

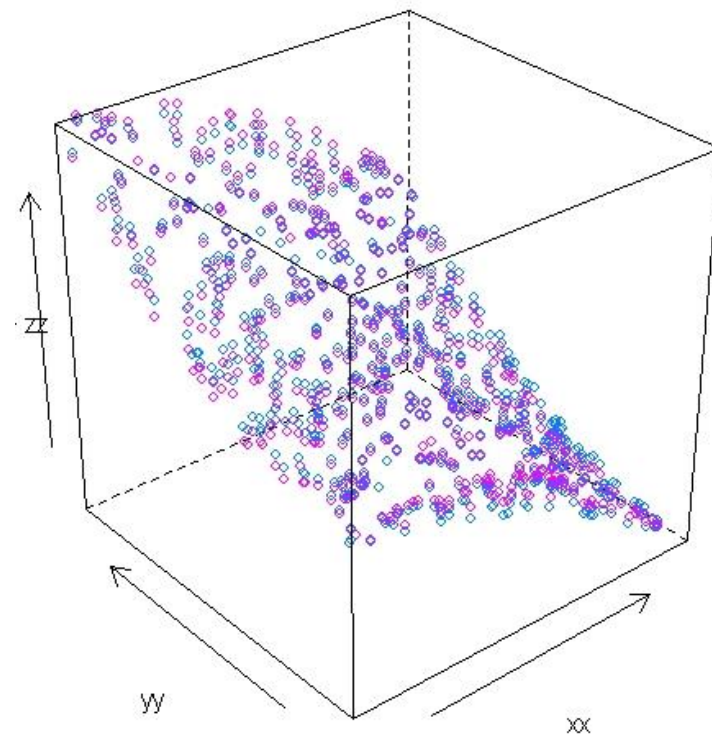
Többváltozós statisztika elméleti alapjai (BMNPS07500M)

Készítette: Soltész-Várhelyi Klára

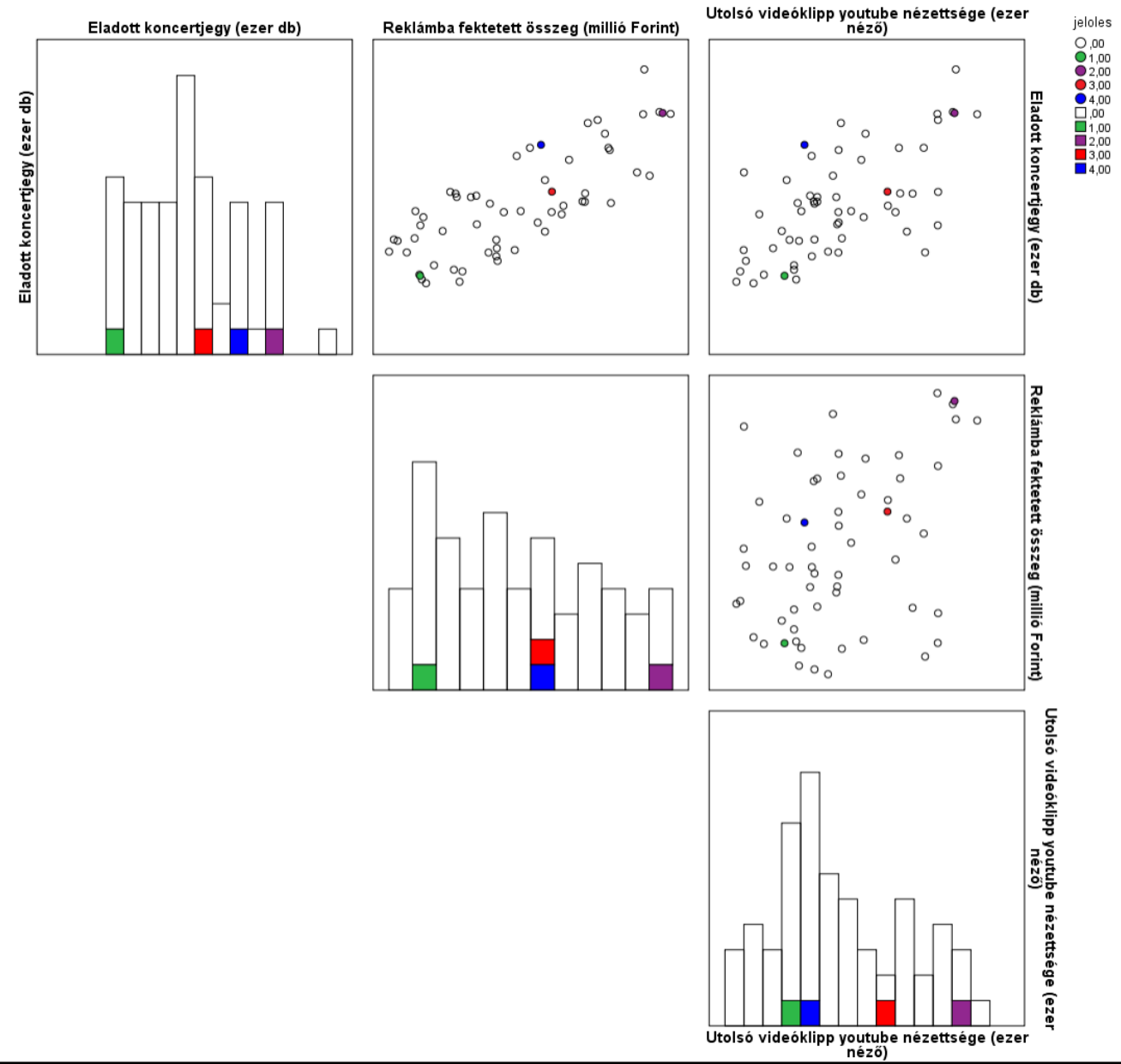
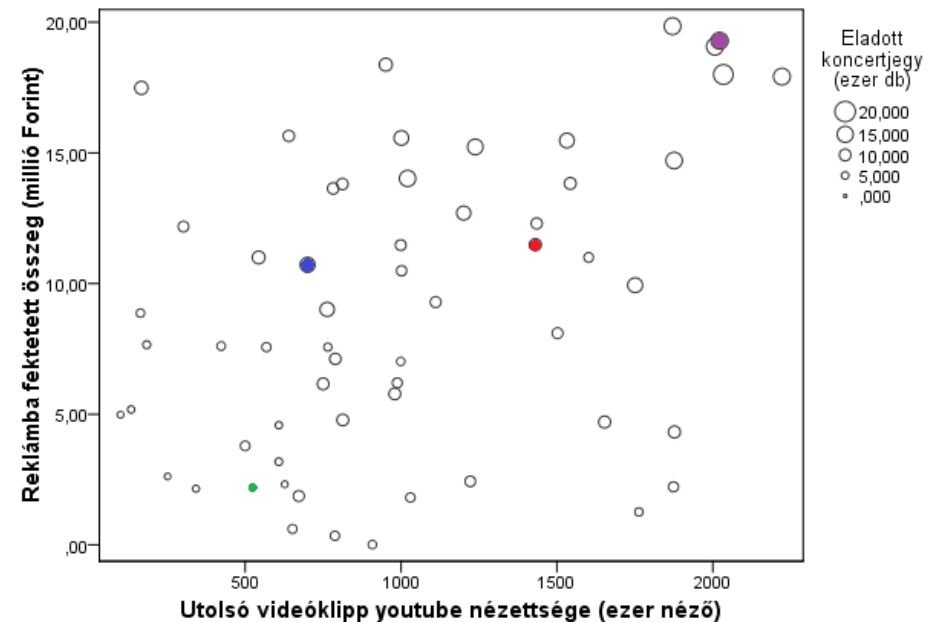
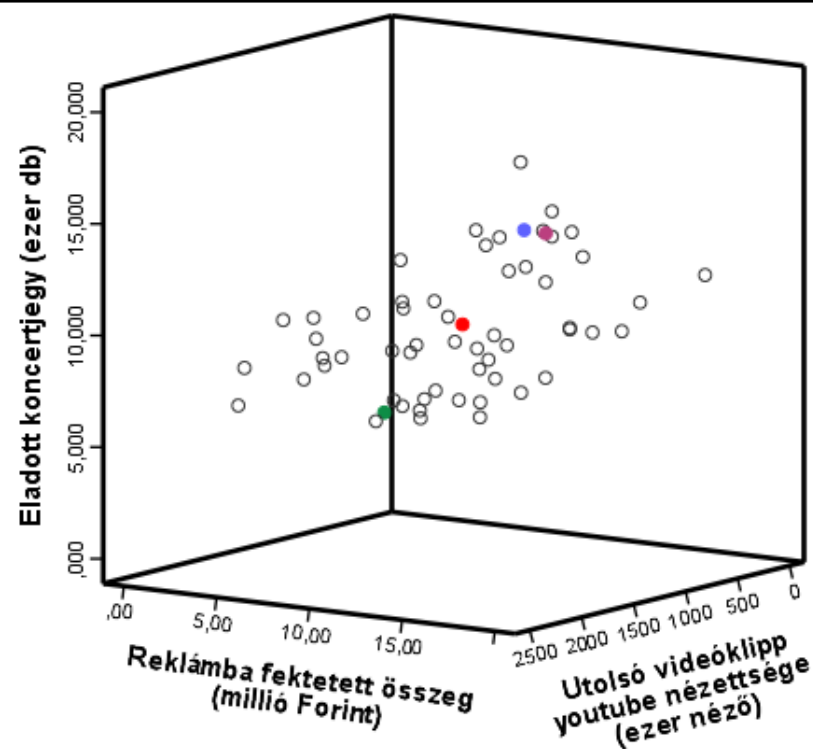
Lineáris regresszió

2. Többszörös lin. reg. elméleti háttére

Többszörös lineáris regresszió elmélet

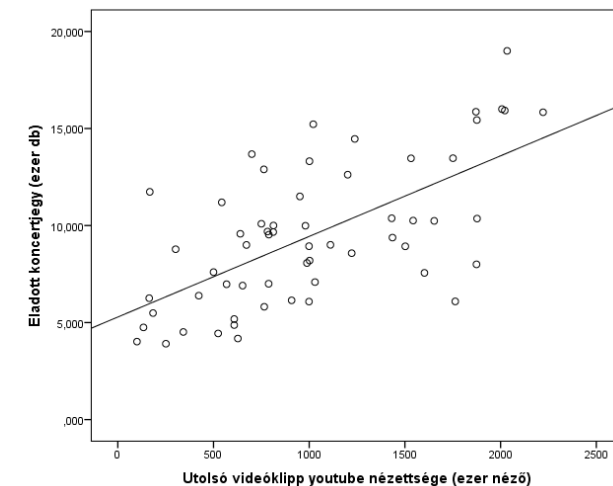
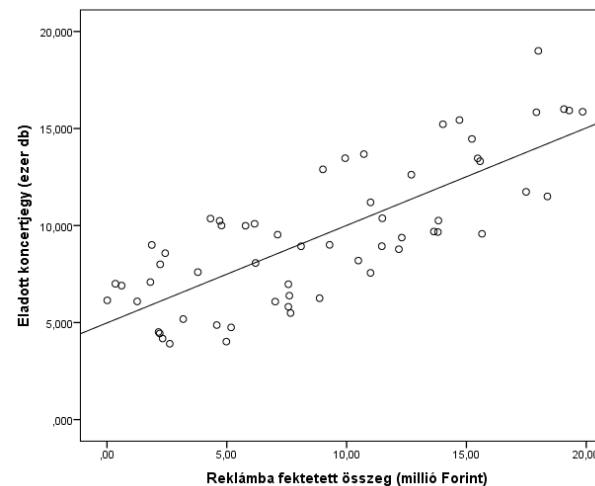
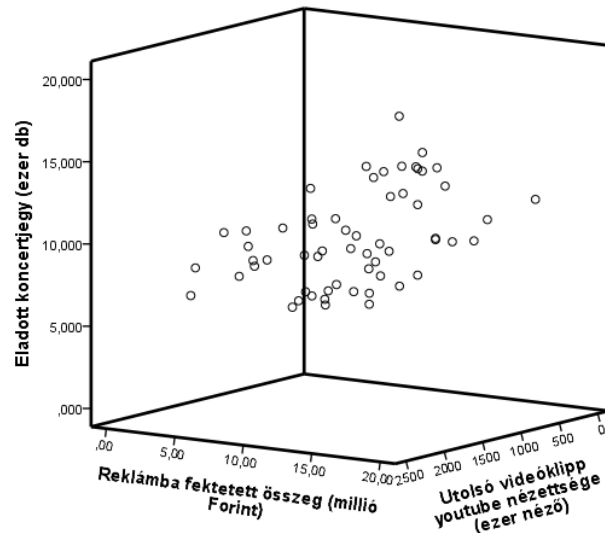


Azaz több prediktor változóval próbáljuk megbecsülni a kimeneti változó értékét



Többszörös regresszió

- K darab független prediktor változó esetén:
 - A pontfelhő **k+1 dimenziós térben** képzelhető el, ahol k a prediktorok száma, a plusz 1 pedig a kimeneti változó tengelye. A regressziós egyenes ebben a többdimenziós térben halad.
- Együtthatók:
 - A regressziós egyenes továbbra is egy pontban metszi az Y tengelyt (ott, ahol minden prediktor értéke 0), tehát **egyetlen b_0 konstans** tartalmaz az egyenlet.
 - A regressziós egyenes meredekségét viszont dimenzióként vizsgálhatjuk, azaz **minden prediktorhoz külön b_1 érték** fog tartozni.



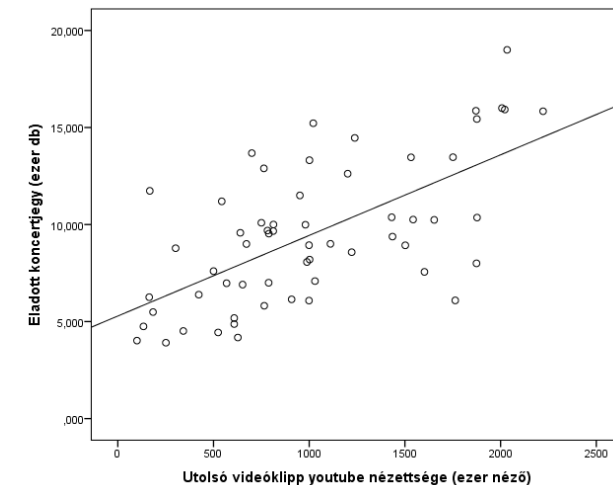
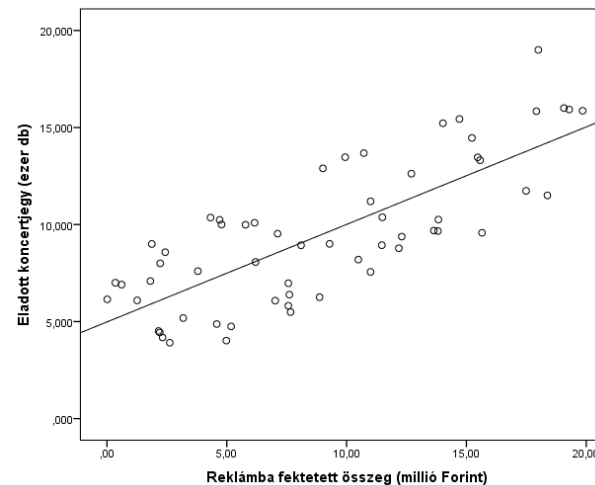
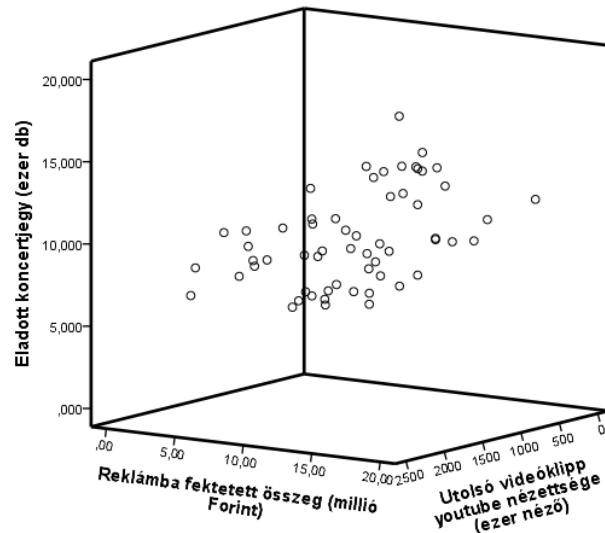
Többszörös regresszió

- A regressziós egyenes képlete:

- $Y_{\text{pred},i} = b_0 + b_1 * X_{1,i} + b_2 * X_{2,i} + \dots + b_k * X_{k,i}$, ahol i az egyedek, k pedig a prediktorok indexe
- Több prediktor változó lineáris kombinációja hozza létre a kimeneti változó predikált értékét
- Példa: Y_i zenekar várható jegyeladása = $b_0 + b_{\text{reklám}} * X_i$ zenekar reklámjába fektetett összeg + $b_{\text{youtube}} * X_i$ koncert nézettsége a youtube-on

- A kimeneti változó értéke:

- $Y_i = Y_{\text{pred},i} + \varepsilon_i = (b_0 + b_1 * X_{1,i} + b_2 * X_{2,i} + \dots + b_k * X_{k,i}) + \varepsilon_i$
- Y_i zenekar koncertjére eladott jegyek = Y_i zenekar várható jegyeladása + tévedés_i zenekar jegyeladásának becslésében



Többszörös regresszió

- A mutatók bonyolultabbak a többdimenziós tér miatt, de jelentésük ugyanaz

- R** - Korreláció a modell által predikált értékek és a valós értékek között

- Ha tökéletesen be tudnánk jósolni a kimeneti változót, akkor nem lenne hiba a jóslatban, a predikált és mért értékek korrelációja 1 lenne

- R²** - Effect-size mutató

- a modellünk mennyit magyaráz a függő változó varianciájából

- F** - A hatás és hiba aránya

- Ehhez rendeljük a szignifikancia értéket

- b₀** – Konstans

- ha minden prediktor értéke nulla lenne, ennyi lenne a kimeneti változó értéke (itt metszi a regressziós egyenes az y tengelyt)

- b₁**-ek és **t-próbák**

- Prediktor változónként meredekségek
- Prediktor változók hatása a kimeneti változóra

R = .662

Prediktor (X)	Kimeneti (Y)	Predikált (Y _{pred})	Hiba (ε)
4.38	10.22	9.78	0.44
6.27	10.09	9.59	0.50
2.65	9.23	9.95	-0.72
14.32	8.57	8.79	-0.22

R = 1

Prediktor (X)	Kimeneti (Y)	Predikált (Y _{pred})	Hiba (ε)
5.50	11.00	11.00	0
6.12	12.24	12.24	0
6.35	12.70	12.70	0
8.46	19.92	19.92	0

Modell felépítése & a modell szerkezete

- A modell kimenetét meghatározza a **prediktor változók**
 - 1. **kiválasztása és száma**
 - 2. egymáshoz való **viszonya**
 - 3. a **modellbe építésüknek sorrendje** (a modell felépítésének módja).

- **1. Prediktorok száma**

- Helytelen **rengetek véletlenszerűen kiválasztott változót** a modellbe dobni

- A **modell összetettsége** miatt

- A változók számának növelésével
 - nő a modell bonyolultsága
 - nő a becsült paraméterek (b1 értékek) száma
 - nő a szükséges elemszám

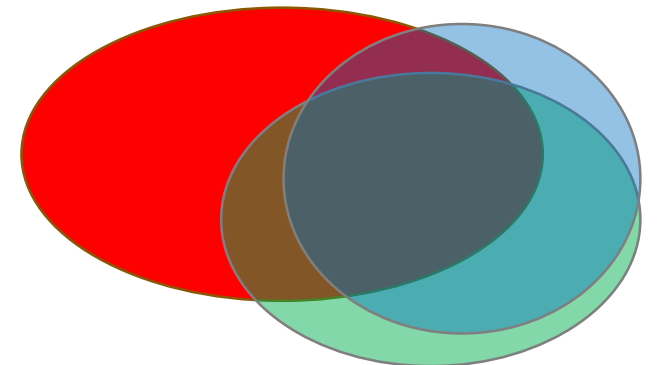
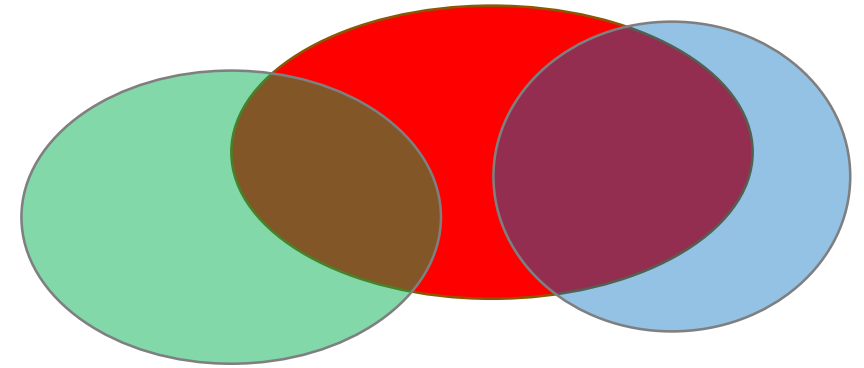
$$F = \frac{\frac{SS_M}{df_M}}{\frac{SS_R}{df_r}}$$

- **Specifititás és generalizálhatóság** problémája miatt

- A túlspecifikált modell lehet nagyon pontos a mintán, de nem feltétlenül az a populációban is (nem generalizálható), mert a modell „rátanul” a mintában lévő zajra is

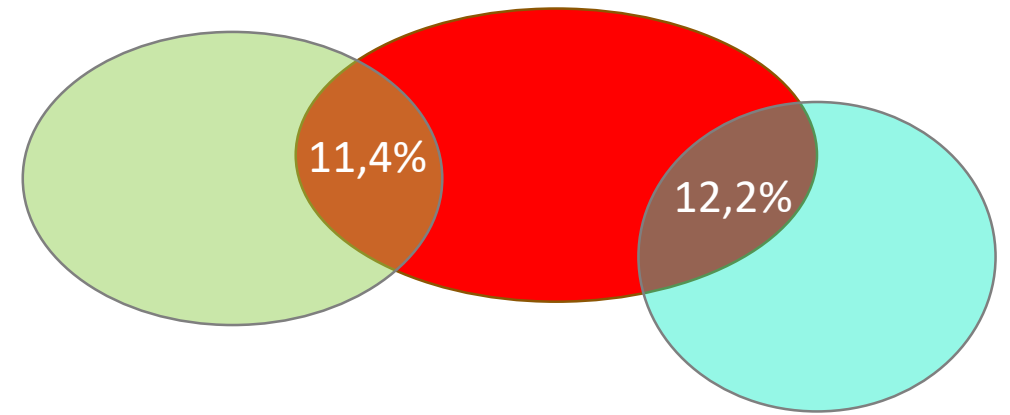
Prediktorok viszonya

- 2. A prediktorok egymáshoz való **viszonya**
 - $Y_{\text{pred},i} = b_0 + b_1 * X_{1,i} + b_2 * X_{2,i} + \dots + b_k * X_{k,i}$
 - Ha a **változókim korrelálatlanok**,
 - akkor minden prediktor más részt magyaráz a kimeneti változó variáciájából,
 - tehát predikált érték a **prediktorok hatásának lineáris kombinációjaként** fog összeállni,
 - azaz az egyéni hatások egyszerűen összeadódnak
 - **Korreláló prediktor változóknál**
 - a prediktorok **részben ugyanazt a részt magyarázzák** a kimeneti változóból,
 - ezért „**versengeni fognak**” a kimeneti változó variáciájáért



Prediktorok viszonya

- Példa **korrelálatlan prediktor változók** viselkedésére
 - Prediktor1** és **Prediktor2** : $r = .000338$ $p = .998$
 - Kimeneti** és **Prediktor1**: $r = -.337$ $p = .017$
 - Kimeneti** és **Prediktor2**: $r = .350$ $p = .013$
- Regressziós modellek:
 - A **magyarázó erők összeadódnak** $R^2_{pred1} + R^2_{pred2} = R^2_{közös}$
 - A **meredekségek ugyanazok maradnak**



Model Summary					
R Square					
.114					
ANOVA ^a					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4363,257	1	4363,257	6,152	.017 ^b
Residual	34044,123	48	709,253		
Total	38407,380	49			
Coefficients ^a					
Model	Unstandardized Coefficients	Standardized Coefficients			
	B	Beta	t	Sig.	
(Constant)	60,138		7,933	.000	
Prediktor 1	-,339	-,337	-2,480	.017	

Model Summary					
R Square					
.122					
ANOVA ^a					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4703,528	1	4703,528	6,699	.013 ^b
Residual	33703,85	48	702,164		
Total	38407,38	49			
Coefficients ^a					
Model	Unstandardized Coefficients	Standardized Coefficients			
	B	Beta	t	Sig.	
(Constant)	24,376		2,903	.006	
Prediktor 2	,349	,350	2,588	.013	

Model Summary					
R Square					
.236					
ANOVA ^a					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	9069,852	2	4534,926	7,265	.002 ^b
Residual	29337,528	47	624,203		
Total	38407,380	49			
Coefficients ^a					
Model	Unstandardized Coefficients	Standardized Coefficients			
	B	Beta	t	Sig.	
(Constant)	40,694		4,055	.000	
Prediktor 1	-,339	-,337	-2,645	.011	
Prediktor 2	,349	,350	2,746	.009	

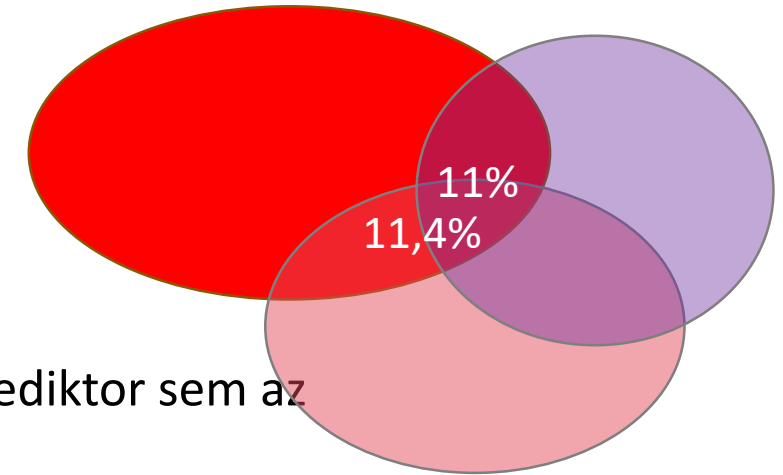
Prediktorok viszonya

- Példa **korreláló prediktor változók** viselkedésére

- Prediktor3** és **Prediktor4** : $r = .258$ $p = .070$
- Kimeneti** és **Prediktor3**: $r = .331$ $p = .019$
- Kimeneti** és **Prediktor4**: $r = .338$ $p = .016$

- Regressziós modellek:

- A **magyarázó erők nem adódnak össze**: $R^2_{pred1} + R^2_{pred2} \neq R^2_{közös}$
- A **meredekségek mérséklődnek**, sőt bár a modell szignifikáns, egyik prediktor sem az



Model Summary

Regresszió Pred3-mal

R Square

,110

ANOVA^a

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4208,463	1	4208,463	5,907	,019 ^b
Residual	34198,917	48	712,477		
Total	38407,380	49			

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
	B	Beta		
(Constant)	24,546		2,795	,007
Prediktor 3	,333	,331	2,430	,019

Model Summary

Regresszió Pred4-gyel

R Square

,114

ANOVA^a

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4387,942	1	4387,942	6,191	,016 ^b
Residual	34019,438	48	708,738		
Total	38407,380	49			

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
	B	Beta		
(Constant)	24,652		2,875	,006
Prediktor 4	,386	,338	2,488	,016

Model Summary

Regresszió
Pred3 és Pred4-gyel

R Square

,178

ANOVA^a

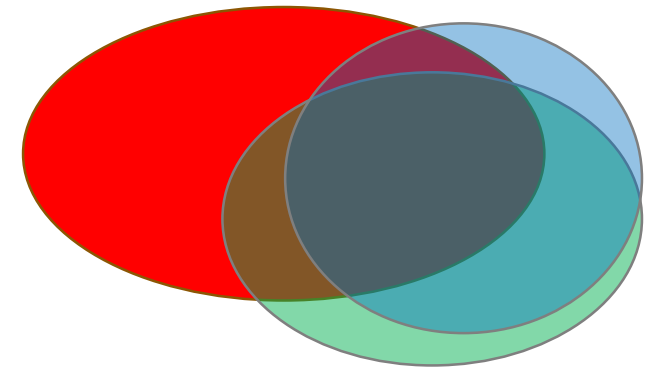
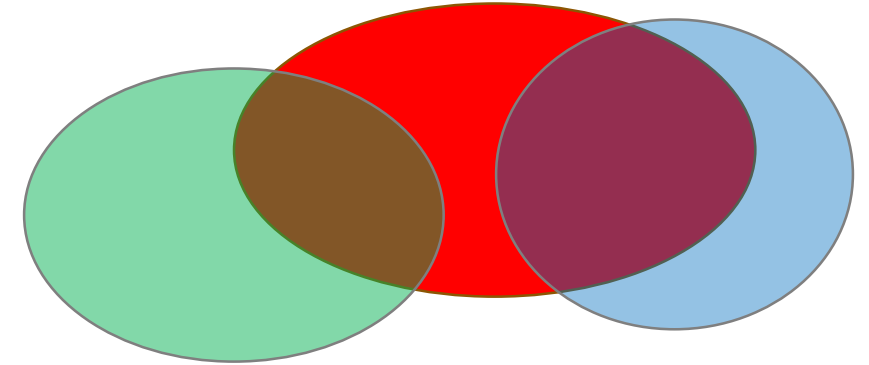
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	6832,121	2	3416,06	5,085	,010 ^b
Residual	31575,259	47	671,814		
Total	38407,380	49			

Coefficients^a

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
	B	Beta		
(Constant)	13,273		1,294	,202
Prediktor 3	,263	,261	1,907	,063
Prediktor 4	,309	,271	1,976	,054

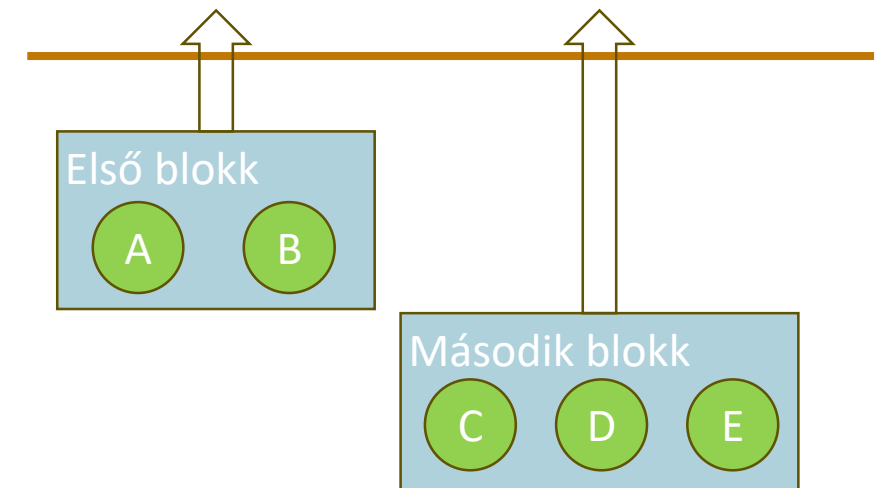
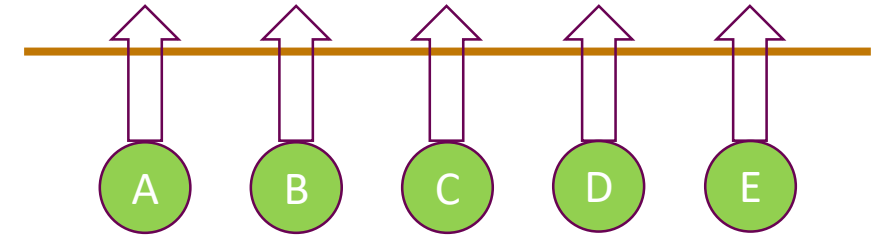
Prediktorok sorrendje

- 3. A prediktorok modellbe való építésének **sorrendje**
 - Ha a **változókim korrelálatlanok**,
 - akkor **nem számít** a modellbe építés **sorrendje**,
 - mert minden prediktor más részt magyaráz a kimeneti változó varianciájából,
 - tehát a hatások összeadódnak,
 - az összeadás elemei pedig felcserélhetők
 - **Korreláló prediktor változóknál**
 - **fontos** a modell kialakításának módja
 - mert a prediktorok **részben ugyanazt a részt magyarázzák** a kimeneti változóból,
 - ezért „**versengeni fognak**” a kimeneti változó varianciájáért



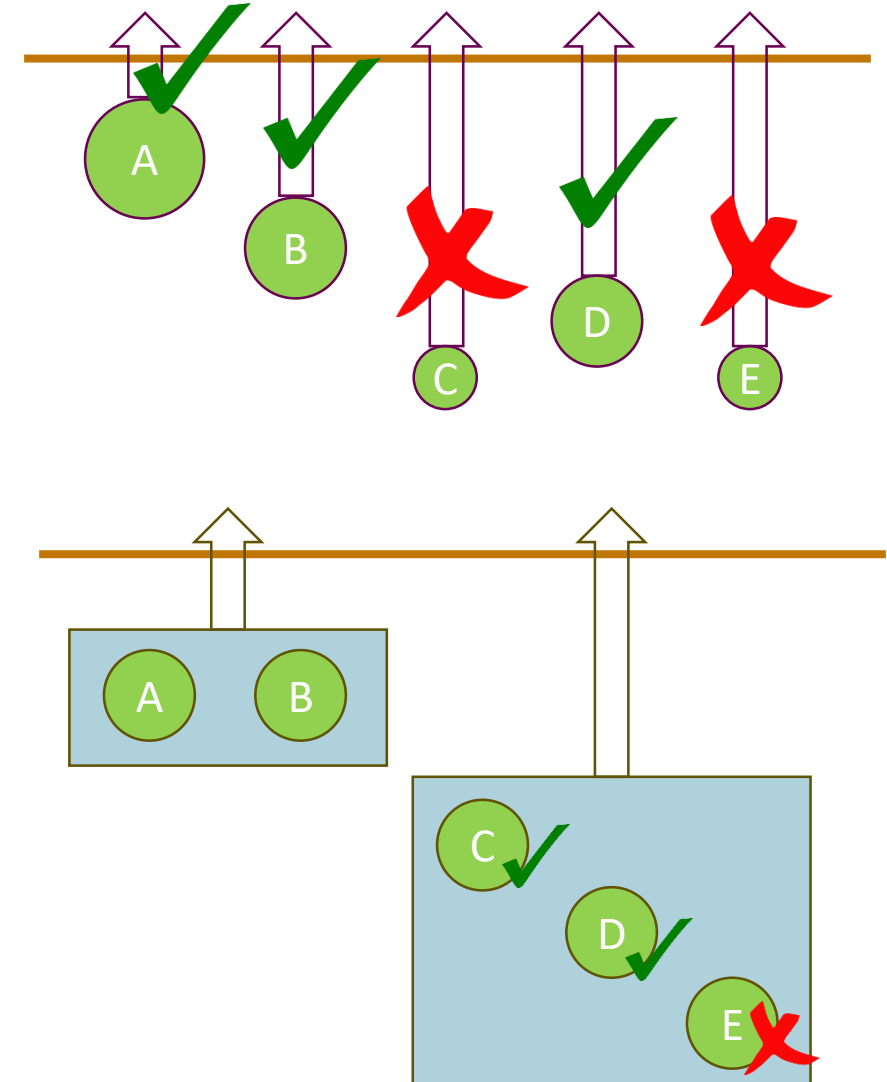
Prediktorok sorrendje

- **A modell felépítésének módja(i):**
 - **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



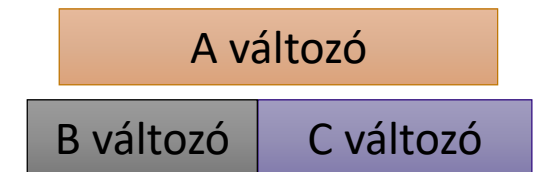
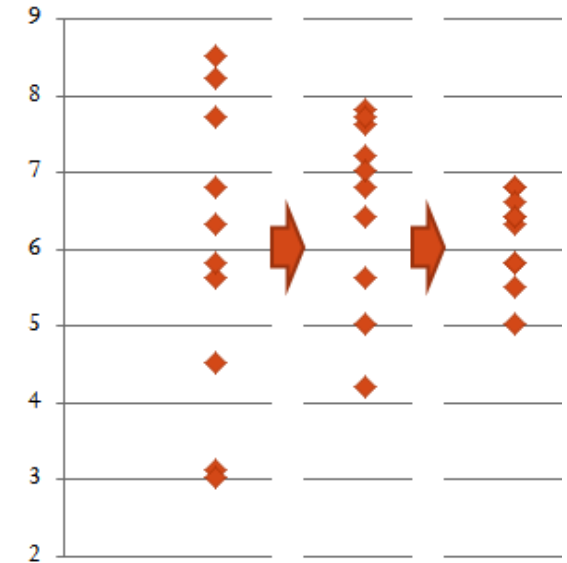
Prediktorok sorrendje

- **A modell felépítésének módja(i):**
 - **Stepwise módszerek**
 - A modellbe való kerülést matematikai kritériumok szabják meg
 - Ha egy prediktor nem tud szignifikáns 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



Modell felépítésének módja(i)

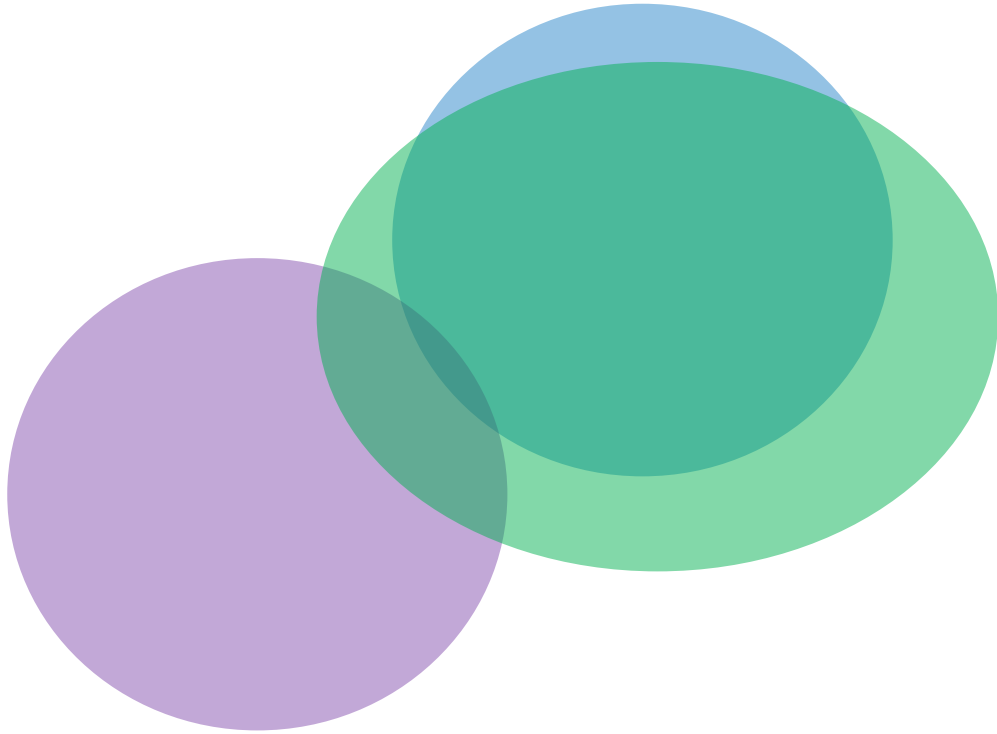
- **Stepwise módszerek**
 - **Forward**
 - Először csak b_0 kerül a modellbe
 - Majd bekerül az a változó, ami a legtöbbet tudja magyarázni a varianciából (legnagyobb a korrelációja a függő változóval), ha ettől szignifikánsan jobb lett a modell, benn marad
 - Majd bekerül a következő változó, mely a maradék varianciát legjobban magyarázza (legnagyobb a parciális korrelációja a függő változóval), ha ettől szignifikánsan jobb lesz a modell, ez is bennmarad
 - Addig folytatódik míg minden változó a modellbe kerül, vagy az új változóval már nem lesz szignifikánsan jobb a modell
 - **Backward**
 - Minden változó bekerül a modellbe
 - Majd amelyik legkevésbé teszi jobbá a modellt, az kikerül
 - **Stepwise**
 - Olyan, mint a Forward modell, de minden új változó bekerülésekor végignézi a már bennlévőket, és ellenőrzi a leggyengébbet, az új változó függvényében is szükség van-e rá -> csökkentve a redundáns információt a modellben



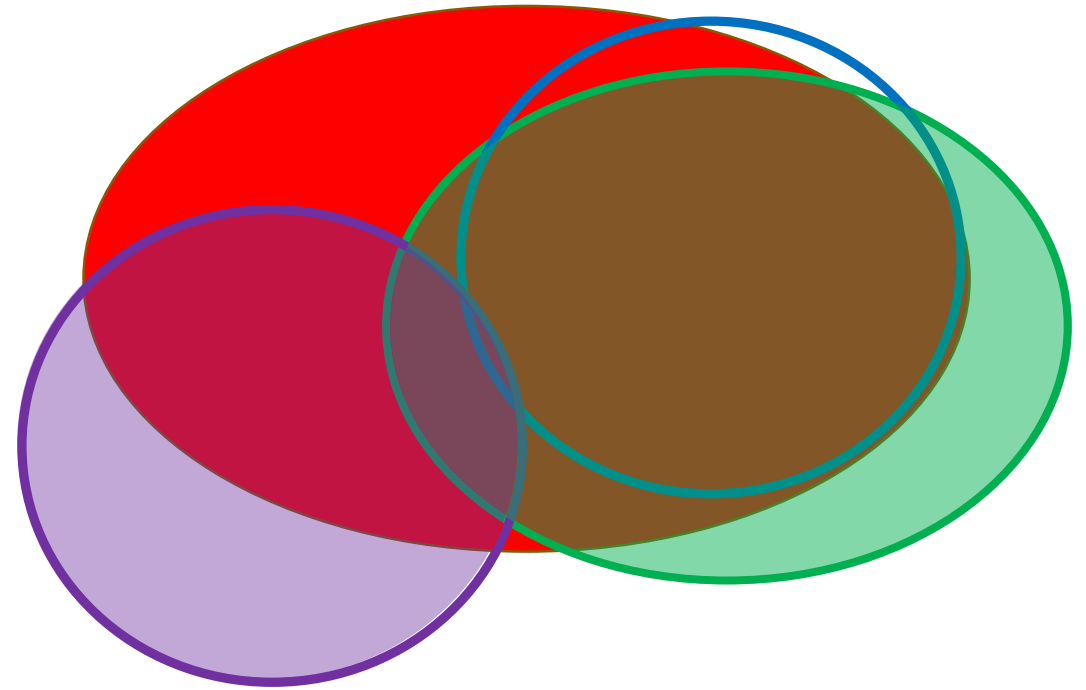
Modell felépítésének módja(i)

- Példa Forward típusú stepwise modellre

PREDIKTOROK



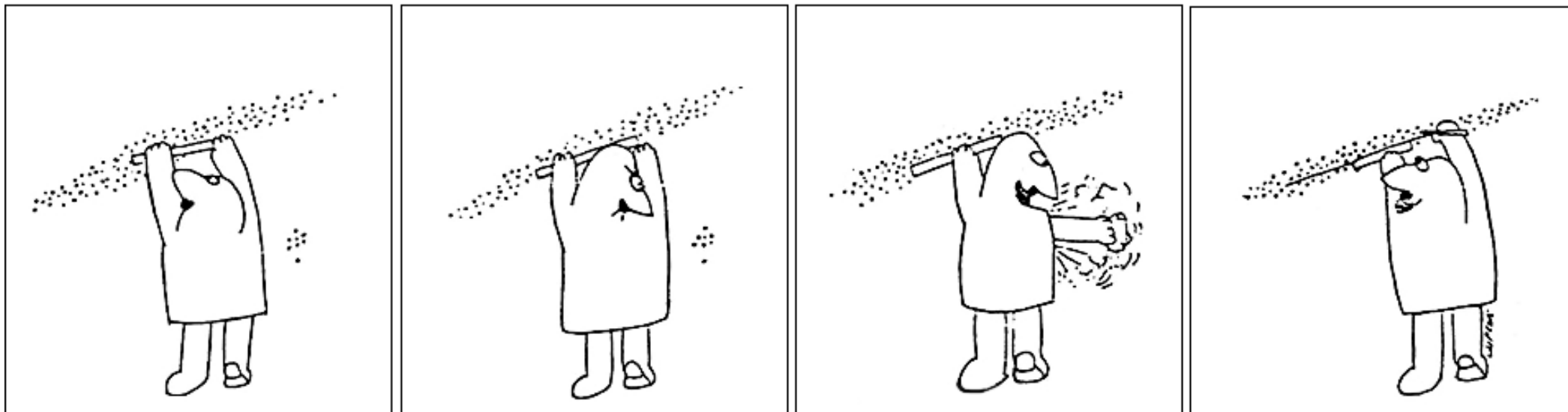
MODELL



Blockwise, enter vagy stepwise?

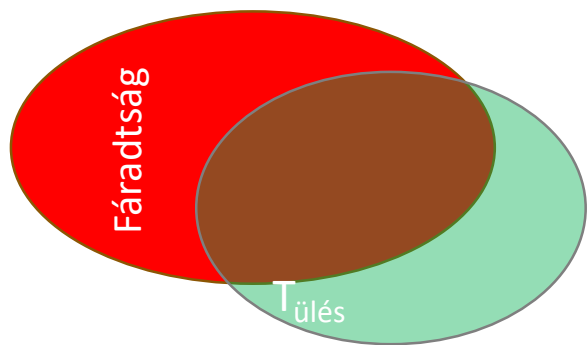
- **Ha van tudományos elképzelésed, használj blockwise (hierarchikus) módszert**
 - A blockwise módszer lehetővé teszi, hogy az előzetes tudásod is számításba vedd, a számodra fontosabb változókat kiemeld
- **Ha feltáró vizsgálatot végzel, használj enter módszert**
 - A változó modellbe való kerülését nem befolyásolják az előzetes elképzeléseid, tudások
 - Pontosabb képet kapsz a prediktor változók és a kimeneti változó kapcsolatáról, plusz a prediktor változók egymással való kapcsolatáról is
- **Ha „csak” a leggazdaságosabb modellt keresed, használj stepwise módszert**
 - Megkeresi azt a modellt, ahol a lehető legkevesebb változóval a lehető legpontosabb jóslatot lehet tenni
 - Például, ha célod, hogy lehetőleg kevés prediktor változóval kapj pontos becslést (mert a prediktor változók felvétele pénzbe kerül)
 - Hátránya, hogy a változókról önmagunkban kevés információt ad
 - Fennáll az overfitting veszélye
 - **Sokat támadják**, mert tisztán matematikai és véletlenül alapuló döntések határozzák meg a módszert – minden fontos döntést „kivesz” a kutató kezéből. Van folyóirat, mely nem fogad el cikket stepwise módszerrel :S

Feltételek

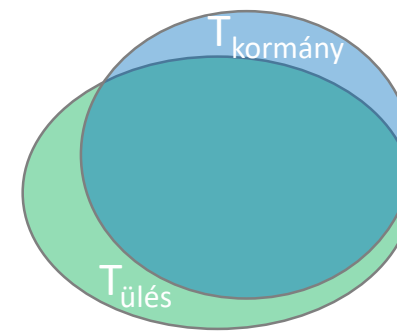
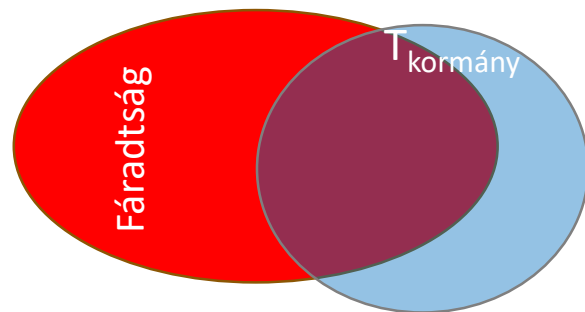


- ① **Változók típusa**
 - Független / prediktor változó: skála típusú vagy dichotóm
 - Függő / kimeneti változó: skála típusú, folytonos
 - Előzetesen ellenőrizendő
- ② **Nincs nulla varianciájú változó**
 - Van a változóknak varianciája
 - A deskriptív statisztikák alapján ellenőrizendő
- ③ **Nincs (multi)kollinearitás**
 - Kollinearitás: a független változók közötti tökéletes /nagyon erős együttjárás
 - Multikollinearitás: sok független változó közötti közepes-erős korreláció
 - Csökkenti egy-egy változó értelmezésének lehetőségét
 - A kollinearitás ellenőrizhető az elemzés során kikért deskriptív statisztikák korrelációs tábláival.
 - A multikollinearitás a kollinearitás analízis tolerancia értékei segítségével
 - Megoldás: ki kell venni a változók közül vagy főkomponens elemzéssel összevonni őket
- ④ **Nincs külső változó**
 - Nincs egy külső változó, amellyel a független változók korrelálnak (sérülése általában kollinearitást vagy multikollinearitást okoz)
 - A kutatás tervezésekor kell átgondolni

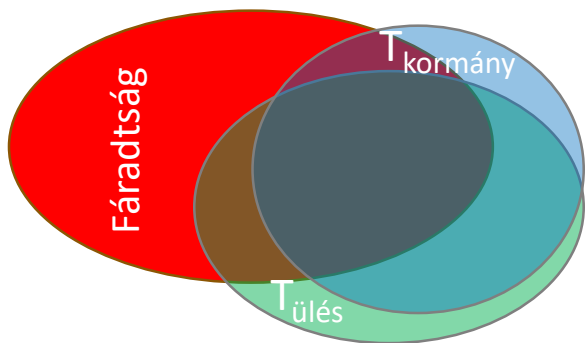
- Kollinearitás



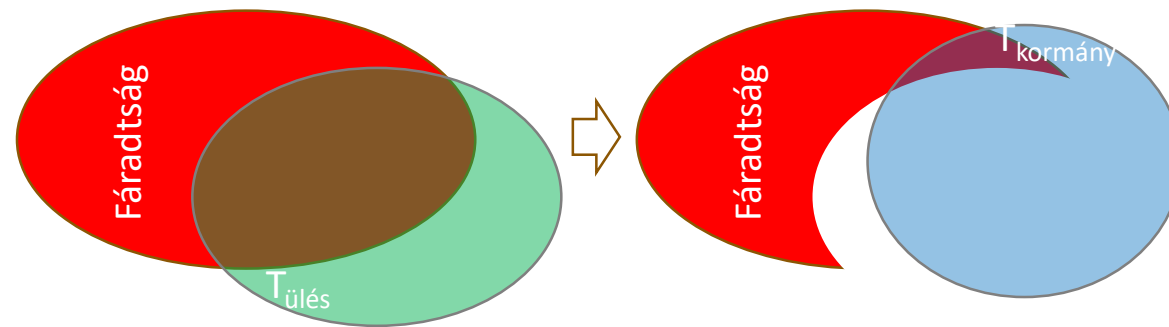
A fáradtságot mind a kormány fogásával töltött idő, mind a vezetőülésben töltött idő viszonylag jól magyarázza



A baj az, hogy a kormányfogással és ülésben töltött idő között erős a korreláció



Ezért a fáradtság varianciájából majdnem teljesen ugyanazt a rész magyarázzák. Ezért a regresszióban tévesen alulbecsülnénk a magyarázóerejüket.



Sőt Stepwise modellnél az is előfordulhat, hogy ha az ülésben töltött idő lép először a modellbe, kormány fogásával töltött időnek már „nem marad további variancia”, amit magyarázhatna, és be sem kerül a modellbe

- **Multikollinearitás**

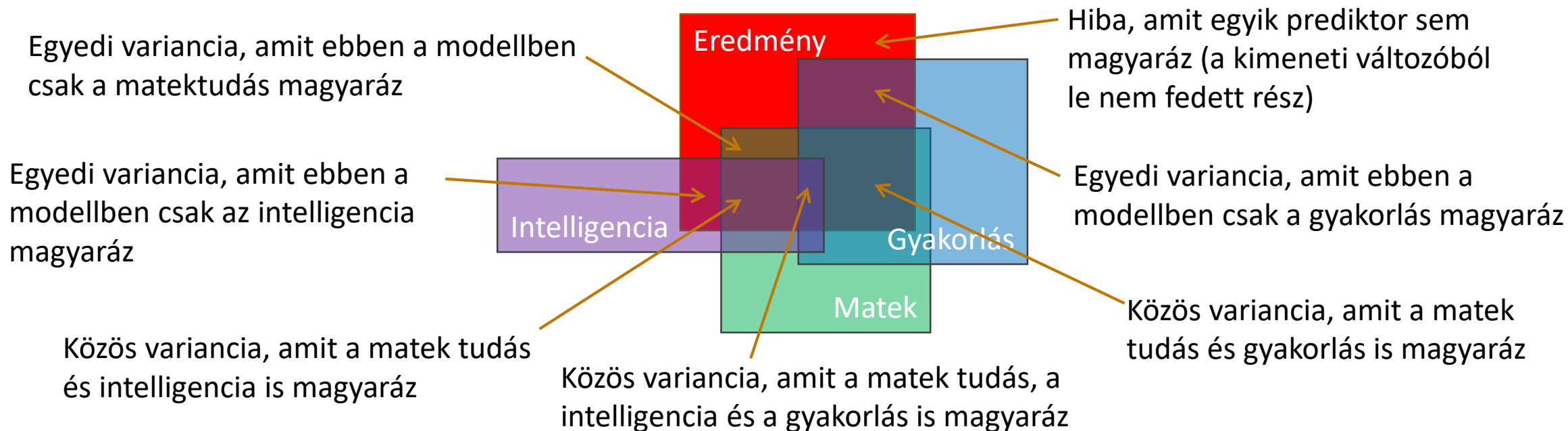
A probléma ugyanaz, csak nehezebb észrevenni, mert nem két változó korrelál egymással nagyon erősen, hanem sok közepesen erősen.

Megmagyarázott variancia: A kimeneti változó (eredmény) varianciájának azon rész, amit a prediktor változók magyaráznak (lefednek).

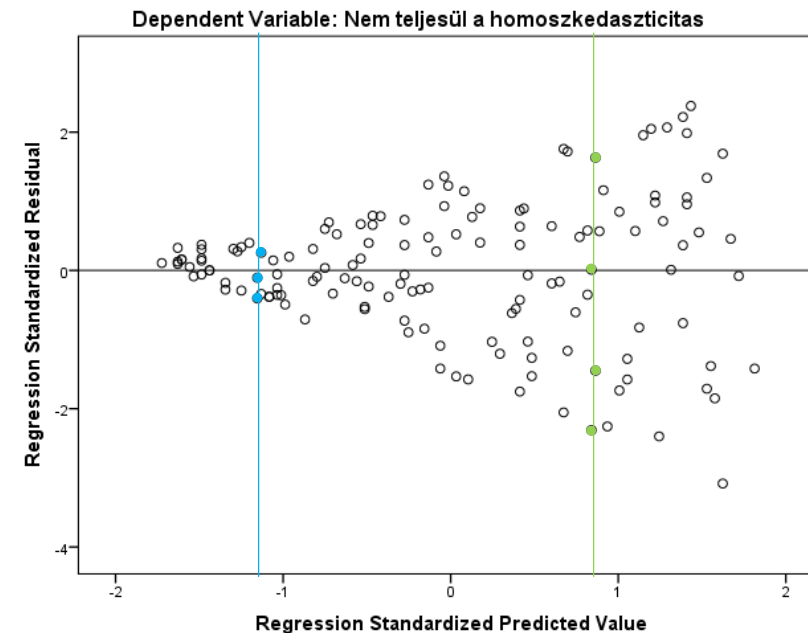
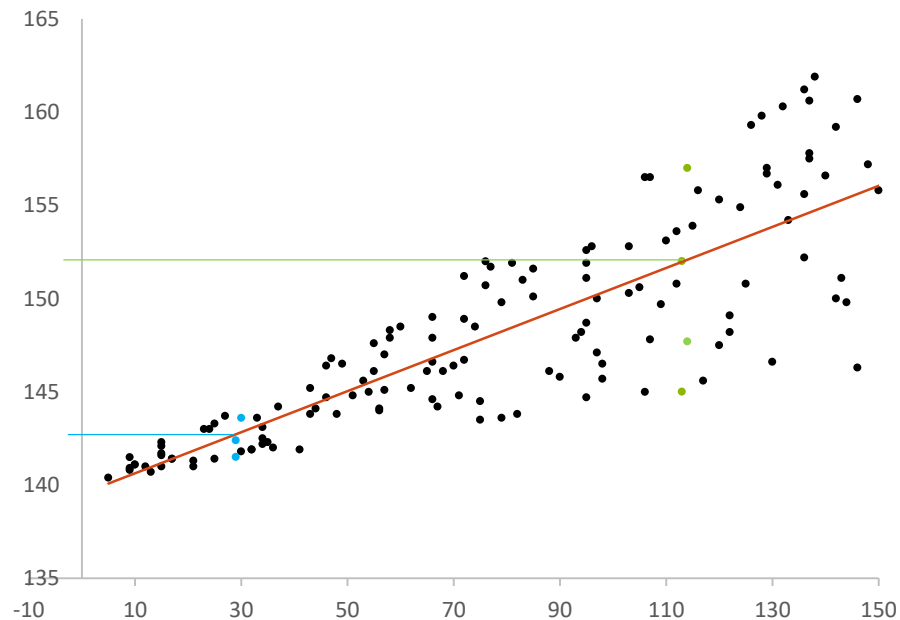
Egyedi variancia: a megmagyarázott variancia azon része, amit csak egyetlen prediktor fed le

Közös variancia: a megmagyarázott variancia azon része, amit kettő vagy több prediktor is magyaráz.

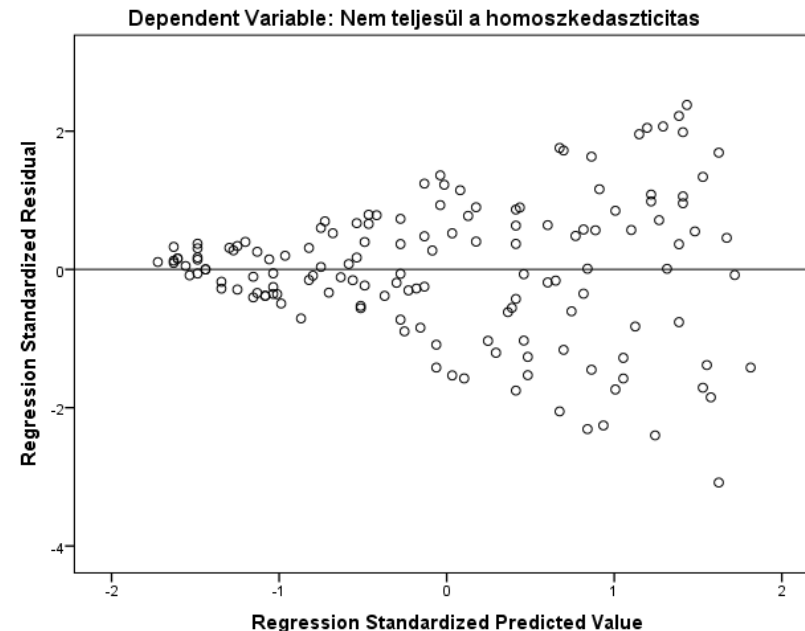
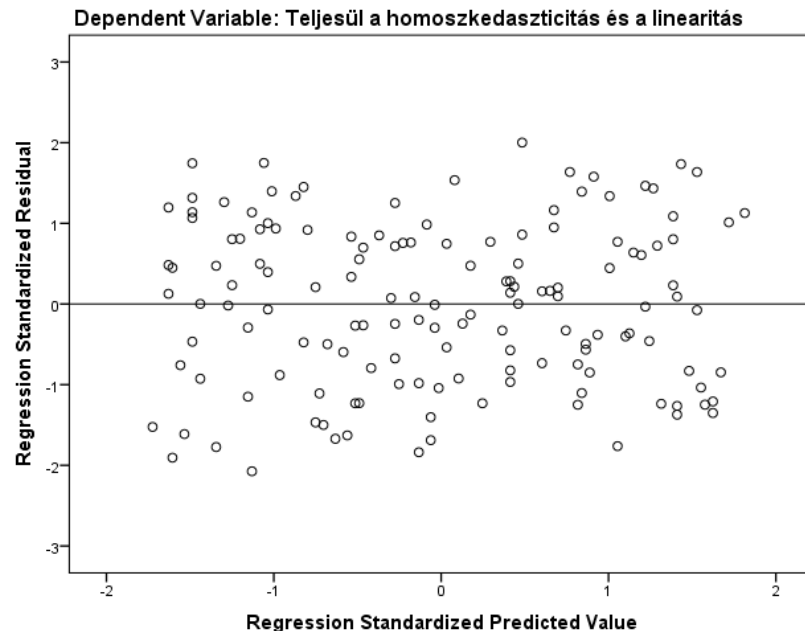
Egy **prediktor toleranciája** a hozzá tartozó egyedi és közös variancia hányadosa



- ⑤ **Homoszkedaszticitás / Varianciák homogenitása / Szóráshomogenitás**
 - Definíció eddig egyetlen prediktor változó (2D) esetén: A prediktor változó minden szintjén hasonló nagyságú a kimeneti változó szórása.
 - Fogalmazzuk át úgy, hogy 3D felett is szemléletes maradjon!
 - A prediktor változó helyett (amiből most több van) a predikált érték mentén haladunk, és azt figyeljük meg, hogy a mért értékek milyen távol vannak a predikáltaktól, azaz mekkora a reziduális hiba mértéke
 - **Definíció most: A predikált értékek minden pontjában hasonló a reziduális hibák nagysága / a hibák nagysága homogén.**

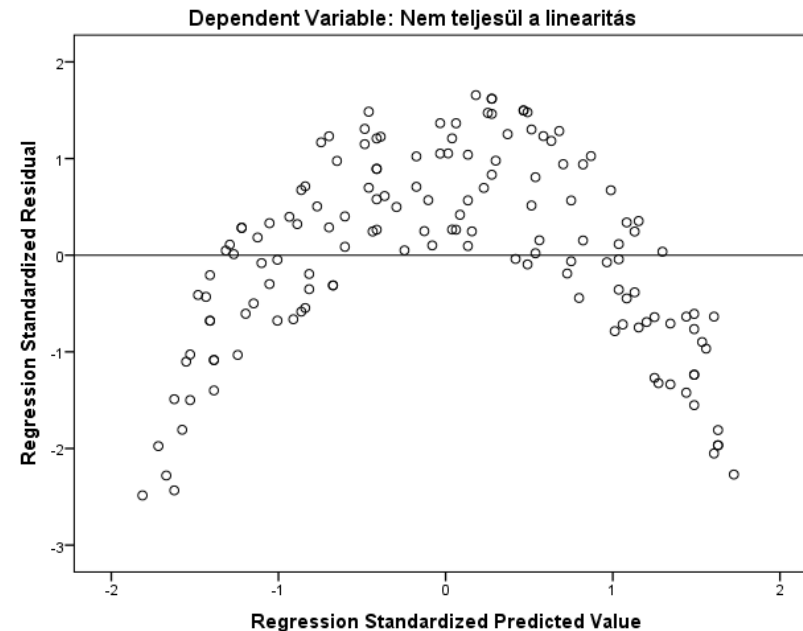
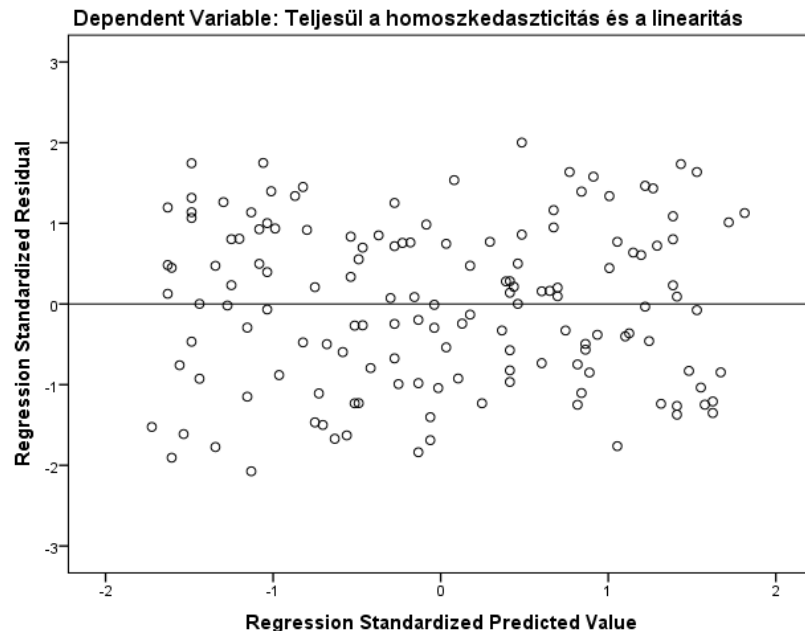


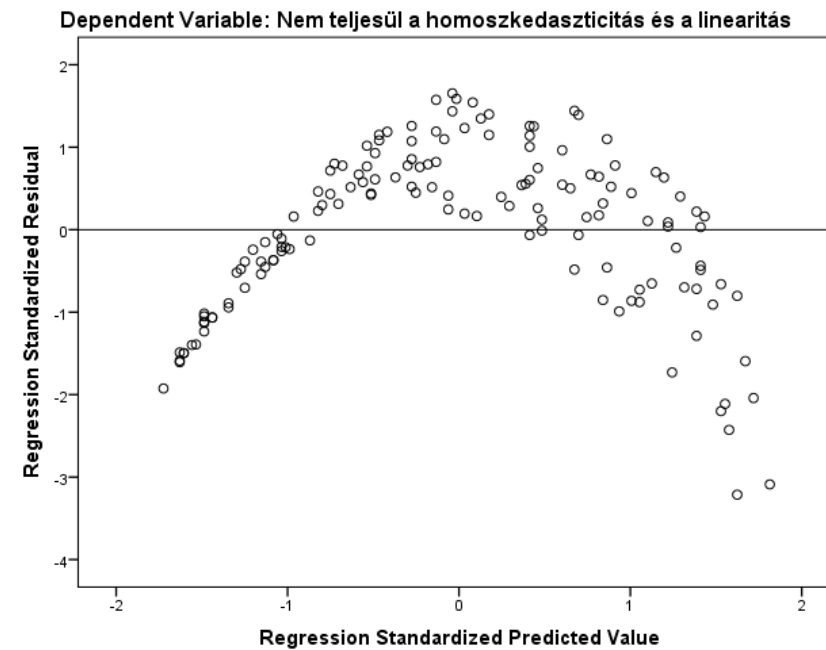
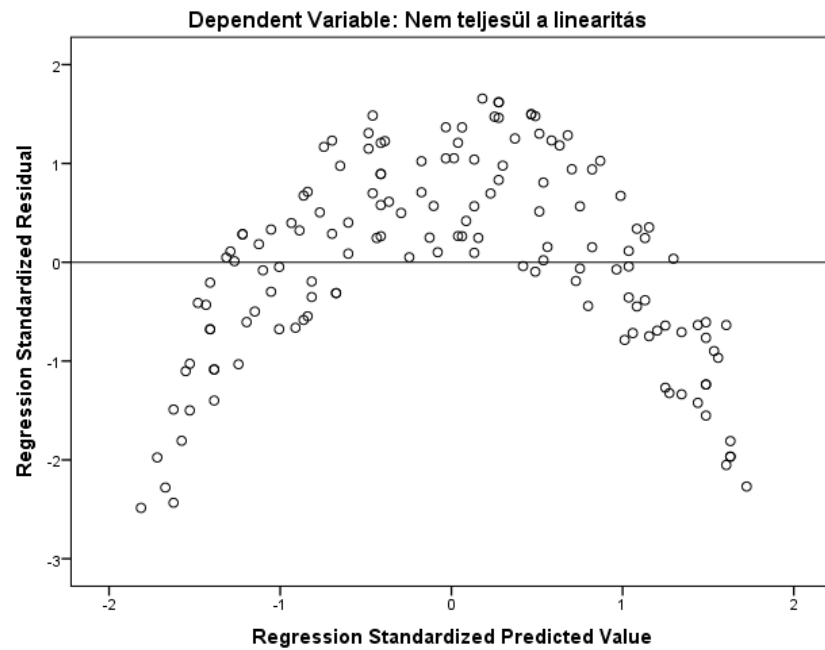
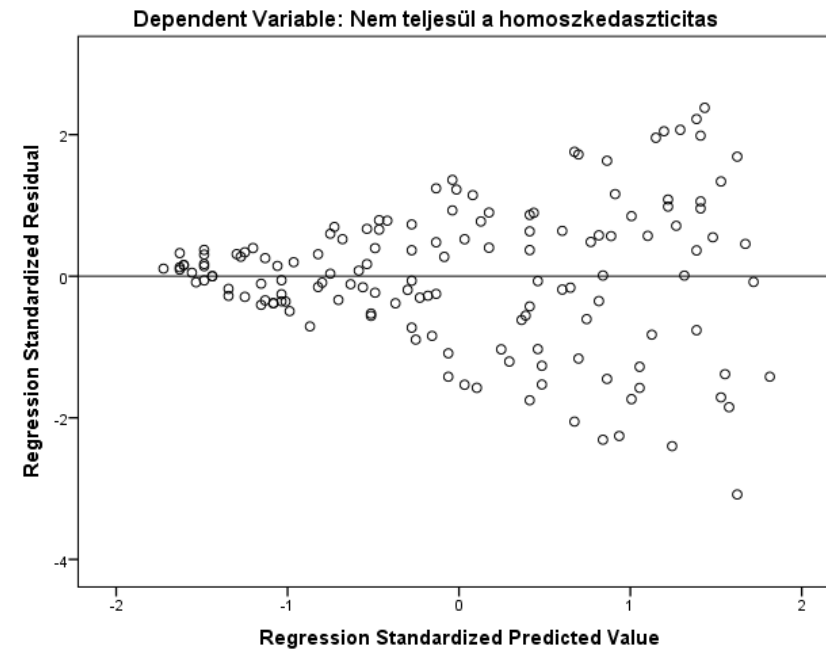
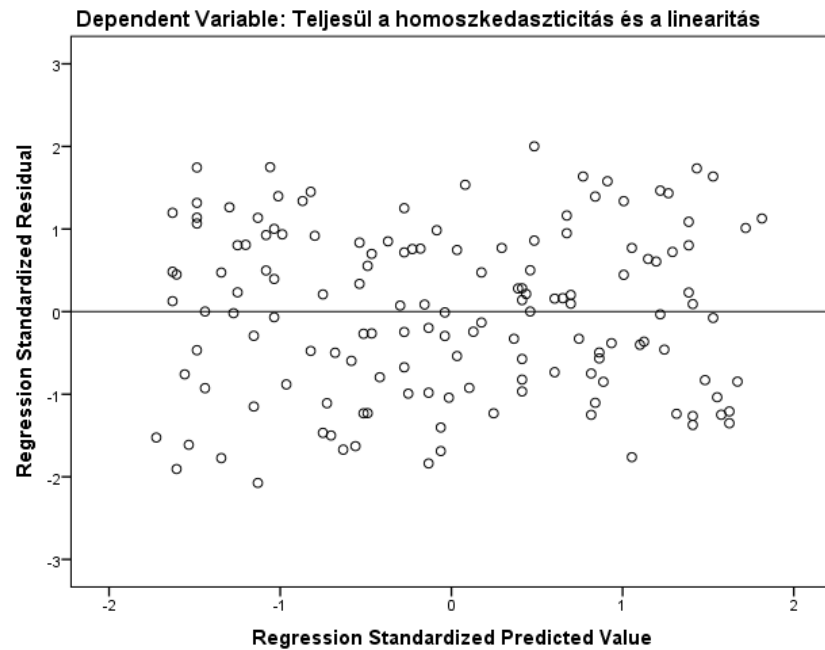
- ⑤ Homoszkedaszticitás / Varianciák homogenitása / Szóráshomogenitás
 - A predikált érték minden szintjén hasonló a reziduális hibák nagysága / a hibák nagysága homogén.
 - Ha nem teljesül a homoszkedaszticitás (ergo heteroszkedaszticitás van), a regresszióval tett predikció nem mindenhol ugyanolyan mértékben bízható meg (ahol nagyon a hibák, ott pontatlanabb)
 - Ellenőrizhető az elemzés során a **predikált** és **reziduális** értékek plottolásával
 - Ha az előző hibát mutat, az megkereshető az elemzés során a **prediktor változók** és a **reziduális értékek** plottolásával prediktoronként egyenként



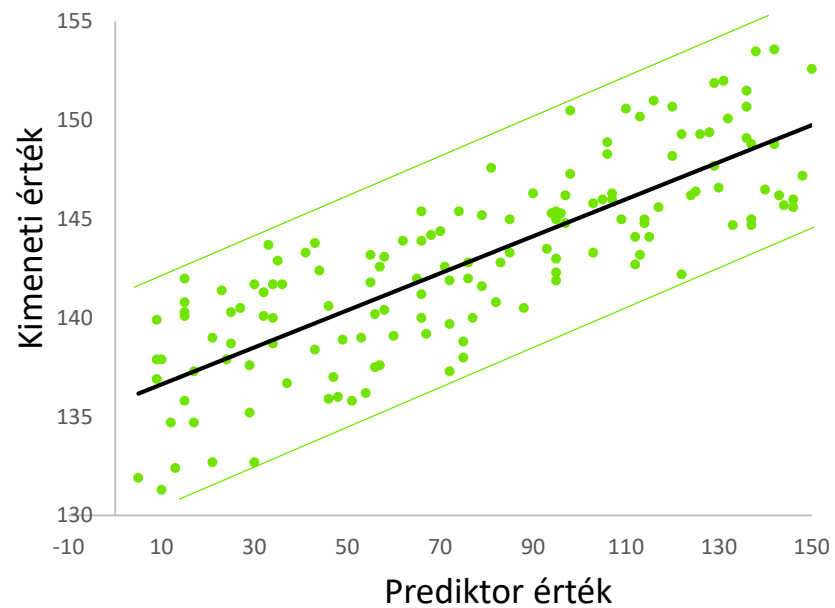
- ⑥Linearitás

- Lineáris kapcsolat van minden prediktor változó és a kimeneti változó között (a többi prediktor változóra kontrollálva)
- Ellenőrizhető az elemzés során a **predikált értékek és reziduális értékek**, vagy a **predikált értékek és kimeneti változó** közötti kapcsolat plottolásával (elég az egyiket kikérni)
- Ha a linearitás sérülésére gyanakszunk, ellenőrizhető előzetesen a **prediktor változó és kimeneti változó** közötti kapcsolat plottolásával prediktoronként egyenként

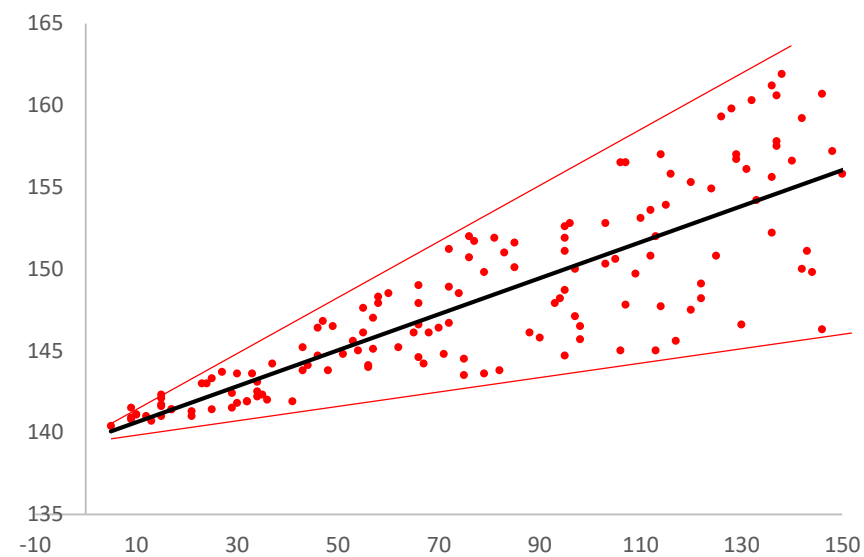




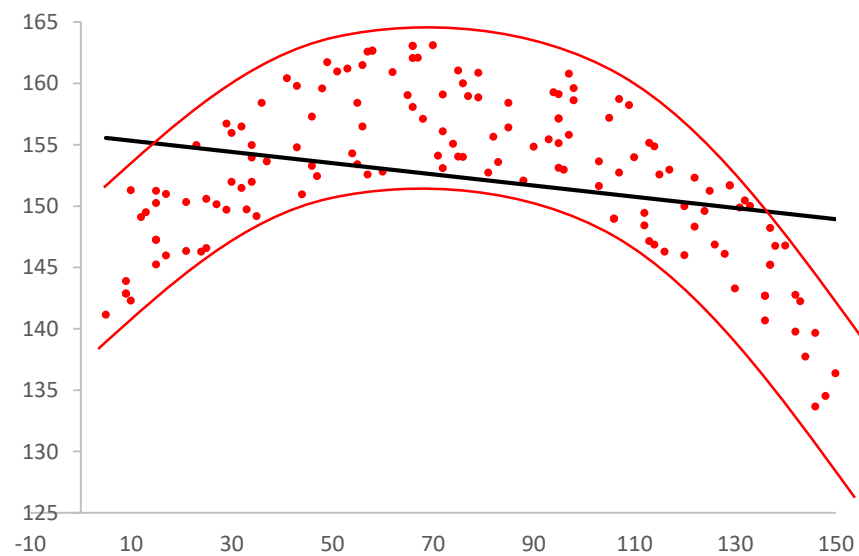
Teljesül a homoszkedaszticitás és linearitás



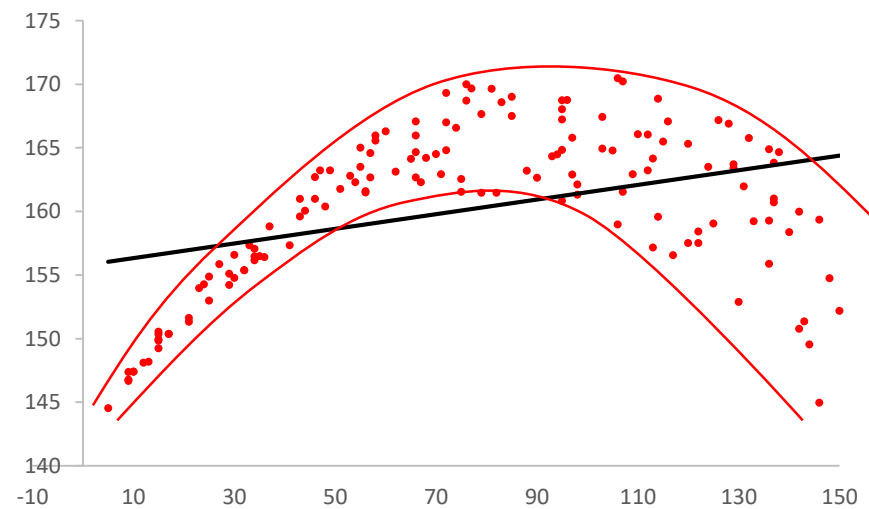
Nem teljesül a homoszkedaszticitás



Nem teljesül a linearitás



Nem teljesül a homoszkedaszticitás és linearitás sem



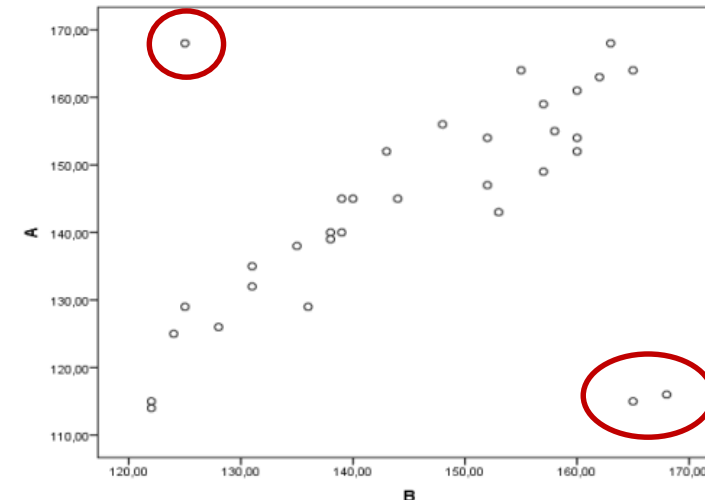
- **⑦Független hibák**

- Feltétel, hogy a reziduális hibák ne korreláljanak
- Idősorok elemzése során könnyen előforduló hiba (pszichológiában ritka)
- Rosszul specifikált modellt jelezhet (rosszul kiválasztott prediktor változók)
- Linearitás sérülésének jele is lehet
- Ellenőrizhető az elemzés során a Durbin-Watson teszttel
 - 2 körül tökéletes, 1 alatt vagy 3 felett jelez bajt

- **⑧Hiba normális eloszlása**

- Mert a hibák a véletlen műve
- A konfidencia intervallumok és a regressziós együtthatókhoz tartozó tesztek kiszámításához szükséges.
- A prediktorok vagy kimeneti változó normál eloszlása NEM feltétele a lineáris regressziónak, de ha ezek nem követnek normál eloszlást, gyakran a hibák sem követnek.
- A linearitás sérülése is okozhatja a hibák nem normális eloszlását
- Ellenőrizhető az elemzés során a hiba hisztogramjának plottolásával vagy a hibaértékek mentésével, és az így létrejövő új változó normalitásának tesztelésével
- Ha nem teljesül, a linearitást és a változók normalitását érdemes ellenőrizni

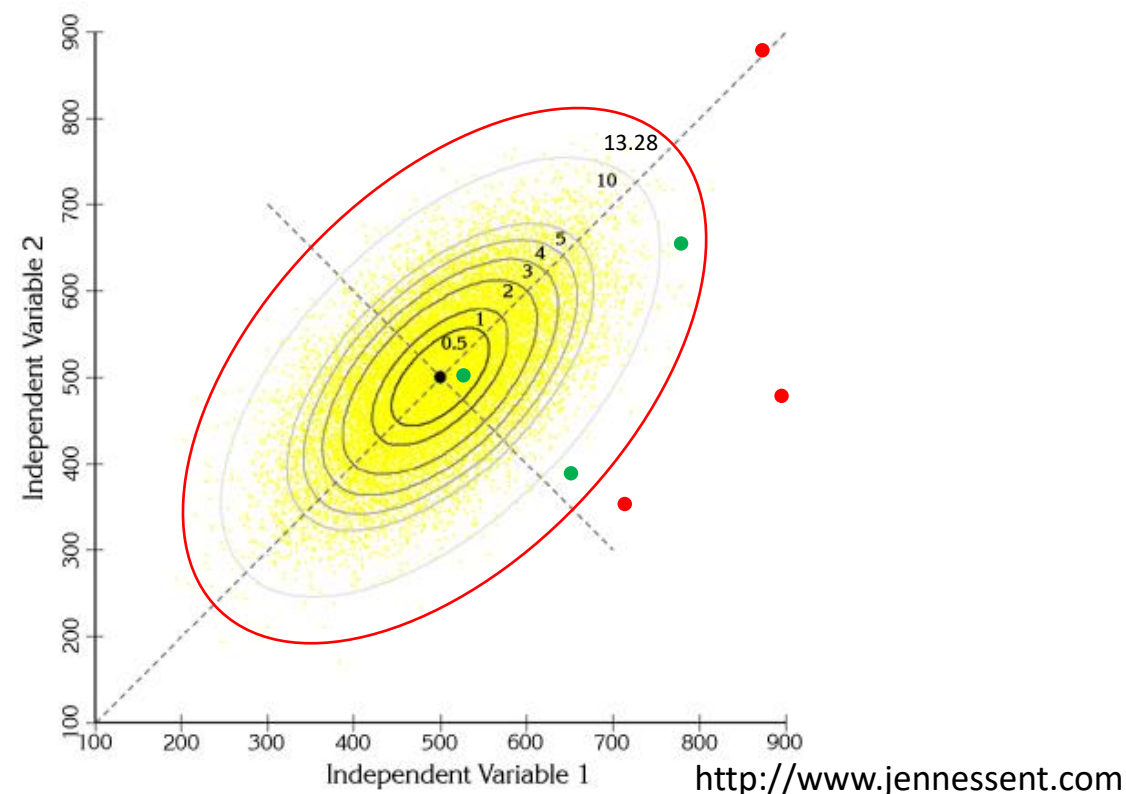
- ⑨ **Elemszám**
 - Függ: (1) változók számától, (2) milyen következtetést akarunk levonni, (3) mekkora a hatás
 - Ökölszabály: változószor 10 vagy 15
 - Ha nagy a hatás várunk: 80 mindig elég (20 változóig)
 - Közepes hatást várunk: 200 mindig elég (20 változóig), de kevés változóval is minimum 60
 - Kicsi hatást várunk: 600 kell (6 változóig)
- ⑩ **Nincsenek (több)dimenziós outlierok**
 - Önmagában nem feltétel, de a többi feltétel teljesülését sérti
 - Többdimenziós outlier külön-külön egyik dimenzión belül sem szélsőséges érték, csak az adott kombináció szélsőséges
 - Például 50kg-os és 190cm magas ember
 - Ellenőrizhető az elemzés során
 - Mahalanobis távolság
 - Cook távolság
 - Leverage values
 - Standardizált df beta értékek



- **Mahalanobis távolság**

- Elmélet az adatfeldolgozás diasor végén
- Outlier egy adat, ha a hozzá tartozó Mahalanobis távolság meghaladja a kijelölt határértéket
- A határ pontos értéke függ az elemszámtól és a változók mennyiségétől, valamint attól, mennyire szigorú alfa kritérium szint mellett dolgozunk.

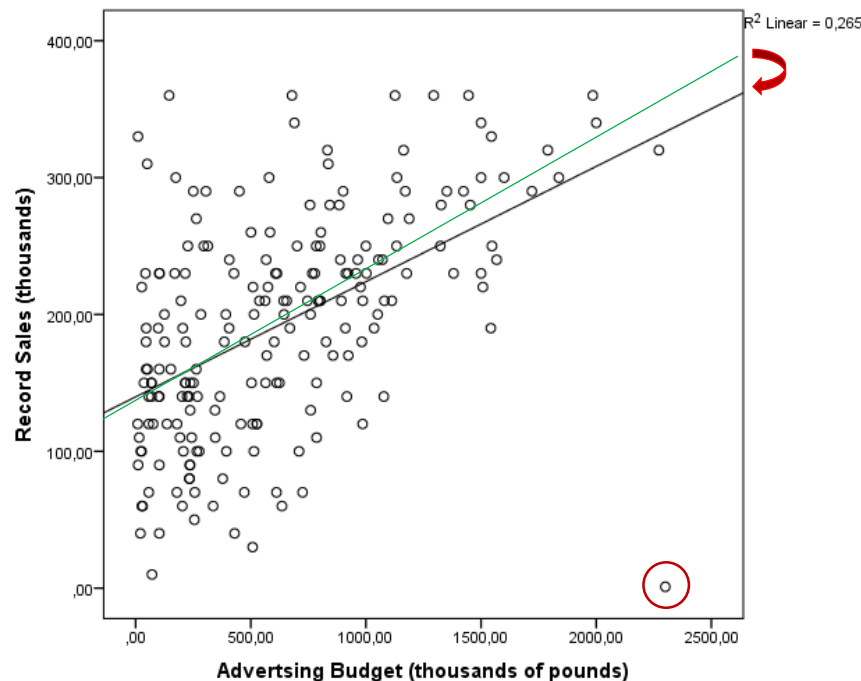
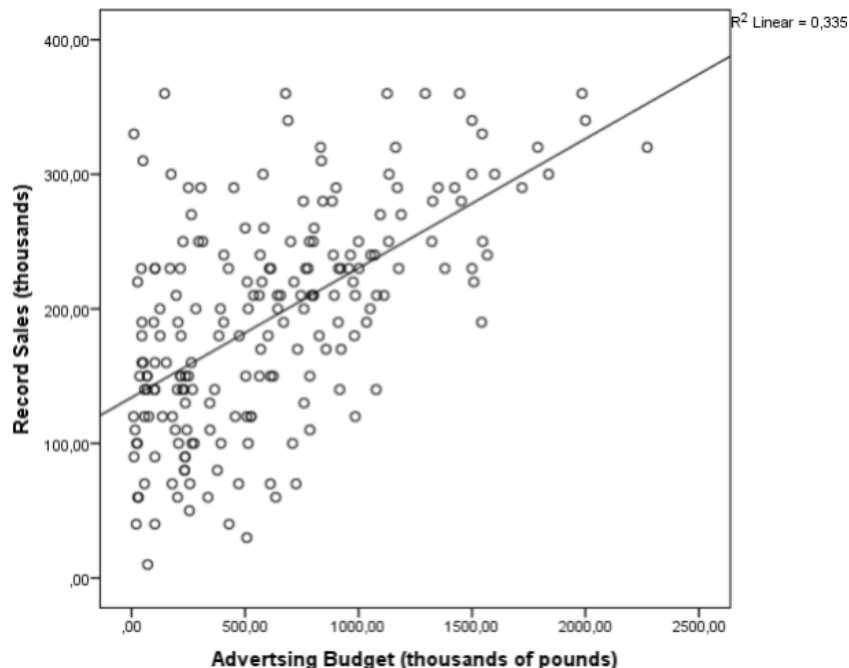
Változók száma	Határérték $p < .001$ szinten
2 (korreláció)	13.28
3	16.27
4	18.47
5	20.52
6	22.46
7	24.32
8	26.13
9	27.88
10	29.59



Többdimenziós outlierek

- **Cook távolság**

- Minden egyes érték regressziós modellre gyakorolt hatását becsli meg.
 - Mennyire lenne más a regressziós modelled, ha az az érték nem lenne benne.
 - Egy a modellnek megfelelő érték hatása viszonylag alacsony, a szélsőségesen értékek viszont el tudják téríteni az egyenest a maguk irányába
- Mindenkihez külön Cook-érték tartozik, mely megadja, az adott személynek mekkora hatása van
 - Többdimenziós outlier az, akinek a Cook-értéke 1 felett van



Ralph Dennis Cook

Többdimenziós outlierek

- **Centered Leverage Values**

- Mennyivel nagyobb a megfigyelt érték hatása a kimeneti változóra, mint a bejósolt érték hatása
 - 0 – az egyednek nincs plusz hatása
 - 1 – teljesen meghatározza a kimeneti értéket
- Rossz *lehet*, ha nagyobb, mint
 - $2 * ([\text{független_változók_száma}] + 1) / [\text{elemszám}]$ (Hoaglin & Welsch, 1978)
 - $3 * ([\text{független_változók_száma}] + 1) / [\text{elemszám}]$ (Stevens, 2012)

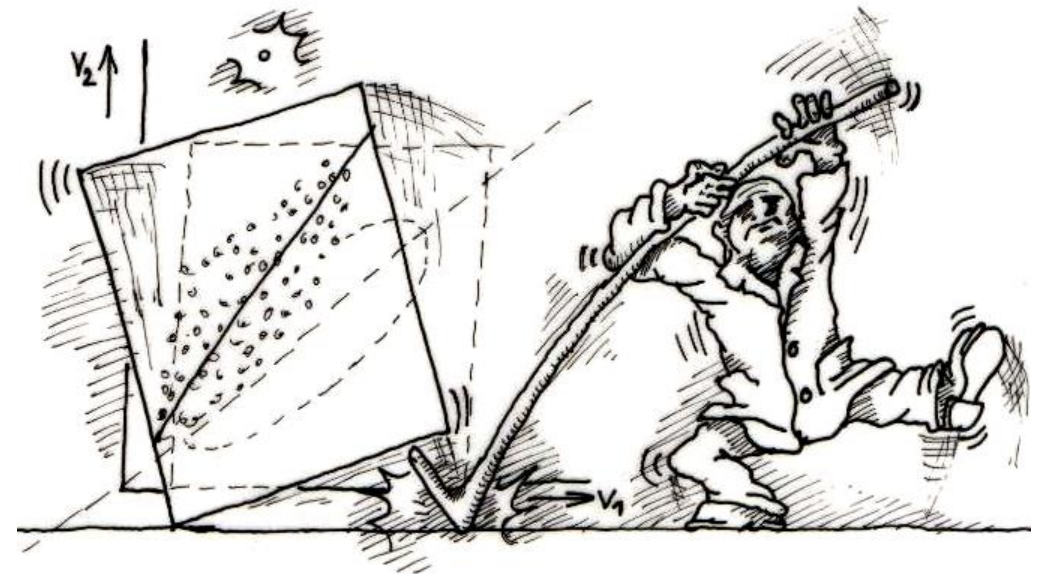


- **Standardizált DfBeta**

- Mennyivel lenne más a b0 és b1 standardizált értéke, ha a változót kivennénk.
- Mindenkinek külön értéke van b0 és minden b1 értékhez
- Valamely $| \text{Standardizált DfBeta} | > 2$ esetben tekintünk valakit outliernek

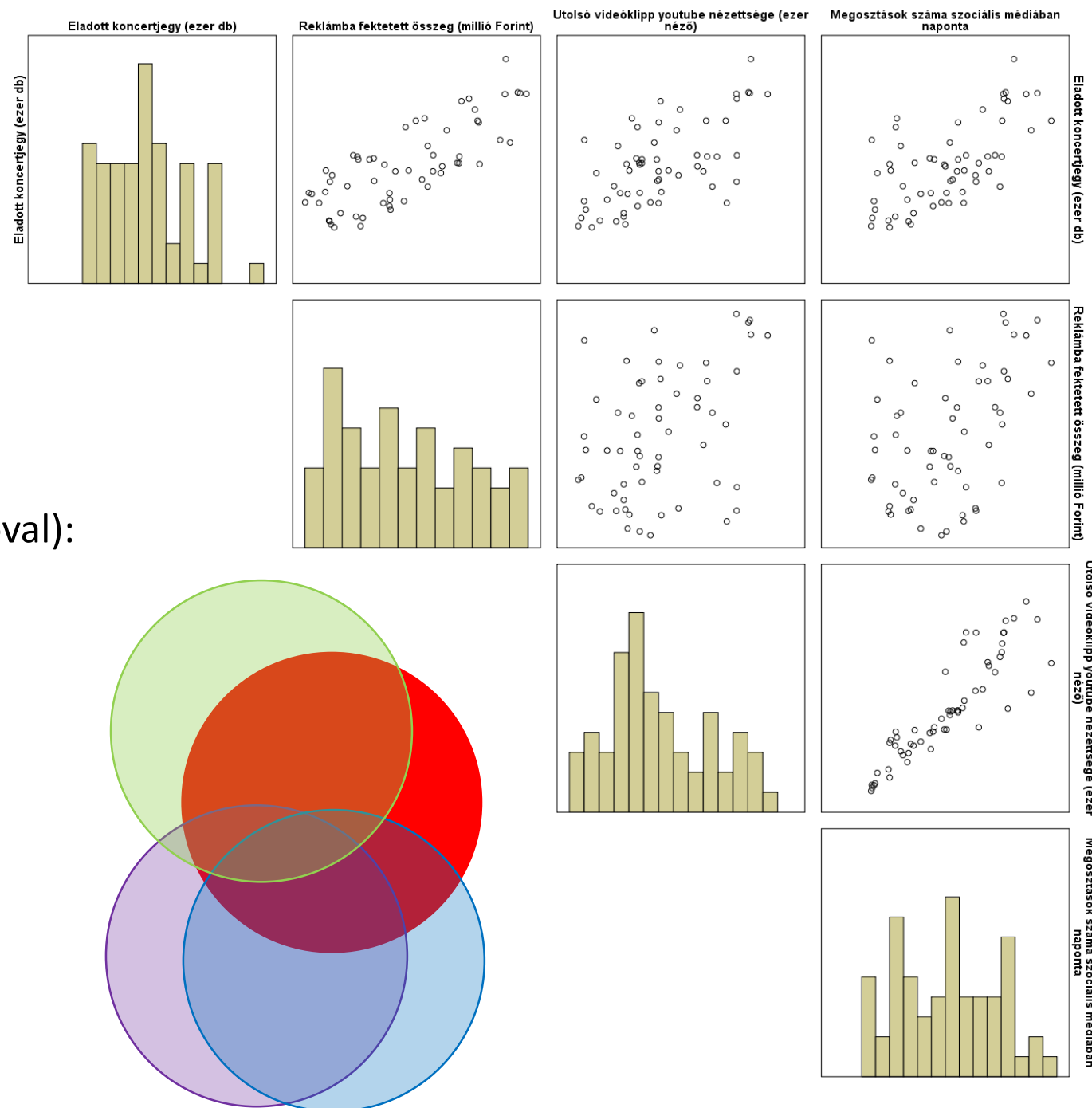
Többdimenziós outlierek

- **4 outlier mutató. Előfordulhat, hogy mindegyik mást mond. Melyiknek higgyek?**
 - Kutatói döntés, de nem használható ürügyként a hipotézisemmel szembemenő adatok törlésére
 - Különbség van a következők között (adatfeldolgozás részben nézd meg):
 - **(Több)dimenziós outlierek** – olyan érték, aki nagyon eltér a többiektől
 - **Befolyásos értékek** – olyan értékek, akik nagyon befolyásolja, hol fut a regressziós egyenes
 - **Nem jól bejósolható értékek** – olyan értékek, akiket a regressziós modell nem tud bejósolni
- Ha valaki „csak” (több)dimenziós outlier, attól még a regresszióban bennmaradhat. Érdekes megvizsgálni, miért más mint a többiek
- Ha nem jól bejósolható, akkor valószínűleg outlier is. Megnöveli az F értékben a hibát, tehát kevésbé lesz szignifikáns a regresszió
- Ha túl befolyásos, akkor viszont megváltoztatja a regressziós egyenest, és ki érdemes venni



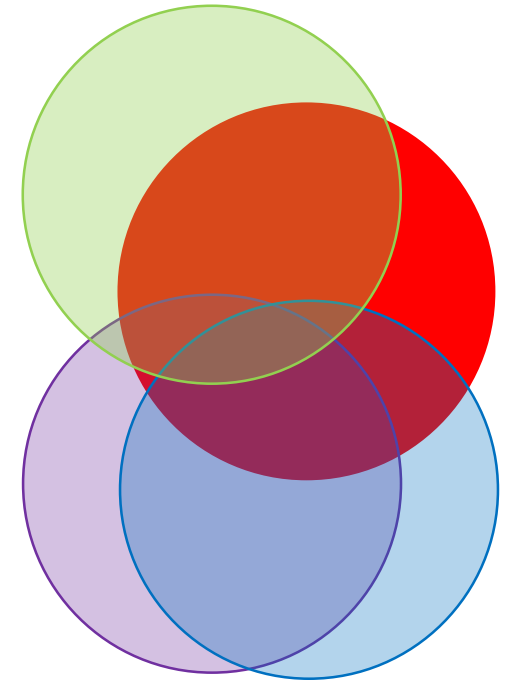
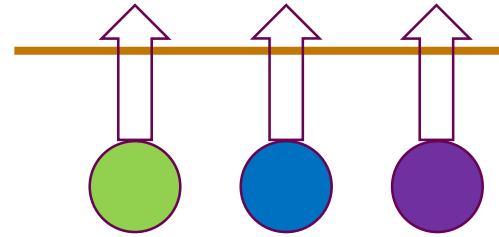
Példa modell felépítésére

- Jósoljuk be az eladott **koncertjegyek** számát (kimeneti változó) abból, hogy
 - mennyit fektetnek a koncert **reklámozás**ába
 - mekkora a zenekar nézettsége **Youtube**-on
 - naponta átlagosan hány **poszt** születik a zenekarról
- Korrelációk a **koncertjegyek**kel (kimeneti változóval):
 - Reklám**: $r = .780$ $p < .001$ (~60%)
 - Youtube**: $r = .643$ $p < .001$ (~41%)
 - Poszt**: $r = .639$ $p < .001$ (~41%)
- Korreláció a prediktorok között:
 - Reklám** x **Youtube**: $r = .393$ $p = .002$ (~15%)
 - Reklám** x **Poszt**: $r = .435$ $p = .001$ (~18%)
 - Poszt** x **Youtube**: $r = .884$ $p < .001$ (~78%)



Példa modell felépítésére

- Mi történik ha **ENTER** módszert választunk?
 - Mindhárom változó egyszerre kerül a modellbe.
 - A **reklám** hatását viszonylag pontosan látni fogjuk, hiszen viszonylag független a többi prediktortól
 - A **youtube** és **poszt** hatását nem tudjuk igazán felmérni, mert túl nagy az átfedés közöttük, az általuk megmagyarázott varianciarész túl nagy részén kell egymással osztozniuk
 - A **poszt**nak, ami 78%-ban osztozik a **youtube**-bal és 18%-ban a **reklámmal**, szinte semmilyen önálló variancia rész nem jut, így nem is lesz szignifikáns – bár önállóan 41%-ot magyarázott volna



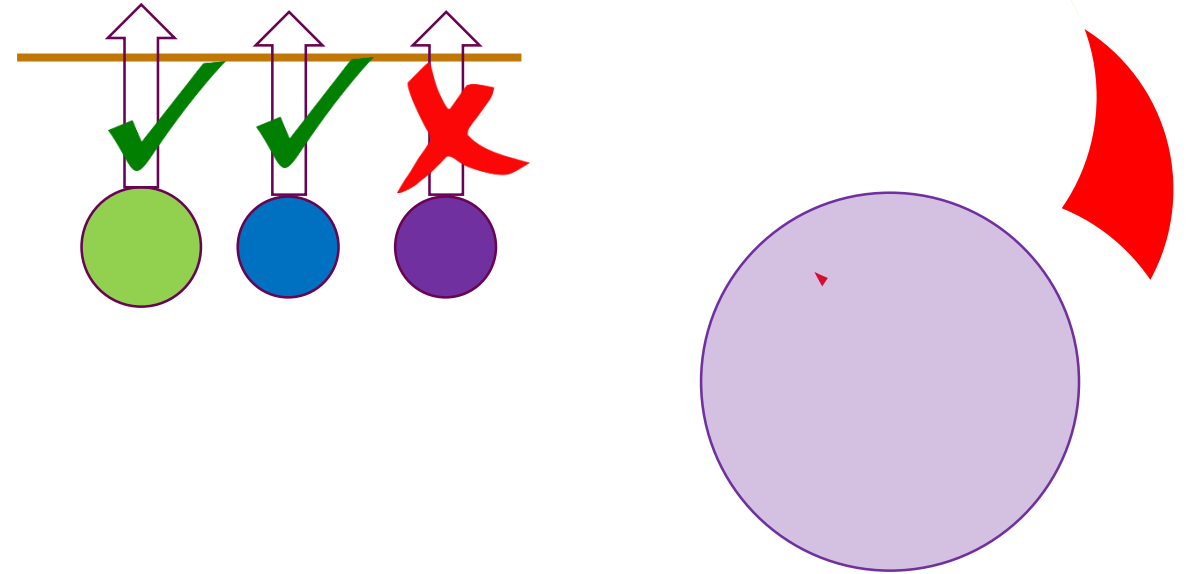
Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,091	,671		4,607	,000
	Reklámba fektetett összeg (millió Forint)	,396	,049	,615	8,102	,000
	Utolsó videóklipp youtube nézettsége (ezer néző)	,002	,001	,329	2,248	,029
	Megosztások száma szociális médiában naponta	,003	,006	,082	,546	,587

a. Dependent Variable: Eladott koncertjegy (ezer db)

Példa modell felépítésére

- Mi történik ha **FORWARD** módszert választunk?
 - Először bekerül a **reklám**, mert ő magyaráz meg legtöbbet a kimeneti változóból
 - Ezt követően megnézzük, a maradék megmagyarázatlan variáciát, melyik változó magyarázza jobban, és ez szignifikánsan hozzá tesz-e a modellhez. Így kerül be a **youtube**
 - Végül megvizsgáljuk, hogy a maradék megmagyarázatlan variációhoz a **poszt** hozzá tud-e tenni. Mivel nem, ezért ő nem kerül be a modellbe
 - Eredményként a **reklám**ot és **youtube**-ot viszonylag pontosan látjuk, a **posztról** viszont semmit nem fogunk tudni.
 - Azt, hogy a második lépésben a **youtube** vagy a **poszt** kerül a modellbe, nem mi döntöttük el, hanem az SPSS



Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,983	,565		8,817	,000
	Reklámba fektetett összeg (millió Forint)	,502	,053	,780	9,405	,000
2	(Constant)	3,290	,560		5,876	,000
	Reklámba fektetett összeg (millió Forint)	,402	,048	,623	8,439	,000
	Utolsó videóklipp youtube nézettsége (ezer néző)	,003	,000	,397	5,379	,000

a. Dependent Variable: Eladott koncertjegy (ezer db)

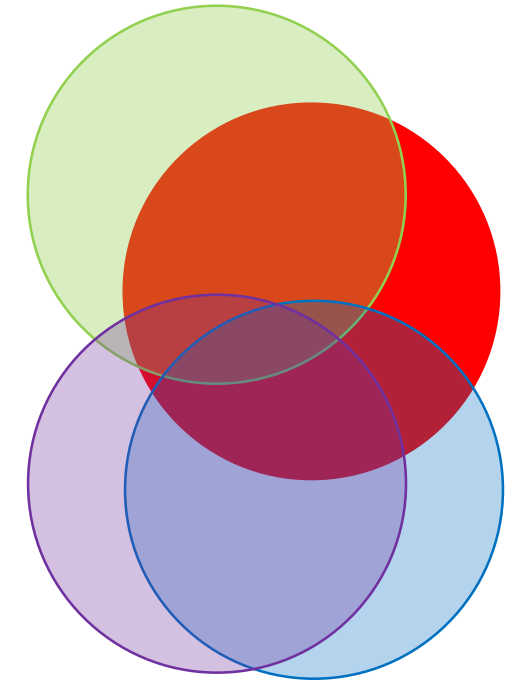
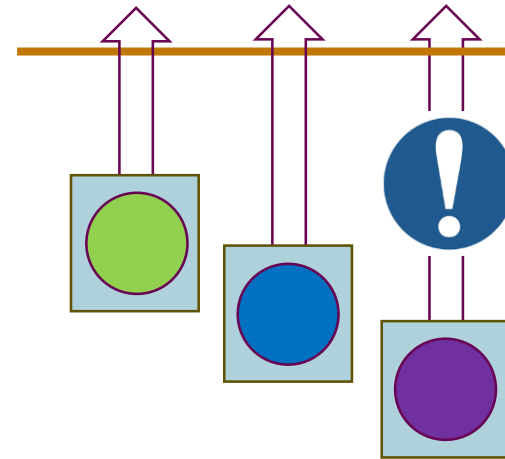
Példa modell felépítésére

- Mi történik ha **BLOCKWISE** módszert választunk?
 - Itt te döntheted el, melyik változók fontosak számodra.
 - Ha azt mondd, hogy a legfontosabb a **reklám**, aztán a **youtube** végül a **poszt**, akkor ebben a sorrendben fog a beléptetés haladni

Coefficients^a

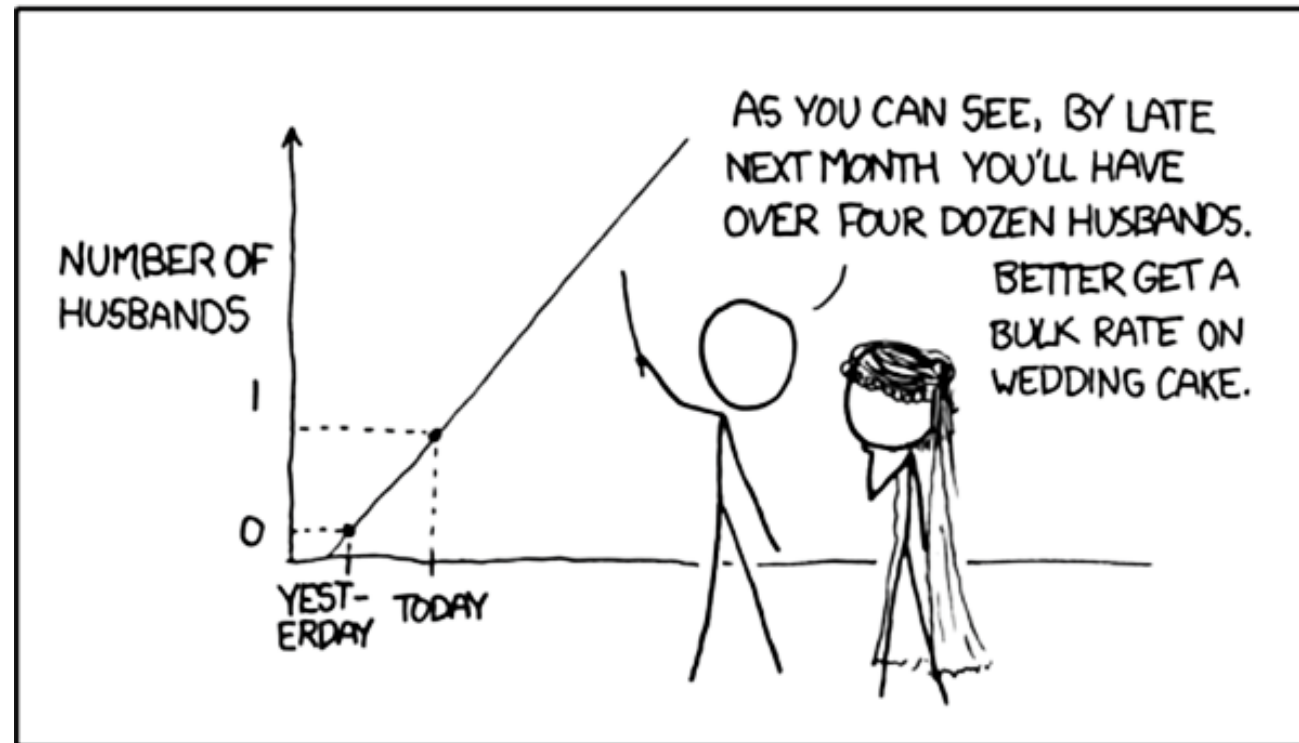
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,983	,565		8,817	,000
	Reklámba fektetett összeg (millió Forint)	,502	,053	,780	9,405	,000
2	(Constant)	3,290	,560		5,876	,000
	Reklámba fektetett összeg (millió Forint)	,402	,048	,623	8,439	,000
	Utolsó videóklipp youtube nézettsége (ezer néző)	,003	,000	,397	5,379	,000
3	(Constant)	3,091	,671		4,607	,000
	Reklámba fektetett összeg (millió Forint)	,396	,049	,615	8,102	,000
	Utolsó videóklipp youtube nézettsége (ezer néző)	,002	,001	,329	2,248	,029
	Megosztások száma szociális médiában naponta	,003	,006	,082	,546	,587

a. Dependent Variable: Eladott koncertjegy (ezer db)



- Meg fogod kapni mind a három modellt, és a R²-et változásából látni fogod, hogy a második modellhez képest a harmadik már nem jelent szignifikáns növekedést.
- Ezért bár látod mind a három modellt, a másodikat fogod értelmezni.
- Így kaptad a legteljesebb képet, és kontrolláltad a változók sorrendjét is

Többszörös lineáris regresszió értelmezés



Többszörös lineáris regresszió értelmezés

- A regressziós táblák értelmezése előtt a feltételek teljesülését ellenőrizzük
 - A feltételek egymásra is hatnak, ezért érdemes azokkal a feltételekkel kezdeni, melyek viszonylag függetlenek a többi teljesülésétől, vagy amelyek sérülése nagyon sok mindent „elronthat”
- Ajánlott sorrend:
 - Nem ellenőrizendő, hanem a vizsgálat tervezése során kialakítandó feltételek:
 - ① Változók típusa
 - ④ Nincs külső változó
 - ⑨ Elemszám
 - Elemzés során ellenőrizhető feltételek (és ajánlott sorrendjük):
 - ② Nincs nulla variáciájú változó
 - ⑩ Nincsenek (több)dimenziós outlierok
 - ③ Nincs (multi)kollinearitás
 - ⑥ Linearitás
 - ⑤ Homoszkedaszticitás
 - ⑦ Független hibák
 - ⑧ Hiba normális eloszlása

Többszörös lineáris regresszió értelmezés

- ⑩ (Több)dimenziós outlierok

- Az egydimenziós outlierok szűrése, kezelése az adattisztítás során megismert módon történik.
- Többdimenziós outlierok ellenőrzésére 4 mutatót használhatunk, melyek kritériumszintjei:
 - Mahalanobis:** a táblázatból kikeresve 16.27 a kritérium
 - Cook's distance:** 1 a kritérium
 - Centered Leverage:** $2 \cdot (2+1)/59 = 0.10$ vagy $3 \cdot (2+1)/59 = 0.15$ a kritérium
 - Dfbeta értékek:** ± 2 a kritérium
- ha valaki az adatbázisban e kritériumnál magasabb pontszámot ér el, az többdimenziós outlier vagy befolyásos érték
- A négy módszer nem mindig hoz azonos eredményt, ekkor mérlegelni kell, kihagyjuk-e az adott elemet
 - Különbség van többdimenziós outlier, befolyásos érték és bejósolhatatlan érték között!

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4,98680	16,23129	9,48331	3,148902	59
Std. Predicted Value	-1,428	2,143	,000	1,000	59
Standard Error of Predicted Value	,250	,719	,411	,114	59
Adjusted Predicted Value	5,04957	16,26439	9,48521	3,142641	59
Residual	-4,262718	4,286459	,000000	1,858773	59
Std. Residual	-2,253	2,266	,000	,983	59
Stud. Residual	-2,295	2,298	,000	1,003	59
Deleted Residual	-4,421225	4,408982	-,001910	1,938978	59
Stud. Deleted Residual	-2,389	2,393	,003	1,019	59
Mahal. Distance	,032	7,388	1,966	1,685	59
Cook's Distance	,000	,106	,014	,019	59
Centered Leverage Value	,001	,127	,034	,029	59

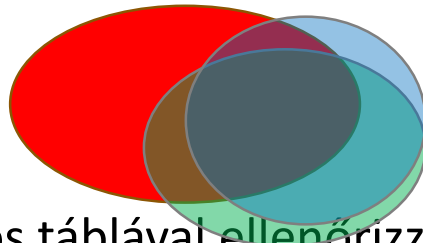
a. Dependent Variable: Eladott koncertjegy (ezer db)

Többszörös lineáris regresszió értelmezés

- ③Nincs (multi)kollinearitás

- Kollinearitás:

- Ha két változó közötti túl erős korreláció, akkor a kimeneti változóból is hasonló részt magyaráznak, mindkettőt a modellbe téve csökkennének a magyarázóerők



- Korrelációs táblával ellenőrizzük: ha a prediktorok között van olyan, mely .8-nál erősebben korrelál, akkor nem teljesül a kritérium.
 - Ekkor vagy ki kell venni a két változó közül az egyiket, vagy főkomponens elemzéssel össze kell őket egyetlen változóba gyúrni

Correlations

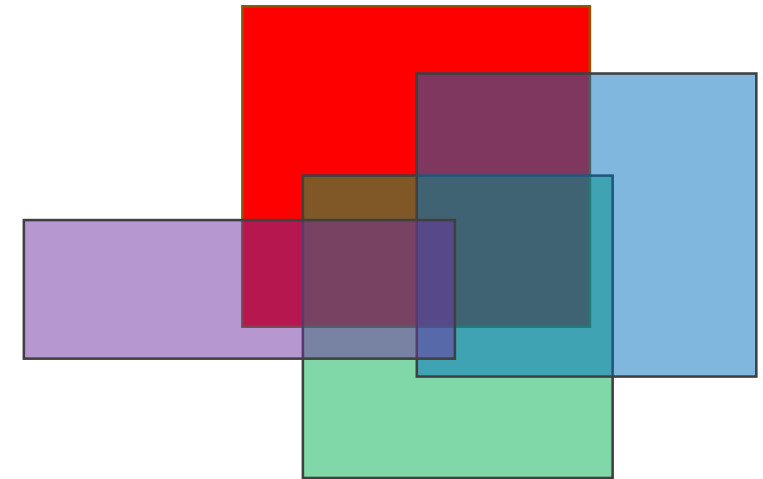
		Eladott koncertjegy (ezer db)	Reklámba fektetett összeg (millió Forint)	Utolsó videóklipp youtube nézettsége (ezer néző)
Pearson Correlation	Eladott koncertjegy (ezer db)	1,000	,780	,643
	Reklámba fektetett összeg (millió Forint)	,780	1,000	,393
	Utolsó videóklipp youtube nézettsége (ezer néző)	,643	,393	1,000
Sig. (1-tailed)	Eladott koncertjegy (ezer db)	.	,000	,000
	Reklámba fektetett összeg (millió Forint)	,000	.	,001
	Utolsó videóklipp youtube nézettsége (ezer néző)	,000	,001	.
N	Eladott koncertjegy (ezer db)	59	59	59
	Reklámba fektetett összeg (millió Forint)	59	59	59
	Utolsó videóklipp youtube nézettsége (ezer néző)	59	59	59

Többszörös lineáris regresszió értelmezés

- ③Nincs (multi)kollinearitás

- Multikollinearitás:

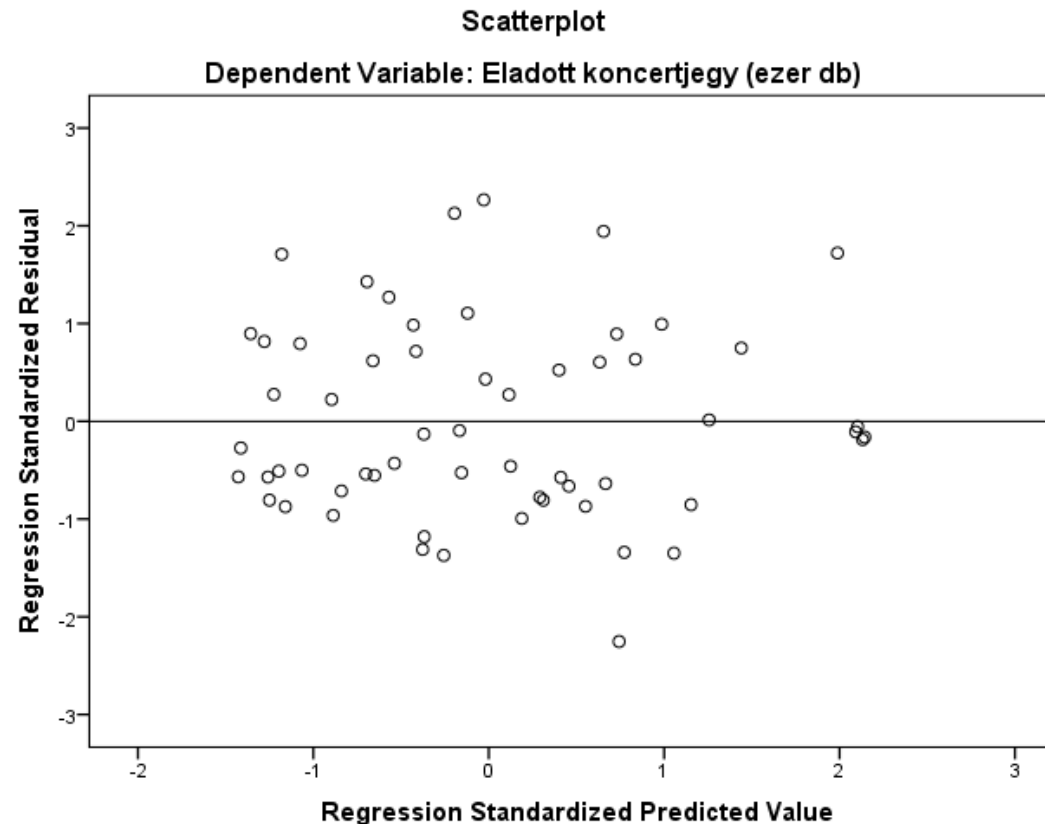
- Előfordulhat olyan változó, melyek több változóval közepesen erősen / erősen korrelál, és az általa megmagyarázott variancia nagy részét valamely más változó is magyarázza
- Az ilyen változó redundáns, a modellben nem lehet a hatását értelmezni
- A multikollinearitást tolerancia értékekkel ellenőrizzük
 - Ha van olyan változó, akinek a toleranciája nagyon alacsony (10% alatt) van, azt a változót ki kell venni a modellből.
 - Már a 40% tolerancia értékeknél is figyelni kell, mert a változó hatását nem látjuk tisztán (az alacsony toleranciájú változó hatása sokszor nem szignifikáns, és nem tudjuk eldönteni, hogy azért, mert tényleg nincs a változónak hatása vagy csak átfedésben van a többi változó hatásával)



		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	Reklámba fektetett összeg (millió Forint)	1,000	1,000
2	(Constant)		
	Reklámba fektetett összeg (millió Forint)	,845	1,183
	Utolsó videóklipp youtube nézettsége (ezer néző)	,845	1,183

Többszörös lineáris regresszió értelmezés

- ⑤,⑥ **Homoszkedaszticitás és linearitás ellenőrzése**
- A standardizált predikált érték és a standardizált reziduális hibák pontdiagramja
 - Egyenletes pontfelhőt várunk – minden prediktor értéknél hasonló mértékű hiba kell hogy legyen
 - Homoszkedaszticitás: nem változik a pontfelhő szélessége
 - Linearitás: nincs a pontfelhőnek iránya, vagy görbülete



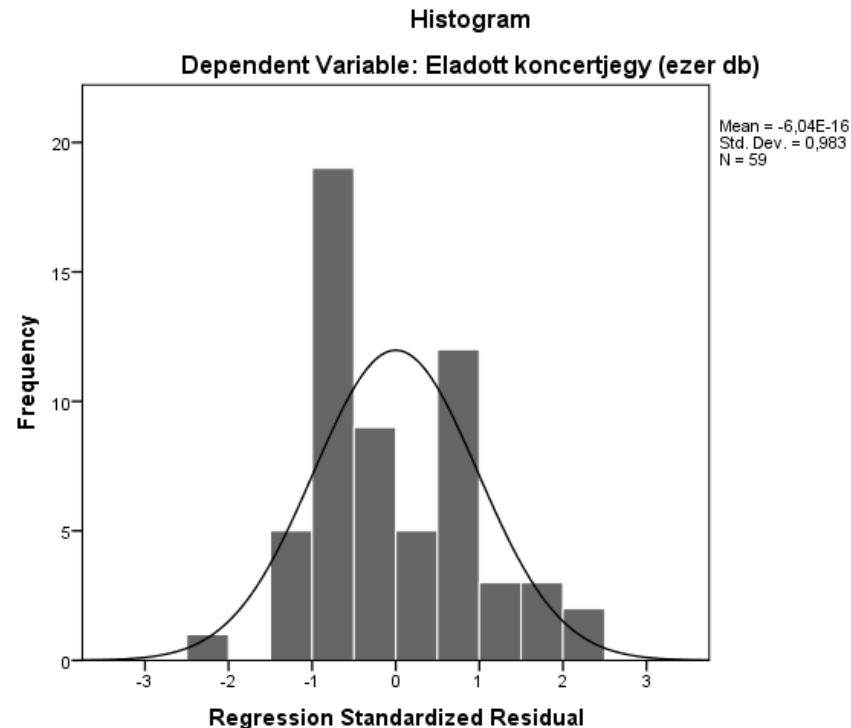
Többszörös lineáris regresszió értelmezés

- ⑦Független hibák

- A Durbin-Watson teszt ellenőrzi a hibák korrelálatlanságát
- ha 2 körül van jó, ha 1 alatt vagy 3 felett, akkor baj van

- ⑧Hiba normális eloszlása

- A prediktorok segítségével a kimeneti változóból meg nem magyarázott varianciarész (a reziduális hiba) a vizsgálat szempontjából már csak zaj, ennek megfelelően az eloszlása normál eloszlást kell kövessen
- Hisztogrammal ellenőrizzük



Többszörös lineáris regresszió értelmezés

- **Model Summary**

- **R Square:** mennyit tudunk a modellel megmagyarázni az adatokból. Az első modell csak 60,8%-ot, a második 74,2%-ot
- **Adjusted R Square:** Mennyire számíthatnánk az egész populációt mérve. Ha közel van a sima R^2 -hez, akkor a mintán létrehozott modell várhatóan hasonlóan jó lenne a populációban is
- **Change Statistics:** Az új változók beemelése mennyiben tette jobbá a modellt (R Square Change), és egy F-próba arra, hogy a javulás szignifikáns-e
- A táblázat alatti megjegyzések segítenek az értelmezésben

Model Summary^c

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson ^⑦
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,780 ^a	,608	,601	2,309072	,608	88,447	1	57	,000	
2	,861 ^b	,742	,732	1,891674	,133	28,929	1	56	,000	1,891

a. Predictors: (Constant), Reklámba fektetett összeg (millió Forint)

b. Predictors: (Constant), Reklámba fektetett összeg (millió Forint), Utolsó videóklipp youtube nézettsége (ezer néző)

c. Dependent Variable: Eladott koncertjegy (ezer db)

Többszörös lineáris regresszió értelmezés

- **ANOVA** tábla
 - **Sum of Squares** és **Mean Square** oszlopok tartalmazzák az SS_M , SS_R , SS_T és MS_M , MS_R értékeket – vesd össze az elmélettel
 - **F-érték és Sig.:**
 - Statisztikai érték, tehát a hatás és hiba aránya
 - A modell általánosítható-e a populációra

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	471,583	1	471,583	88,447	,000 ^b
	Residual	303,913	57	5,332		
	Total	775,496	58			
2	Regression	575,104	2	287,552	80,357	,000 ^c
	Residual	200,392	56	3,578		
	Total	775,496	58			

a. Dependent Variable: Eladott koncertjegy (ezer db)

b. Predictors: (Constant), Reklámba fektetett összeg (millió Forint)

c. Predictors: (Constant), Reklámba fektetett összeg (millió Forint), Utolsó videóklipp youtube nézettsége (ezer néző)

Többszörös lineáris regresszió értelmezés

- Az ANOVA és modell Summary táblát együtt kell értelmezni.
 - A regresszió két lépésben állt fel: első modell még csak a reklámot tartalmazta, a második a reklámot és youtube-ot.
 - Az **ANOVA** táblából látod, hogy mindkét modell szignifikáns, tehát mindkettőt szabad értelmezni (a populációra értelmezni)
 - A **Model Summary** táblából látod, hogy az első 60.8%, a másik 74.2%-ot magyaráz, mindkettő modellnek elég magas a magyarázó ereje, hogy legyen értelme őket magyarázni
 - A **Change Statistics**-ből látod, hogy a 60.8%-ról 74.2%-ra való növekedés a magyarázóerőben szignifikáns, tehát érdemesebb ezt a második, bonyolultabb, több prediktort használó modellt használni – a youtube modellbe léptetése szignifikánsan növelte a megmagyarázott varianciát.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	471,583	1	471,583	88,447	,000 ^b
	Residual	303,913	57	5,332		
	Total	775,496	58			
2	Regression	575,104	2	287,552	80,357	,000 ^c
	Residual	200,392	56	3,578		
	Total	775,496	58			

Model Summary^c

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,780 ^a	,608	,601	2,309072	,608	88,447	1	57	,000
2	,861 ^b	,742	,732	1,891674	,133	28,929	1	56	,000

a. Predictors: (Constant), Reklámba fektetett összeg (millió Forint)

b. Predictors: (Constant), Reklámba fektetett összeg (millió Forint), Utolsó videóklipp youtube nézettsége (ezer néző)

c. Dependent Variable: Eladott koncertjegy (ezer db)

Többszörös lineáris regresszió értelmezés

- **Coefficients (Regressziós együtthatók)**

- **B:** b_0 -át (konstans kezdőérték) és b_1 -et (meredekség) adja meg minden változóra úgy, hogy mi lenne a hatásuk, ha a többi kontrollálnánk
 - b_0 - ha a független változók értéke 0, mennyi lenne a függő változó értéke
 - b_1 – a független változó egységnyi változása mekkora változást idéz elő a függőben
- **Beta:** a b_1 standardizált értékei – így már összehasonlíthatóak egymással
- **t és sig:** b_0 és b_1 értéke szignifikánsan eltér-e a 0-tól (a 0 azt jelentené, hogy az adott változónak nincs hatása a függő változóra)

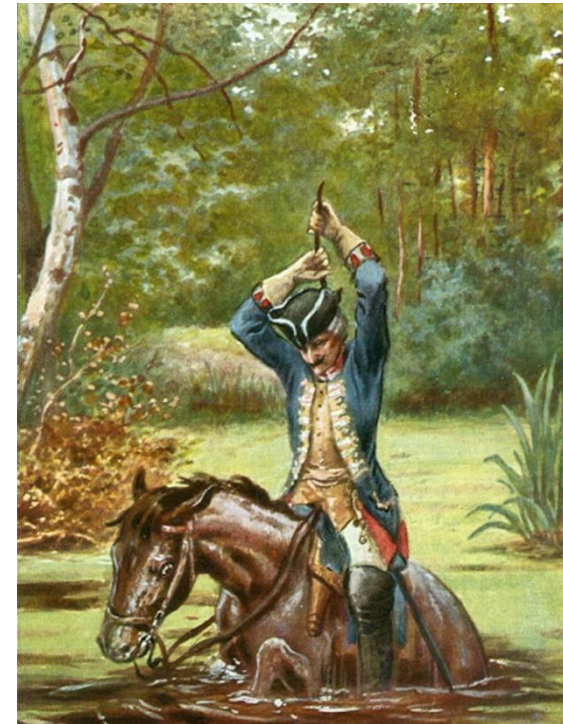
Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,983	,565		8,817	,000		
	Reklámba fektetett összeg (millió Forint)	,502	,053	,780	9,405	,000	1,000	1,000
2	(Constant)	3,290	,560		5,876	,000		
	Reklámba fektetett összeg (millió Forint)	,402	,048	,623	8,439	,000	,845	1,183
	Utolsó videóklipp youtube nézettsége (ezer néző)	,003	,000	,397	5,379	,000	,845	1,183

a. Dependent Variable: Eladott koncertjegy (ezer db)

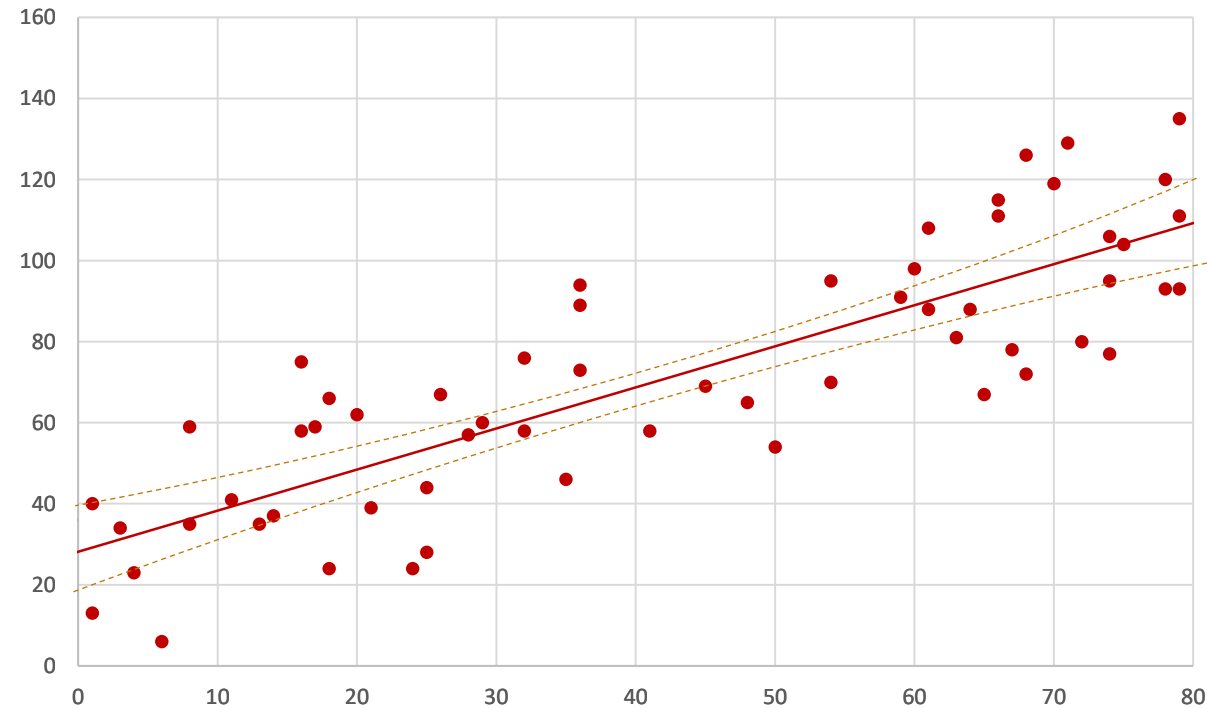
Béták összehasonlíthatósága

Meredekséghez tartozó konfidencia intervallumok számítása Bootstrappinggel



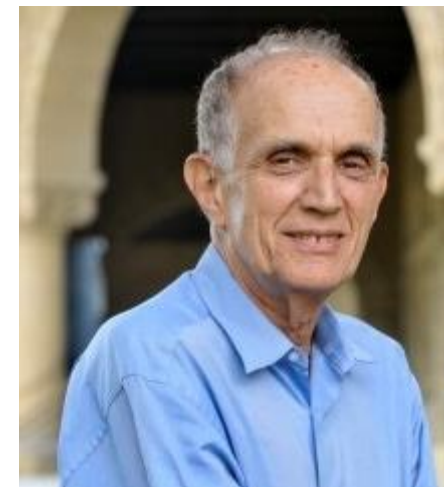
Bootstrapping

- A mintából becsült paramétereknek (legyen az átlag vagy a regressziós egyenes meredeksége) van valamekkora bizonytalansága
- Paraméter bizonytalanságának becslésére használható a **Bootstrapping**



Bootstrapping

- A **Bootstrapping** egy a mintából való ismételt mintavételezésen alapuló módszer
 - Az N méretű mintából kiválaszt sok (5-10ezer) N méretű mintát úgy, hogy egy elemet többször is választhat. Mind a sok ezer új mintán kiszámolja a becsülni kívánt paramétert (például regressziós együtthatót), így felrajzolhatjuk annak eloszlását, számolhatjuk bizonytalanságát (pl. az együttható szórását vagy konfidencia intervallumát)
- Becsülhető vele mintából becsült paraméter bizonytalansága
 - Becsülhető vele a végzett statisztika megbízhatósága
- Használható hipotézis tesztelésre torzító hatás jelenlétében vagy feltétel sérülése esetén is
 - Csökkenthető az outlierok torzító hatása
 - Csökkenthető a normalitás sérüléséből adódó torzítás
- Fun fact:
 - Egyesek szerint Münchhausen báró után nevezték el, aki önmagát saját hajánál vagy csizmájánál fogva húzta ki a mocsárból :D
 - A Bootstapping elődje az ún. jackknife (zsebkés) módszer
- Bootstrap csak az újabb SPSS-ekben van



Bradley Efron



Bootstrapping

Bootstrap for Coefficients

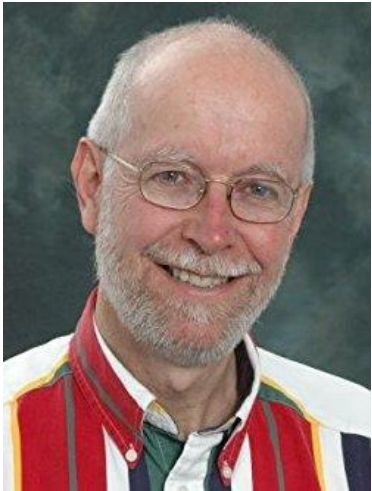
Model	B	Bootstrap ^a				
		Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
					Lower	Upper
1 (Constant)	1,040E-15	,000	,069	1,000	-,125	,137
Zscore: Reklámba fektetett összeg (millió Forint)	,623	-,001	,060	,001	,507	,739
Zscore: Utolsó videóklipp youtube nézettsége (ezer néző)	,397	,000	,062	,001	,272	,513

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Standardizált regressziós együttható (értéke megegyezik a „rendesen” kikért regresszió Béta értékével. Ezt mértük a mintában.

Standardizált regressziós együtthatóhoz tartozó 95%-os CI alsó és felső határa. A populációban a regressziós együttható 95%-os valószínűséggel e kettő érték között helyezkedik el.

Különbség béták között

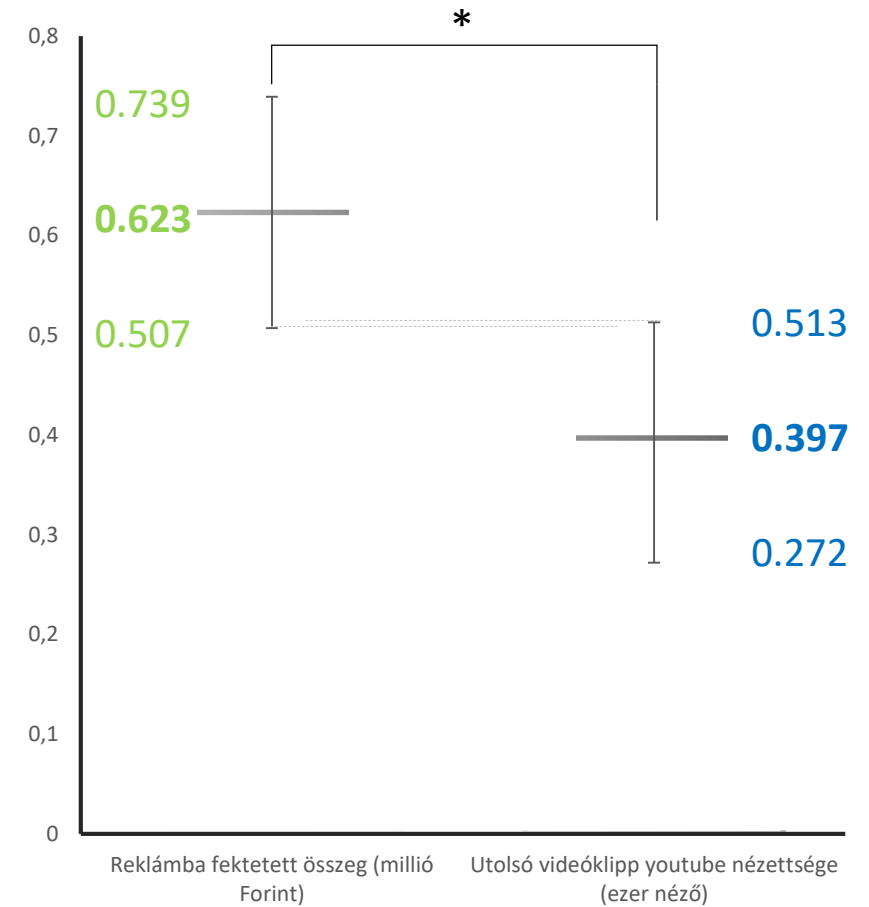


Geoff Cumming

- Ha két változó regressziós együtthatójához tartozó 95CI maximum 50%-ban fed át egymással, akkor a két regressziós együttható szignifikánsan különbözik egymástól $p < .05$ szinten.
- Ha nincs átfedés, akkor szignifikánsan különböznek egymástól $p < .001$ szinten (Cumming és Finch, 2005)
- Szemre is látszik, hogy a két regressziós együttható konfidencia intervalluma alig fed át egymással, az átfedés bőven 50% alatt van, tehát a REKLÁM hatása $p < .05$ szinten szignifikánsan nagyobb, mint a YOUTUBE hatása.
- Ha az eredmény szemre is egyértelmű, nem kell a pontos értékeket kiszámolni.



Sue Finch

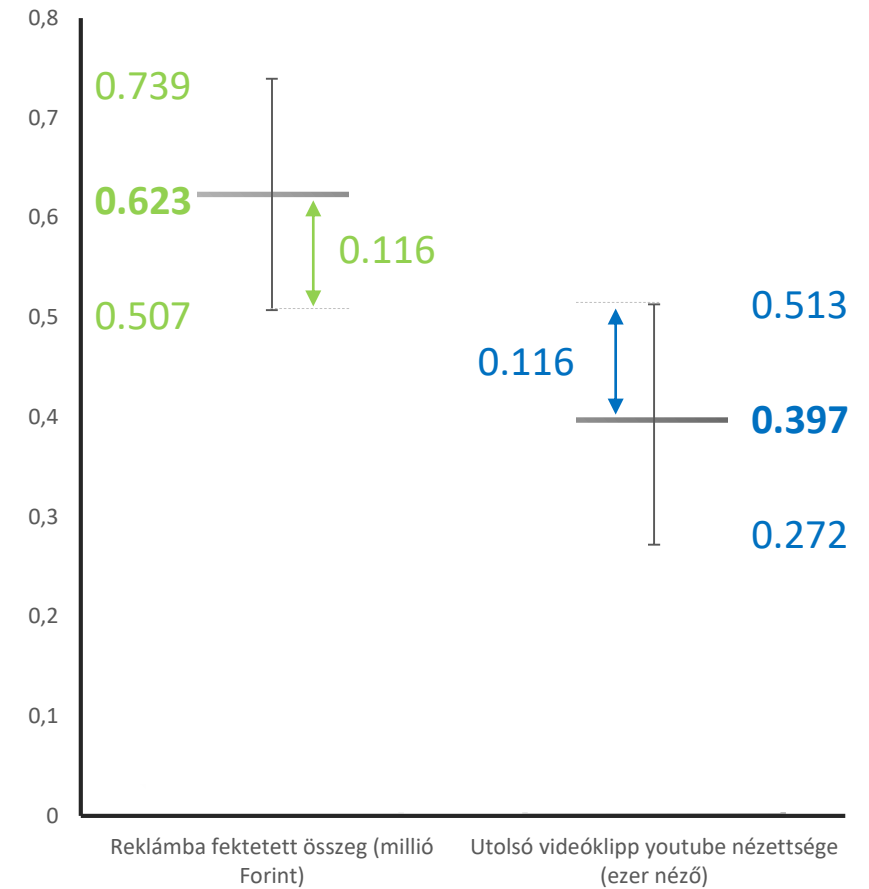


Különbség béták között

- Ha szemre nem egyértelmű, ki kell számolni a két CI átfedését.
- A számítás szemléltetéséhez használjuk a következő értékeket:

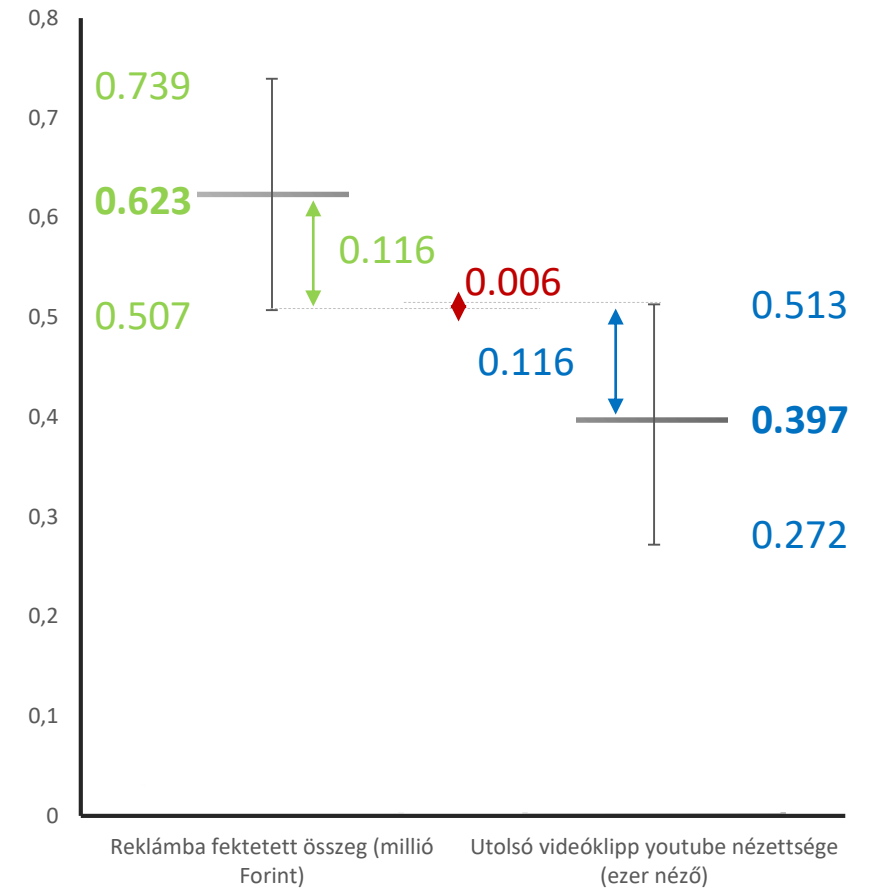
	95CI alsó határa	β értéke	95CI felső határa
REKLÁM	0.507	0.623	0.739
YOUTUBE	0.272	0.397	0.513

- 1. lépés:
számítsuk ki a magasabb bétához tartozó konfidencia intervallum alsó felének szélességét, és az alacsonyabb béta értékhez tartozó konfidencia intervallum felső felét szélességét.
(a bootstrappinggel kapott CI-k nem feltétlenül szimmetrikusak)
 $\text{Ciszélesség}_X = 0.623 - 0.507 = 0.116$
 $\text{Ciszélesség}_Y = 0.513 - 0.397 = 0.116$ (véletlen azonos a két érték)



Különbség béták között

- 2. lépés
Kiszámoljuk a két szélesség átlagát:
 $M_{CI} = (0.116 + 0.116)/2 = 0.116$
- 3. lépés
Kiszámoljuk a két intervallum átfedő részét
Átfedés = $0.513 - 0.507 = 0.006$
- 4. lépés
Kiszámoljuk, hogy az átfedés hány százaléka a két CI átlagos szélességének.
 $\% = \text{Átfedés} / M_{CI} = 0.006 / 0.116 = 0.0517$
- Az átfedés a két konfidencia intervallum átlagos szélességének 5.17%-át teszi ki, tehát a két regressziós együttható $p < .05$ szinten szignifikánsan különbözik egymástól.



Különbség béták között - gyakorlás

Bootstrap for Coefficients

Model		B	Bootstrap ^a				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	-1,287E-16	,001	,039	1,000	-,082	,087
	ZPREDA	,512	,001	,043	,001	,425	,597
	ZPREDB	,512	,001	,046	,001	,423	,609
	ZPREDC	,189	-,003	,040	,001	,110	,260

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Különbség béták között - gyakorlás

Bootstrap for Coefficients

Model		B	Bootstrap ^a				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	-1,841E-15	-,002	,035	1,000	-,067	,064
	ZPRED1	,704	-,002	,046	,001	,609	,784
	ZPRED2	,112	,000	,042	,013	,037	,194
	ZPRED3	,266	,002	,046	,001	,168	,364
	ZPRED4	,044	-,001	,036	,232	-,024	,110

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Elemzési feladat

Elemyszám becslése regresszióban

Elemzési becslés regresszióban

Elemzési becslés természetesen működik folytonos változók esetében is. A regresszióval kapcsolatos elemzési becslők a mért hatásnagyságra, a modell bonyolultságára (prediktorok számára), statisztikai erőre és a szignifikancia szintre fognak rákérdeni.

Egy jó kalkulátor: <http://www.danielsoper.com/statcalc/>

Please enter the necessary parameter values, and then click 'Calculate'.

Anticipated effect size (f^2):

0.117



Desired statistical power level:

0.8



Number of predictors:

2



Probability level:

0.05



Calculate!

Minimum required sample size: 85

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,343 ^a	,117	,052	19,618

a. Predictors: (Constant), PRED1, PRED2

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1382,501	2	691,251	1,796	,185 ^b
	Residual	10391,666	27	384,877		
	Total	11774,167	29			

a. Dependent Variable: KIMENETI

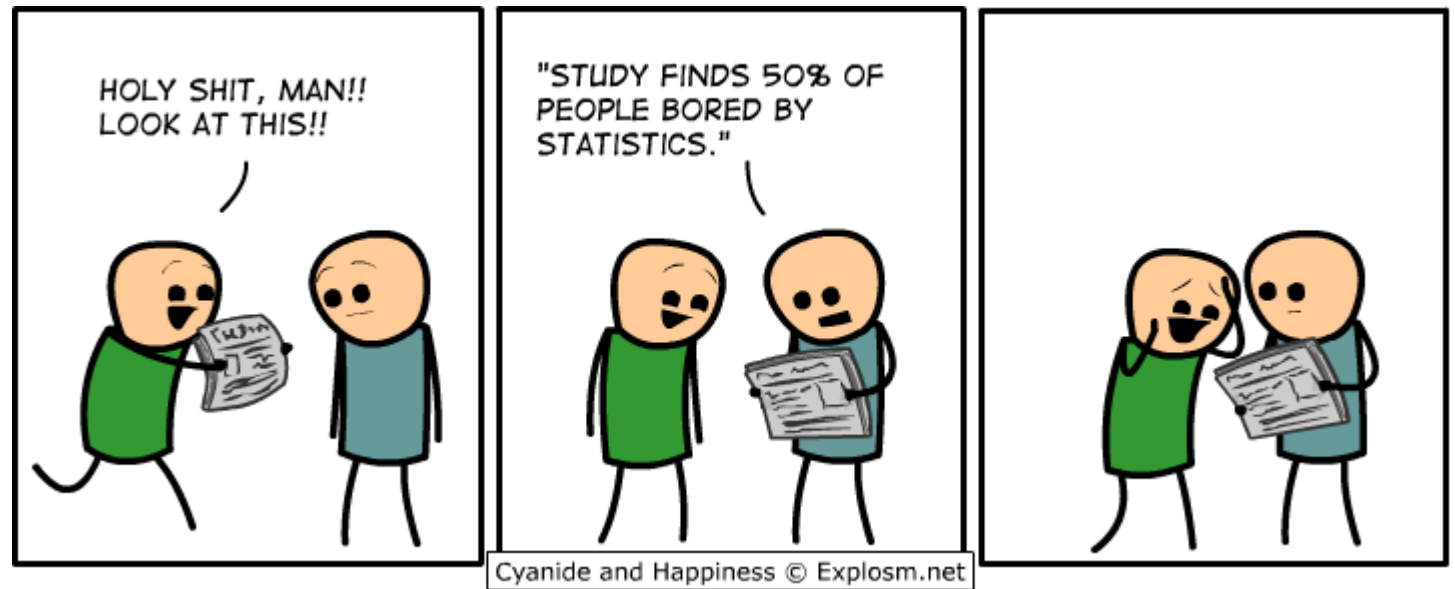
b. Predictors: (Constant), PRED1, PRED2

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-13,973	58,843		-,237	,814
	PRED2	7,747	4,766	,298	1,626	,116
	PRED1	,397	,326	,223	1,216	,234

a. Dependent Variable: KIMENETI

Többszörös lineáris regresszió publikálása



Többváltozós regresszió publikálása

Kezdd a választott statisztika, a kimeneti változó és a prediktor változók megnevezésével!

A publikálás példájában a gyakorlaton elsőként elhangzott, és a gyakorlati pdf-ben részletesen értelmezett modell egy lehetséges publikálását adom meg, tehát a kimeneti változó a koncertjegy, a prediktorok a reklámba fektetett összes és a zenekar utolsó videoklipjének nézettsége Youtube-on.

Hierarchikus regresszió segítségével elemeztük, hogy egy koncertre eladott jegyek számát (továbbiakban KONCERTJEGY) milyen mértékben jósolja be a koncert reklámozásába fektetett pénz (továbbiakban REKLÁM) és a zenekar utolsó videoklipjének nézettsége a Youtube-on (továbbiakban YOUTUBE).

Többváltozós regresszió publikálása

Ha az adattisztítás során vettél ki értékeket az elemzésből, az itt leírhatod. Az egydimenziós adattisztítást, lehet, hogy leírtad már korábban, ebben az esetben nem kell itt is megismételni.

A adatokban az outlier labelling rule (Tukey, 1977) alapján nem található egydimenziós outlier.

A többdimenziós outlier szűrésre azonban csak most került sor, tehát azt mindenképp itt kell közölni. Csak azokban a mutatóknak a kritériumszintjét közöld, amit használtál is (pl. ha csak Mahalanobis és Cook távolságot használtál, akkor csak azokat. Ha nincsen outlier, vagy úgy döntesz, nem veszed ki őket, akkor nem kell erre túl sok sort áldozni, ha viszont kivettél értéket, akkor érdemes azt a grafikonon is megmutatni.

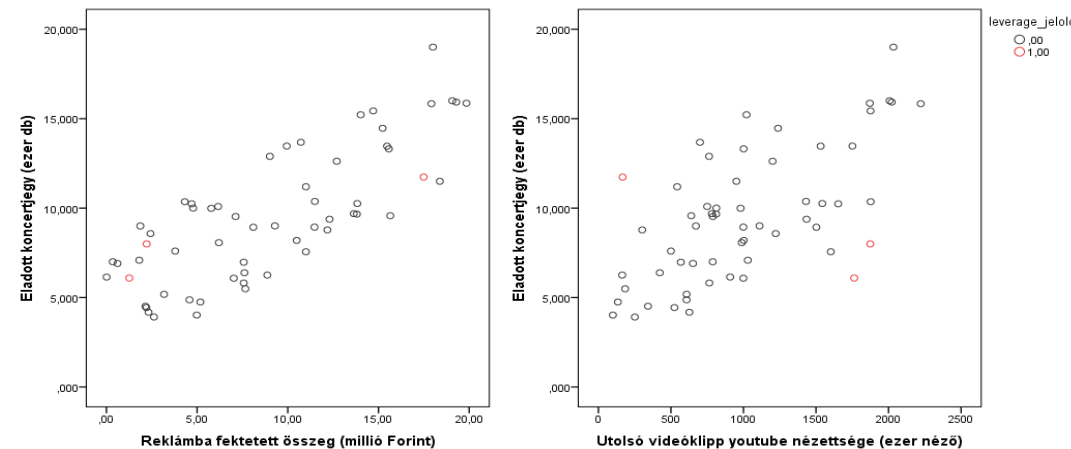
A többdimenziós outlierok szűrésére Mahalanobis és Cook távolságot, Centered Leverage valamint Standardizált DfBeta értékeket számoltunk és az adatokat a következő kritériumszintek mellett ellenőriztük: Mahalanobis távolság $p < .001$ mellett 16.27 (Barnett és Lewis, 1994), Cook távolság 1 (Cook és Weisberg, 1982), Centered Leverage érték 0.10 (Hoaglin és Welsch, 1978), Standardizált DfBeta 2 (Stevens, 2012).

Az eredmények azt mutatják, hogy az adatokban a Centered Leverage értékek alapján 3 többdimenziós outlier található. A többi módszer eredményei alapján nem található szélsőséges érték az adatbázisban.

Többváltozós regresszió publikálása

Ha vettél ki outliert, akkor érdemes őket grafikonon is megmutatni.

A Centered Leverage érték alapján többdimenziós outliernek tekintendő értékek helyét ábrázolja a következő két grafikon. A grafikonokon eltérő színnel jelöltük a Hoaglin és Welsch (1978) javaslata alapján számolt szigorúbb, 0.10-es kritérium szintet meghaladó értékeket. A Stevens (2012) alapján számolt megengedőbb, 0.15-ös kritériumot egyik potenciális outlier sem haladja meg.



Végül írd le, hogy kivetted-e őket vagy sem!

A grafikonokat értékelve valamint figyelembe véve a többi diagnosztikus statisztika eredményét is Stevens (2012) útmutatása alapján az értékek nem kerültek a mintából kivételre.

Többváltozós regresszió publikálása

Érdemes a változót bemutatni, vagy ha már korábban a módszertani résznél megtetted ezt, akkor legalább visszautalni arra a táblára.

Az adatok megtisztítását követően a regresszióba kerülő változók jellemzői a következők:

	Minimum	Maximum	Átlag	Szórás	Elemszám
Eladott koncertjegy	3.910	19.000	9.483	3.657	59
Reklámba fektetett összeg	0.010	19.85	8.958	5.676	59
Videoklip nézettsége	100	2222	1010.458	565.769	59

Eladott koncertjegy ezer darabban, a befektetett összeg millió Forintban, a videoklipp nézettsége ezer nézőben értendő

Többváltozós regresszió publikálása

Kollinearitás vizsgálatakor csak a prediktorok közötti korrelációkat figyeljük, és fölösleges betenni a teljes korrelációs táblát, elegendő csak a legkisebb és legnagyobb korrelációs együtthatót megnevezni, vagy azt írni például, hogy a prediktorok közötti korrelációs együtthatók gyengék / közepesek.

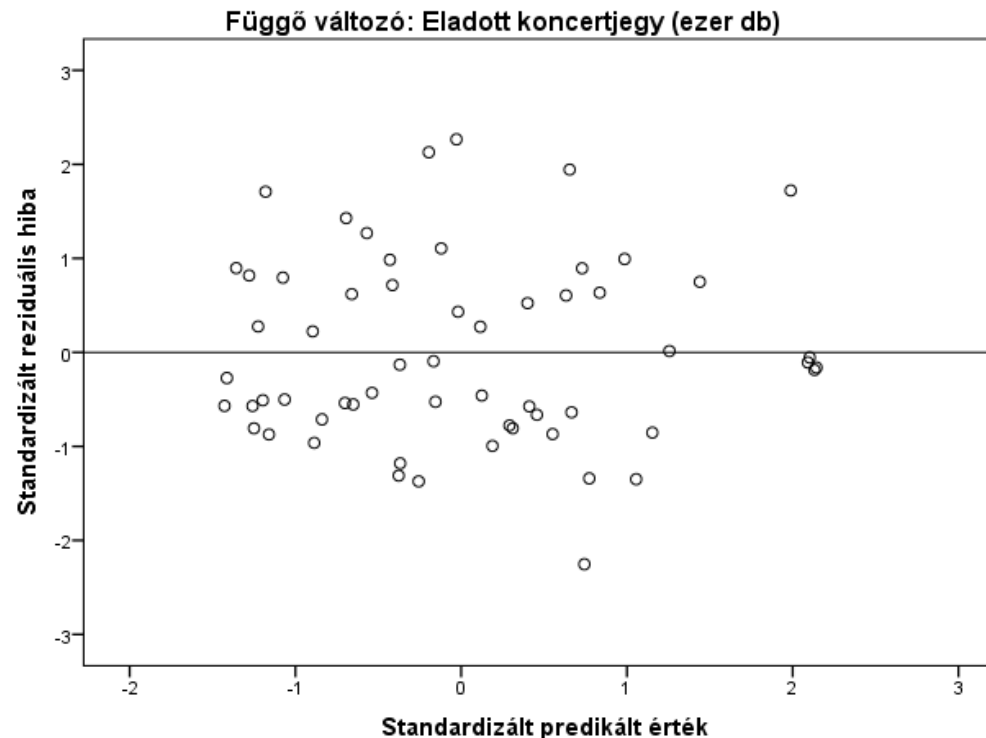
Multikollinearitáshoz használt tolerancia és VIF értékek közül elég az egyiket publikálni. Toleranciát százalékban adjuk meg, a VIF-et általában $VIF = [\text{számérték}]$ formában. Ha valamely prediktor toleranciája nagyon alacsony lenne, már itt fel lehet hívni az olvasó figyelmét, hogy azt a prediktort majd csak óvatosan szabad értelmezni.

A kollinearitás feltételének ellenőrzésére alkalmas tesztek alapján a változók között sem kollinearitás, sem multikollinearitás nem áll fenn. A prediktor változók közötti korreláció $r = .393$, és a tolerancia értékek minden prediktorra magasak: REKLÁM 84.5%, YOUTUBE 84.5%. A prediktorok nagymértékű függetlensége lehetővé teszi, hogy hatásukat tisztán értelmezzük a regresszióban.

Többváltozós regresszió publikálása

A szóráshomogenitás (aka homoszkedaszticitás) és linearitás ellenőrzéséhez használt pontdiagrammot nem feltétlen kell a dolgozatba betenni, elég a mellékletbe. Szövegesen azonban írd le, hogy teljesül a feltétel!

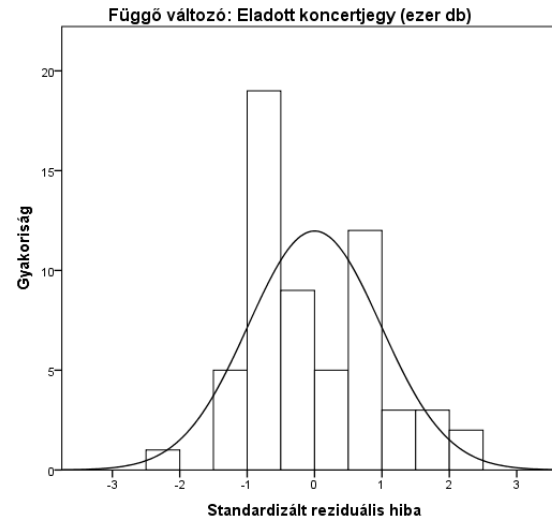
A standardizált predikált értékeket és standardizált reziduális hibákat megjelenítő pontdiagramm alapján megállapítható, hogy a mintán teljesül a homoszkedaszticitás és linearitás feltétele.



Többváltozós regresszió publikálása

A hibák normalitását elég hisztogramon ellenőrizni, és a hisztogramot elég a függelékbe betenni.

A hisztogram alapján a standardizált reziduális hibák normalitásának feltétele teljesül.



Ha viszont bizonytalan voltál a hisztogramban, és kimentetted a standardizált reziduálisokat, és elvégezted rajtuk a S-W tesztet, azt az annak megfelelő módon kell publikálni.

A standardizált reziduális hibák normalitás-feltételének teszteléséhez Ghasemi és Zahedi-Asl (2012) útmutatását követve az ismertebb Lilliefors korrigált Kolmogorov-Smirnov teszt helyett Shapiro-Wilk normalitás tesztet használtam. A teszt eredményei alapján a hiba eloszlásgörbéje nem tér el szignifikánsan a normál eloszlástól, $W(59) = .968$ $p = .128$.

Többváltozós regresszió publikálása

Hibák korrelátlanságáról a Durbin-Watson értéket kell betenni.

Az adatokon teljesül a reziduális hibák függetlenségének feltétele, a Durbin-Watson érték = 1.891 (Durbin és Watson Durbin, 1951).

Többváltozós regresszió publikálása

Írd le, hogyan épült fel a modelled. Ez segít értelmezni az olvasónak, hogy a különböző modellek mit takarnak. A *hierarchikus* szó azt jelenti, hogy több blokkod is volt. Ha csak egy blokkod volt, és ebbe az egyetlen blokkba tetted be ENTER-rel az összes változód, akkor *többszörös* regressziót írn!

Hierarchikus regresszió első szintjén a REKLÁM változót léptettük be a modellbe (ide írhatod, hogy ENTER módszerrel, de fölösleges, mert egyetlen változó van csak). A második szinten a YOUTUBE került az elemzésbe.

Többváltozós regresszió publikálása

Az eredményeket inntől lehet szövegesen vagy táblázatosan is közölni, de a kettő közül általában csak az egyiket használjuk. Ha egyszerű a modell, akkor érdemes szövegesen írni ki az eredményeket, de egy összetettebb modell esetében átláthatóbb lehet a táblázatos leírás.

Példaként először leírom szövegesen, majd megmutatom azt is, hogy néz ki táblázatosan.

Az első modellben a REKLÁM a KONCERTJEGY varianciájának 60,8%-át magyarázza ($r^2_{Adj} = .601$), a modell szignifikáns, $F(1,57) = 88.447$ $p < .001$. A REKLÁM hatása szignifikáns ($\beta = .780$ $t(59) = 9.405$ $p < .001$), egy millió Forinttal több befektetés az eladási mennyiségben 502db növekedést jelent ($B = 0.502$ $SE_B 0.053$).

A második modellben a REKLÁM és YOUTUBE prediktorok a KONCERTJEGY varianciájának 74,2%-át magyarázzák ($r^2_{Adj} = .732$). Mind a modell, mind a megmagyarázott variancia előző modellhez viszonyított növekedése is szignifikáns, $F(2,56) = 80.357$ $p < .001$ és $F(1,56) = 28.929$ $p < .001$. Mindkét prediktor változó szignifikánsan bejósolja az eladott koncertjegyek számát, és mindkettőnek pozitív hatása van arra. REKLÁM: $\beta = .623$ $t(59) = 8.439$ $p < .001$; YOUTUBE: $\beta = .397$ $t(59) = 5.379$ $p < .001$.

Ez volt a szöveges publikálás. Egy ilyen egyszerű modellnél még lehet a szöveges verziót választani, de ennél bonyolultabbnál már előnyösebb a táblázat, mert sokkal átláthatóbb. Ekkor a táblázatban megjelenő értékeket nem kell külön szövegesen is leírni, minden mást, az értelmezést viszont igen!

Többváltozós regresszió publikálása

Publikálás táblázatos formában:

Nyers reg.
együtthetők

B0-hoz és B1-
hez tartozó SE

Standardizált reg.
együtthetők

A t-próba df értéke az SS_T df
értékével azonos

	B	SE B	β	t
Modell 1				
Konstans	4.983	0.565		$t(58) = 8.817$ $p < .001$
REKLÁM	0.502	0.053	.780	$t(58) = 9.405$ $p < .001$
Modell 2				
Konstans	3.290	0.560		$t(58) = 5.876$ $p < .001$
REKLÁM	0.402	0.048	.623	$t(58) = 8.439$ $p < .001$
YOUTUBE	0.003	0.0005	.397	$t(58) = 5.379$ $p < .001$

Modell1: $R^2 = .608$ $R^2_{Adj} = .601$ $F(1,57) = 88.447$ $p < .001$.

Modell2: $R^2 = .742$ $R^2_{Adj} = .732$ $F(2,56) = 80.357$ $p < .001$. Változás: $F(1,56) = 28.929$ $p < .001$

A REKLÁM a KONCERTJEGY varianciáját szignifikánsan, 60.8%-ban magyarázza. A REKLÁM és a YOUTUBE közösen 74.2%-ot magyaráz, mely szintén szignifikáns, és az előző modellhez képest szignifikáns növekedés a magyarázóerőben. E második modellben mind a REKLÁM, mind a YOUTUBE szignifikáns, pozitív hatással van a KONCERTJEGY értékére.

Többváltozós regresszió publikálása

A hatások összehasonlításához először ki kellett számolnunk a béták konfidencia intervallumait. Ha a Bootstrappinget nem ezer mintán végezted, akkor ne felejtse el itt is átírni.

A konfidencia intervallumok olvasásának mikéntje nem annyira ismert, mint a szignifikancia értéké, ezért érdemes röviden leírni.

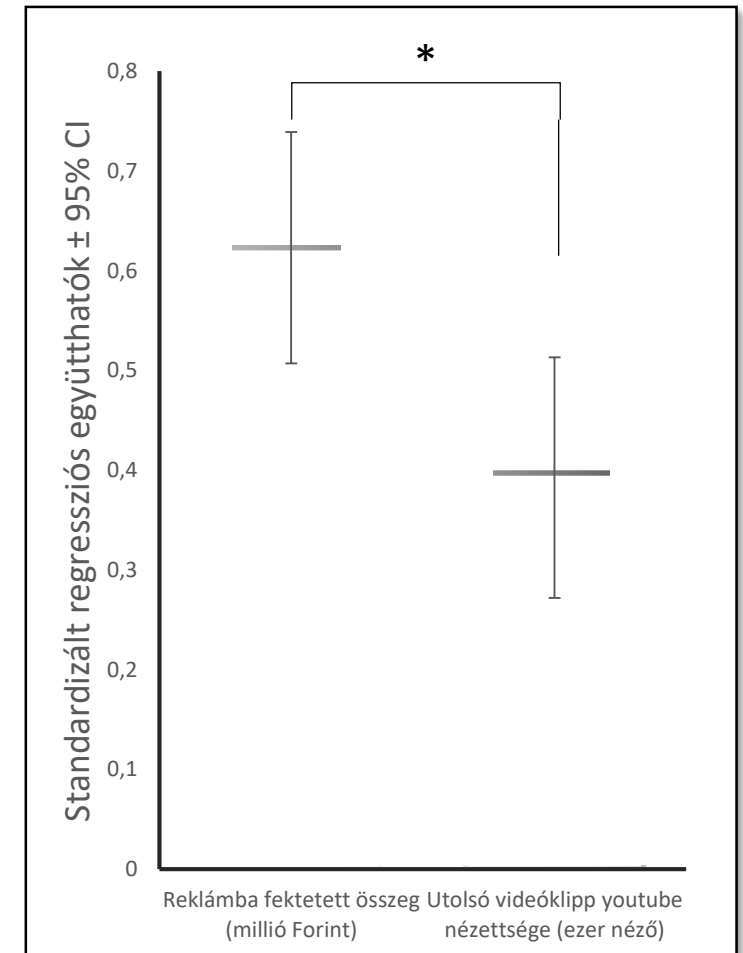
Annak tesztelésére, hogy a két prediktor változóhoz tartozó standardizált regressziós együttható szignifikánsan különbözik-e egymástól, a hozzájuk tartozó 95%-os konfidencia intervallumok kerültek kiszámításra bias korrigált Bootstrapping módszerrel 1000 mintavételezéssel. Abban az esetben ha két konfidencia intervallum maximum 50%-ban fed át egymással, a két regressziós együttható szignifikánsan különbözőnek tekinthető $p < .05$ szinten (Cumming 2005).

A következő dián lévő grafikont a dolgozatba mindenképp érdemes betenni, ha a hatásnagyságokat összehasonlítod. Én Excellel szoktam őket megrajzolni, a béta értéket teszem fel a diagramra, a béta és alsó határából kiszámolom a hiba nagyságát, és egyéni hibásávként teszem ezeket fel. Választhatsz bármilyen más rajzolási módszert is!

Többváltozós regresszió publikálása

A REKLÁM-hoz ($\beta = .623$) és YOUTUBE-hoz ($\beta = .397$) tartozó standardizált regressziós együtthatók 95%-os konfidencia intervallumai közötti átfedés szemmel láthatóan kevesebb, mint 50%, tehát feltételezhető, hogy a REKLÁM hatása szignifikánsan erősebb, mint a YOUTUBE-é.

Ha szerencsés vagy, és ránézésre meg tudod állapítani az átfedés mértékét, elég ennyi. Ha olyan eredményt közölsz, ahol nem csak ránéztél, hanem kiszámoltad az értékeket, ott a következőt tudod írni: A REKLÁM és YOUTUBE változó standardizált regressziós együtthatójához tartozó konfidencia intervallumok közötti átfedés az átlagos intervallumnagyság 5.17%-át teszi ki. Ez alapján kijelenthető, hogy a két regressziós együttható $p < .05$ szinten szignifikánsan különbözik egymástól.



Többváltozós regresszió publikálása

- Hivatkozott cikkek:
 - Tukey, J. W. (1977). Exploratory data analysis.
 - Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. Journal of the American Statistical Association, 82(400), 1147-1149.
 - Barnett, V., & Lewis, T. (1994). Outliers in statistical data (Vol. 3). New York: Wiley.
 - Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression.
 - Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32(1), 17-22.
 - Stevens, J. P. (2012). Applied multivariate statistics for the social sciences. Routledge.
 - Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression. II. Biometrika, 159-177.
 - Ghasemi, A., & Zahedi-Asl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.
 - Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. American Psychologist, 60(2), 170.