Empirikus elemzés 1.

Elmélet és adatok:

- megfigyelés vagy tapasztalatok -> kép, hogy hogyan működik a világ egy szelete -> átgondoljuk + feltevések, egyszerűsítések -> következtetések (pl. ha X és Y fennáll, akkor Z történik -> ha nő az egyének fizetése és nem változnak az árak, akkor nő a fogyasztásuk) => közgazdasági elmélet
- a jó elmélet egyik ismérve, hogy adatokon ellenőrizhető, azaz tesztelhető előrejelzéseket (testable predictions) ad
- kidolgozott elmélet -> adatgyűjtés (hiányzó adatokat részben pótló, kifejező adatok)
- adatokon teszteljük az elmélet helyességét (/versengő elméletek közül melyik)
- az elmélet és az adatelemzés adja a teljes közgazdasági elemzést (alternatív elméletek)
- a világ sokkal bonyolultabb annál, hogy tökéletesen meg tudjuk magyarázni, hibák, hiányok

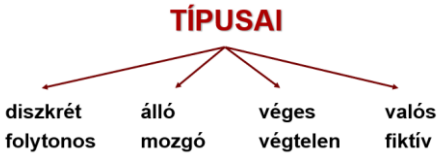
Alapfogalmak:

Statisztika: adatok elemzésével foglalkozó tudomány

- A valóság tényeit tömören, számszerűen jellemző, modellezni törekvő gyakorlati (adatgyűjtő, elemző) tevékenység és tudományos módszertan
- Tömegesen előforduló jelenségek tömör, számszerű jellemzése adatokkal és mutatószámokkal

Ökonometria: közgazdasági modellek alapján, adatok elemzésével foglalkozó tudomány

Sokaság:



- A vizsgálat tárgyát képező egységek összessége, halmaza
- Típusai: (4x2)
  - Diszkrét vs. folytonos:
    - *Diszkrét:* jól elkülönülő egységekből áll
    - *Folytonos:* csak önkényesen elkülöníthető egységekből áll
  - Álló vs. mozgó:
    - *Álló:* időpontra vonatkozó (állományjellegű, stock)
    - *Mozgó:* időtartamra vonatkozó (folyamatjellegű, flow)
  - Véges vs. végtelen: számosságra utal
  - Valós vs. fiktív:
    - *Valós:* létező, ténylegesen előforduló egységekből áll
    - *Fiktív:* elképzelt egységekből áll

Ismérv

- Olyan vizsgálati szempont, amely alapján a sokaság részekre bontható
- A valamely adott szempont szerint lehetséges tulajdonságokat *ismérvváltozat*oknak nevezzük
- Az ismérvváltozatok által adott információ természete alapján négyféle ismérvfajtát különböztetünk meg:
  - <u>Területi:</u> Az egységek térbeli (földrajzi) elhelyezkedésére utal
  - <u>Időbeli:</u> Az egységek időbeli elhelyezkedésére utal
  - <u>Mennyiségi:</u> A sokaság egységeihez valamilyen mérés vagy számlálás eredményét rendeli hozzá
    - Kvantitatív → szám (pl. infláció, jövedelem, munkanélküliség)
  - <u>Minőségi</u>: A sokaság egységeit verbálisan jellemzik
    - Kvalitatív → nem szám (– számokká alakítjuk)
    - Férfi/nő - kétértékű változó (dummy/binary): 0-1
    - Iskolai végzettség: alapfok – 1, középfok – 2, felsőfok – 3 -> pl. a középfok 2 száma nem azt jelenti, hogy az az alapfok kétszerese

Mérési skálák/szintek:

- Mérés: sokasági egységek tulajdonságainak szám formájában történő rögzítése.
- Névleges (nominális): az egységekhez rendelt számérték *egyező* vagy *különböző* voltát engedi meg az egységek ténylegesen is jellemző tulajdonságaként elfogadni (pl. név, nem, adószám)
- Sorrendi (ordinális): nem csupán a skálaértékek azonos vagy nem azonos volta, hanem azok *sorrendisége* is vonatkoztatható a vizsgált egységekre (pl. vizsgajegyek, iskolai végzettség, katonai rangok)
- Különbségi (intervallum): A skálaértékek *különbségei* is értelmezhetőek: mennyivel több/nagyobb/stb? A skála szerves tartozéka valamilyen mértékegység (pl. hőmérséklet, naptári idő)
- Arány: Van természetes kezdőpontja. Két skálaérték egymáshoz viszonyított aránya meghatározható és értelmezhető. (pl. életkor, népesség)

Adatfelvétel:

Adatszerzés:

Statisztikai alapműveletek:

1. A sokaság nagyságának meghatározása
   - diszkrét - számlálás
   - folytonos - mérés
2. Összehasonlítás (két vagy több sokaság egészét jellemző adatok felsorolása, különbsége, hányadosa)
3. Osztályozás (ismérv szerinti tagolás, csoportosítás)
   - egy ismérv szerint: csoportosító sor
   - több ismérv szerint: kombinációs tábla

C - ismérvváltozat, f - gyakoriság

X ismérv szerinti osztályok	Y ismérv szerinti osztályok						
	C_1^Y	C_2^Y	⋯	C_j^Y	⋯	C_c^Y	$\sum j$
C_1^X	f_{11}	f_{12}	⋯	f_{1j}	⋯	f_{1c}	$f_{1.}$
C_2^X	f_{21}	f_{22}	⋯	f_{2j}	⋯	f_{2c}	$f_{2.}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
C_i^X	f_{i1}	f_{i2}	⋯	f_{ij}	⋯	f_{ic}	$f_{i.}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
C_r^X	f_{r1}	f_{r2}	⋯	f_{rj}	⋯	f_{rc}	$f_{r.}$
$\sum i$	$f_{.1}$	$f_{.2}$	⋯	$f_{.j}$	⋯	$f_{.c}$	N

Osztály	Egységek száma
C_1	f_1
C_2	f_2
⋮	⋮
C_i	f_i
⋮	⋮
C_k	f_k
Összesen	N

Életkor	f_i
0 - 14	1 421 739
15 - 64	6 461 058
65 -	1 889 959
Összesen	9 772 756

Életkor	Nem		
	Nő	Férfi	Összesen
0 - 14	729 954	691 785	1 421 739
15 - 64	3 228 776	3 232 282	6 461 058
65 -	717 091	1 172 868	1 889 959
Összesen	4 675 821	5 096 935	9 772 756

Empirikus elemzés 2.

Adattípusok

Idősoros adatok – *pl. GDP, kamatláb, pénzkínálat … az idő mentén*

- időben rendezett változók (– változó: a mérés alapjául szolgáló tényező, pl. GDP)
- sorrend számít
- megfigyelési gyakoriság (meghatározott, egyforma időközök): évi, negyedévi, havi, heti, napi
- jelölés: Y_t (t/T az időre utal) -> *Y t-edik időszaki értéke*
- pl.: a Ft/CHF árfolyam, a GDP, a népességszám alakulása
- pénzügyi adatok nagyon gyakoriak (makroökonómia)
- nagyon sok adat (egyre több), folytonos adatok
- a grafikus ábrázolás segít tömören áttekinteni az adatokat
- vonaldiagram, kirajzoljuk, ránézünk!

Keresztmetszeti adatok - *pl. egy iparág dolgozóinak a bére 2001 augusztusában*

- egy adott időpontban a gazdaság (egyéni) szereplőiből (egyén, vállalat, ország stb.) vett minta – pillanatfelvétel az időben minket érdeklő szereplőkről
    - sok dolgot nézünk egy időben
- megfigyelések sorrendje nem számít, nem lényeges
- jelölés: Y_i (i=1; 2; … N) -> i. egyedre vonatkozó adat (N: megfigyelési egységek száma)
- sokszor véletlen minta - általában lehetetlen mindenkit megfigyelni
    - pl. a bérek érdekelnek egy adott szektorban -> véletlenszerűen kiválasztok X személyt az adott szektorból és megnézem a bérüket
- máskor minden egység megfigyelhető
    - pl. országok esetén, hogy diktatúrák-e? milyen az urbanizációs szintjük?
- outlierek (kívüleső adatok, „elronthatják” a mutatót)

Hisztogram (keresztmetszeti adatok ábrázolására)

- hisztogram: diszjunkt halmazokba tartozó elemeket összeszámoló függvény
- pl.: egy főre jutó jövedelem megoszlása (eloszlása)
- egyenlő osztályközök (rekesznagyság) – Excelben célszerű megadni adatoktól függően – hisztogram érzékeny lehet a rekesznagyságra
- gyakoriság egyes osztályközökben
- egycsúcsú, kétpólusú (=kétmóduszú)

Panel adatok: - *pl. GDP/fő alakulása 1950 és 2020 között a világ országaiban*

- keresztmetszeti mintáról megfigyelés több időpontban (idősoros+keresztm. együttesen)
- 2 fajta dimenzió -> plusz elemzési lehetőségek
- jelölés: Y_{it} (i,t/N,T) -> *i. megfigyelés a t. időszakban*
- összesen TxN megfigyelés, ahol N a megfigyelt egységek (pl. egyének, országok) száma
- pl.: GDP európai országokban, egyéni keresetek alakulása, urbanizáció alakulása a nagyvilágban
- ha kevés megfigyelt egység, akkor lehet ábrázolni normálisan (nehéz átlátni), ha azonban sok megfigyelt egység van (N=1000), akkor nem lehet szépen ábrázolni
- pontdiagram? (– az összefüggést láttatja)

Szint és változás

- sokszor nem a szint, hanem a változás érdekel minket
- szintet változással nem érdemes összehasonlítani
- pl.: az árfolyam szintje és a GDP változás kapcsolata becsapós, mert pl. az árfolyam folyamatosan nőtt Magyarországon, amíg a GDP változása 0 körül ingadozik

- szintből egyszerű növekedési rátát (változást) számolni: *növekedés/jelen*100*

$$\%\text{változás} = \frac{(Y_{t+1} - Y_t)}{Y_t} \cdot 100$$

$= 100*(\ln Y_t - \ln Y_{t-1})$

- %-os változás ≠ … százalékponttal változik!

Adatok közti összefüggés

- két változó összefüggését úgy vizsgálhatjuk, hogy megnézzük, hogy együtt mozognak-e (eszközei: korreláció, regresszió)
- pl.: A jegybanki alapkamat tényleg hat az inflációra?
  Az országok közötti beruházásban tapasztalt különbségek magyarázzák a GDP-változásbeli különbségeket?
  Magasabb-e a GDP ott, ahol magasabb a felsőfokú végzettségűek aránya?
  Diktatúrákban magasabb-e a városiasodás?
- az együttmozgás nem feltétlenül jelent oksági viszonyt! (pl.: lehet, hogy a sok magasan kvalifikált miatt magas egy ország GDP-je, de az is lehet, hogy mivel magas az ország GDP-je annál többen mehetnek egyetemre)
- közvetett vs, közvetlen okság
- az összefüggés jellemzően nem determinisztikus, a tendenciák érdekelnek minket
- sokszor vannak kiugró értékek (outlierek), amik nem illenek a tendenciába (nem baj)
- legalapvetőbb módja az együttmozgás megvizsgálásának a pontdiagram
- egyenes illesztése – trendvonal – segíthet a kapcsolat „láthatóvá válásában”
- igazán jó megoldás: számszerű leírás (kapcsolatvizsgálat)

Leíró statisztikák:

- Leíró statisztika: számszerűen és tömören összefoglalni változó jellemzőit
  - Szintje? – átlag, medián (percentilisek), módusz => elhelyezkedésmutatók
  - Változékonysága? – szórás, terjedelem, interkvartilis terjedelem

Átlag:

- N: minta elemszáma, megfigyelések száma
- Probléma: két csúcsú eloszlásnál nem biztos, hogy informatív

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Módusz: (= leggyakoribb érték)

- probléma: nem mindig létezik (pl. minden értékből egy), ill. több módusz is lehet
- megoldás lehet: hisztogram legmagasabb pontja (függ osztályközöktől) – osztályköz közepe

Medián: (=középső érték)

- megfigyelések fele alatta, fele felette van (ha páros, 2 középső átlaga)
- X-edik **percentilis:** megfigyelések X%-a kisebb értéket vesz fel *–> =PERCENTILIS.TARTALMAZ()*
- **Kvartilis / Decilis / Kvintilis**: negyedeli / tizedeli /ötödöli az adatokat
  - 1. kvartilis: 25% alatta, 2. kvartilis = medián
  - 1. decilis: 10% alatta, 5. decilis = medián
  - 1. kvintilis: 20% alatta

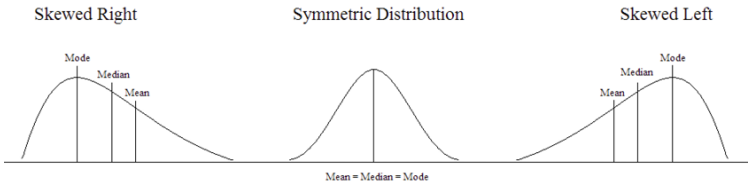
Átlag, módusz medián

- a változó jellemző értékéről adnak számot különbözőképpen (Central tendencies)
- adatelemzéskor érdemes mindegyikre figyelni, ne ragadjunk le az átlagnál, hiszen láttuk, hogy az átlag nem mindig jellemző érték

- egymáshoz viszonyított helyzetük (főleg az átlag és a medián viszonya) a változó eloszlásáról nyújt képet
- normális eloszlás esetén egybeesnek
- Pl.: emberek magassága, vérnyomás, zh-n szerzett eredmények

ferdeség (skewness)

- Szimmetrikus: átlag = medián = módusz
- Pozitív ferdeség (jobbra hosszan elnyúló): módusz < medián < átlag
- Negatív ferdeség (balra hosszan elnyúló): módusz > medián > átlag



Szóródás:

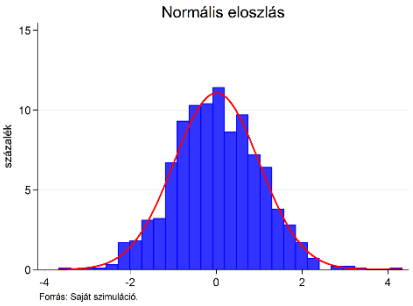
- **Terjedelem** (range, disperse): maximum és minimum közti eltérés
- Nem megbízható (kiugró értékek) -> interkvartilis terjedelem = 3. – 1. kvartilis
  -> interdecilis terjedelem = 9. – 1. decilis
- **Variancia:** átlagos négyzetes eltérés, azaz szórásnégyzet
- **Szórás:** (dispersion)
  - Önmagában nehezen értelmezhető
  - Négyzet szerepe: megoldja a negatív értékek problémáját

$$s = \sqrt{Var} = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}}$$

- **Relatív szórás:** szórás/átlag

Normális eloszlás esetén az adatok

- 68%-a az $[\bar{Y} - s; \bar{Y} + s]$
  (átlag-szórás; átlag+szórás)
- 95%-a az $[\bar{Y} - 2s; \bar{Y} + 2s]$
- 99,7%-a az $[\bar{Y} - 3s; \bar{Y} + 3s]$ intervallumba esik.
-



Empirikus elemzés 3.

Indexek = indexszámok:

- Adott ország inflációja, árainak a változása érdekel. Mit tegyünk? Nem egyes árak alakulása, hanem az országos árszínvonal megfigyelése a célunk.
- Képezünk egy tipikusnak tekinthető fogyasztói kosarat, amely egy átlagos fogyasztó által vásárolt termékeket és szolgáltatásokat tartalmazza. A termékeket fontosságuk szerint súlyozzuk.

Indexek alaptulajdonságai:

- általában idősorosak
- választunk egy bázisidőszakot (b) és ennek az időszaknak az értékét 100-nak vesszük
- a többi időszak értékét a bázishoz viszonyítjuk
- változást mér, szint jellemzésére nem alkalmas - arról nem tudunk meg semmit, hogy hol magasabbak az árak, sokszor a drágább országban kisebb az infláció (= drágulás)
- *Mit jelent, ha Y_1=100, Y_2=105, Y_3=112?*
- *Mit mondhatunk a 2. és 3. év közötti változásról? - 7 százalék**pont**tal nőtt*
- az index(szám) két aggregátum hányadosa: $I_t = \frac{A_t}{A_{t-1}}$
    - ahol A (az aggregált sokaság): $A_t = \sum_{i=1}^{n} q_{i,t} p_{i,t}$
    - ahol q (quality) a mennyiséget, p (price) az árat jelöli
    - az aggregátum állhat egy elemből is (n=1) - *Mi a jelentése akkor?*
- az index vonatkozhat árra, mennyiségre és értékre – a probléma az utolsóval az, hogy két dolog változik egyszerre, önmagában nem informatív

Árindexek – *rögzített mennyiségek (kosár)*

- árindex: bázisidőszak %-ában kifejezett árszínvonal (átlagos ár nehezen értelmezhető)
- éves infláció: évenként változó bázis
- egy termék esetén az árindex: $\frac{p_t}{p_{t-1}}$ (tárgyidőszaki ár/bázisidőszaki ár)*100

  több terméknél a termékeket súlyozni kell, a súlyok fontosságot tükröznek

-> Fogyasztói árindex

- a súlyok megválasztása alapján két fajta index (b bázisévet jelöl, t pedig tárgyévet)
    - **Laspeyres - bázisidőszaki súlyozás,** azaz a bázisév mennyiségei a súlyok

      LPI= $\frac{\sum_{i=1}^{n} q_{i,b} p_{i,t}}{\sum_{i=1}^{n} q_{i,b} p_{i,b}}$
    - **Paasche - tárgyidőszaki súlyozás,** azaz a tárgyév mennyiségei a súlyok

      PPI= $\frac{\sum_{i=1}^{n} q_{i,t} p_{i,t}}{\sum_{i=1}^{n} q_{i,t} p_{i,b}}$

Példa: Árindex

	2009		2010	
termék	p	q	p	q
A	10	0,3	12	0,4
B	4	0,7	5	0,6

Laspeyres árindex

(ha ugyanazt a mennyiséget amit tavaly drágábban fogyasztanánk idén)

$$\frac{\sum_{i=1}^{n} q_{i,b} p_{i,t}}{\sum_{i=1}^{n} q_{i,b} p_{i,b}} = \frac{(0,3 * 12) + (0,7 * 5)}{(0,3 * 10) + (0,7 * 4)} = 1,224$$

Paasche árindex

(ha ugyanazt a mennyiséget amit idén olcsóbban fogyasztottuk volna tavaly)

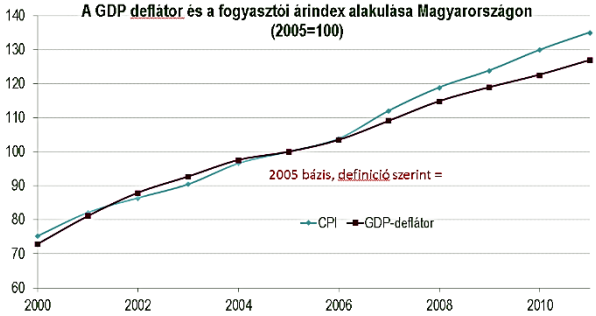
$$\frac{\sum_{i=1}^{n} q_{i,t} p_{i,t}}{\sum_{i=1}^{n} q_{i,t} p_{i,b}} = \frac{(0,4 * 12) + (0,6 * 5)}{(0,4 * 10) + (0,6 * 4)} = 1,219$$

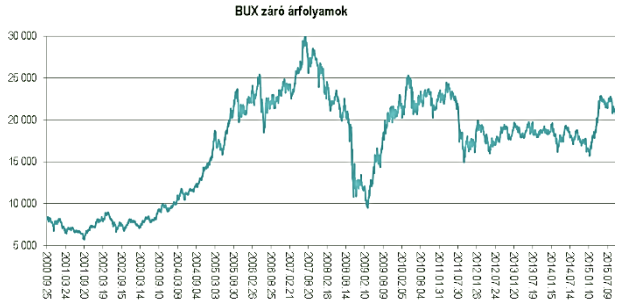
Legfontosabb árindexek: *fogyasztói árindex, GDP-deflátor*

- **fogyasztói árindex** (=consumer price index, **CPI**):
    - jellemzőnek vélt fogyasztói kosárban szereplő javak és szolgáltatások árainak a változását méri
    - bázisév változik, de módszertanilag nem probléma (átváltható)
- Infláció - árak nőnek
- Dezinfláció - infláció mértéke csökken
- Defláció - árak csökkennek
- **GDP deflátor**= reál GDP = GDP mérőszáma (nominál) / alkalmas árindex
    - a nemzetgazdaság összes termékének és árának a változását méri

- szigorúan véve nem árindex, mert nem egy rögzített termékkosáron alapul (ha az ár változik a fogyasztott mennyiség is változik a GDP-ben, tehát a kosár is változik)
- **nominál GDP** - kiinduló vátozó **reál GDP=nominál GDP / alkalmas árindex**
  reál GDP **=> kiszűri az inflációt**

- a legtöbb esetben nincs szignifikáns különbség a GDP deflátor, és a CPI között, de a fókuszuk eltér - más termékekre és szolgáltatásokra terjednek ki *beruházási (GDP deflátor) fogyasztási (CPI)*





- BUX index: példa
  - a Budapesti Értéktőzsde (BÉT) hivatalos árindexe a BUX
  - az index a BÉT részvény szekciójában szereplő legnagyobb tőkeértékű és forgalmú részvények árának átlagos változását mutatja (valós időben, 5 másodpercenként frissül az aktuális piaci árak alapján)
  - az index kosarába kerülő részvények súlyának meghatározása: a piacon ténylegesen forgó állományt jobban megragadó közkézhányad alapú súlyozás (a tiszta kapitalizáción alapuló súlyozás egy kifinomult formája)
  - BUX-index fontos mutatója a magyar gazdasági folyamatoknak, a változás az érdekes, de a szintet is számontartják
  - hasonló indexek: S&P, Dow Jones, FTSE, NASDAQ, CAC, DAX, Nikkei.

Mennyiségi indexek:

- árakat rögzítjük, így a mennyiségi változásokat tanulmányozhatjuk
- nominál GDP nem kifejező
- (reál) GDP-növekedés: $\frac{\sum_{i=1}^{n} q_{i,t} p_{i,b}}{\sum_{i=1}^{n} q_{i,b} p_{i,b}}$

Koncentrációs mutatók:

- **Koncentráció:** egy adott változó értékének jelentős része vagy egésze, kevés egységre összpontosul - *egyenlőtlenség*
- pl. adott iparágban az összkibocsátás jelentős részét pár vállalat adja (tesco)
  lakossági jövedelem eloszlása egyenlőtlen (**80-20 szabály:** a társadalom 20%-a jut az összjövedelem 80 %-ához)
- ideális tökéletes versenyben sok kicsi árelfogadó vállalat van
- valóságban nem ez a tipikus: sok piacon csak néhány vállalat van, más piacokon jóval több, de néhány meghatározó
- a nagy vállalat visszaélhet piaci erejével (nem biztos, hogy tud!) és a fogyasztók rosszul járna - versenyhivatalok ezt árgus szemmel figyelik

Egyszerű mutatók:

- állítsunk fel egy csökkenő rangsort a vállalatok között nagyságuk egy bizonyos (minket érdeklő) mérőszáma (pl. piaci részesedés) szerint
- majd **kumulálás**sal könnyen kiszámítható, hogy mekkora a piaci részesedése a legnagyobb, az első két legnagyobb, első három legnagyobb stb. vállalatnak

- ha az előző számolás eredményeit ábrázoljuk, akkor a **koncentrációs görbét** kapjuk
- alapötlet: piaci részesedés alapján a teljes kibocsátás mekkora hányada koncentrálódik a legnagyobb vállalatoknál
- **CR_n:** az n legnagyobb vállalat piaci részesedése (leggyakoribb a CR_4)

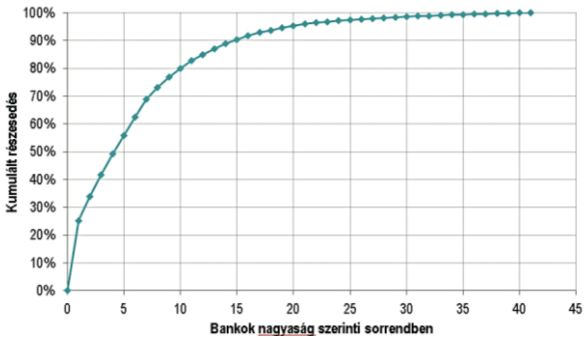
Példa: pénzügyi intézetek

Magyarországon 2015-ben 40 részvénytársasági formában működő hitelintézet volt. (Adatok: PSZÁF, Aranykönyv 2015). Eszközállomány alapján rangsorolva.

	Név	Piaci részesedés	Kumulált részesedés
1	OTP Bank Nyrt.	25.24%	25.24%
2	K&H	8.60%	33.83%
3	UniCredit Bank Hungary Zrt.	7.83%	41.67%
4	Raiffeisen Bank Zrt.	7.44%	49.11%
5	ERSTE BANK HUNGARY Zrt.	6.68%	55.79%
6	MKB Bank Zrt.	6.67%	62.46%
7	CIB Bank Zrt.	6.28%	68.74%
8	OTP Lakás	4.40%	73.13%
9	MFB Magyar Fejlesztési Bank Zrt.	3.74%	76.88%
10	Budapest Bank	2.99%	79.87%

CR_4= 49.5%

Itt is érvényesül majdnem a 80-20 szabály. (CR_8= 73.3%)



- az előző mutatókkal az a baj, hogy az egyik (CR4) szerint az egyik iparág koncentráltabb, míg a másik (CR8) alapján a másik - csak a koncentrációs görbe egy pontját ragadják meg
- olyan mutató kell, ami egy számban fejezi ki a koncentráltságot, megragadva a koncentrációs görbe egészét

Herfindahl-Hirschman index:

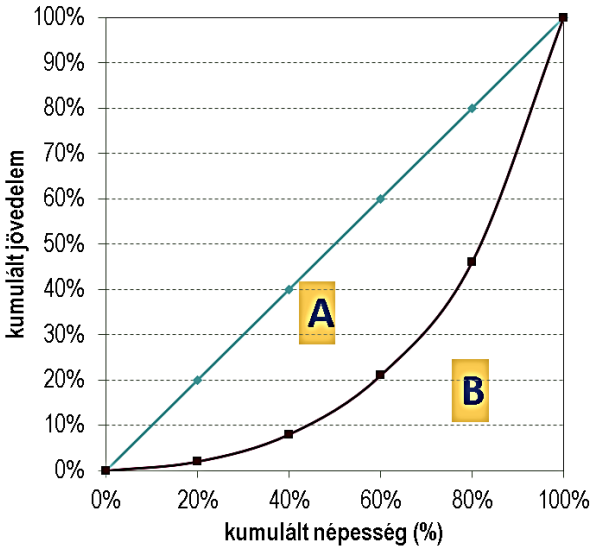
- piaci koncentráció gyakori mérőszáma
- legyen **s** az **i**. vállalat részesedése egy piacon, ahol **n** szereplő van - Herfindahl index ekkor: $HI = \sum_{i=1}^{n} s_i^2$
- $\frac{1}{n} < HI < 1$ - *A határértékek milyen piacszerkezetre utalnak?*
  - 1/n: ha mindegyik részesedése ugyanakkora - n db (*) $1/n^2$
  - 1: teljesen koncentrált szétosztás
- önmagában a szám nem mond sokat, időbeli alakulása érdekesebb
- példa: A" és „B" országban 5-5 autógyár működik a következő részesedésekkel: 10, 14, 16, 28 és 32 %, illetve 12, 13, 14, 30, 31 %. Ha a CRx mutatókat kiszámoljuk, akkor x függvényében hol az egyik, hol a másik ország autógyártása koncentráltabb. A Herfindahl-indexek a következők lesznek:
  $HI = 0{,}1^2 + 0{,}14^2 + 0{,}16^2 + 0{,}28^2 + 0{,}32^2 = 0{,}236$
  $HI = 0{,}12^2 + 0{,}13^2 + 0{,}14^2 + 0{,}30^2 + 0{,}31^2 = 0{,}237$
- (Néha a 0-10000 intervallumban adják meg a Herfindahly indexet, azaz HIA=2360 és HIB=2370)

Lorenz görbe:

- a koncentráció egyenlőtlenségét gyakran a Lorenz görbe segítségével jelenítjük meg
- leggyakrabban jövedelmi egyenlőtlenségek bemutatására használják
- a görbe pontjai azt mutatják, hogy az egyének (országok) x%-a az összjövedelem hány százalékát birtokolja - minél kevesebb százalék birtokol minél nagyobb hányadot, annál jobban koncentrálódik a jövedelem
- (felfelé) kumulálás: egyre több osztályközhöz tartozó értéket adunk össze
- példa: jövedelem (figyelem! a lakosságot rendeztük a jövedelem szempontjából
- Átló: egyenletes eloszlás
- Minél nagyobb az átló és a görbe közötti terület, annál egyenetlenebb az eloszlás (annál jobban koncentrálódik a jövedelem).

- **Gini együttható: G=A/(A+B)**
  0≤G≤1 - kisebb Gini, egyenletesebb eloszlásra utal, 1: teljes egyenlőtlenség

lakosság	jövedelem	kumulált jöv.
1. kvintilis	0,02	0,02
2. kvintilis	0,06	0,08
3. kvintilis	0,13	0,21
4. kvintilis	0,25	0,46
5. kvintilis	0,54	1



Empirikus elemzés 4.

Korreláció: = együttmozgás

- **két változó közti kapcsolat számszerűen** - mérhetővé tudjuk vele tenni
- jelölés: X és Y közti korreláció $\mathbf{r_{XY}}$
- NEM ok-okozati kapcsolat
- intuíció:
  - ha az egyik változó magas (alacsony) értékeivel a másik változó magas (alacsony) értékei „járnak együtt", akkor **pozitív** a korreláció
  - ha az együttjárás ellentétes, akkor **negatív**
  - ha pedig nem mozognak együtt, akkor **0** (nincs korreláció, 0 közeli)
- *Mihez képest?* - átlaghoz képest magas/alacsony

$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- számláló: i. megfigyelés – átlag
- nevező: x és y szórása

-> az átlagtól való eltérések szorzata a szórások négyzetének gyökével normálva

- *Előjel*? Ha x és y is magasak +*+ + vagy -*- pozitív, ha ellentétes előjel akkor negatív
- *Mikor lesz az értéke nullához közeli?* Ha nincs kapcsolat

Korreláció tulajdonságai:

- -1 és 1 közötti érték -> tényleges jelentőséggel, értelemmel bír az érték
- nagyobb pozitív érték – erősebb pozitív kapcsolat
- szimmetrikus kapcsolat -> X és Y közti korreláció = Y és X közti korreláció
- változó korrelációja önmagával = 1
- konstanssal való korreláció = 0
- **Korreláció négyzete (r_{XY}^2): Y varianciájának mekkora hányadát magyarázza X varianciája = X varianciájának mekkora hányadát magyarázza Y varianciája** – *van értelme*
- nincs korreláció ≠ nincs kapcsolat – hanem nincs lineáris kapcsolat
- pl. Munkanélküliségi ráta és GDP/fő közti korreláció = -0,51
  - ez egy általános tendencia, lehetnek eltérések
  - közepes erősségű negatív kapcsolat (magasabb GDP/fő – alacsonyabb munkanélk)
  - GDP/fő varianciáját 26%-ban magyarázza a munkanélküliségi ráta varianciája (vagy fordítva) – azaz a korreláció négyzete, r^2=0,51*0,51=0,26

Okság:

- *Egyik változó „okozza-e" a másikat közvetlenül?* – közvetlen vs. közvetett oksági kapcsolat
- korreláció önmagában nem árulkodik az okság irányáról
- sokszor intuitív, sokszor nehezen eldönthető pl.
  - *a telek nagysága növeli az ingatlan árát, de fordítva nem igaz*
  - *magasabb végzettség okozza a magasabb bért és nem fordítva, de lehet, hogy nem az egyetem miatt nagyobb a bér, azaz az okozat nem közvetlen*
  - *a vállalat többet költhet reklámra, mert a magasabb bevételek miatt megengedheti*
- lehet egy harmadik mögöttes ok is -> közvetett okság (pl. *gólyák, dohányzás*)

Több változó közötti korreláció:

- M változó – M(M-1)/2 korreláció
- Korrelációs mátrix 3 változóra (X, Y, Z)

	X	Y	Z
X	1		
Y	r_{XY}	1	
Z	r_{XZ}	r_{ZY}	1

főátló

	Ingatlan ár	*telekméret*	*#hálószoba*
Ingatlan ár	1		
telekméret	0.535796	1	
#hálószoba	0.366447	0.151851	1

Ha ez nem lenne picit pozitív, megkérdőjelezhető lenne

Regresszió:

- Pontdiagram két változó között – a trendvonal „görbeillesztés" alapvetően egyenesillesztés
- A pontdiagramon néha nehezen meglátható a kapcsolat.
- *„Mennyire egyszerű egyenest húzni (görbét illeszteni) a pontokra?" És milyen a meredeksége?*

Korreláció vs. regresszió:

- Változók közötti kapcsolat számszerűsítése
- Korreláció: 2 változó között; Oksági kapcsolat?
- Regresszió: Komplex összefüggések (több változó és hatásuk); Lehet mögötte gazdasági modell – okság

Egyváltozós regresszió:

pl. Ingatlanár és telekméret kapcsolata Windsorban:

- *második okozza az elsőt, ezért az ingatlanár alakulását magyarázzuk telekméret alakulásával -> ingatlanár: függő változó (Y), telekméret: magyarázó (vagy független) változó (X)*

1. *feltesszük, hogy létezik lineáris kapcsolat a két változó között: pl. egy négyzetlábbal nagyobb telekméret ß kanadai dollárral növeli az ingatlanárát -> minket pontosan ß érdekel*

- *ez eddig egy elméleti modell, amiről úgy gondoljuk, hogy a windsori ingatlanok ára és telekmérete között fennáll*
- *ez a modell önmagában nem teljes, mert a telekméreten túl más is hat az ingatlanárakra, azonban a hiba ellenére érdemes a modellel foglalkozni, mert többet tudunk meg az ingatlanárak alakulásáról a telekméret segítségével, mint nélküle*

2. *nem ismerjük az összes windsori ingatlan aktuális árát és telekméretét, csak azokat, amelyeket adott évben eladtak vagy felértékeltettek*

- *az így rendelkezésre álló adatok segítségével próbáljuk megbecsülni az előzőekben felvázolt elméleti modellt*
- *azaz csak becsüljük ß-t és reméljük, hogy a becslés valós képet ad, hogy az összes ingatlant tekintve mi az összefüggés az ár és a méret között*

3. *úgy becsüljük a ß-t, hogy a telekméret – ingatlanár pontdiagramban szereplő pontokra ráillesztünk egy olyan egyenest, amely minimalizálja az összes pont (négyzetes) eltérését magától az egyenestől*

- *ugyanis a pontok biztosan nem fekszenek egy adott egyenesen, hanem szóródnak körülötte*
- *minden egyes pontra kiszámoljuk, hogy mennyire tér el az egyenestől (reziduum) és az összeget minimalizáljuk*

- Y függő változó, X magyarázó változó
- Feltevés: lineáris kapcsolat (ha nem az, akkor matematikai transzformációval linearizálható)
- Regressziós egyenes: $Y = \alpha + \beta X$
- **β: ha egy egységgel változik X, hány egységgel változik Y –>** amire kíváncsi vagyok
- α: y tengelymetszet, mennyi az Y ha X=0 - *„mekkorat ér az ingatlan ami mérete 0"*
- lineáris regresszió - közelítés, azaz hibát követünk el
  pl. kihagyott, meg nem figyelhető változók, nem lineáris kapcsolat
- Regressziós modell hibataggal: $Y = \alpha + \beta X + e$
- Hiba - e**: adatpont (ha az összes ingatlanról lenne teljes információnk) és a (valódi) regressziós egyenes közti távolság**
- együtthatók értékeit nem ismerjük, rendelkezésre álló adatokból becsüljük
- Becsült együtthatók: legjobban illeszkedő egyenes együtthatói, jelölés: $\hat{\alpha}, \hat{\beta}$
- Azaz: $Y = \hat{\alpha} + \hat{\beta} X + u$
- Reziduum (maradéktag**):** $\boldsymbol{u \neq e}$
- meglévő adatpontok nem illeszkednek az egyenesre, egyenes és pont távolsága = reziduum

OLS becslés – hogyan kapjuk meg az egyenest?

- Legjobban illeszkedő egyenes, ha a maradéktagok négyzetösszege (SSR – sum of squared residual) minimális $\min SSR = \min \sum_{i=1}^{N} u_i^2 = \min \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$
- Legkisebb négyzetes becslés = ordinary least squares (OLS)

pl - ingatlan

- *Becsült együtthatók:*
- *α = 34136.2 – tengelymetszet „0 m^2-es telek ára 34136 dollár"*
- *6.6 – telekméret együtthatója „ha 1 m^2-rel növeljük a telekmértetet, az igatlan ára 6,6 dollárral emelkedik"*

- Meredekség:
- Y átlagos változása X egységnyi növekedése esetén – általános tendenciát ragad meg (nem igaz minden egyes pontra)
- Marginális hatás: $\frac{\mathrm{dY}}{\mathrm{dX}} = \beta$
- $\hat{\beta}$: X Y-ra gyakorolt marginális hatásának becslése, azaz a legpontosabb becslés szerint, X 1 egységgel való megváltoztatása, Y-ra $\hat{\beta}$-nyi hatással van

Empirikus elemzés 5.

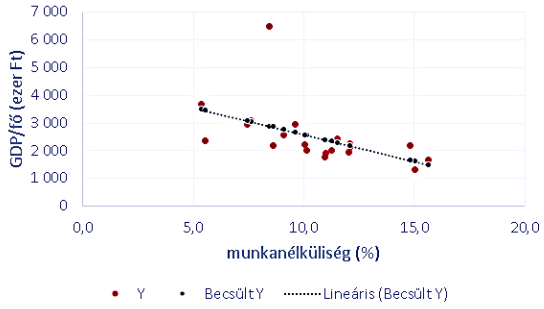
Ismétlés: egyváltozós regresszió

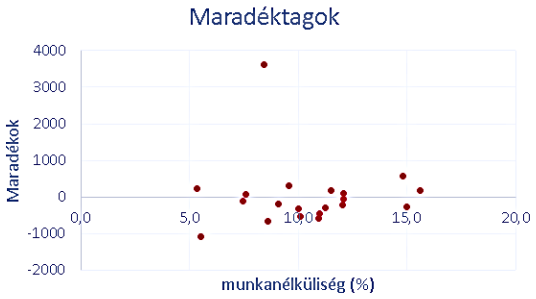
- Feltesszük, hogy van lineáris kapcsolat 2 változó között
- Regressziós modell: $Y_i = \alpha + \beta X_i + e_i$
  - elméleti hiba a jobb és bal oldal között (elvárásaink vannak)
- Becsült modell: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
  - becsült X_i-vel megadhatjuk Y_i-t
- Becsült modell hibával: $Y_i = \hat{\alpha} + \hat{\beta} X_i + u_i$
  - eltérés az általunk adott becslés és a valóság között
  - u_i: valós eredmény, teljesülnek rá bizonyos feltételek (e_i: tipp, feltételezés)
- Becslés: OLS – általános legkisebb négyzetek módszere: ötlet $\hat{\alpha}$; $\hat{\beta}$ – ra

pl. függő változó: GDP/fő (e Ft-ban)

r^2 =0.26 – *a GDP/fő varianciájának 26%-át lehet magyarázni a munkanélküliség var.-jával*

Munkanélküliségi ráta együtthatója: -194.3 – *ha egy százalékponttal növeljük a munkanélküliséget, a GDP/fő kb. 193 300 forinttal csökken*





Illeszkedés mérése:

- OLS: legjobban illeszkedő egyenes megtalálása - maradéktagok négyzetét minimalizáljuk
- Mennyire jó az illeszkedés?
- Mérőszám: R^2
- Egyváltozós regresszió esetén a korreláció négyzete= R^2

Becsült érték (=fitted value)

- Regressziós egyenlet: $Y_i = \alpha + \beta X_i + e_i$
- Becsült/illesztett/előrejelzett érték: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- Kettő összehasonlítása – illeszkedés jósága
- A kulcs: maradéktag $u_i = Y_i - \hat{Y}_i$
- Azt már láttuk, hogy a regresszió becslésekor a maradéktagok négyzetösszegét (SSR - sum of squared residuals) akarjuk minimalizálni, azaz $\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$
- deriválni kell és nullával egyenlővé tenni (2 változós függvény ($\hat{\alpha}$; $\hat{\beta}$) – nem csináljuk meg)
  - ➔ eredmény: $\hat{\beta} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y}_i)(X_i - \bar{X}_i)}{\sum_{i=1}^{N} (X_i - \bar{X}_i)^2}$ *(x varianciájával osztunk le)*
  - ➔ $\hat{\alpha} = \hat{Y}_i - \hat{\beta} X_i$

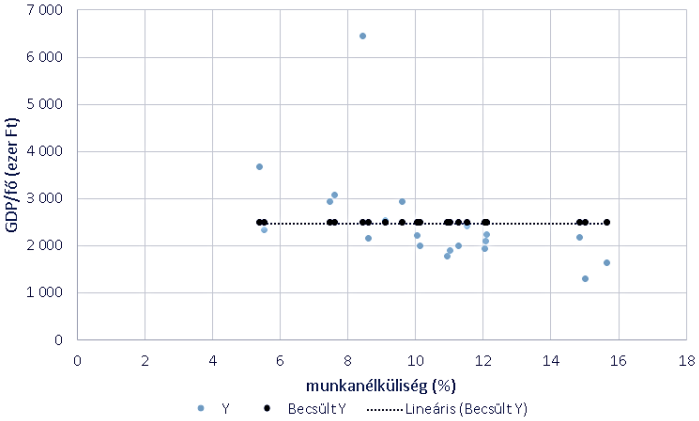
- Mire emlékeztet ez? – kapcsolatot mérnek mindketten (csak bétát nem szorítjuk be -1 és 1 közé (osztás azért kell, hogy -1 és 1 közé essen): $r = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$

R^2:

- Alapból, ha a GDP/fő érdekel minket és semmilyen magyarázó változóval nem rendelkezünk, akkor a legjobb, amit tehetünk az az átlag használata.
- Azt figyelhetjük meg ekkor, hogy az egyes megfigyelések eltérnek az átlagtól (szórás mutatja az eltérés mértékét)
- az eltéréseket akarjuk megérteni a magyarázó változó segítségével
- **TSS** (=total sum of squares)**: teljes szórásnégyzet** – *SS oszlop „összesen”*
  - Y szóródásának mérőszáma
  - a megfigyelések szóródása az átlaghoz viszonyítva

$$TSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

  - *A varianciához az kellene, hogy (N-1)-gyel osszuk el (akár meg is tehetnénk (ha az összes későbbi változóban RSS, SSR is megtennénk), de nem változtatna az eredményen, inkább mindegyiknél elhagyjuk)* - *Variancia = TSS/(N-1)*



- **SSR** (=sum of squared residuals)**: maradéktagok négyzetösszege** – *SS oszlop „maradék”*
  - a megfigyelések szóródása a regressziós egyeneshez képest – a regresszió által meg nem ragadott rész (reg. hiba)

$$SSR = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \text{ - } u_i^2$$

- **RSS** (=regression sum of squares): **regressziós négyzetösszeg**– *SS oszlop „regresszió”*
  - a TSS és az SSR között teremti meg a kapcsolatot
  - regresszió segítségével becsült pontok szóródása az átlaghoz képest (az eltérések négyzetösszegét adja)
  - mennyit magyaráz a regresszió az összes szóródásból

$$RSS = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$$

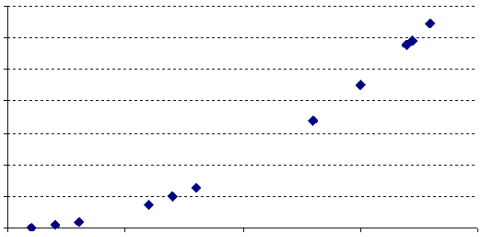
- *Átlaghoz képest a megfigyelések szóródnak.*
- *Regresszióhoz képest is szóródnak a megfigyelések (maradékok), de jóval kevésbé, mint az átlaghoz képest.*
  - ➔ *Azaz a regresszió „megfogja” a megfigyelések szóródásának egy részét.*
  - ➔ *Úgy számszerűsítjük, hogy megnézzük a megfigyelések átlaghoz viszonyított szóródását és a megbecsült értékek átlaghoz mért szóródását. Ha a kettő közel van egymáshoz, akkor a megfigyelések szóródását a regresszió jól magyarázza.*
- belátható, hogy **TSS=RSS+SSR**
- **R^2 azt mutatja, hogy a regresszió a teljes variancia hány százalékát magyarázza meg.**

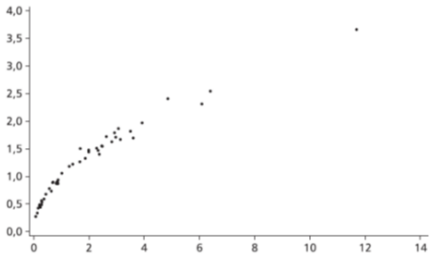
  Számszerűleg: $R^2 = 1 - \frac{SSR}{TSS} = \frac{RSS}{TSS}$ $0 \leq R^2 \leq 1$
- ha R^2=1, akkor tökéletes az illeszkedés

- nullához közeli R^2 rossz illeszkedésre utalhat, de alacsony R^2 nem jelenti azt, hogy egy regresszió rossz (általában azt szeretjük, ha r^2 magas)
- feladhatjuk, hogy szimmetriát várunk e

NEMLINEARITÁS:

- ha nem lineáris kapcsolatot használunk X és Y között, négyzeteset vagy logaritmikusat





- Megoldás nemlin. esetén: olyan matematikai transzformációt alkalmazni, mely linearizálja a kapcsolatot. (pl. X helyett X^2 alkalmazása a regresszióban)
  - Nemlinearitás lehetséges közgazdasági oka: csökkenő határhaszon
  - Függvények: kvadratikus, logaritmus
  - pl. a bér hogyan függ a munkatapasztalattól (években megadva)
    bér=5,25+0,48tap-0,008tap² -> négyzetesnél elvész az együtthatók konkrét jelentése
  - érdemes kirajzoltatni a grafikonját
  - a másik lehetőség, hogy r négyzeteket vizsgálunk

Logaritmikus forma: *(log=ln - angolszász)*

- Lineáris összefüggést eredményezhet.
- Könnyű értelmezhetőség - pl. rugalmasság esetén $\ln Y = \alpha + \beta \ln X$
  $\beta = \frac{d \ln Y}{d \ln x}$ -> az együttható jelentése megmarad
- X adott értékének nincs szerepe (nemlinearitás nem gond)
- Mértékegységnek nincs szerepe
- Ha x értéke r %-kal nő, akkor x-ből x(1+r) lesz
  ez a változás log-ban kifejezve: *log(x(1+r))-log(x)= log(x)+log(1+r)-log(x)=log(1+r)*
  log(1+r)≈r, ha r kicsi.

Rugalmasság: $\ln Y_i = \alpha + \beta \ln X_i$
értelmezés: *%-os változás % (ha vesszük a logaritmusát)*
Félrugalmasságok: $Y_i = \alpha + \beta \, ln \, X_i$
értelmezés: *egység (ha nem) %*
$ln \, Y_i = \alpha + \beta X_i$
értelmezés: *% egység*

Klasszikus példa: árrugalmasság – ha egy %-kal nő az ár, hány %-kal változik a kereslet

- Szint-szint regresszió (level-level): $\boldsymbol{Y_i = \alpha + \beta X_i + e_i}$
  „Ha X egy egységgel változik, akkor azt várjuk, hogy Y β–val változzon."
- Log-szint regresszió (log-level): $\boldsymbol{\ln Y_i = \alpha + \beta X_i + e_i}$
  „Ha x egy egységgel változik, akkor azt várjuk, hogy Y 100*β százalékkal változzon."
- Szint-log regresszió (level-log): $\boldsymbol{Y_i = \alpha + \beta \, ln \, X_i + e_i}$
  „Ha X egy százalékkal változik, akkor azt várjuk, hogy Y β/100–al változzon."
- Log-log regresszió (log-log): $\boldsymbol{ln \, Y_i = \alpha + \beta \, ln \, X_i}$
  „Ha X egy százalékkal változik, akkor azt várjuk, hogy Y β százalékkal változzon."

Empirikus elemzés 6.

Bizonytalanság:

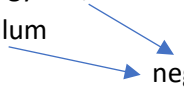
- regressziós együtthatók (α; **β**) valódi értéke nem ismert
- legtöbbször **minta** alapján becslünk (kalap) - így a legkézenfekvőbb
- becsült érték nem pontosan azonos a valódi értékkel
- **pontbecslés** (=legkisebb négyzetes becslés)-ünk lesz -> ez a β-ra adható legjobb közelítés, de a bizonytalanságot nem tükrözi *($\hat{\beta}$ = β pontbecslése)*
- nem mindegy, hogy melyik pontbecslésnek mekkora a bizonytalansága
  -> cél: **bizonytalanság mérése** (megnézzük, hogy a becsült érték mennyire pontos, mennyire különbözik nullától):
  1. Konfidencia-intervallumok kiszámítása –> pontbecslést átalakítjuk intervallumbecslésre
  2. Hipotézis vizsgálat –> felállítunk egy hipotézist β-val kapcsolatban, megpróbáljuk megcáfolni
- Bizonytalanságot befolyásoló tényezők:
  1. **megfigyelések száma:** egyre több pont, egyre pontosabb képet ad
  2. Kisebb **hibatagok** (/maradéktagok) –> pontosabb becslés (Y minél kevésbé szóródik annál biztosabb)
  3. **X** nagyobb **szóródása** –> pontosabb becslés
- *Pl.* *nemek hatása a jövedelemre, dummy változó, kevéssé tudok kapcsolatot becsülni végzettség hatásának becslése jövedelemre (kiket vizsgálunk? - változatosság)*

Konfidencia intervallum: -> intervallumbecslés

- képlet: $(\hat{\beta} - t_\beta s_\beta; \hat{\beta} + t_\beta s_\beta)$
- szám -> kivonjuk és hozzáadjuk a pontbecsléshez -> szimmetrikus intervallum $\hat{\beta}$ körül
- mekkora bizonyossággal teljesül: $\hat{\beta} - t_\beta s_\beta < \beta < \hat{\beta} + t_\beta s_\beta$?
- valószínűséget fejez ki -> az adott intervallumban milyen valószínűséggel helyezkedik el β
- megbízhatósági szint: a konfidenciaintervallumban tükröződő bizonyosság mértéke (95%)
- leggyakoribb: 95%-os konfidenciaintervallum
  - *„95% a valószínűsége, hogy az együttható valódi értéke az adott intervallumba esik"*
  - *„Ha 100 mintát vennénk, akkor 95 esetben a becsült együttható benne lenne az intervallumban"*
  - *„Ha egy képlet alapján újra és újra kiszámolnánk a konfidenciaintervallumot, az így keletkező intervallumok 95%-a tartalmazná béta valódi értékét"*
- ahogy növeljük az intervallumot, úgy nő a megbízhatósági szint
- ugyanahhoz a megbízhatósághoz minél kisebb (szűkebb) intervallum, annál jobb a becslés

standard hiba - s_β: (=standard deviation, standard error)

- bizonytalanság mérőszáma
- $\hat{\beta}$ is egy változó, értéke mintáról mintára változik -> értéke szóródik => **$\hat{\beta}$ szóródása**
- Képlet: $s_\beta = \sqrt{\frac{SSR}{(N-2)\sum_{i=1}^{N}(X_i - \bar{X})^2}}$ -> függ: SSR (hibatagok), N (mintaelemszám), X varianciája
- kisebb SSR -> kisebb s_β -> szűkebb konfidenciaintervallum
  - azaz a konfidenciaintervallum szélessége pozitívan függ az SSR nagyságától (nagyobb maradéktagok nagyobb bizonytalanságra utalnak)
- Nagyobb N -> kisebb s_β -> szűkebb intervallum (több megfigyelés, kisebb bizonytalanság)
- Nagyobb X változékonyság -> kisebb s_β -> szűkebb intervallum
- intuíciónknak megfelelő



negatív kapcsolat

t_β:

- egy bizonyos eloszlásból (Student-féle eloszlás - sok adatnál student eloszlás = normális eloszlás) jön
- A t_β szerepe az, hogy a **kívánt megbízhatósági szint függvényében felnagyítja a standard hiba hatását** -> „büntetőfaktor”
- Ha 95% megbízhatósági szinttel dolgozunk, akkor t_β más (kisebb), mint amikor a megbízhatósági szint 99% -> **a választott megbízhatósági szint emelésével** t_β **nő**
- a standard hiba felnagyítása a megbízhatósági szint függvényében (nem lineáris, bonyolultabb)
- **Ha N nagy, 95%: t=1.96**, N növelésével t csökken (több adat, kisebb intervallum)
- excel: megbízhatósági szint megadható

Hipotézisvizsgálat:

- *Nő-e a végzettséggel a kereset?*
- *Hirdetés pozitívan hat-e az értékesítésre?*
- *Több tanulás növeli-e a várható jegyet?*

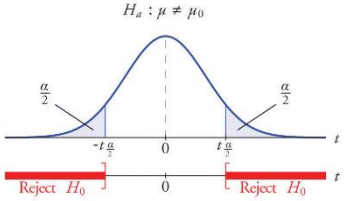
} közgazdaság kérdések, x hat-e y-ra?

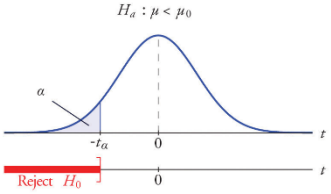
1. meghatározunk egy nullhipotézist (H_0), ami általában az összefüggés hiányára utal (β=0)
   - *pl. „a vállalat reklámra költött költsége nem hat a hirdetésre, β=0”*
2. ezt szembeállítjuk az alternatív hipotézissel (H_1: β≠0) - *„DE van hatás, kapcsolat”*

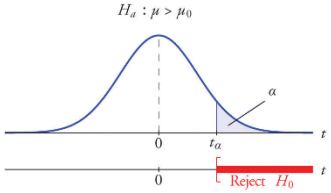
- ezek diszjunkt kategóriak (csak az egyik igaz, de 1 mindig igaz)

3. elvégzünk egy statisztikai próbát, tesztet (vagy a konfidenciaintervallumot hívjuk segítségül), hogy eldöntsük, hogy elvethetjük-e a nullhipotézist
4. Meghatározzuk a szignifikanciát, azaz megvizsgáljuk, mekkora a valószínűsége annak, hogy a nullhipotézis igaz
   - ha H0-nak kicsi a valószínűsége a megfigyelt adatok alapján, elvetjük (szignifikáns)
   - ha nem, nem vetjük el

- <u>kétoldali hipotézis</u> (β=0) vs. egyoldali hipotézis (β >/< 0 - később inkább)







- **feltesszük, hogy H0 igaz -** azt várjuk, hogy a mintából számított β nulla közelében legyen
- mivel a **mintavétel miatt bizonytalanság van a β valós értékét illetően, ezért figyelembe vesszük a szórását** (standard hibát)
- a kívánt bizonyosságnak megfelelő t-értékkel (ami a minta elemszámától és a megbízhatósági szinttől függ) megszorozzuk a standard hibát és levonjuk/hozzáadjuk a nullhipotézisben feltett β értékhez, azaz **megszerkesztjük a konfidencia-intervallumot**, de most nem a becsült, hanem a feltételezett β körül
- **a nullhipotézis alapján azt várjuk, hogy a valós / becsült β ebbe a tartományba essen, ha nem, akkor elutasítjuk a nullhipotézist**
- **β=0 vizsgálata: a becsült β konfidenciaintervalluma tartalmazza-e a 0-t** (csak 0 esetén)

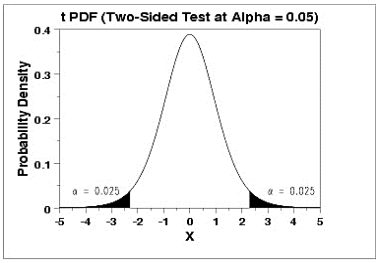
rajz:

- β eloszlása t-eloszlást követ (majdnem olyan, mint a normál eloszlás), közepén a nullhipotézisben megfogalmazott értékkel

- az előbbi módon meghatározzuk az alsó és a felső határokat, és a görbe alatti terület lefedi a lehetséges β értékek 95%-t (ha a megbízhatósági szint 95%), azaz ha a nullhipotézis igaz, akkor 5% az esélye, hogy a β kívül esik ezen intervallumon
- ha a mintából számított β nincs benne az intervallumban, elutasítjuk a nullhipotézist
- A H0-t elvetjük, illetve nem tudjuk elvetni („elfogadni" helytelen mert, ha a H0 bekövetkezésének a valószínűsége 11%, az nem jelenti azt, hogy β=0, az erő az elvetésben van, azaz arról tudunk erős kijelentéseket tenni, ami eléggé valószínűtlen)
- **Ha H0-t elvetjük**, akkor levonhatjuk a következtetést, hogy:
  „X szignifikánsan magyarázza Y-t"
  „β statisztikailag szignifikáns"
  „β szignifikánsan különbözik 0-tól"
  } szignifikanciáról, szignifikanciaszintről beszélünk
- **konfidenciaszint**: bizonyosság szintje (95%-os bizonyossággal állítjuk)
- **szignifikanciaszint**: „tévedés valószínűsége" (5% esélye van, hogy H0 igaz legyen)
  szignifikanciaszint = 100% – megbízhatósági szint
- ha 5%-os szignifikanciaszinten elvetjük H0-t, akkor a 95%-os konfidenciaintervallumban nincs benne a 0

t próba:

- $\left(t = \frac{\hat{\beta} - \beta_0}{s_\beta},\ ha\ H_0\colon \beta = \beta_0\right) \quad \rightarrow \quad t = \frac{\hat{\beta}}{s_\beta}$
- t-értéket kiszámoljuk és megnézzük benne van-e az intervallumban
- a t-próbával kapott t-érték más, mint a konfidenciaintervallum kiszámításához használt! *(ott a mintaelemszám és a kívánt megbízhatósági szint függvényében egy táblázat adta meg, hogy a nullhipotézisben feltett β-tól hány szórásnyi távolságra van az intervallum alsó és felső határa)*
- **a t-próba azt adja meg, hogy a mintából számított β** milyen messze (azaz hány standard hibányira – **hány szórás távolsága) van a feltételezett** (nullhip-ben megfogalmazott) **βtól**
- ha nagy t-értéket ad a t-próba, akkor az arra utal, hogy a rendelkezésre álló adatok tükrében kicsi a valószínűsége, hogy a feltevésünk (nullhipotézis) igaz legyen
- azért standard hibában mérjük az eltérést, mert így figyelembe vesszük, hogy alapból mekkora az adott változóval kapcsolatos bizonytalanság
- a t-érték és a konf.int. számításakor használt „t" (=kritikus t) közti összefüggés: ha t-érték nagyobb, mint a kritikus t, akkor a becsült β kívül esik a feltételezett β0 (β=0) konf.intervallumán -> elutasítjuk a H0-t, β szignifikánsan különbözik nullától
- a kritikus érték (amelynél nagyobb t esetén elvetjük H0-t) t eloszlásától függ
- t-érték egy adott eloszlást követ a H0-t igaznak feltéve, ha a kapott tesztstatisztika értéke elég nagy (|t |>1,96), akkor elvethetjük a β=0-t
- *Az x-tengelyen nem β, hanem t-statisztika van!*



p érték – *azt méri hogy t nagy-e*

- p (=probablity): **annak a valószínűsége, hogy az együttható nulla**, azaz, hogy a kritikus értékeken kívül esik
- pontos definíció: annak a valószínűsége, hogy egy olyan tesztstatisztikát (esetünkben t-értéket) kapjunk, amit kaptunk, amennyiben a nullhipotézis igaz
- használt szignifikanciaszintek: 1%; 5% és (10%)
  pl. 5%-os szignifikanciaszinten, ha a p-érték 5%-nál (0,05-nél) kisebb (nagyobb), akkor elvetjük (nem tudjuk elvetni), hogy β=0

rajz:

- kiszámítottuk a t-értéket a t-próbából, megmutatja, hogy a kapott érték milyen messze van a feltételezettől -> Mekkora a valószínűsége, hogy ilyen (vagy extrémebb) eredményeket kapjunk? -> a görbe alatti terület 1, a görbe szimmetrikus - ki lehet számolni, hogy mekkora a kapott t-érték által kijelölt terület nagysága => ez a p-érték.
- a nagy t-érték kis p-értékkel jár együtt és fordítva, a két érték ugyanazt fejezi ki
- az előzőkhöz hasonlóan bármilyen H_0: $\beta = c$ nullhipotézist tesztelhetünk, a t-próba a korábban leírt módon módosul (standard hiba ugyanaz!)

A szignifikancia indikátorai:

- **becsült együttható és standard hibája**
- **t-érték** (nagy – szignifikáns, azaz elvethetjük) egymásból kiszámíthatóak
- **p-érték** (kicsi – szignifikáns) – van intuitív magyarázata
- **konfidenciaintervallum** (tartalmazza-e nullát)

Hipotézisvizsgálat menete, összefoglalás:

1. Vizsgálandó hipotézis (H0 és H1 megfogalmazása)
2. Statisztikai próba (szignifikanciaszint megválasztása → kritikus érték, teszt statisztika = próba értékének kiszámítása)
3. Döntés (a próba értékét összehasonlítjuk a kritikus értékkel: elvetés vs. nem tudjuk elvetni)

- Excel regressziós tábla: t-értéket, P-értéket is közli
  - ha p-érték<5%: 5%-os szignifikanciaszinten elutasítjuk β=0 hipotézist
  - ha p-érték<1%: 1%-os szignifikanciaszinten elutasítjuk β=0 hipotézist

F próba

- R^2=0 hipotézis vizsgálata (H0: R^2=0 vs. H1: R^2≠0)
- Van-e magyarázóereje a regressziónak?
- Egyváltozós regresszió esetén F próba ekvivalens β=0 tesztelésével
- $F = \frac{(N-2)R^2}{1-R^2}$
- p-érték („F szignifikanciája”) alapján nullhipotézis elfogadása vagy elvetése

példa 1

Regressziós statisztika	
r-négyzet	0.298

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikancia*
Regresszió	1	4.500	4.500	22.735	0.0000033

Regressziós statisztika	
r-négyzet	0.065

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	1	15.988	15.988	1.254	0.277
Maradék	18	229.400	12.744		
Összesen	19	245.388			

példa 2

	Koeff.	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelym.	9.621	1.172	8.208	1.7E-07	7.159	12.084
X változó	-1.2E-05	1.04E-05	-1.120	0.277	-3.35E-05	1.02E-05

r négyzet: mennyivel ad jobb becslést a regresszió, mint ha az átlagot vennénk figyelembe