(Lineáris) Regresszió

Adatfájlok

- AlbumSales.csv
- GSS2018happiness.sav
- Mothering.csv

A regresszióelemzés céla

- Ha két változó együttjár, az egyiknek az értékéből meg lehet jósolni a másiknak az értékét.

- A modell egy vagy több **prediktor** (magyarázó) változóból jósolja meg az **eredmény** (outcome) értékét
  - Egyváltozós regresszió: egy prediktor
  - Többváltozós regresszió: több prediktor
  - Eredmény is és prediktorok is skála vagy folyamatos ordinális típusúak (Nominális változókhoz logisztikus regresszióelemzés kell)

Egyváltozós regresszió

- A modell alapja:

  Eredmény = (modell) + hiba

- A modell lineáris.

- A regressziós egyenes írja le, amit a **legkisebb négyzetek** módszerrel illesztünk.

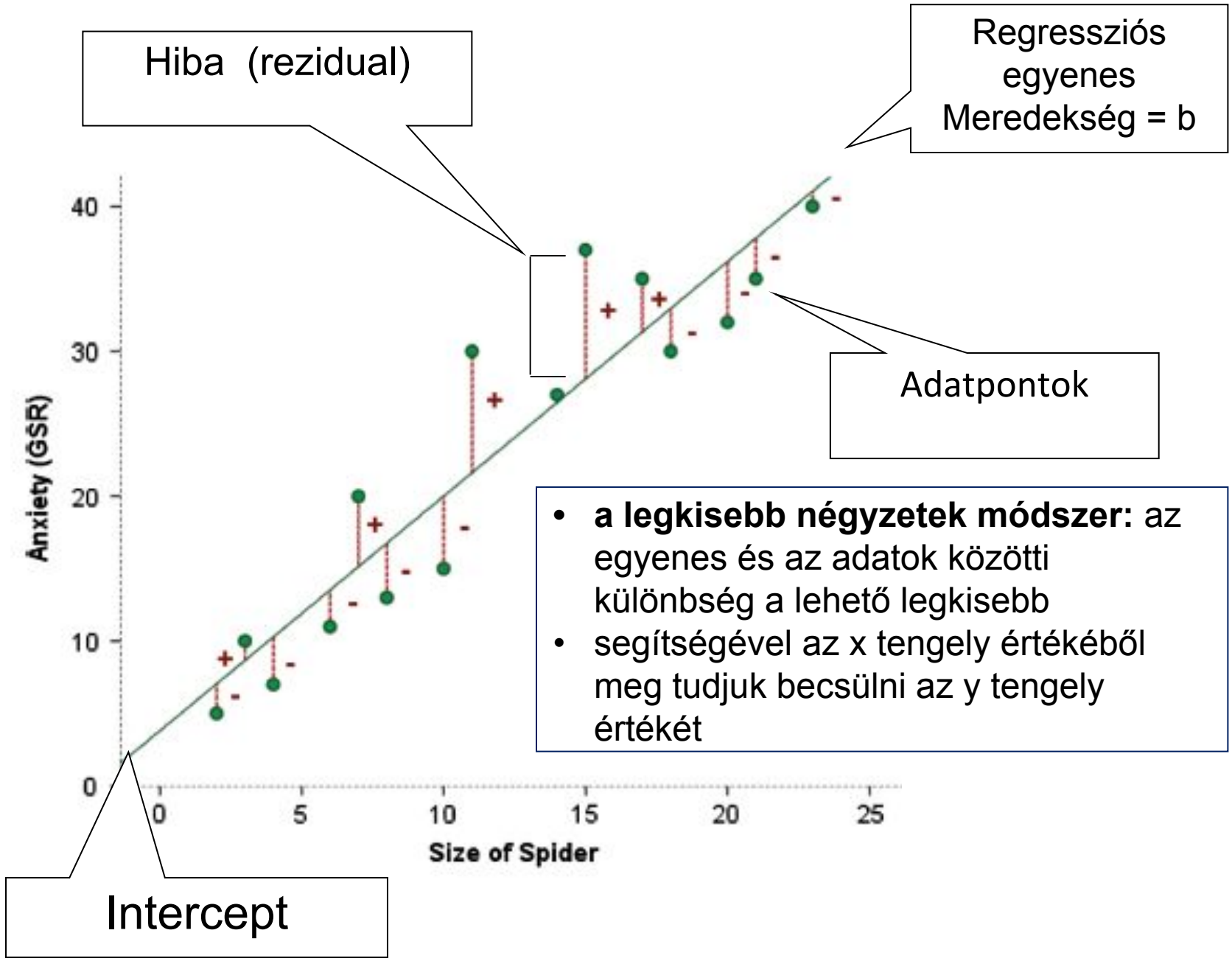
A regressziós egyenes

- Egy egyenes definiálásának elemei:
  - Az egyenes meredeksége (b)
  - A pont ahol az egyenes az y tengelyt átszeli: **intercept** (a)

Eredmény = (modell) + hiba ==>

$$Y_i = (\text{intercept} + bX_i) + \varepsilon_i$$

- az intercept és b a **regressziós együtthatók**
  - b írja le a modell meredekségét
  - az intercept helyezi el a modellt a térben
  - ε_i a hiba: az *i* személy megjósolt és a valóságban megfigyelt értékei közötti különbség.


Hiba (rezidual)
Regressziós egyenes
Meredekség = b
Adatpontok
Intercept
Anxiety (GSR)
Size of Spider
0
10
20
30
40
0
5
10
15
20
25
• **a legkisebb négyzetek módszer:** az egyenes és az adatok közötti különbség a lehető legkisebb
• segítségével az x tengely értékéből meg tudjuk becsülni az y tengely értékét

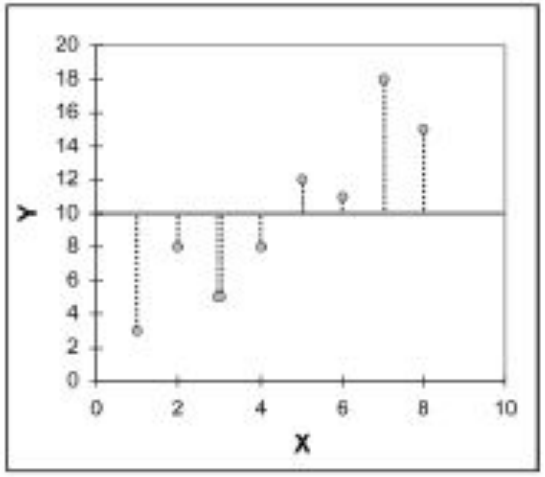
Milyen pontos a modell?

- **goodness of fit:** mennyivel jósolja meg jobban a modell az y értékét, mintha egyszerűen csak az átlagot használnánk = mennyivel csökkenti a modell a teljes varianciát.

- Üzleti példa: mivel lehet növelni az eladások számát? (AlbumSales.csv):
  - Mennyivel több lemezt ad el a cég, ha 100,000 fonttal növeli a reklámra költött pénzt?
  - Ha nincs információnk a reklámköltés és a lemezeladások kapcsolatáról, a legjobb becslésünk az eladások átlagos száma.

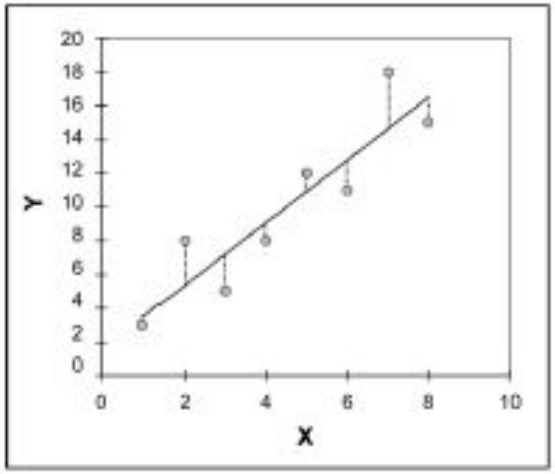
Goodness of Fit

SS_T



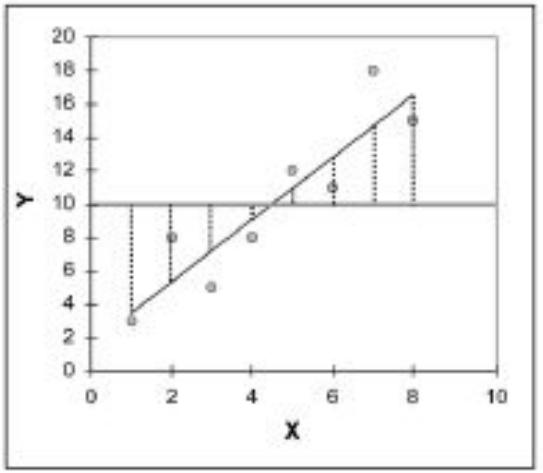
SS_T = az átlagtól való összes négyzetre emelt eltérés. Összes variancia. (SS: Sum of Squares, T: total)

SS_R



SS_R = a regressziós egyenestől való összes négyzetre emelt eltérés (R: residual)

SS_M



SS_M = mennyivel csökkent az eltérés a regressziós egyenesnek köszönhetően (M: model)

Goodness of Fit

Ha SS_M sok, akkor sokat javított a becslésen a regressziós modell az átlaghoz képest. Mit nevezünk “sok”-nak? Két mérőszám: (1) R^2 és (2) az *F teszt*..

$$R^2 = \frac{SS_M}{SS_T}$$

$$F = \frac{SS_M/df_M}{SS_R/df_R} = \frac{MS_M}{MS_R}$$

A teljes varianciának mekkora részét magyarázza a modell

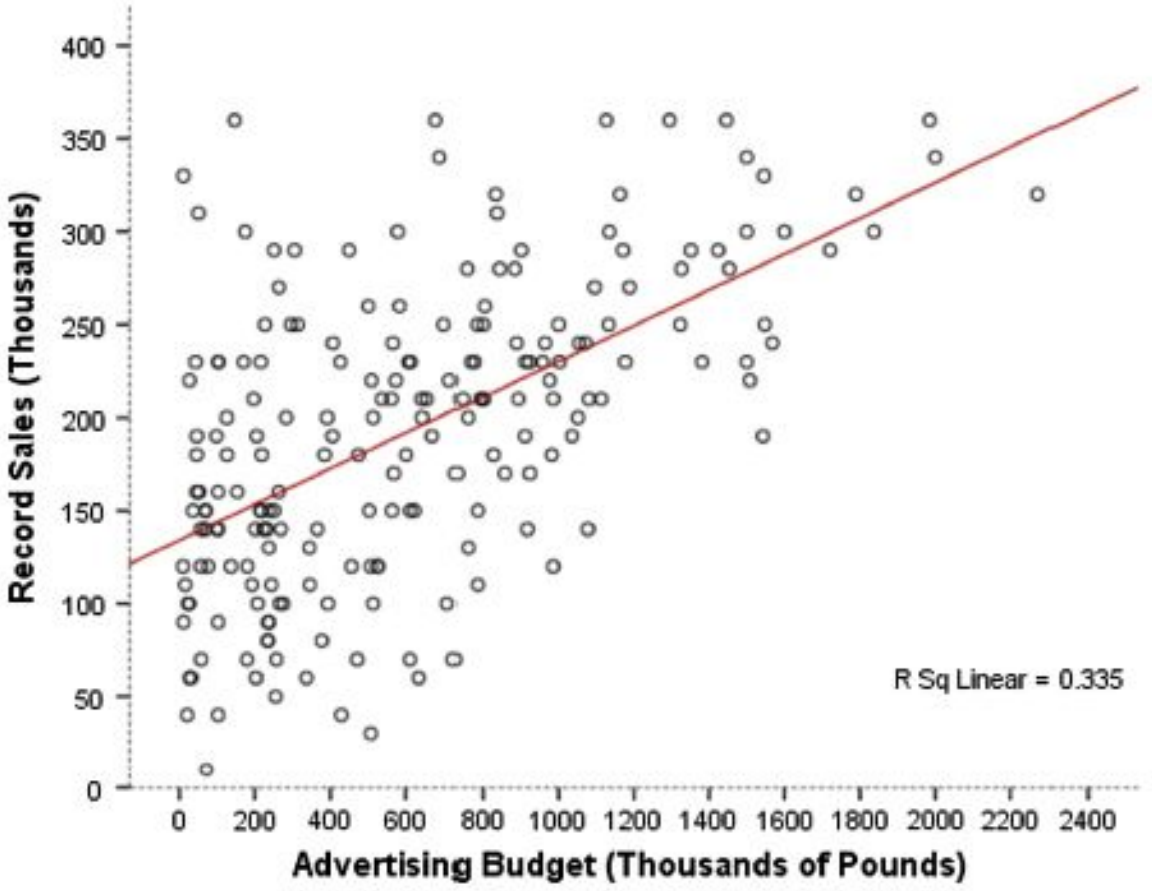
df_M = prediktorok száma

df_R = egyének száma - prediktorok száma - 1

(MS: Mean Squares)

Eladások száma (Y) és Reklámkiadás (X)

AlbumSales.sav



JASP

Regression > Linear Regression

Dependent variable: Eredményváltozó

Covariates: Prediktorok

Statistics:

Estimates: együtthatók

Model fit: F érték (ANOVA)

A reklámköltés az eladások varianciájának a 33%-át magyarázza (SS_M/SS_T)

Regressziós modell standard hibája: MS_R négyzetgyöke

Model Summary ▼

Model	R	R^2	Adjusted R^2	RMSE
1	0.58	0.33	0.33	65.99

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	433687.83	1	433687.83	99.59	< .001
	Residual	862264.17	198	4354.87		
	Total	1.30e +6	199			

SS_M

SS_R

SS_T

p: annak az esélye, hogy a kapott F érték a véletlen műve.
Szignifikáns p érték: a modell szignifikánsan jobban jósolja be az eladásokat, mint az átlag.

Regressziós együtthatók

Itt szeli át a regressziós egyenes az y tengelyt: ha 0 fontot költünk reklámra, 134 (ezer) lemezt adunk el

az együtthatók szignifikánsan különböznek a 0-tól

a b esetében érdekes: a reklámköltés szignifikánsan hat az eladásra

Coefficients ▼

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	134.14	7.54		17.80	< .001
	Adverts	0.10	9.63e −3	0.58	9.98	< .001

b: az egyenes meredeksége

Minden újabb egység (1000 font) reklámköltés után 0,1 egységgel (100-zal) több lemezt adunk el.

Beta: az egyenes meredeksége standardizálva

Minden újabb szórásnyi reklámköltés után 0,58 szórással több lemezt adunk el.

A modell használata a gyakorlatban

Mivel a modell szignifikáns (megbízható), megjósolhatjuk belőle az eladások számát a reklámköltés ismeretében:

eladás = intercept + b x reklámköltés

= 134,14 + (0,1 x reklámköltés)

Hány lemezt fogunk eladni, ha 500 ezer fontot költünk reklámra?

Többváltozós lineáris regresszió

Többváltozós regresszió

- AlbumSales.csv - további magyarázó változók:
  - reklámköltés
  - hány órát játszák a számokat a rádióban
  - milyen vonzó a zenész (1-től 10-ig terjedő skálán)

A többváltozós modell

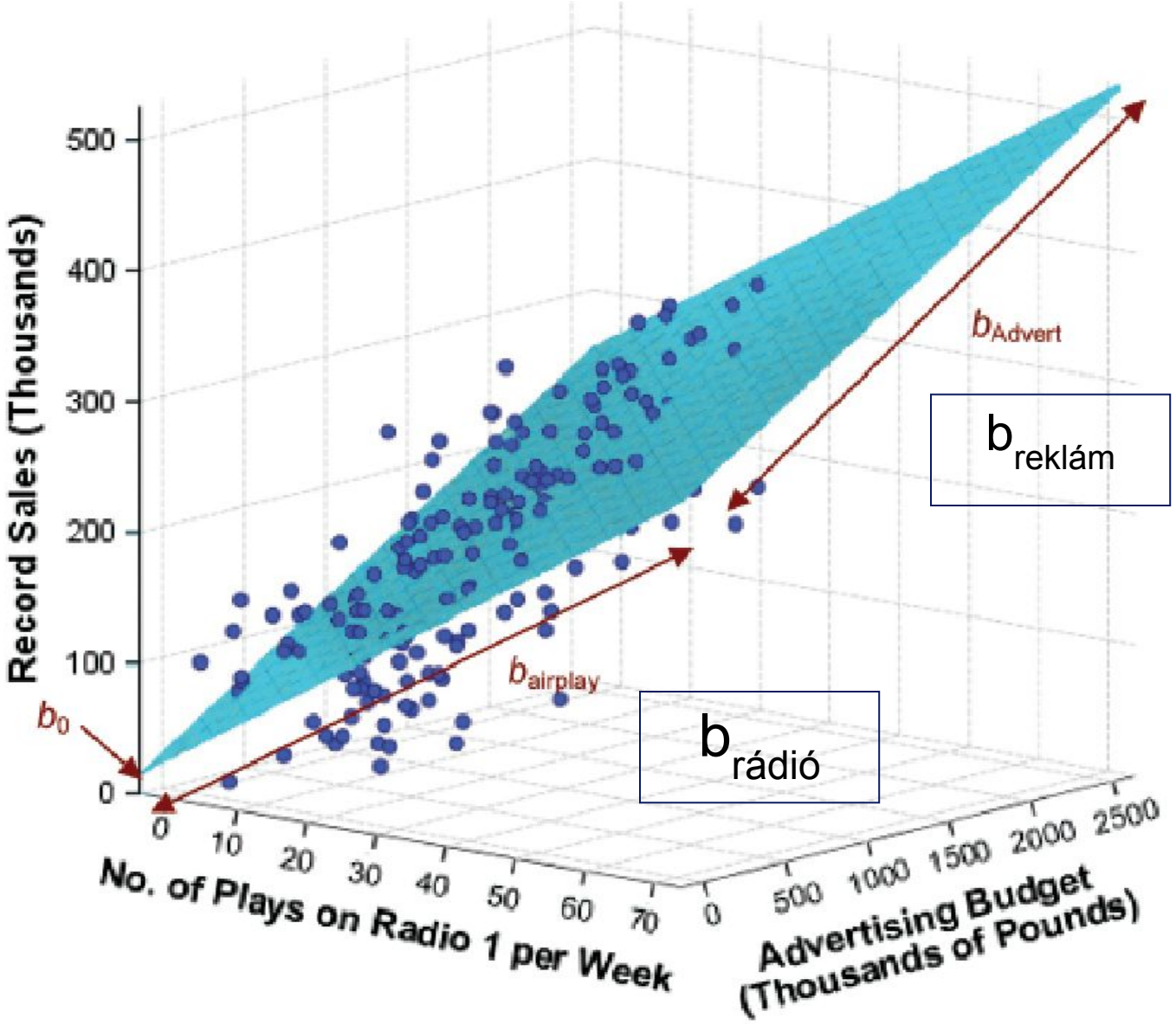
- Egy eredmény változó (skála vagy ordinális)
- Két vagy több prediktor (skála vagy ordinális)
- A modell:

  eredmény = (modell) + hiba

  $Y = (\text{intercept} + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon$

  $\text{eladás} = \text{intercept} + b_1\textit{reklám}_i + b_2\textit{rádió}_i + b_3\textit{szépség}_i + \varepsilon$

Eladás, reklám és rádió



A variancia felosztása

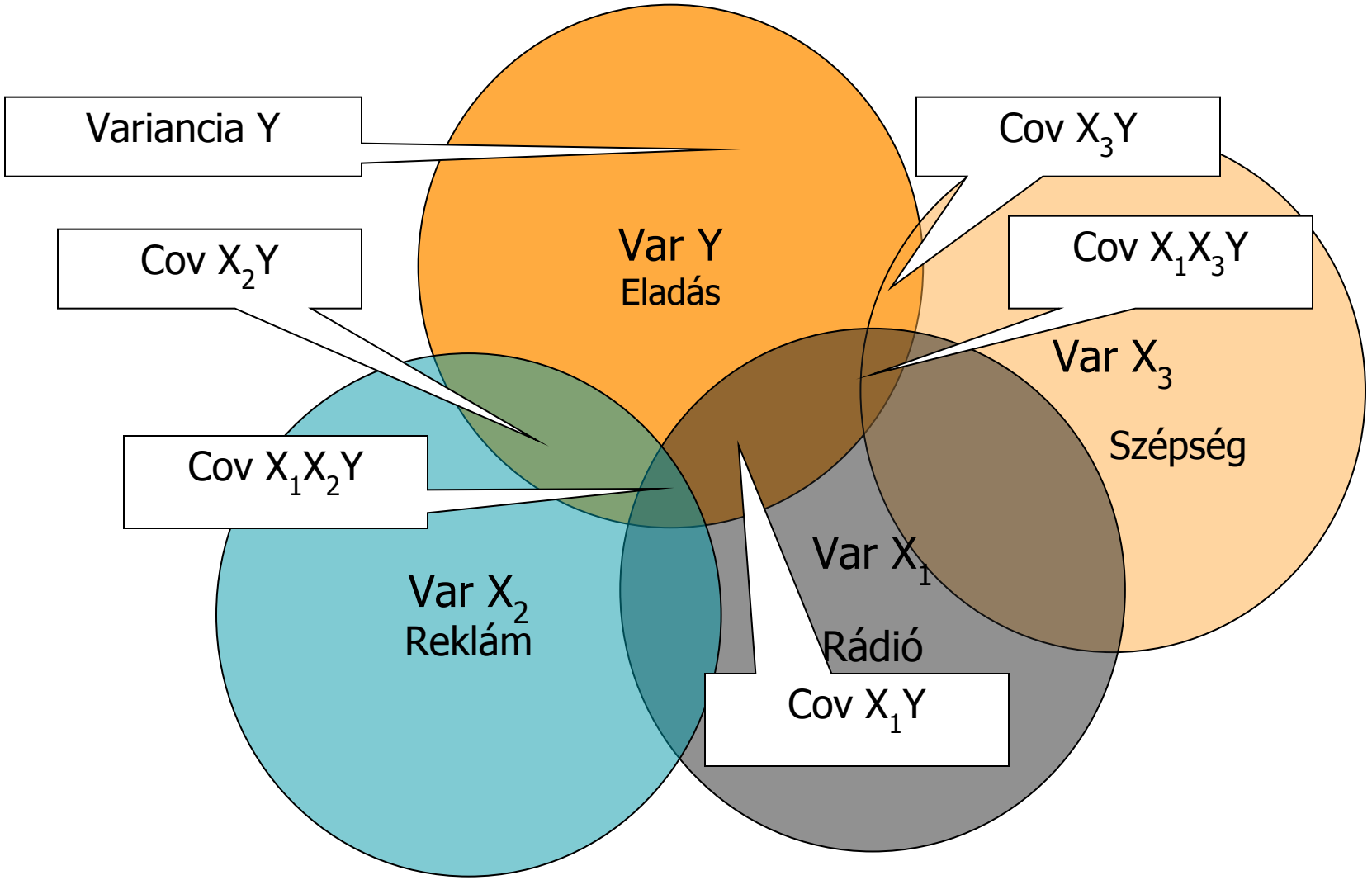
Az eredményváltozó varianciája

- a prediktorok hatásából
- + a megmaradt hibából

tevődik össze.

A probléma: a prediktorok hatásai nem feltétlenül különülnek el egymástól (egymással is korrelálnak): **Kovariancia**

Kovariancia



Parciális és szemiparciális korreláció

- Parciális: Egy prediktor saját hatása az eredményváltozóra **minden más prediktor hatásának kiszűrésével.**
- Szemiparciális: Egy prediktor saját hatása **a többi prediktor hatásán felül.**
  - Többváltozós regresszióban: amikor új prediktort adunk a modellhez, az előző prediktoron hatásán felüli hatást látjuk.

Modellépítés

- Prediktorok kiválasztása
  - Hipotézis szerint (nem összevissza!)
  - Ha túl sok a prediktor, értelmezhetetlen lesz a modell.
  - Ha a prediktorok erősen korrelálnak egymással, használhatatlan lesz a modell.
- A modell felépítése
  - Jobb kisebb modellel kezdeni
  - Fokozatosan bővíteni

Modellépítés módszerei

- Enter
  - Minden prediktor egyszerre kerül a modellbe.
- Hierarchikus Enter
  - Több modell egymás után, egyre több prediktorral.
  - Először a hipotézis szerint a legerősebb hatású prediktor, azután a várhatóan kevésbé erős hatásúak

Forward módszer (lényegében a hierarchikus Entry módszer automatikus változata)

1. A szoftver kiválasztja azt a prediktort, ami a legerősebben korrelál a outcome-mal.
2. További prediktorokat aszerint választja, hogy milyen erős a szemiparciális korrálációjuk az outcome-mal. Erősebb előbb.
3. Addig folytatja, amíg nem marad olyan prediktor, ami szignifikánsan korrelál az outcome-mal.

Backward módszer

1. A szoftver az összes prediktort beteszi a modellbe, és kiszámolja a hatásukat.
2. Kiveszi azokat a prediktorokat, amiknek a hatása egy adott szint alatt van.
3. Addig folytatja, amíg a valamennyi modellben maradt prediktornak szignifikáns hatása van.

Stepwise módszer

A Forward és Backward kombinációja:

1. A szoftver egyenként adja a modellhez a prediktorokat.
2. Minden lépés után leteszteli, hogy a modell valamennyi prediktora szignifikáns-e.
3. Ha valamelyik prediktor már nem szignifikáns, kiveszi a modellből.

Feltételek

- **Kolinearitás:** A prediktorok ne korreláljanak egymással túl erősen
  - Pontdiagramokkal, korrelációval ellenőrizhető
  - A VIF (variance inflation factor) értékkel is tesztelhető: Ne nagyon legyen 8 fölött (Statistics > Collinearity diagnostics)
- **A residuumok normál eloszlást mutassanak**
  - JASP: Plots > Residuals vs. histogram
  - Vagy Residuals > Casewise diagnostics: Kiadja a megadott z értéknél távolabb lévő reziduumokat
  - Kiugró értékeket törölni vagy együtthatók bootstrapping módszerrel (Regression Coefficients > Estimates > From [1000] bootstraps)

- **Homoszkedaszticitás:** a reziduumok ne kövessen mintázatot
  - Plots > Residuals vs. predicted
- **Független hibák**: A reziduumok ne korreláljanak
  - Residuals > **Durbin-Watson**, értéke 0 és 4 között. 2: nincs korreláció, 4 erős negatív korreláció, 0 erős pozitív korreláció

0. modell: reklám
1. modell: reklám, rádió, szépség

Model Summary ▼

Model	R	R^2	Adjusted R^2	RMSE	Durbin-Watson Autocorrelation	Durbin-Watson Statistic	Durbin-Watson p
0	0.58	0.33	0.33	65.99	−0.04	2.03	0.82
1	0.82	0.66	0.66	47.09	2.70e −3	1.95	0.72

Note. Null model includes Adverts

A 3 prediktor együtt az eladások varianciájának 66 %-át magyarázza

2 körül van, a hibák függetlenek

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	433687.83	1	433687.83	99.59	< .001
	Residual	862264.17	198	4354.87		
	Total	1.30e +6	199			
1	Regression	861377.42	3	287125.81	129.50	< .001
	Residual	434574.58	196	2217.22		
	Total	1.30e +6	199			

Note. Null model includes Adverts

Mindkét modell szignifikánsan jobban jósolja meg az eladást, mint az átlag

Ha 1000 fonttal többet költünk reklámra, 100-zal több lemezt adunk el. Vagy mégsem?

Az együttható szórásegységben: 1 szórással több reklám vagy rádió, fél szórással több eladás.Ha egy szórással szebbek a zenészek, 0,19 szórással nő az eladások száma.

Messze van a 8-tól: nem korrelálnak egymással a prediktorok.

Coefficients ▼

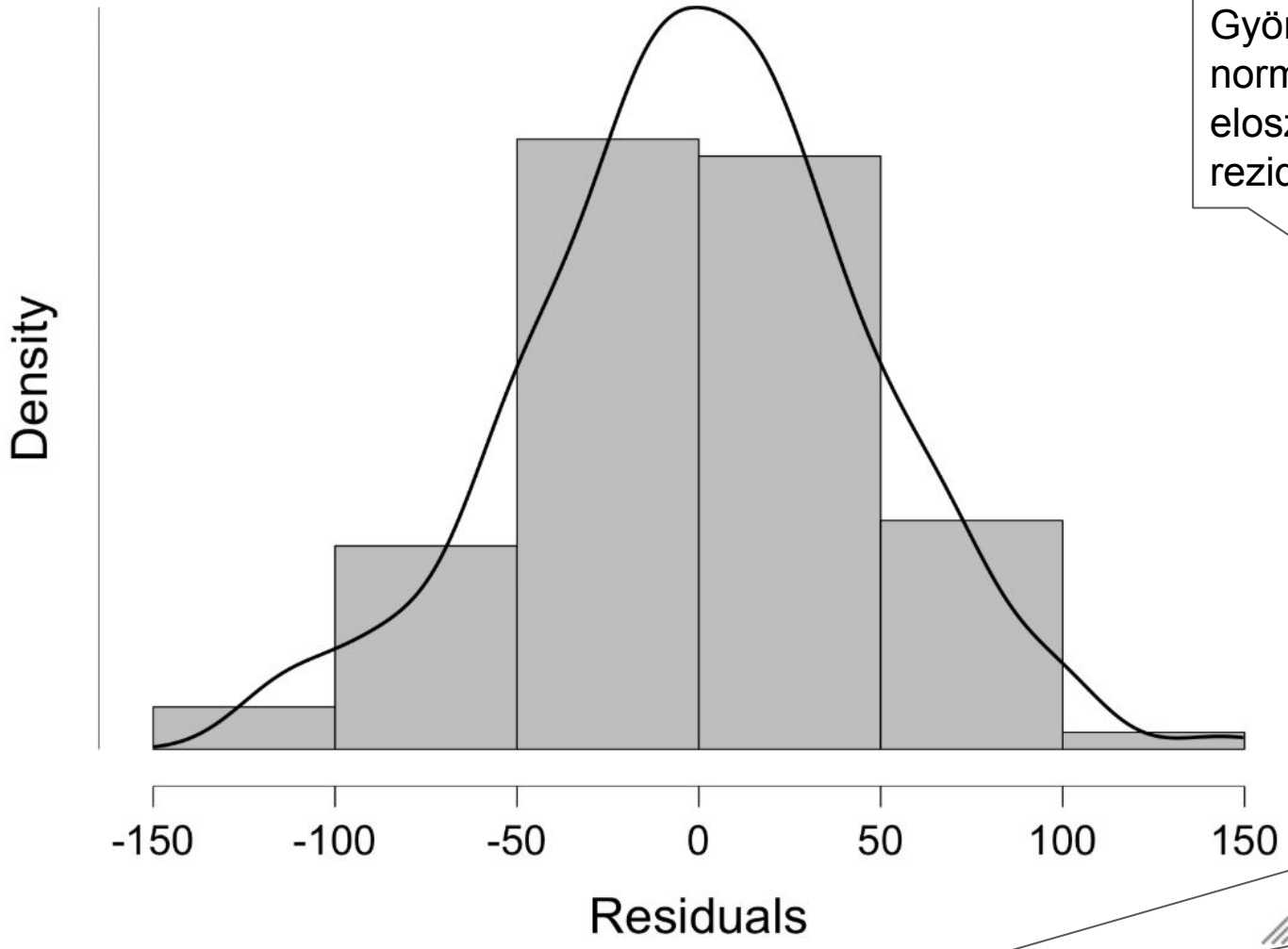
Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
0	(Intercept)	134.14	7.54		17.80	< .001		
	Adverts	0.10	9.63e −3	0.58	9.98	< .001	1.00	1.00
1	(Intercept)	−26.61	17.35		−1.53	0.13		
	Adverts	0.08	6.92e −3	0.51	12.26	< .001	0.99	1.01
	Airplay	3.37	0.28	0.51	12.12	< .001	0.96	1.04
	Attract	11.09	2.44	0.19	4.55	< .001	0.96	1.04

Ami a reklám hatásának tűnt, annak egy része valójában a másik két prediktor hatása. **Azokon felül**, 1000 fontnyi reklámtól csak 80-nal növekszik az eladások száma.

Ha 1 órával többet játszák a számokat a rádióban, 3370-nel nő az eladások száma.

Mindegyik prediktor hatása szignifikáns.

Residuals Histogram ▼

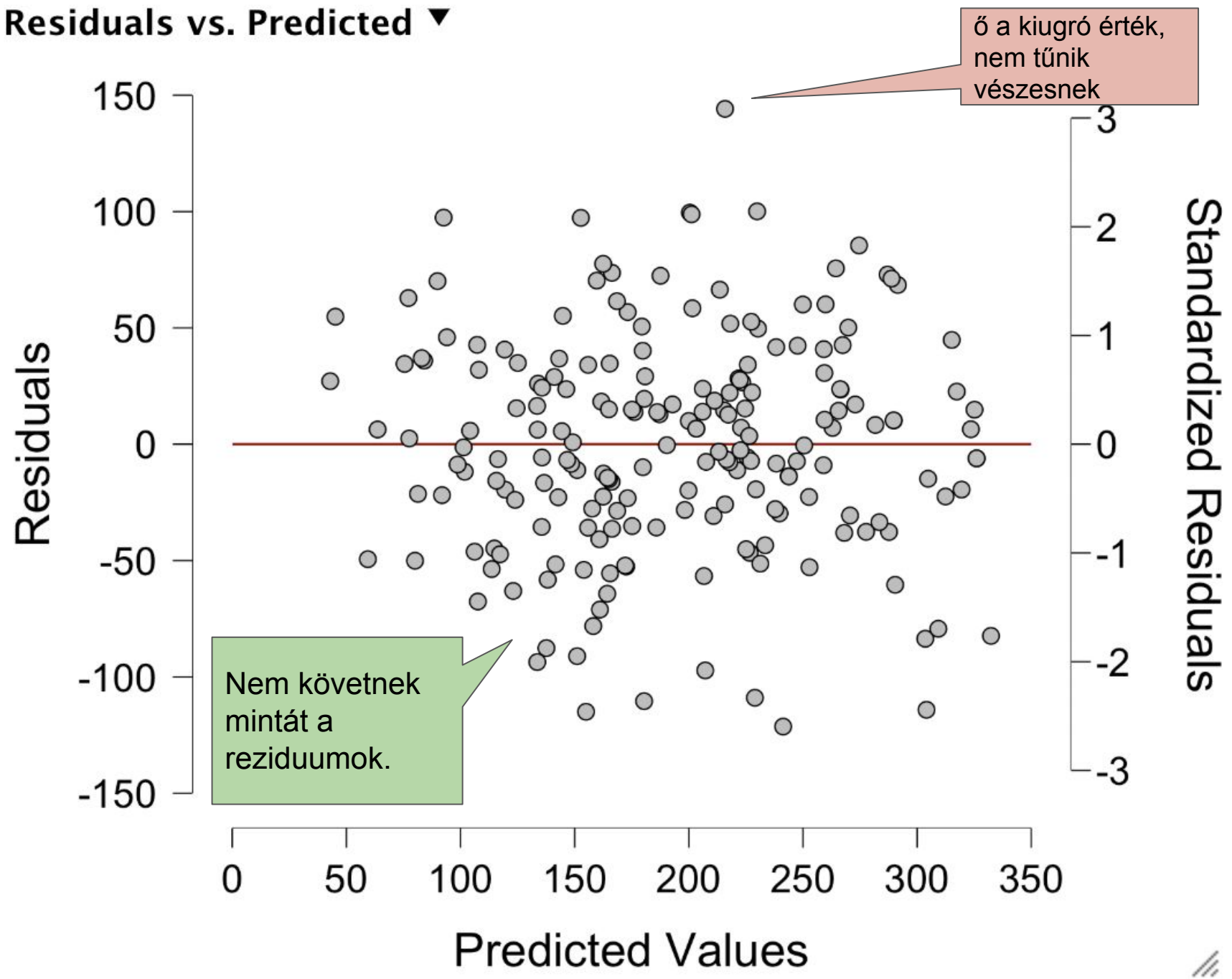


Gyönyörű normál eloszlású reziduumok.

Bár van egy kicsit kiugró értékű reziduum: 3,09 szórásra az átlagos reziduumtól.

Casewise Diagnostics

Case Number	Std. Residual	Sales	Predicted Value	Residual	Cook's Distance
169	3.09	360.00	215.87	144.13	0.05


Residuals vs. Predicted
ő a kiugró érték, nem tűnik vészesnek
Nem követnek mintát a reziduumok.
Residuals
Standardized Residuals
Predicted Values
150
100
50
0
-50
-100
-150
3
2
1
0
-1
-2
-3
0
50
100
150
200
250
300
350

Writing up the results

Table + description:

	b	Beta	t	p
Step 1				
Intercept				
Advert				
Step 2				
Intercept				
Advert				
Radio				
Attract				

A linear regression model was built to predict record sales from advertising budget, hours of radio play of songs and attractiveness of the band. Advertising budget was entered first followed by the remaining two predictors. The first model with just advertising budget entered gave a significant result ($R^2 = .335$, $F(1, 198) = 99.59$, $p < .001$). The second model with all three predictors was also significant ($R^2 = .665$, $F(3, 196) = 129.5$, $p < .001$) with each of the predictors having a significant contribution. The contributions of advertising budget and radio play were the highest (about .5 SD increase in record sales), while the attractiveness of the band had a smaller effect (.19 SD increase in record sales). The three predictors together explained about 66% of the variance in record sales.

Példa: Mothering.csv

A kutatás azt vizsgálja, hogy a gyermekkori kapcsolat az anyával (Maternal Care) milyen hatással van arra, hogy egy nő mennyire bízik a saját anyai képességeiben (Confidence). A kutatók a nők általános önértékelését is vizsgálták (SelfEsteem) (Leerkes and Crockenberg 1999).

Mindegyik változónál: magasabb érték jobb helyzetet jelöl.