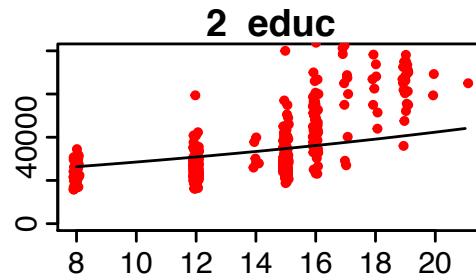
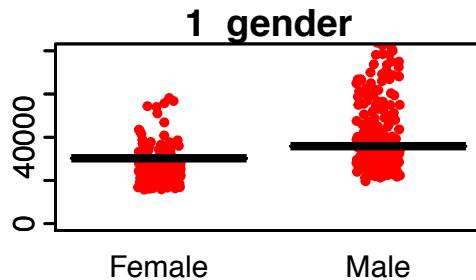
- a. Prediktorok parciális hatása
- b. Prediktorok fontossága
- c. Modell jósága
- d. Korreláció vagy ok-okozat?

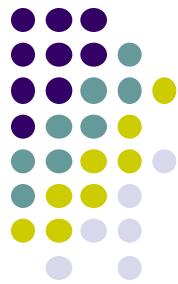


a. Mi a prediktorok parciális hatása?

	β_0	β_j (x_j numerikus)	β_j (x_j dummy)
	Ha minden $x_j = 0$, akkor...	Ha x_j cet.par. egységnnyivel nő, akkor...	Az adott kategóriában cet.par...
bal oldal: y	... y várható értéke β_0 y várhatóan β_j -vel nő.	... y várhatóan β_j -vel több, mint a ref.kat.-ban.
bal oldal: $\ln y$... $\ln(y)$ várható értéke β_0 y várhatóan e^{β_j} - szeresére nő.	... y várhatóan e^{β_j} - szer több, mint a ref.kat.-ban.

a. Mi a prediktorok parciális hatása?





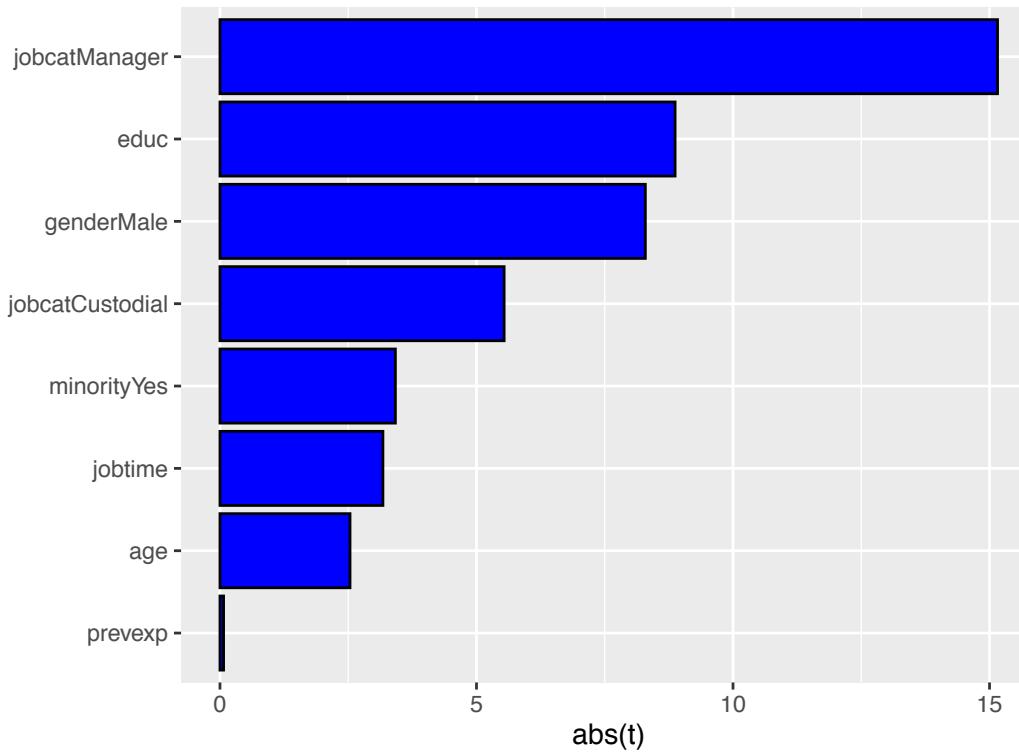
a. Mi a prediktorok parciális hatása?

- A *parciális regressziós diagram* (előző dia) a becsült kimenetet ábrázolja egyesével a prediktorok függvényében.
- minden részdiagram feltételezi, hogy az összes többi prediktort átlagos értékükön rögzítjük (ceteris paribus elemzés).



b. Melyik prediktorok a legfontosabbak?

- A t -statisztikák abszolút értékei alapján rangsorolható a prediktorok fontossága.





c. Mennyire jó a modell?

- R^2 : a tényleges és becsült kimenet közötti korreláció négyzete, 0 és 1 között méri a modell jóságát. Nincs univerzális kritérium arra, hogy mi számít jó értéknek.
- *Sztenderd hiba*: a hibatagok becsült szórása, a „tipikus” tévedés nagysága.



d. Korreláció vagy ok-okozat?

- Nagy a kísértés arra, hogy azt mondjuk, a prediktor változása “okozza” a kimenet változását (és nem csak “együtt jár” vele).
- Hárrom eset, amikor egy együttható nem értelmezhető így (*torzított*):
 - kihagyott változó,
 - fordított okság,
 - mérési hiba.
- Ezen esetek összefoglaló neve *endogenitás*.

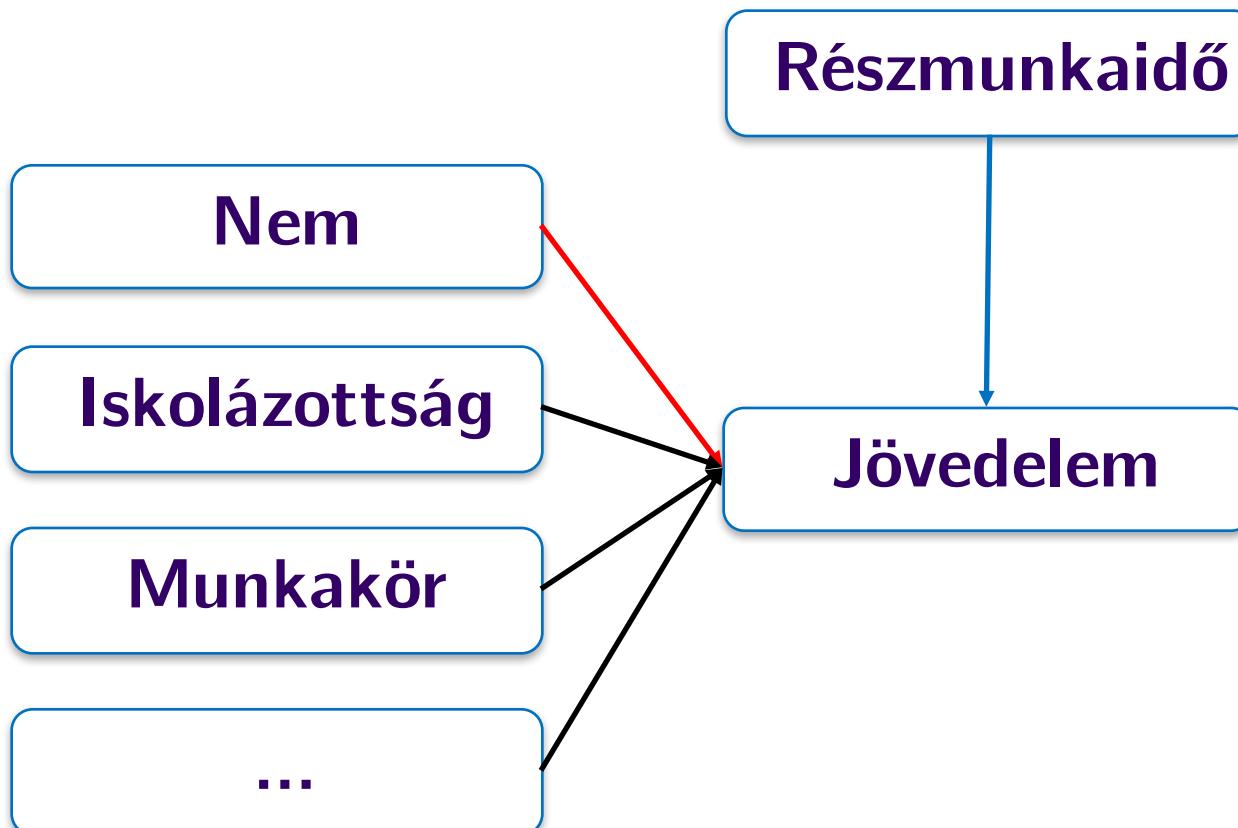
d. Korreláció vagy ok-okozat: kihagyott változók



- Ha x és y változókkal korreláló prediktort kihagyunk, az oksági interpretáció elromlik.
- Példa: alkalmazottak,
 y : jövedelem,
 x : nem és egyéb változók.
- Kismamáknál gyakori a részmunkaidő.
- Ezt a változót kihagyva a nem együtthatója torzul: belekeveredik a részmunkaidő hatása is.



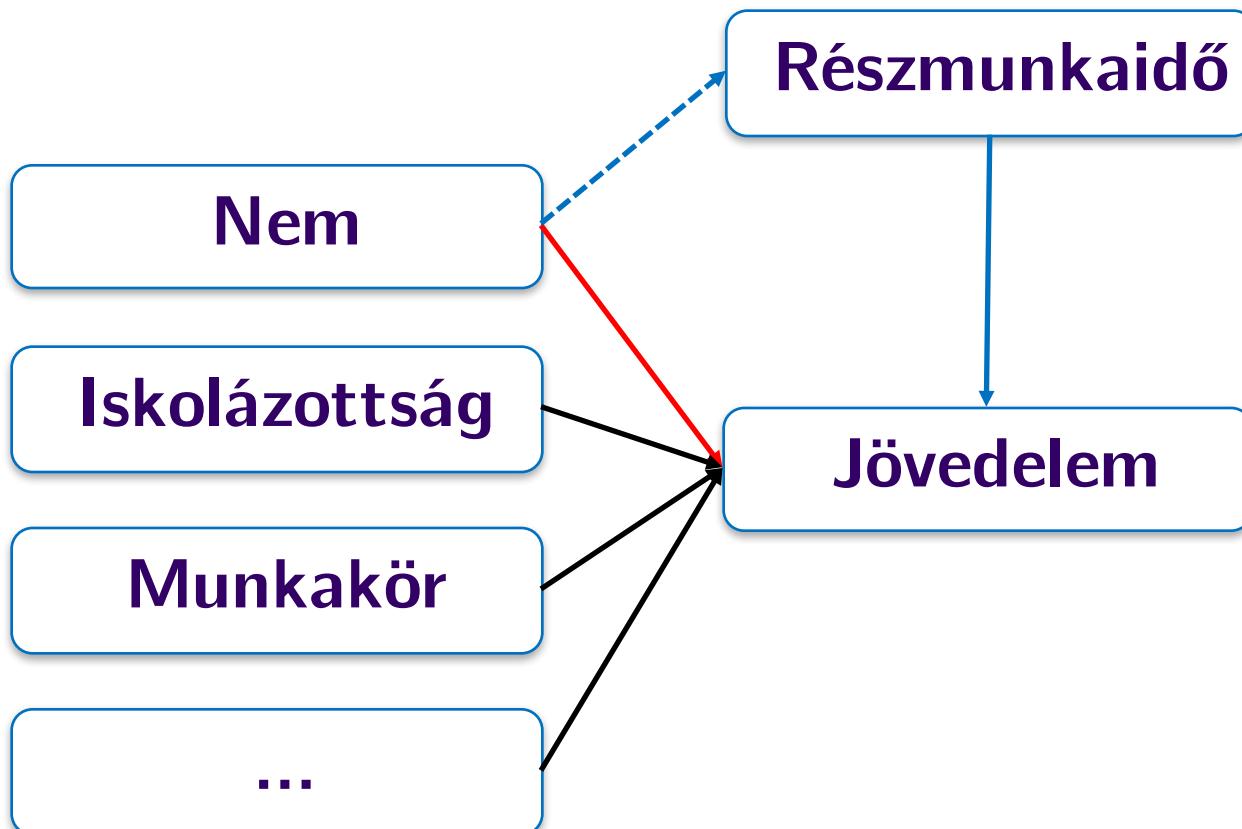
Helyes modell



A modellben külön szerepel a **nem** és a **részmunkaidő** hatása is.



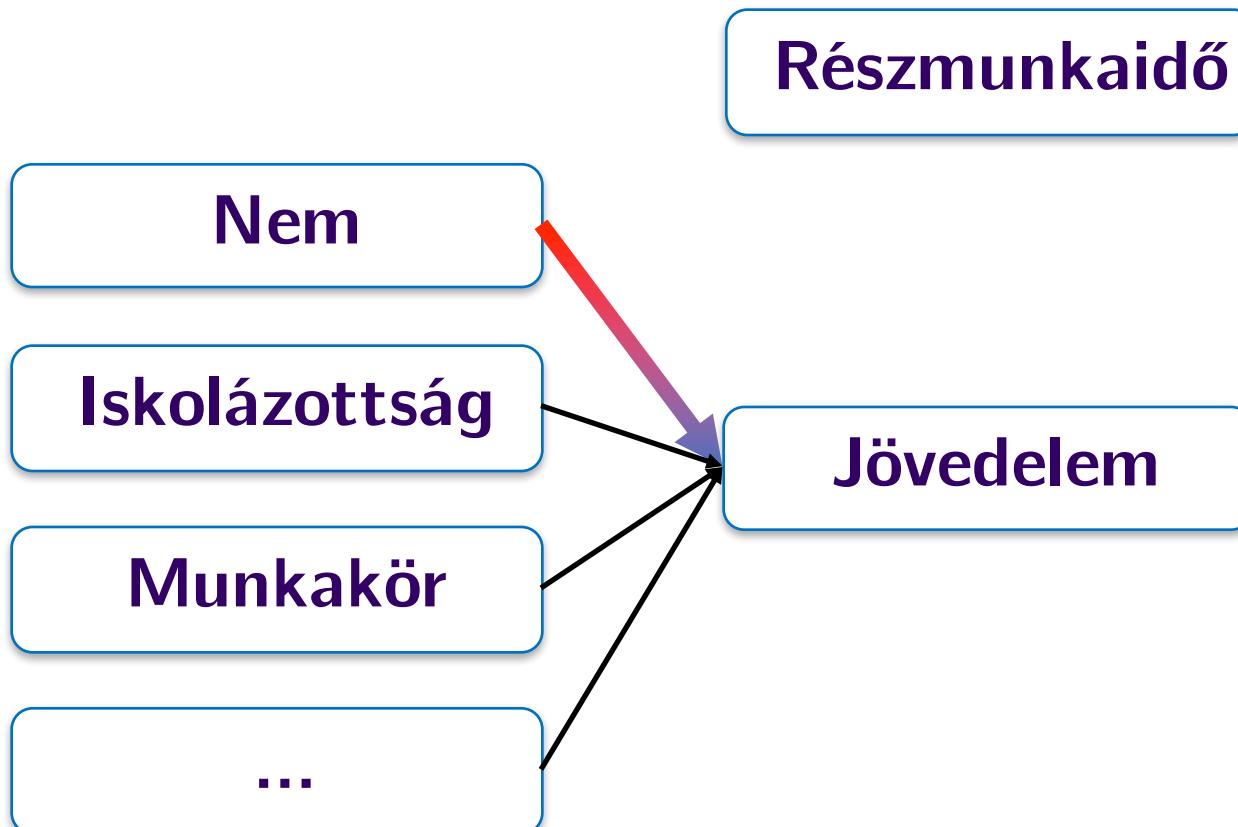
Helyes modell



A nem összefügg a részmunkaidővel, így közvetetten is hat a fizetésekre.



Kihagyott változóval



A részmunkaidőt kihagyva a nem valódi oksági és (rész munkaidőn keresztüli) közvetett hatása összem osódik.

d. Korreláció vagy ok-okozat: fordított okság

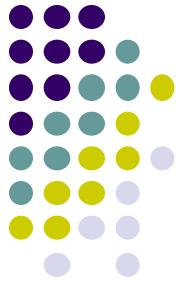


Forrás: <https://tinyurl.com/hvnayu7>

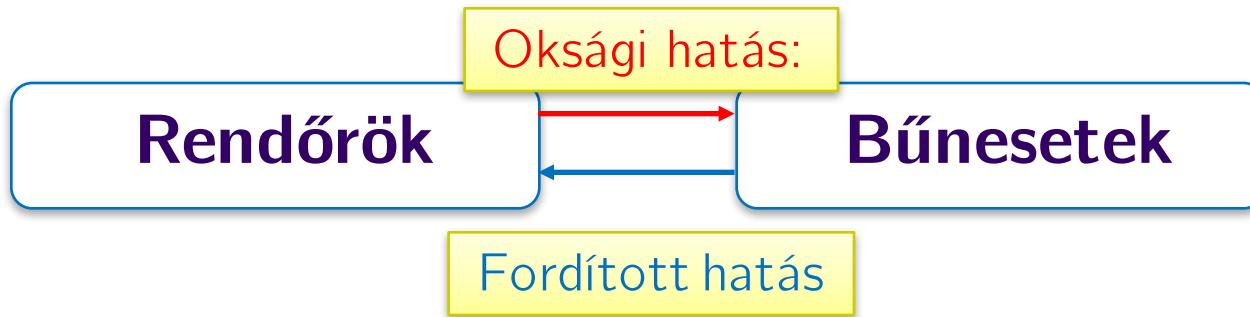
d. Korreláció vagy ok-okozat: fordított okság



- Ha y is hat x -re, az oksági interpretáció elromlik.
- Példa: városok,
 y : 1000 főre jutó bűncselekmények,
 x : 1000 főre jutó rendőrök.
- A rendőrök csökkentik a bűncselekményeket, de a bűncselekmények “vonzzák” a rendőröket!
- A rendőrök együtthatója torzul, mert a két irányú hatás összekeveredik benne.

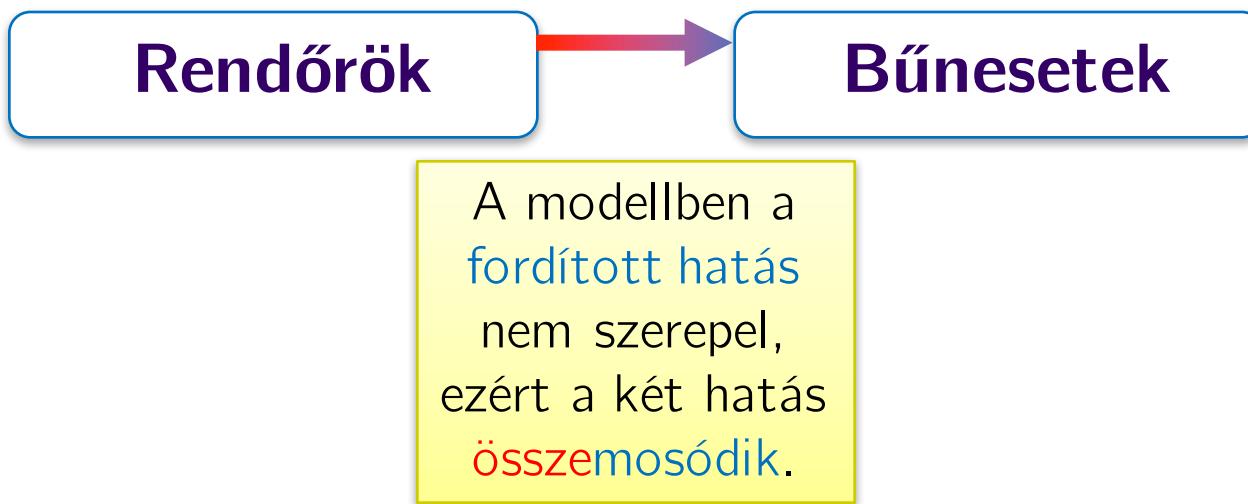


Fordított okság





Fordított okság



d. Korreláció vagy ok-okozat: mérési hiba

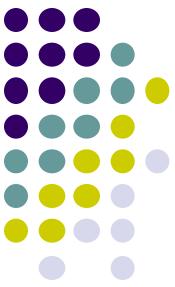


- Ha x -et hibával figyeljük meg, az együtthatója torzul (elromlik az oksági értelmezése).
- Például: magán egészségbiztosítás, y : éves egészségügyi kiadások, x : cigareták száma naponta (önbevallás).
- Dohányosok díja magasabb: kevesebbet cigit vallanak be, torzul a hatás (de akár a véletlen, nem szisztematikus irányú hibák is torzítanak).
- Ilyenkor fontos az értelmezésben hangsúlyozni, hogy a “mért” x hatásáról van szó.

d. Korreláció vagy ok-okozat: endogenitás



- Az endogenitást nem teszteljük, hanem az adatok és a probléma ismeretében mérjük fel.
- Instrumentális változók módszerével eltüntethető a torzítás (a kurzuson nem).
- Ha fennáll, óvatosan értelmezzünk:
 x_j egységnnyi növekedése cet. par. az y változó β_j egységnnyi növekedésével “jár együtt” (nem “okozza”)!



Köszönöm a figyelmet!