Konfidencia intervallum, hipotézisvizsgálat

Bevezetés az empirikus elemzésbe – 8. hét

Budapesti Corvinus Egyetem
Corvinus University of Budapest

Tartalom

1. negyedéves dolgozat

Statisztikák

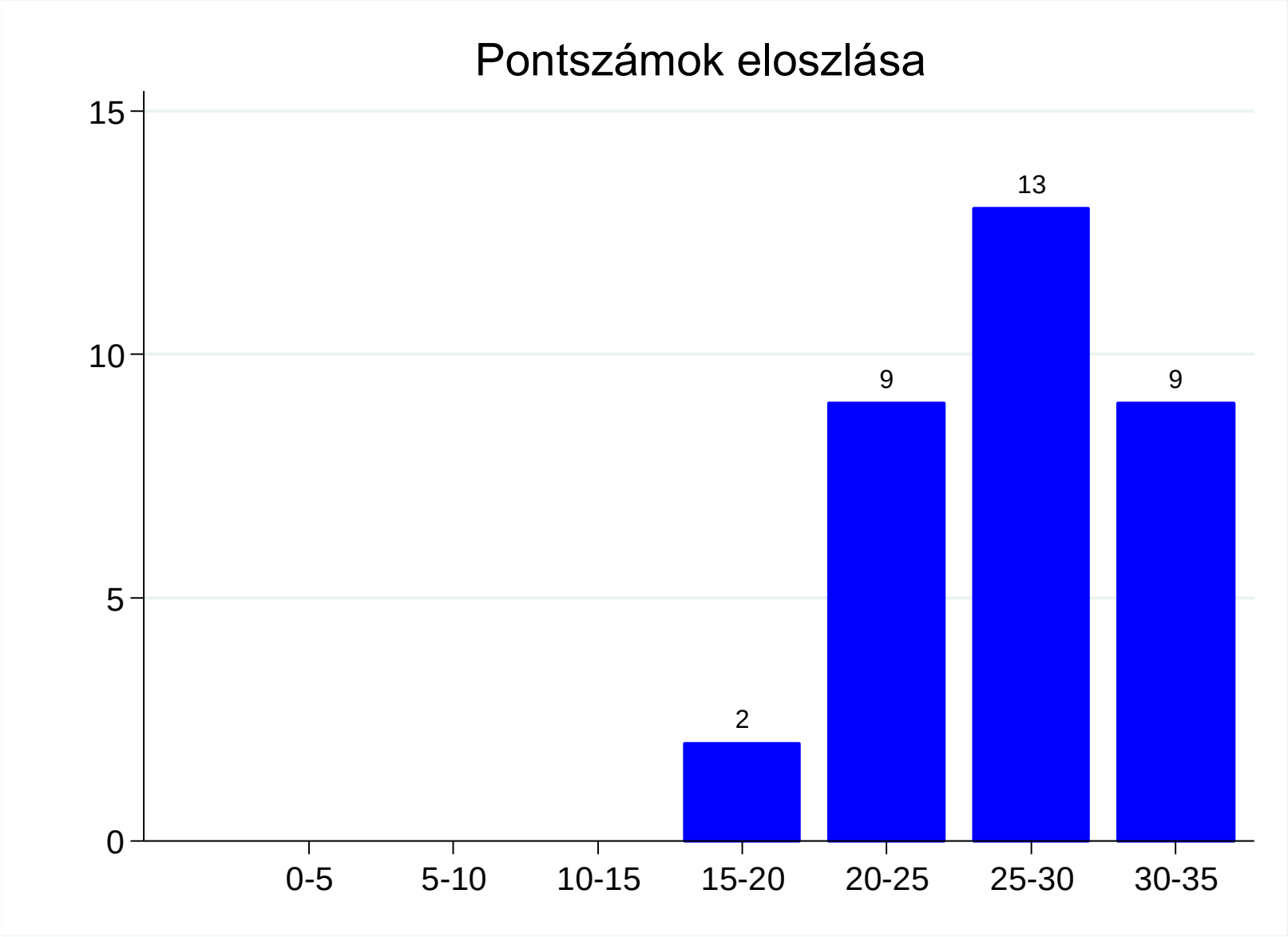
Átlag: 27,1

Minimum: 18,5

Maximum: 33,5

Szórás: 4,1

Medián: 27,5



Problémás kérdések: A csoport

- 2. feladat, a) pont: Cím ne legyen nagyon hosszú, csak az alapinformációt tartalmazza (a részletek legyenek a tengelyeken, illetve a megjegyzésben).
  - Rossz: A 2011-es PPP-dollárban mért egy főre jutó GDP és a százalékos felsőoktatási beiratkozási ráta közötti kapcsolat a világ 182 országában
- 2. feladat, e) pont: -403 értelmezése!
- 3. feladat, b) pont: minőségi ismérv fogalma és példa.
- 4. feladat, d) pont: korrelációs együttható és becsült β közötti kapcsolat.
- 4. feladat, e) pont: Micimackó és a munkapiac.
- 5. feladat.

Problémás kérdések: B csoport

- 1. feladat, d) kérdés: Interkvartilis terjedelem mikor hasznos?
- 2. feladat, a) pont: Cím ne legyen nagyon hosszú, csak az alapinformációt tartalmazza (a részletek legyenek a tengelyeken, illetve a megjegyzésben).
- 2. feladat, c) pont: OLS illesztésének módszere.
- 2. feladat, e) pont: -2471,3 értelmezése!
- 3. feladat, a) pont: mozgó sokaság fogalma és példa.
- 5. feladat.

6. egyéni házi feladat

Gyakori hibák

- Azt az adatot használjuk, amit kell! GDP/fő vs. GDP
- Hiányzó adat: ne pótoljuk 0-kkal!
- Ábrák kinézete: cím, tengelyfeliratok, mértékegységek, forrás stb.
- Még mindig ábrák:
  - munkanélküliségi ráta tengely minek megy 100 fölé (120)?
  - GDP/fő tengely minek megy negatívba?
- Értelmezés: % vs. %pont.
- Értelmezés: ~~egységgel~~
- Értelmezés: ha x 1 egységgel ~~változik~~ növekszik
- **Egyéni** házi feladat

Konfidencia intervallum

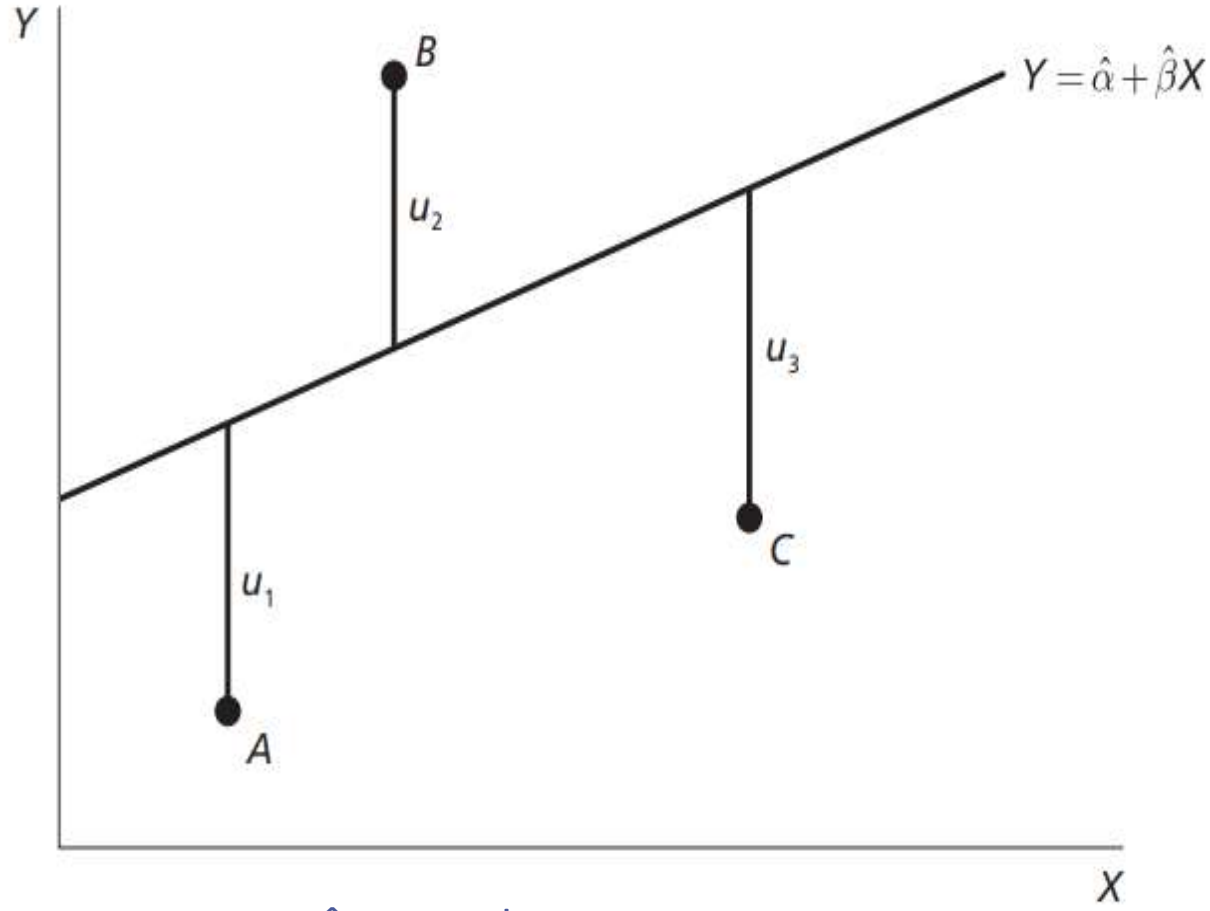
(Ismétlés)

Egyváltozós regresszió – ismétlés: Lakásár és teleknagyság

- X és Y közötti lineáris kapcsolat:
$$Y = \alpha + \beta \cdot X + e$$
  - Probléma: α és β értékeit nem ismerjük, azokat csak becsüljük.

- A becsült modell: $Y = \hat{\alpha} + \hat{\beta} \cdot X + u$

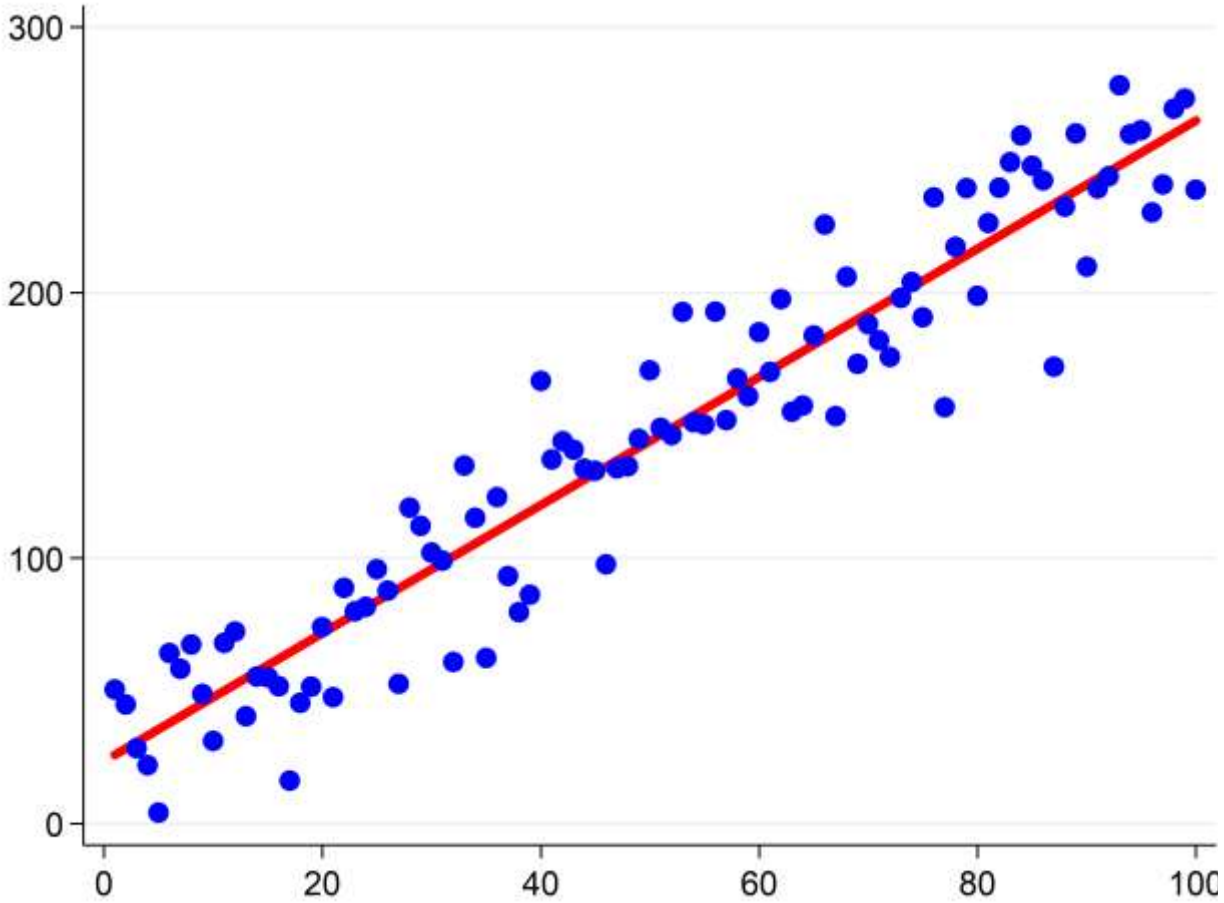
- Az OLS becslés: a $\sum u^2$-t minimalizáljuk.

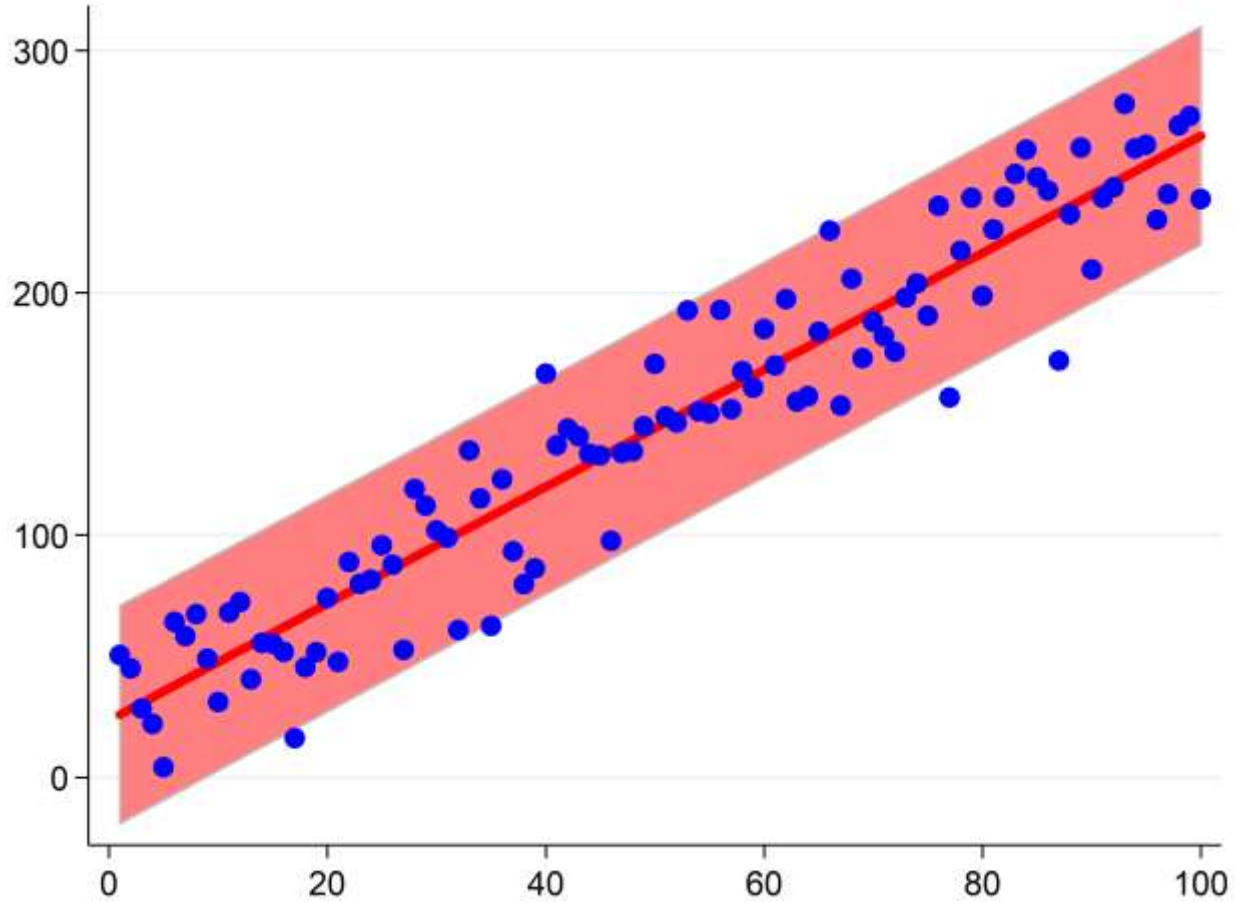


$\hat{\alpha}$ = tengelymetszet

$\hat{\beta}$ = meredekség

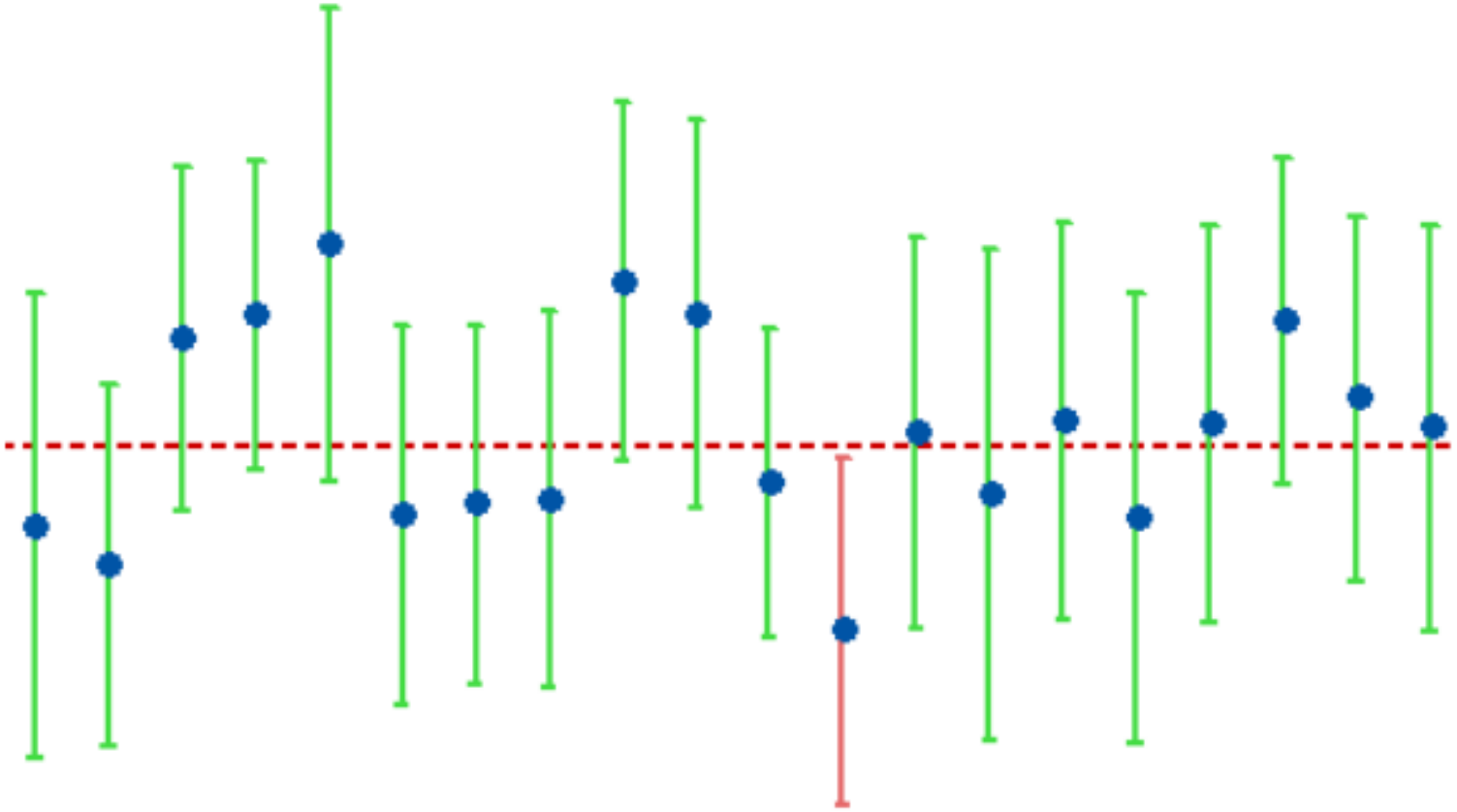
Konfidencia intervallum





Ötlet: Adjunk meg egy értéksávot, ami nagy valószínűséggel tartalmazza a tényleges β értékét!

Konfidencia intervallum: precízebben



Ha újra és újra kiszámítanánk a 95%-os konfidencia intervallumot, akkor az így keletkező intervallumok 95%-a tartalmazná a β valódi értékét.

Forrás: Jim Frost (2019): Regression Analysis – An Intuitive Guide. 55. old.

Konfidencia intervallum

Képlet: $[\hat{\beta} - t_\beta s_\beta, \hat{\beta} + t_\beta s_\beta]$

- Közepe a becsült β (azaz a $\hat{\beta}$).
- Két tényező határozza meg az intervallum nagyságát: s_β ($\hat{\beta}$ szórása) és t_β (Student-féle eloszlásból jön).

Adott megbízhatósági szinthez, minél szűkebb, annál pontosabb a becslés.

Ahogy növeljük az intervallumot, úgy nő a megbízhatósági szint.

Hipotézisvizsgálat

Alapok

Alapok: kérdés megfogalmazása

Kérdések, hipotézisek:

A Föld lapos.
A vádlott ártatlan.
A jelenlegi pontok alapján a csoport 10%-a fog megbukni.
Több tanulás növeli-e a várható jegyet?
Ugyanazon termékek online eladási ára alacsonyabb, mint az offine áruk.
A c.p. magasabb iskolai végzettség nagyobb keresethez vezet.
Hirdetés pozitívan hat-e az értékesítésre?

Alapok: hipotézisvizsgálat

Hipotézisvizsgálat: annak mérlegelése, hogy egy *sokaság*ra vonatkozó adott állítás mennyire hihető a *mintá*n végzett számítások alapján.

Alapok: a döntés

- Döntés: a nullhipotézist elvetjük, vagy nem vetjük el.
  - Elvetjük, ha elég bizonyíték van ellene.
  - Nem vetjük el, ha nincs elég bizonyíték ellene.

- A hipotézistesztelés lépései:
  1. Egy kérdésünkre választ adó statisztika megfogalmazása (ha 1000$-ral többet költünk reklámra, 200$-ral növekszik a bevétel).
  2. Nullhipotézis megfogalmazása (H_0: $\beta = 0{,}2$).
  3. Alternatív hipotézis megfogalmazása a nullhipotézissel szemben (H_1: $\beta \neq 0{,}2$).
  4. Kritikus érték meghatározása.
  5. Tesztstatisztika kiszámítása az adatokból.
  6. Döntéshozatal (statisztikai próba elvégzése) a tesztstatisztika és a kritikus érték összevetése alapján. Két lehetséges döntés:
     a) Elvetjük a nullhipotézist (ha a H_0 valószínűsége kicsi a megfigyelt adatok alapján)
     b) Nem tudjuk elvetni a nullhipotézist.

Alapok: a döntés helyessége

	H_0 igaz	H_0 hamis
H_0-t nem vetjük el	helyes döntés	másodfajú hiba
H_0-t elvetjük	elsőfajú hiba	helyes döntés

Példa:
H_0: ártatlan
H_1: bűnös

A H_0-t védjük, csak akkor vetjük el, ha erős ellene a bizonyíték.

	H_0 igaz	H_0 hamis
H_0-t nem vetjük el	ártatlan felmentése	bűnös felmentése
H_0-t elvetjük	ártatlan elítélése	bűnös elítélése

Hipotézistesztelés a regressziónál: a β=0 hipotézis

Kétoldali hipotézis:

H_0: β=0

H_1: β≠0

Egyoldali hipotézis (egy lehetséges példa)

H_0: β≤0

H_1: β>0

A t-teszt

A t-teszt

- Feltesszük, hogy a nullhipotézis igaz, és bizonyítékot keresünk ellene/mellette.
- A bizonyíték keresése: megnézzük, hogy a kapott statisztika (β) milyen messze van a hipotézisben megfogalmazott értéktől (0-tól).
- Ha a távolság nagy, a nullhipotézist elvetjük, ha kicsi, akkor nem tudjuk elvetni.
- Mi legyen a távolság mérőszáma? A t-statisztika.

$$t = \frac{\hat{\beta} - \beta_0}{s_\beta}, ha\ H_0\colon\ \beta = \beta_0$$

A β=0 hipotézis tesztelése: $t = \frac{\hat{\beta}}{s_\beta}$

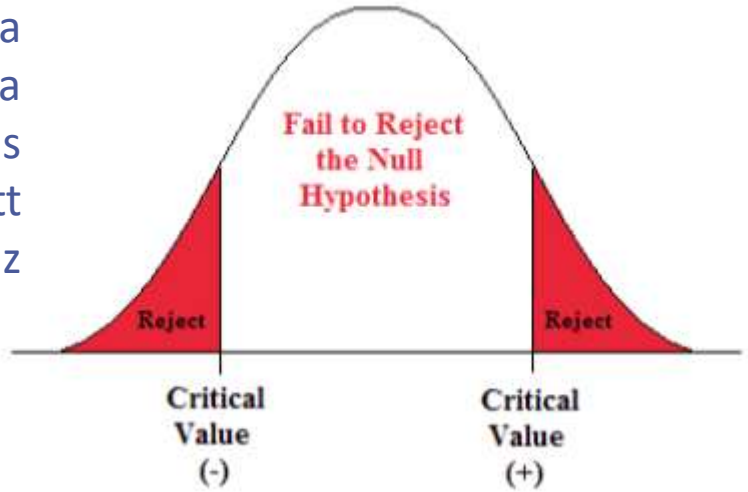
Tehát a **t-próba azt adja meg, hogy a mintából számított β milyen messze** (azaz hány standard hibányira) **van a feltételezett β-tól. Ha nagyon messze** (nagy t-értéket ad a t-próba), **akkor** az arra utal, hogy a rendelkezésre álló adatok tükrében **kicsi a valószínűsége, hogy a feltevésünk (nullhipotézis) igaz legyen**.

A t-teszt

Ha a t-próbából kapott t nagy (nagyobb, mint a **kritikus t érték**), akkor β szignifikánsan különbözik nullától.

A kritikus érték (amelynél nagyobb t esetén elvetjük H0-t) t eloszlásától függ.

Úgy szerkesztjük meg a tartományokat, hogy a próbafüggvény a nullhipotézis fennállása esetén előre megadott nagy valószínűséggel (1-α) az elfogadási tartományba essen.



Ábra forrása: https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-3-hypothesis-testing/

Ha $|t| > 1{,}96$, az elsőfajú hiba elkövetésének valószínűsége 5%.

A p-érték

A p-érték

- Szignifikanciaszint: az elsőfajú hiba elkövetési valószínűségének maximálisan tolerált értéke.
  - Szokásos szintek: 1; 5 és 10%.

- A p-érték (pontos definíció): annak a valószínűsége, hogy egy olyan tesztstatisztikát (esetünkben t-értéket) kapjunk, amit kaptunk, amennyiben a nullhipotézis igaz.
- A p-érték intuitíven: az elsőfajú hiba valószínűségéről informál.
- A p-érték: az a legkisebb szignifikanciaszint, amin H_0 már éppen elvethető.

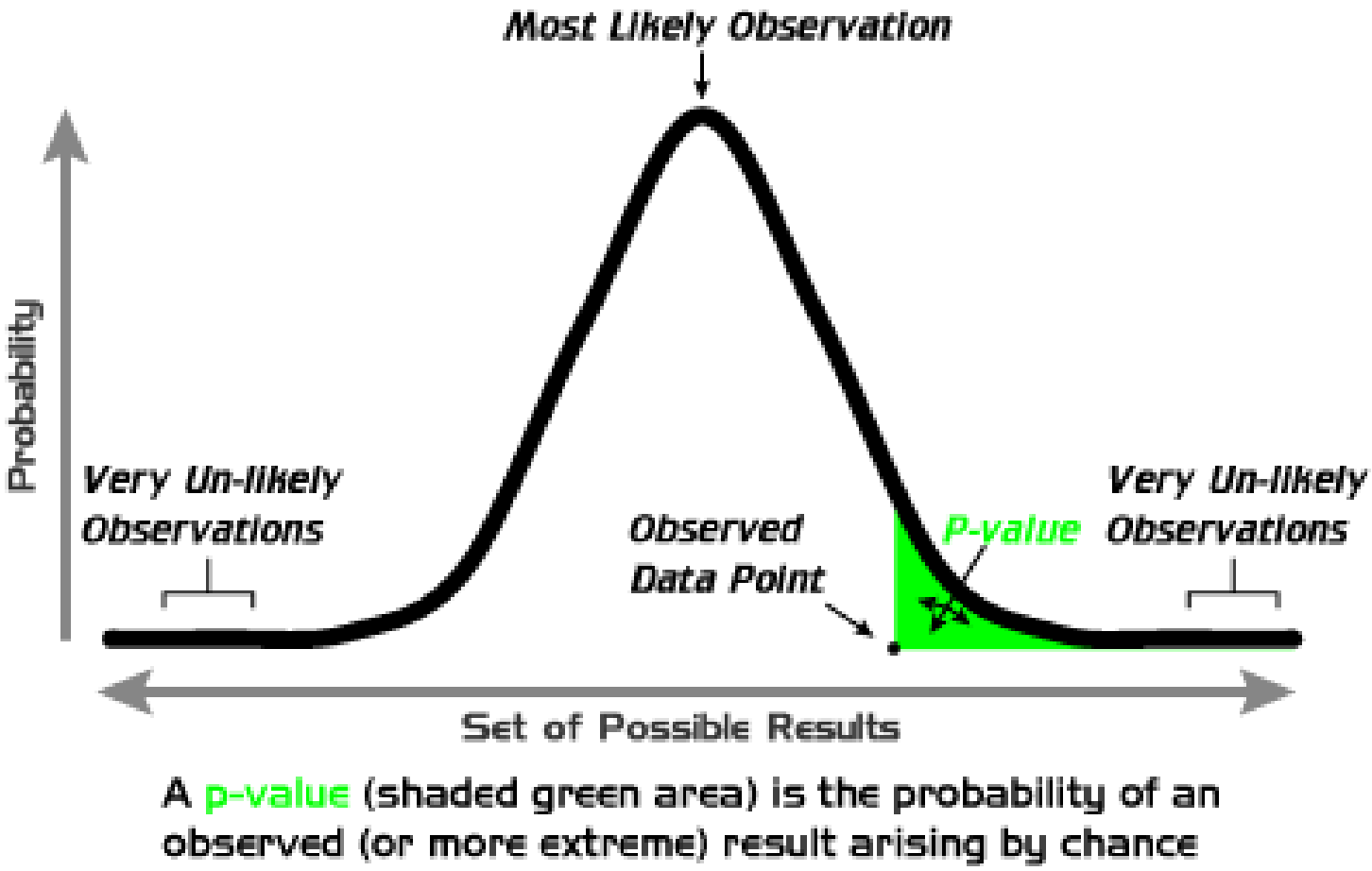
- Ha a p-érték a választott szignifikanciaszintnél kisebb (nagyobb), akkor elvetjük (nem tudjuk elvetni), hogy $\beta=0$.

Hipotézisvizsgálat – p-érték

Ábra forrása: https://med.stanford.edu/news/all-news/2016/03/misleading-p-values-showing-up-more-often-in-journals.html

Rajzon:

Kiszámítottuk a t-értéket a t-próbából feltéve, hogy a nullhipotézis igaz. Megmutatja, hogy a kapott érték milyen messze van a feltételezettől. Kérdés: mekkora a valószínűsége, hogy ilyen (vagy extrémebb) eredményeket kapjunk. A számítás alapja az, hogy a görbe alatti terület 1 és a görbe szimmetrikus, továbbá ki lehet számolni, hogy mekkora a kapott t-érték által kijelölt terület nagysága. Ez nem más, mint a p-érték.



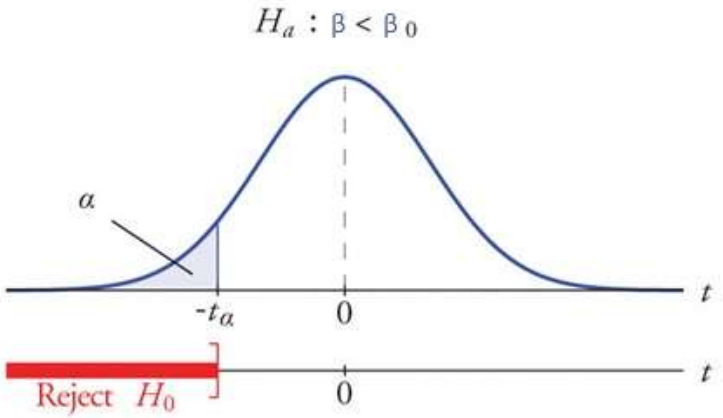
A konfidencia intervallum

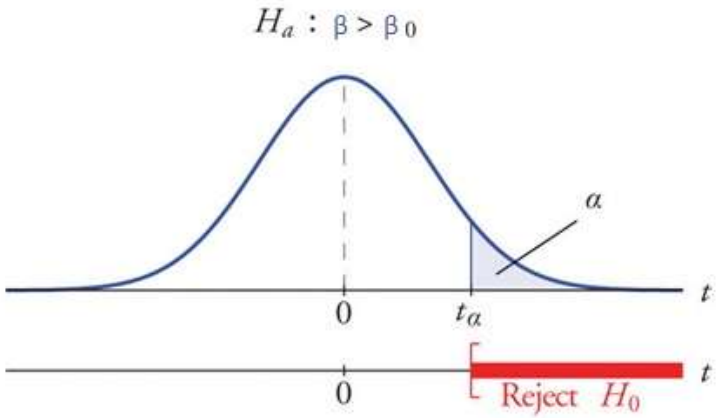
Hipotézisvizsgálat: konfidencia intervallum

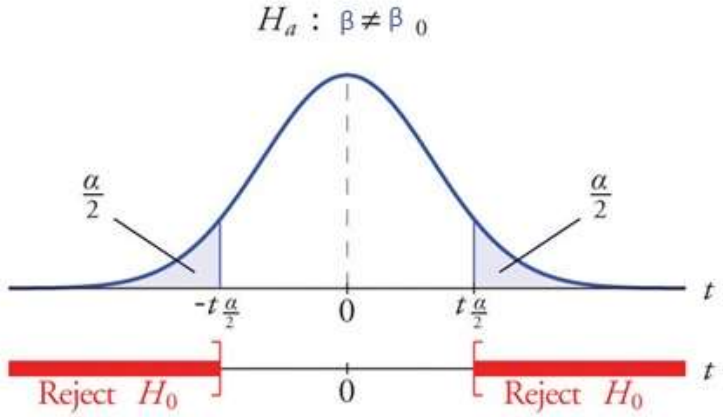
Feltesszük, hogy H0 igaz (vagyis, H0: β=0). Ekkor azt várjuk, hogy a mintából számított β nulla közelében legyen. Mivel a mintavétel miatt **bizonytalanság van a β valós értékét illetően, ezért figyelembe vesszük a szórását** (standard hibát). Nem csinálunk mást, mint a kívánt bizonyosságnak (megbízhatóságnak) 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 ez a helyzet, akkor elutasítjuk a nullhipotézist.**

Rajz: A β eloszlása t-eloszlást követ, 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 (amennyiben a megbízhatósági szint 95%), azaz ha a nullhipotézis igaz, akkor 5% az esélye, hogy a valós / becsült β kívül esik ezen az intervallumon.

Hipotézisvizsgálat: konfidencia intervallum







where α is the level of significance (e.g. 5%)

Hipotézisvizsgálat : konfidencia intervallum

Ha azt látjuk, hogy a mintából számított β nincs benne az intervallumban, akkor elutasítjuk (vagy elvetjük) a nullhipotézist.

A H0-t elvetjük, illetve nem tudjuk elvetni. („Elfogadni” helytelen, bár a könyv engedi. Azért 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.

Következtetések

A szignifikancia indikátorai

Hipotézisek: H_0: β=0 és H_1: β≠0

Ha H_0-t elvetjük, akkor „X szignifikánsan magyarázza Y-t”, vagy „ β” statisztikailag szignifikáns”, vagy „ β szignifikánsan különbözik 0-tól”.

A szignifikancia indikátorai:

- t-érték nagy
- p-érték kicsi
- β konfidencia intervalluma nem tartalmazza a nullát

Kapcsolat a megközelítések között

Hipotézisvizsgálat – kapcsolat a megközelítések között

A β=0 vizsgálata egyenértékű azzal, hogy a becsült β konfidenciaintervalluma tartalmazza-e a 0-t.

Ha 5%-os szignifikanciaszinten elvetjük H0-t, akkor az arra utal, hogy a 95%-os konfidenciaintervallumban nincs benne a 0.

Konfidenciaszint – bizonyosság szintje (95%-os bizonyossággal állítjuk).

Szignifikanciaszint – „tévedés valószínűsége” (5% esélye van, hogy H0 igaz legyen).

Nagy (kis) t-érték kis (nagy) p-értékkel jár együtt és fordítva. A két érték ugyanazt fejezi ki.

Példa – GDP növekedési ráta (%) és népességnövekedési ráta (%)

Adatok és számolás: lásd 6. heti Excel fájl

	Koefficiensek	Standard hiba	t érték	p-érték	Alsó 95%	Felső 95%
Tengelymetszet	3,028042515	0,443752603	6,823718	6,26E-10	2,1479639	3,9081211
popgrow	0,454244166	0,179104813	2,536192	0,012708	0,099032	0,8094563

	Koefficiensek	Standard hiba	t érték	p-érték	Alsó 99,0%	Felső 99,0%
Tengelymetszet	3,028042515	0,443752603	6,823718	6,26E-10	1,863455185	4,192629846
popgrow	0,454244166	0,179104813	2,536192	0,012708	-0,01579972	0,924288048

Az F-próba

F-próba

R2=0 hipotézis vizsgálata (H0: R2=0 vs. H1: R2≠0)

Van-e magyarázóereje a regressziónak?

Egyváltozós regresszió: ekvivalens β=0 tesztelésével

F-próba:

$$F = \frac{(N-2)R^2}{1-R^2}$$

P-érték („F szignifikanciája”) alapján nullhipotézis elfogadása vagy elvetése.

Nagy F olyan, mint nagy t-érték → elvethetjük a nullhipotézist.

Példa – GDP növekedési ráta (%) és népességnövekedési ráta (%)

Adatok és számolás: lásd 6. heti Excel fájl

Regressziós statisztika	
r értéke	0,24244286
r-négyzet	0,05877854
Korrigált r-négyzet	0,049640468
Standard hiba	1,83204093
Megfigyelések	105

F	F szignifikanciája
6,432269	0,012707721

	Koefficiensek	Standard hiba	t érték	p-érték	Alsó 95%	Felső 95%
Tengelymetszet	3,028042515	0,443752603	6,823718	6,26E-10	2,1479639	3,9081211
popgrow	0,454244166	0,179104813	2,536192	0,012708	0,099032	0,8094563

Feladat

Az órai Excel fájlban oldja meg a feladatokat!

Házi feladat

A Moodle platformra feltöltve megtalálja a 8. házi feladatot, kövesse az ott megadott instrukciókat.

A számításai mellé írjon rövid értelmező szövegeket is. Eredményeit a Moodle-re töltse fel.

Köszönöm a figyelmet

Corvinus

www.uni-corvinus.hu

Többváltozós regresszió

Bevezetés az empirikus elemzésbe – 9. hét

Budapesti Corvinus Egyetem
Corvinus University of Budapest

Tartalom

Alapok

Potenciális problémák a regressziós becslésnél: kihagyott változók

Tankönyv releváns része: 109-119. oldal

1. csoportos hf. visszajelzések

Lásd Moodle…

Többváltozós regresszió: alapok

Egyenlet és becslés

Több magyarázó változó

- Regressziós modell *k* darab magyarázó változóval:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + e_i \qquad \forall i$$

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \ldots - \hat{\beta}_k X_{ik})^2$$

- OLS: maradéktagok (SSR) négyzetösszegének minimalizálása

Becsült együtthatók értelmezése

Több magyarázó változó – Példa: ár vs. telekméret

Regressziós statisztika	
r értéke	0,535796
r-négyzet	0,287077
Korrigált r-négyzet	0,285766
Standard hiba	22567,05
Megfigyelések	546

Telekméret: négyzetláb
Ár: dollár

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	1	1,12E+11	1,12E+11	219,0558	6,77E-42
Maradék	544	2,77E+11	5,09E+08		
Összesen	545	3,89E+11			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	34136,19	2491,064	13,70346	6,28E-37	29242,91	39029,47
telekméret	**6,598768**	0,445847	14,80053	6,77E-42	5,722976	7,474559

Több magyarázó változó – Példa: ár vs. telekméret és hálószobák száma

Regressziós statisztika	
r értéke	0,608498
r-négyzet	0,370269
Korrigált r-négyzet	0,36795
Standard hiba	21229,05
Megfigyelések	546

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	2	1,44E+11	7,19E+10	159,6367	2,95E-55
Maradék	543	2,45E+11	4,51E+08		
Összesen	545	3,89E+11			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	5612,6	4102,819	1,367986	0,171882	-2446,74	13671,94
telekméret	**6,053022**	0,424333	14,26479	1,94E-39	5,219487	6,886558
#hálószoba	**10567,35**	1247,676	8,469625	2,31E-16	8116,488	13018,22

Több magyarázó változó – Példa: ár vs. telekméret és hálószobák és fürdőszobák száma

Regressziós statisztika	
r értéke	0,697347
r-négyzet	0,486292
Korrigált r-négyzet	0,483449
Standard hiba	19191,61
Megfigyelések	546

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	3	1,89E+11	6,3E+10	171,0248	5,22E-78
Maradék	542	2E+11	3,68E+08		
Összesen	545	3,89E+11			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	-2418,29	3779,412	-0,63986	0,522534	-9842,38	5005,798
telekméret	**5,4112**	0,38797	13,94749	5,33E-38	4,649092	6,173309
#hálószoba	**5826,802**	1206,571	4,829226	1,79E-06	3456,675	8196,93
#fürdőszoba	**19750,21**	1785,083	11,06403	8,54E-26	16243,68	23256,74

Ceteris paribus

- Ceteris paribus =
  - minden más magyarázó változó értékét rögzítjük
  - a többi változó változatlansága esetén.

CETERIS PARIBUS → OKSÁG

- Előző regresszióra tekinthetünk úgy, mintha pl. rögzítették volna a hálószobák és fürdőszobák számát és megnézték volna a telekméret hatását külön-külön az összes lehetséges esetre, majd valahogy súlyozták volna a kapott eredményeket. Fontos: így mindig hasonló ingatlanokat mérnek össze, csak egy dolog változik – a telekméret.
- „Ha hasonló ingatlanokat (ahol a hálószobák és fürdőszobák száma megegyezik) hasonlítunk össze, akkor átlagosan 5,41 dollárral nagyobb az olyan ingatlanok ára, amelyek 1 négyzetlábbal nagyobbak.”

Regressziós statisztikák értelmezése: t és p

Regressziós statisztikák értelmezése: t és p

- A t-próbát egyenként elvégezhetjük az összes magyarázó változóra ugyanúgy, ahogy azt korábban is tettük. Az értelmezés is azonos. (Így persze a p-érték értelmezése is ugyanaz.)

- A konfidenciaintervallumértelmezése is ugyanaz, mint egyváltozósnál.

Több magyarázó változó – Példa: ár vs. telekméret és hálószobák és fürdőszobák száma

Regressziós statisztika	
r értéke	0,697347
r-négyzet	0,486292
Korrigált r-négyzet	0,483449
Standard hiba	19191,61
Megfigyelések	546

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	3	1,89E+11	6,3E+10	171,0248	5,22E-78
Maradék	542	2E+11	3,68E+08		
Összesen	545	3,89E+11			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	-2418,29	3779,412	-0,63986	0,522534	-9842,38	5005,798
telekméret	5,4112	0,38797	**13,94749**	**5,33E-38**	**4,649092**	**6,173309**
#hálószoba	5826,802	1206,571	**4,829226**	**1,79E-06**	**3456,675**	**8196,93**
#fürdőszoba	19750,21	1785,083	**11,06403**	**8,54E-26**	**16243,68**	**23256,74**

Regressziós statisztikák értelmezése: R^2

Regressziós statisztikák értelmezése: R^2

- R^2 = 1-SSR/TSS = RSS/TSS
  - Illeszkedés mérőszáma.
  - A *modell* a függő változó szóródásának hány százalékát magyarázza.

- R2=0 tesztelése: F-teszt

$$F = \frac{(N-k-1)R^2}{1-R^2}$$

Több magyarázó változó – Példa: ár vs. telekméret és hálószobák és fürdőszobák száma

Regressziós statisztika	
r értéke	0,697347
r-négyzet	**0,486292**
Korrigált r-négyzet	0,483449
Standard hiba	19191,61
Megfigyelések	546

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	3	1,89E+11	6,3E+10	**171,0248**	**5,22E-78**
Maradék	542	2E+11	3,68E+08		
Összesen	545	3,89E+11			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	-2418,29	3779,412	-0,63986	0,522534	-9842,38	5005,798
telekméret	5,4112	0,38797	13,94749	5,33E-38	4,649092	6,173309
#hálószoba	5826,802	1206,571	4,829226	1,79E-06	3456,675	8196,93
#fürdőszoba	19750,21	1785,083	11,06403	8,54E-26	16243,68	23256,74

t-próbák és az F-próba

- Az F-próba azt nézi meg, hogy a magyarázóváltozók *együtt* statisztikailag szignifikáns mértékben magyarázzák-e meg a függő változót. Lehetséges, hogy az R-négyzet alacsony, azonban az F-próba azt mutatja, hogy a magyarázó változók összességében segítenek megérteni a függő változó változékonyságát.

- Az is lehetséges, hogy egy (vagy több) magyarázó változó nem szignifikáns, de az F-próba azt mutatja, hogy összességében a magyarázó változók jól magyaráznak. (Felveti azt a kérdést, hogy melyik magyarázó változókat szerepeltessük a regresszióban. Válasz később.)

1. feladat

Az órai excel fájlban oldja meg az **alapok** munkalapokon található feladatokat.

Regressziós statisztikák értelmezése: az R^2 „problémája”

- $R^2 = 1\text{-}SSR/TSS = RSS/TSS$
- Vegyük észre, hogy új magyarázó változó bevonásával az R^2 értéke biztosan növekszik (legalábbis nem csökken).
- Tekintsük az „alapok” munkalap utolsó feladatát! Az R^2 növekszik, de mi a sztori?

- Megoldás: korrigált (adjusztált) R^2
  - Jutalmazza az illeszkedésé javulását.
  - Bünteti az új magyarázó változók bevonását.

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}$$

2. feladat

Az órai excel fájlban oldja meg a **becsles** munkalapokon található feladatokat.

Egyváltozós vs. többváltozós regresszió

Több magyarázó változó: pontosabb következtetések

	Koefficiensek	*p-érték*
Konstans	34 136,192	0,000
telekméret	**6,599**	**0,000**

	Koefficiensek	*p-érték*
Konstans	5 612,600	0,172
telekméret	**6,053**	**0,000**
#hálószoba	10 567,352	0,000

	Koefficiensek	*p-érték*
Konstans	-2 418,293	0,523
telekméret	**5,411**	**0,000**
#hálószoba	5 826,802	0,000
#fürdőszoba	19 750,210	0,000

- A telekméret másként hat az árra egyváltozós regressziónál (6.6), mint a többváltozósnál (5.4).
- Ok: először csak egy ismérv szerint hasonlítjuk össze az árakat, míg utóbb több változót rögzítünk és úgy nézzük meg a telekméret hatását.
- Mivel a változók korrelálnak (nagyobb telek, nagyobb ház, több szoba), így az egyváltozós regresszió felveszi ezen egyéb változók hatásait is, azokat nem tudjuk elkülöníteni az egyváltozós regresszióban.
- Többváltozós regresszió informatívabb, pontosabb következtetések vonhatóak le a segítségével.

3. feladat

Az órai excel fájlban oldja meg a **tobbvaltozo** munkalapokon található feladatokat.

A több változó jobb?

- Tekintsünk egy szimulált adatbázist, amelyben 100 nő és 100 férfi van. A nők bére 100 ezer és 200 ezer, a férfiak bére pedig 150 ezer és 250 ezer között van.
- Létrehoztunk továbbá egy tapasztalat változót a következőképp:

Tapasztalat = Bér/10.000 * r,

ahol r egy 0 és 1 közötti véletlen szám.

- Az órai excel fájlban oldja meg a **Gender wage gap - artifical** munkalapokon található feladatokat.

A több változó jobb?

- Látszik, hogy az egyváltozós regresszióban a nem felvette a tapasztalat hatását is, ezért kaptuk azt, hogy a nők szignifikánsabban kevesebbet keresnek.

- Azonban, ha **kontrollálunk** arra, hogy kinek mekkora a tapasztalata, akkor ez a különbség eltűnik. (A valós életben nem, csak a példában.)

- Azaz ha ugyanakkora tapasztalattal rendelkező férfiakat és nőket hasonlítunk össze, akkor nem igaz, hogy a férfiak számottevően többet keresnek.

Többváltozós regresszió: kihagyott változók problémája (OVB)

Több magyarázó változó – Példa: ár

	Koefficiensek	*p-érték*
Konstans	34 136,192	0,000
telekméret	**6,599**	0,000

	Koefficiensek	*p-érték*
Konstans	5 612,600	0,172
telekméret	**6,053**	0,000
#hálószoba	10 567,352	0,000

	Koefficiensek	*p-érték*
Konstans	-2 418,293	0,523
telekméret	**5,411**	0,000
#hálószoba	5 826,802	0,000
#fürdőszoba	19 750,210	0,000

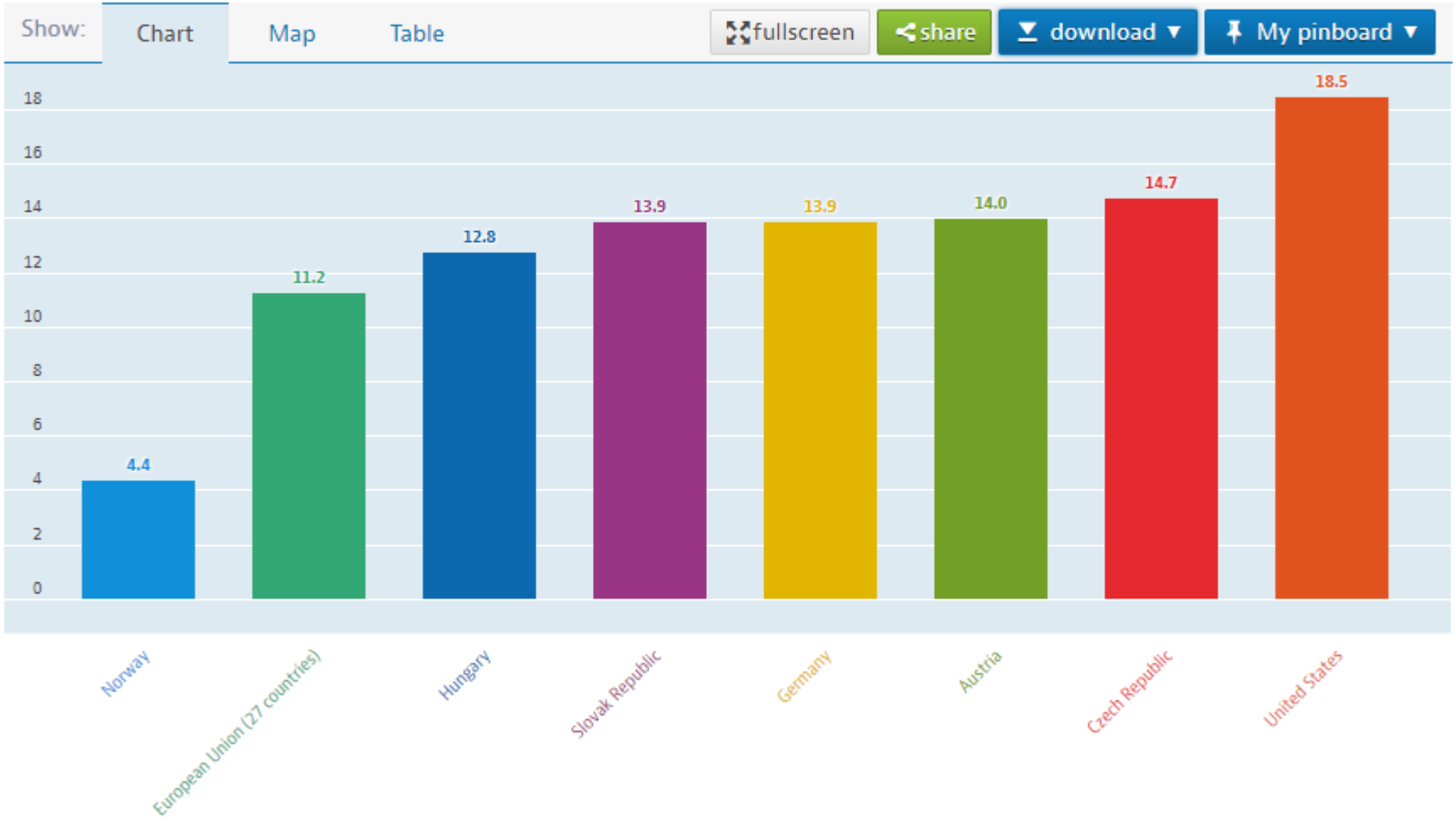
6.3. táblázat. *A házárakról szóló példa változóinak korrelációs mátrixa*

	Eladási ár	Telekméret	Hálószobák száma	Fürdőszobák száma	Emeletek száma
Eladási ár	1				
Telekméret	0,535795	1			
Hálószobák száma	0,366447	0,151851	1		
Fürdőszobák száma	0,516719	0,193833	0,373768	1	
Emeletek száma	0,421190	0,083674	0,407973	0,324065	1

Egy példa: nemi bérkülönbségek

Kép forrása: https://www.unicef.ie/itsaboutus/cards/unicef-itsaboutus-gender-sexism.pdf

A férfiak és a nők medián bére közötti különbség a férfiak mediánbérének arányában



Forrás: OECD.

Kihagyott változók miatti torzítás: a probléma

- Omitted variable bias.
- Kihagyott változók miatti torzítás:
  - A becslés hibás (torzított), ha releváns változót kihagyunk, ami korrelál az egyenletben szereplő változókkal.
- Megoldás: a magyarázóerővel bíró változókat szerepeltessük!

- Óhatatlanul kihagyunk fontos változókat. Például: ingatlanáraknál hogyan mérjük a szomszédok kedvességét?

Kihagyott változók miatti torzítás: a megoldás okozta probléma

- Azonban az se jó, ha sok bába között elvész a gyerek és nincs világos üzenet. Szólnak érvek a kevés változó mellett.
- A felesleges változók csökkentik a becslés pontosságát. Pontosabban: lényegtelen változók bevonása növeli az együttható p-értékét (és konfidencia-intervallumát), a fontos változókét is.
  - Példa: a falak színe a házáras regresszióban.

Kihagyott változók miatti torzítás

- Egyensúly: kihagyott vs. felesleges változók?
- **Mindig legyen egy elmélet!!!**
  - Kezdjünk számos releváns(nak vélt) változóval.
  - Majd dobjuk ki azokat, amelyek nem szignifikánsak.
  - Először a legkevésbé szignifikánsat hagyjuk el (de csak akkor ha nem feltétlenül szükséges megtartani), majd újrafuttatjuk a regressziót. Az új regresszióból aztán megint elhagyjuk a legkevésbé szignifikánsat stb. Nem mindig vezet egyértelmű eredményre. (Az is előfordulhat, hogy egy változó korábban szignifikáns volt, aztán amikor elhagyunk egy másik változót, akkor már nem lesz szignifikáns.)
  - Mindig legyen egy elmélet, ami irányít minket!
- Gyakorlat, szakterületi tudás és józan ész kell a jó regresszióhoz.

5. feladat

Az órai excel fájlban oldja meg a **gender_wage** munkalapokon található feladatokat.

Kihagyott változók miatti torzítás: összefoglalás

- Elméleti összefüggés: $Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot Z + e$
- Tegyük fel, hogy Z nem mérhető, így a becsült modell: $Y = \hat{\alpha} + \hat{\beta}_1 \cdot X + u$

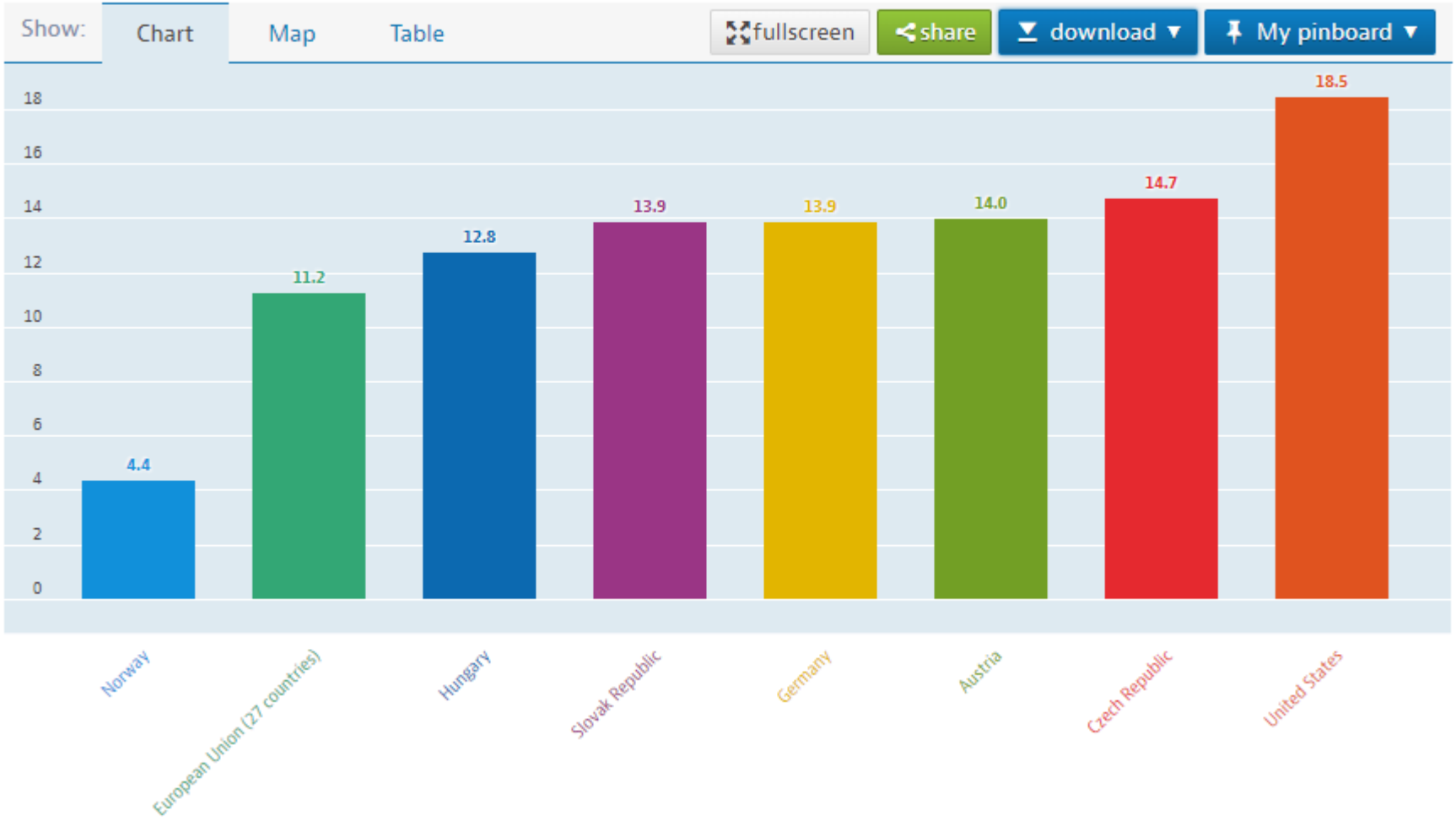
- Ha $Corr(X, Z) \neq 0$, akkor a $\hat{\beta}_1$ torzított.
- A torzítás iránya:

	$Corr(X, Z) > 0$	$Corr(X, Z) < 0$
$\beta_2 > 0$	Pozitív torzítás	Negatív torzítás
$\beta_2 < 0$	Negatív torzítás	Pozitív torzítás

Egy példa: nemi bérkülönbségek

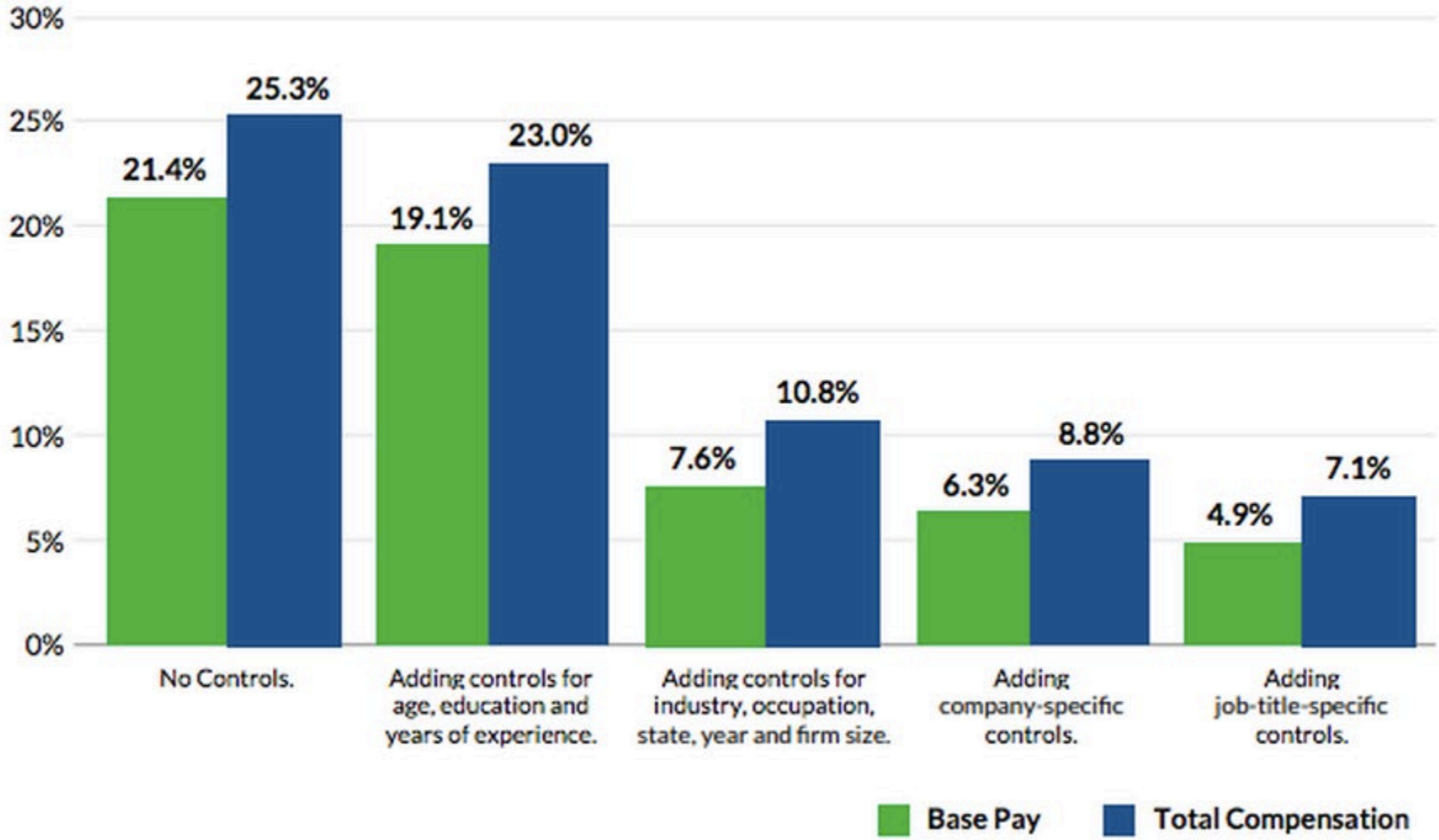
Kép forrása: https://www.unicef.ie/itsaboutus/cards/unicef-itsaboutus-gender-sexism.pdf

A férfiak és a nők medián bére közötti különbség a férfiak mediánbérének arányában



Forrás: OECD.

U.S. Gender Pay Gap, Before and After Adding Statistical Controls



Source: Glassdoor Economic Research (Glassdoor.com/research).

Egy másik példa: roma és nem-roma tanulok tesztpontszámai

TABLE 3—THE ETHNIC GAP IN READING AND MATHEMATICS: UNCONDITIONAL AND CONDITIONAL ON CONTROL VARIABLES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Reading							
Gap	−0.97	−0.87	−0.38	−0.25	−0.16	−0.11	−0.05
[S.E.]	[0.05]**	[0.05]**	[0.05]**	[0.06]**	[0.07]*	[0.05]*	[0.07]
Observations	9,056	9,056	9,056	9,056	9,056	9,056	9,056
R^2	0.06	0.09	0.25	0.53	0.66	0.33	0.68
Panel B. Mathematics							
Gap	−1.05	−0.94	−0.51	−0.33	−0.28	−0.22	−0.15
[S.E.]	[0.05]**	[0.05]**	[0.05]**	[0.05]**	[0.07]**	[0.05]**	[0.07]*
Observations	8,335	8,335	8,335	8,335	8,335	8,335	8,335
R^2	0.07	0.10	0.23	0.54	0.67	0.32	0.69
Control variables							
Health		Yes	Yes	Yes	Yes	Yes	Yes
Home environment			Yes	Yes	Yes	Yes	Yes
School FE				Yes	Yes		Yes
School × Class FE					Yes		Yes
Family background						Yes	Yes

Notes: OLS estimates of the Roma coefficient in seven specifications. Standard errors in brackets are clustered at the school level.

**Significant at the 1 percent level.

*Significant at the 5 percent level.

Forrás: Kertesi, G, Kézdi, G (2011): The Roma/Non-Roma Test Score Gap in Hungary. Amerian Economic Review 101(3): 519-525.

6. feladat

Az órai excel fájlban oldja meg az **osszefoglalo** munkalapokon található feladatokat.

Házi feladat

Használja a Moodle felületen megosztott, hasznaltauto.hu honlapról származó valós adatokat tartalmazó adatbázist és válaszolja meg a kérdéseket!

Köszönöm a figyelmet!

Corvinus

Corvinus

www.uni-corvinus.hu

Többváltozós regresszió - 2

Bevezetés az empirikus elemzésbe – 10. hét

Budapesti Corvinus Egyetem
Corvinus University of Budapest

Tartalom

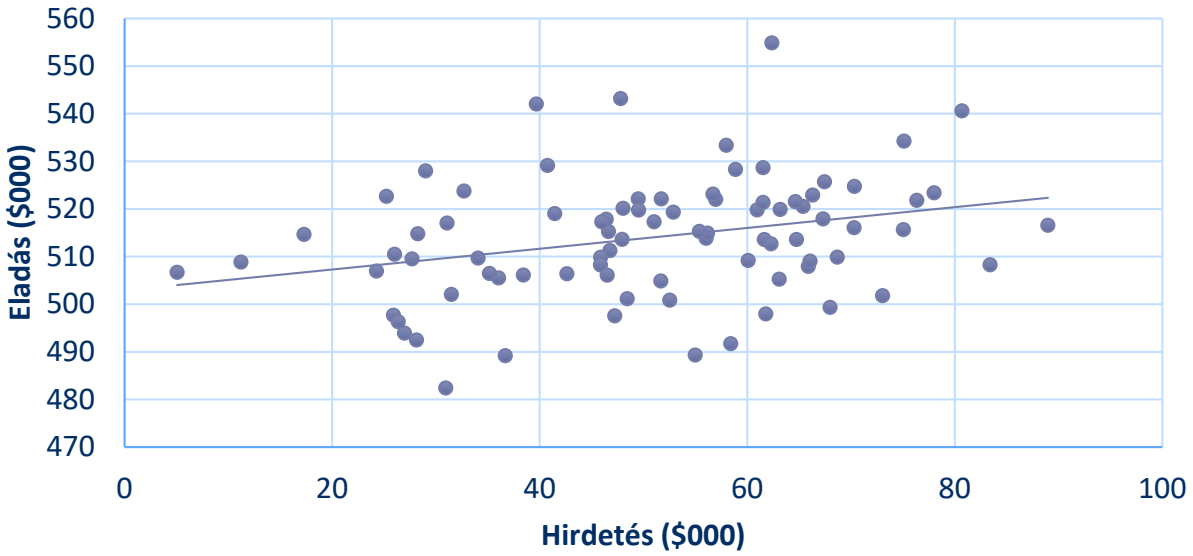
Tankönyv releváns része: 119-125.,
valamint 127-138. oldal

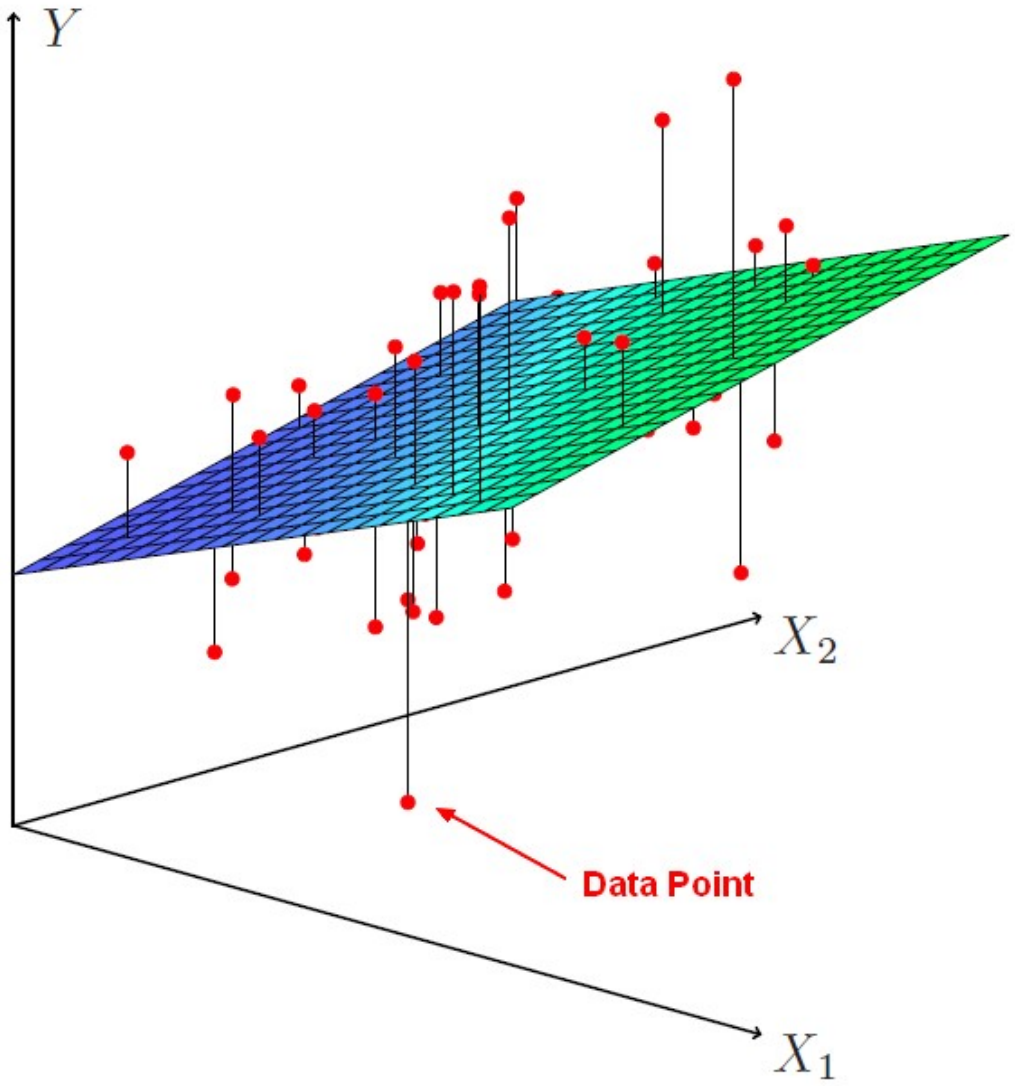
Ismétlés

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + e_i \qquad \forall i$$

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \ldots - \hat{\beta}_k X_{ik})^2$$

Hirdetések és eladások kapcsolata





Forrás: https://datacadamia.com/data_mining/multiple_regression

Ismétlés: többváltozós regresszió outputjának értelmezése

Regressziós statisztika	
r értéke	0,843374
r-négyzet	0,71128
Korrigált r-négyzet	0,709827
Standard hiba	1672460
Megfigyelések	1000

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	5	6,84954E+15	1,36991E+15	489,7557	3,6E-265
Maradék	994	2,78034E+15	2,79712E+12		
Összesen	999	9,62988E+15			

	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelymetszet	-4,08E+08	25440331,65	-16,02490772	1,39E-51	-4,6E+08	-3,6E+08
évjárat	204057,2	12675,99512	16,09792319	5,48E-52	179182,4	228932
lóerő (W)	32316,07	2055,543183	15,72142424	6,61E-50	28282,36	36349,77
hengerűrtartalom (cm3)	-324,866	187,6650983	-1,731094652	0,083745	-693,131	43,3992
megtett kilométerek száma (km)	-9,051603	0,724151456	-12,49959913	2,1E-33	-10,4726	-7,63056
extrák száma (darab)	-140583,9	21998,18719	-6,390702022	2,53E-10	-183752	-97415,6

Ismétlés: releváns kihagyott változó okozta torzítás

	Koefficiensek	*p-érték*
Konstans	34 136,192	0,000
telekméret	**6,599**	0,000

	Koefficiensek	*p-érték*
Konstans	5 612,600	0,172
telekméret	**6,053**	0,000
#hálószoba	10 567,352	0,000

- A két specifikációban a telekméretre eltérő együtthatót kapunk.

- Az első specifikációból kihagytunk egy fontos változót, a hálószobák számát.
- A telekméret és a hálószobák száma között van kapcsolat:
  Corr(telekméret, hálószobák száma) = 0,367.

Az első specifikációban a telekméret „felveszi” a hálószobahatás egy részét, vagyis a telekméret együtthatója *torzított*.

1. feladat

Az órai excel fájlban oldja meg az **ismetles_OVB** munkalapon található feladatokat!

Potenciális problémák a regresszióban: multikollinearitás

Multikollinearitás: A probléma

- Magyarázó változók némelyike erősen korrelál, és így nehéz megállapítani, hogy melyik változó befolyásolja leginkább a függő változót. Egyes változók hatása nehezen elkülöníthető.

- Speciális eset: egzakt multikollinearitás → az egyik érintett változóra nem becsül együtthatót.

Multikollinearitás: Tünetek

Alacsony t-, magas p-értékek, ugyanakkor R2 magas.

- Azaz az együtthatók együttes magyarázó ereje magas, de egyes változók nem szignifikánsak.
- A regresszió nem képes eldönteni, hogy a magyarázó erőt melyik változókhoz rendelje.

Együtthatók nagyon érzékenyek újabb (kollineáris) változó bevonására.

Várttól jelentősen eltérő (akár értelmezhetetlen) együtthatók.

Multikollinearitás: Megoldás

- Fő probléma: a multikollinearitás adatprobléma és nem modellprobléma.
- Egyes változók elhagyása, a korreláló változók közül a fontosabb megtartása.
  - De ez nem mindig kívánatos. Házáras példában magyarázó változók erősen korrelálnak, de együttes szerepeltetésük indokolt, hiszen józan ész alapján hatnak a házárra.
  - Munkavállaló bére: kor/tapasztalat/iskolázottság? Mi érdekel, mi a sztori?
- Változócsoportok létrehozása.
- Ha előrejelzés a cél, akkor nem igazán releváns a probléma.

2. feladat

Az órai excel fájlban oldja meg a **no multicollinearity** és a **multicollinearity** munkalapokon található feladatokat!

Dummy változók és eltérő meredekségek

Ismétlés: mennyiségi vs. minőségi ismérvek

Ismétlés: mennyiségi vs. minőségi ismérvek (1. hét)

- Mennyiségi (kvantitatív) ismérv:
  - A sokaság egységeihez valamilyen mérés vagy számlálás eredményét rendeli hozzá.
  - Számokkal *leírható*.
  - Pl.: infláció, jövedelem, munkanélküliség

- Minőségi (kvalitatív) ismérv:
  - A sokaság egységeit verbálisan jellemzi.
  - Számokat *rendelünk hozzá* (önkényesen).
  - Pl.: nem (0=nő és 1=férfi), legmagasabb iskolai végzettség (1=alapfok, 2=középfok, 3=felsőfok).

Eddigi példáink

- Reklámra költött összeg (ezer $) → Árbevétel (ezer $)
- Népsűrűség (fő/ezer hektár) → Erdőirtási ráta (%)
- Kor (év) → Jövedelem (HUF)
- Népességnövekedési ráta (%) és GDP-arányos beruházás (%) → GDP növekedési ráta (%)
- Munkanélküliségi ráta (%) → GDP/fő (ezer HUF)
- Telekméret (négyzetláb) és fürdőszobák száma (darab) és hálószobák száma (darab) → Ingatlan ára (CAD)
- Stb…

Közös az összesben: a változóink mennyiségi változók!

Értelmezés megegyezik: Ha 1 egységgel növekszik X, akkor β egységgel változik Y (c.p.).

Minőségi változók esete

- Nem (0=nő és 1=férfi) → Jövedelem (HUF)
- Iskolázottság (1=alapfok, 2=középfok, 3=felsőfok) → Jövedelem (ezer HUF)
- Iparág (3=pénzügyi szektor, 5=élelmiszeripar, 10=nehézipar) → Jövedelem (ezer HUF)

Közös az összesben: a magyarázó változóink *minőségi* változók!

Alkalmazzuk az eddigi értelmezésünket a három fenti példára: Ha 1 egységgel növekszik X, akkor β egységgel változik Y (c.p.). Értelmesnek tűnik?

3. feladat

Az órai Excel fájlban oldja meg az **onkenyes** munkalapokon található feladatokat!

Minőségi ismérvek kezelése a regresszióban: bináris változók

Kétértékű (bináris) kvalitatív változók esete

- A kétérétkű kvalitatív változót számokkal írjuk le: 0 és 1
- Példák:
  - Házárak: van-e garázs, van-e légkondicionáló
  - Bérek: férfi – nő
  - Egészségügyi kiadások: van-e biztosítása?
  - stb.

Esetek

1. Egyváltozós regresszió kétértékű magyarázó változóval:

$$Y_i = \alpha + \beta_1 D_{i1} + e_i$$

2. Többváltozós regresszió kétértékű változókkal:

$$Y_i = \alpha + \beta_1 D_{i1} + \ldots + \beta_k D_{ik} + e_i$$

3. Többváltozós regresszió kétértékű és nem kétértékű változókkal: legegyszerűbb eset

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + e_i$$

4. Interakció:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i + e_i$$

5. Regresszió, amiben a függő változó kétértékű

Egyváltozós regresszió kétértékű magyarázó változóval

(1) Egyváltozós regresszió kétértékű változóval – Becslés, együtthatók

OLS módszer változatlan, együtthatók értelmezése kissé más

Egyváltozós regresszió:

$$Y = \alpha + \beta D + e$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta} D$$

$$\hat{Y} = \hat{\alpha}, \text{if D} = 0$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}, \text{if D} = 1$$

Tengelymetszet értelmet nyer, béta értelmezése hasonló, mint korábban.

D=0 a benchmark (vagy referenciakategória, vagy baseline, vagy alapcsoport), amelyhez hasonlítunk.

Példák

1. Házárak

$$\hat{P} = 59\,885 + 25\,996 aircond$$

Nem / légkondicionált ház átlagár: 59 885 / 85 881 CAD
Mennyibe kerülhet egy légkondi? Mi a furcsa az eredményben és miért?

2. Bérek (Bértarifa, 2003, részminta, HUF)

$$\hat{W} = 159\,289 + 66\,854 male$$

Férfiak átlagkereset: 226 142 Ft
Nők átlagkereset: 159 289 Ft

Többváltozós regresszió kétértékű változókkal

(2) Több dummy változó

$$Y_i = \alpha + \beta_1 D_{i1} + \ldots + \beta_k D_{ik} + e_i$$

Csoportok száma: 2^k.

Csoportátlagok: megfelelő együtthatók összege.

Együttható értelmezése: parciális hatás.

(2) Több dummy változó: példa - 4. feladat

Az órai Excel fájlban oldja meg az **HPRICE_only dummy** munkalapon található feladatokat.

(2): Több dummy változó – egy speciális eset: kvalitatív változók és a dummy készlet

Az órai Excel fájlban oldja meg a **region** munkalapon található feladatokat!

Kétértékű és nem kétértékű magyarázó változók

(3) Kétértékű és nem kétértékű magyarázó változók

Csak kétértékű: eltérő átlagok

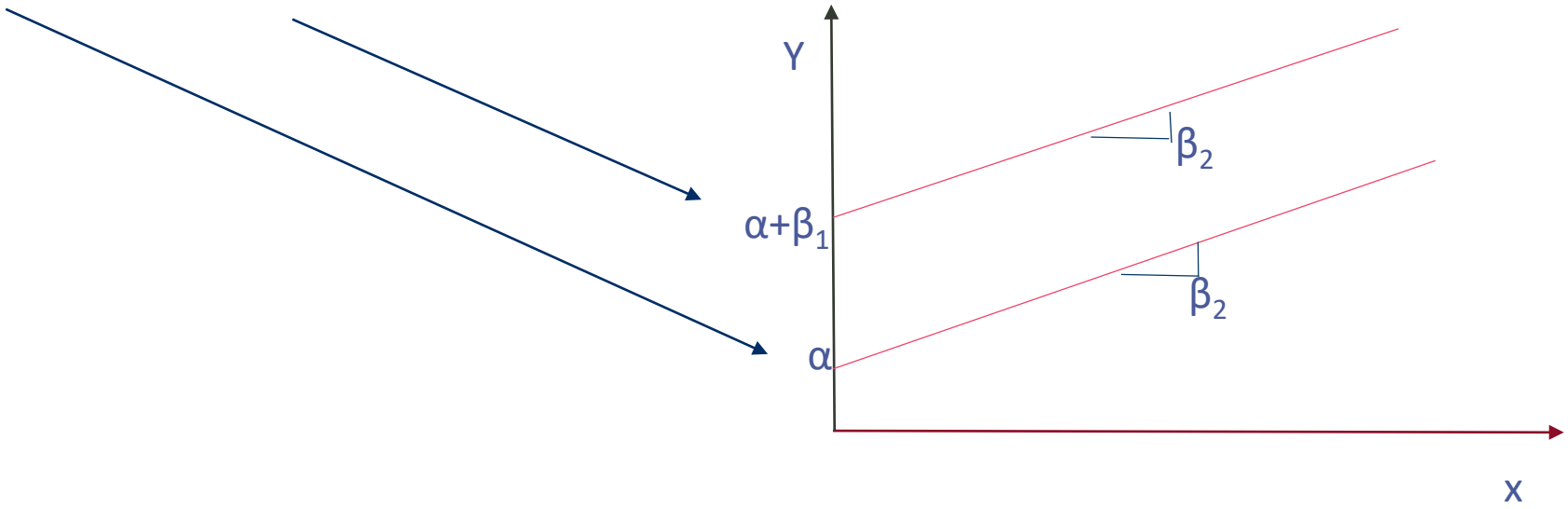
Kétértékű és nem kétértékű: eltérő tengelymetszet, de meredekség nem változik (ez mit jelent?).

Legegyszerűbb modell :

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + e_i$$

$$\text{Intercept} : \alpha \text{ or } \alpha + \beta_1$$

Ábra, ha D=0 és D=1.



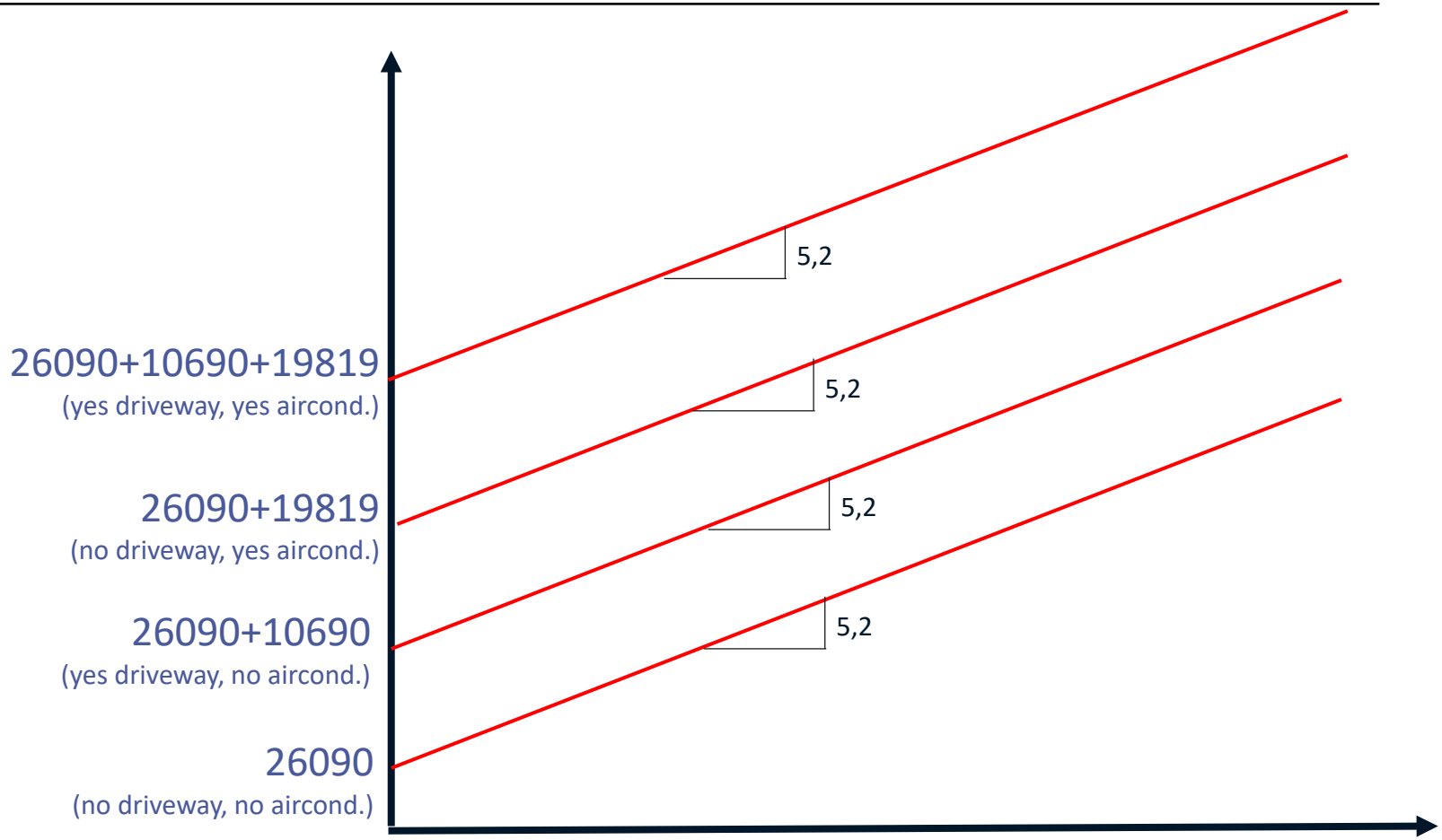
Példa

Házárat magyarázzuk telekmérettel, kocsibeállóval és légkondival.

	Coefficient	*Standard error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	26090.2	2770.2	9.4	0.0	20648.6	31531.8
lot size	5.2	0.4	12.0	0.0	4.3	6.0
driveway	10690.4	2615.9	4.1	0.0	5551.9	15829.0
air cond	19819.2	1921.4	10.3	0.0	16044.8	23593.5

Hogy néz ki a rajz? Hány tengelymetszet van?

	Coefficient	*Standard error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	26090.2	2770.2	9.4	0.0	20648.6	31531.8
lot size	5.2	0.4	12.0	0.0	4.3	6.0
driveway	10690.4	2615.9	4.1	0.0	5551.9	15829.0
air cond	19819.2	1921.4	10.3	0.0	16044.8	23593.5



Interakció

(4) Interakció

Két változó hatása összefügg – szorzat (interakció)

Folytonos X és bináris D interakciója:

- X hatása eltér a két csoportban
- Eltérő meredekség és eltérő tengelymetszet :

$$Y = \alpha + \beta_1 X + \beta_2 D + \beta_3 (DX) + e$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X \text{ if } \mathrm{D} = 0$$

$$\hat{Y} = \left(\hat{\alpha} + \hat{\beta}_2\right) + \left(\hat{\beta}_1 + \hat{\beta}_3\right) X \text{ if } \mathrm{D} = 1$$

(Megjegyzés: Lehet két dummy között is interakció.)

Interakció – Miért fontos?

Az egyes változók elkülönített hatásán túl lehet, hogy létezik egy együttes hatás is.

Példa 1: A bérek magyarázatában fontos lehet, hogy valaki nő-e. A nőknél más lehet a tapasztalat hatása, mert kevesebb van nekik a szülés miatt: így egy év plusz tapasztalat többet érhet.

Példa 2: A házáraknál lehet, hogy más a hatása a telekméretnek, ha városi vagy nem-városi ingatlanokról beszélünk. Ezen hatásokat az interakciós taggal tanulmányozhatjuk.

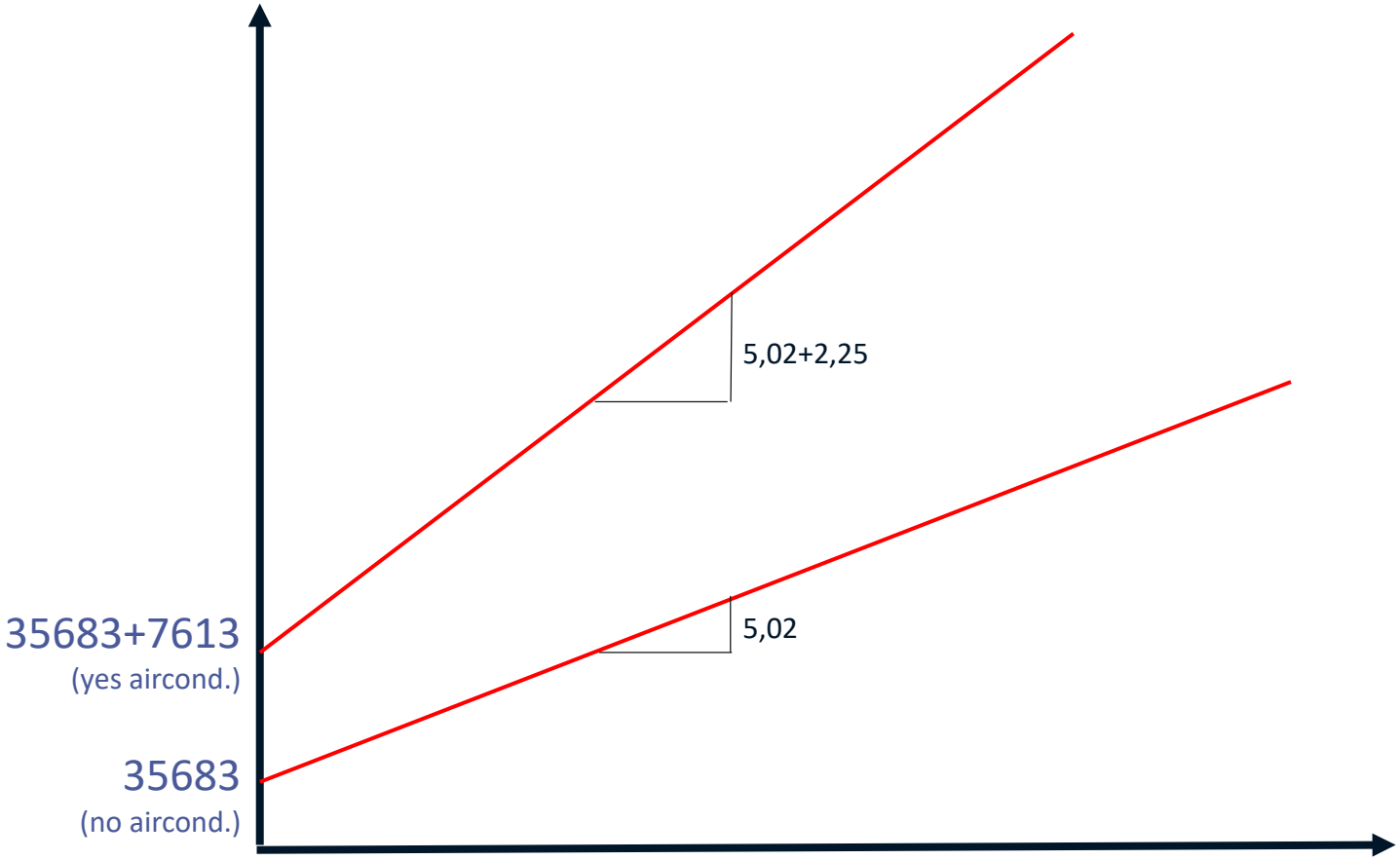
Példa

Házárat magyarázzuk telekmérettel, légkondival és interakcióval :

	Coefficient	*Standard error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
intercept	35683.87	2587.21	13.79	0.00	30601.69	40766.05
lot size	5.02	0.49	10.26	0.00	4.06	5.98
air cond	7613.35	5544.33	1.37	0.17	-3277.65	18504.36
lot size x air cond	2.25	0.93	2.42	0.02	0.42	4.09

Értelmezés? Légkondi hatása?

	Coefficient	*Standard error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
intercept	35683.87	2587.21	13.79	0.00	30601.69	40766.05
lot size	5.02	0.49	10.26	0.00	4.06	5.98
air cond	7613.35	5544.33	1.37	0.17	-3277.65	18504.36
lot size x air cond	2.25	0.93	2.42	0.02	0.42	4.09



6. feladat

Az órai Excel fájlban oldja meg a **WAGE** munkalapon található feladatokat.

Bináris függő változó

(5) Kétértékű függő változó

Példák:

- Van-e saját autója?
- Vállalat nyereséges-e?
- Átmegy a vizsgán vagy sem a tanult órák függvényében?

Az OLS lehetséges, de :

- Becsült érték nem 0-1, hanem [0,1]
- További problémák – más típusú becslés jobb módszer (probit, logit)

Részletek: későbbi tárgyakon…

Házi feladat

Használja a Moodle felületen megosztott, hasznaltauto.hu honlapról származó valós adatokat tartalmazó adatbázist és válaszolja meg a kérdéseket!

+ Már aktív a 2. csoportos házi feladat is!

Köszönöm a figyelmet!

Corvinus

www.uni-corvinus.hu

Idősorelemzés

Bevezetés az empirikus elemzésbe - 12. hét

Budapesti Corvinus Egyetem
Corvinus University of Budapest

Tartalom

1. Bevezetés

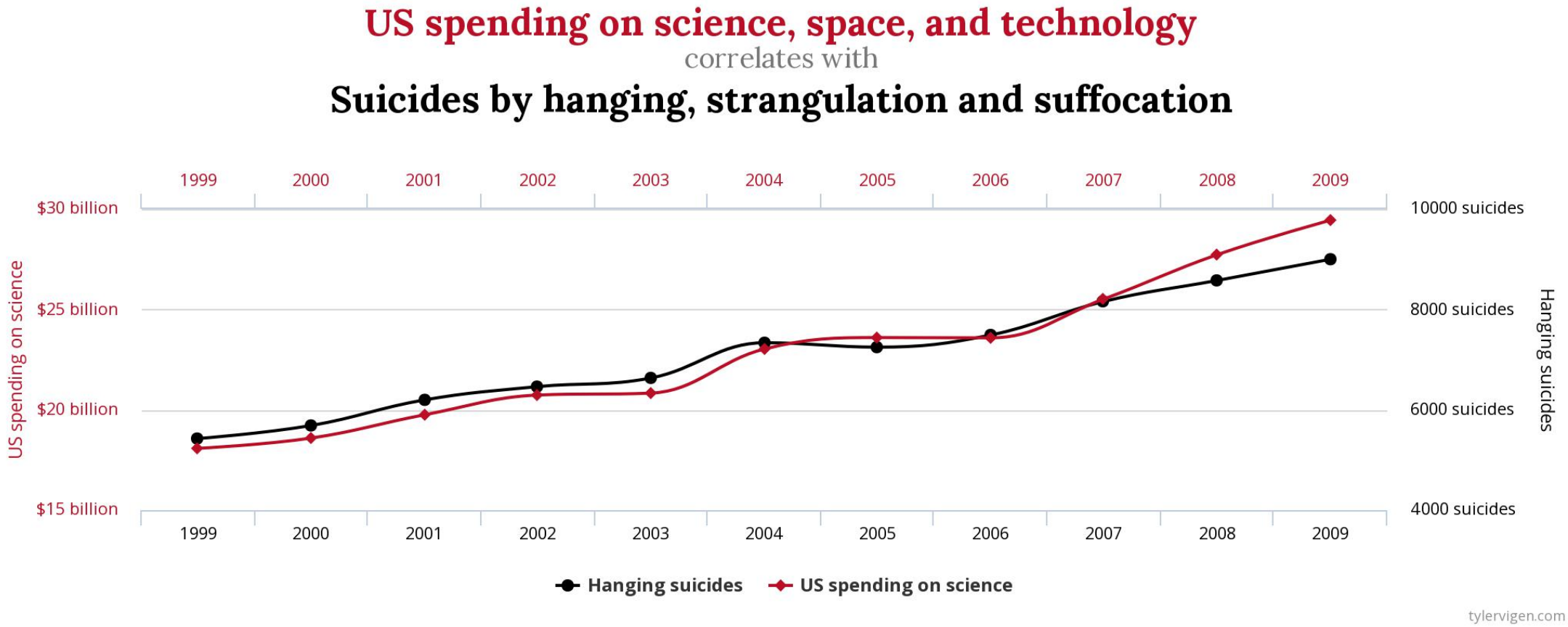
Idősorelemzés: bevezetés

- Idősoros adatok
  - Megfigyelések, amelyek ugyanazon egységre vonatkoznak de különböző időpontokban.
  - Tipikus jelölés Y_t, jellemzően t = 1, 2, ...T
  - A megfigyelések közötti különbség az idősoros adat gyakorisága, pl. perc, nap, hónap, év…
- Cél továbbra is: függő változó alakulásának magyarázata.
- Nehézségek
  - Késleltetett hatások
  - Nem stacioner változók → hamis regresszió (spurious regression)
- Kulcsfontosságú témák, melyekkel foglalkozni kell:
  - Trend, autokorreláció, szezonalitás
- Előrejelzési feladatoknál elkerülhetetlen.

Hamis regresszió

- Olyan, mint a korrelációnál volt a nem igazi oksági összefüggés két változó között. Ott sokszor a korrelációt egy harmadik ok okozta, amely mindkét változóra hatott.
- Idősoroknál a problémát az idő okozza.
  - Ha két változó az idő függvényében változik, akkor a két változó korrelálhat azért, mert időben mindketten nőnek. De ez sokszor hülyeség, a változók között nincs okozati összefüggés.
- R-négyzet magas, F-érték is az, azaz a regresszió jó. Sőt a t-érték is magas és a változók szignifikánsak, de még sincs értelme az egésznek.
- Stacionaritás az ilyen problémákat kiszűri.

Hamis regresszió - példa



További példák innen: http://www.tylervigen.com/spurious-correlations

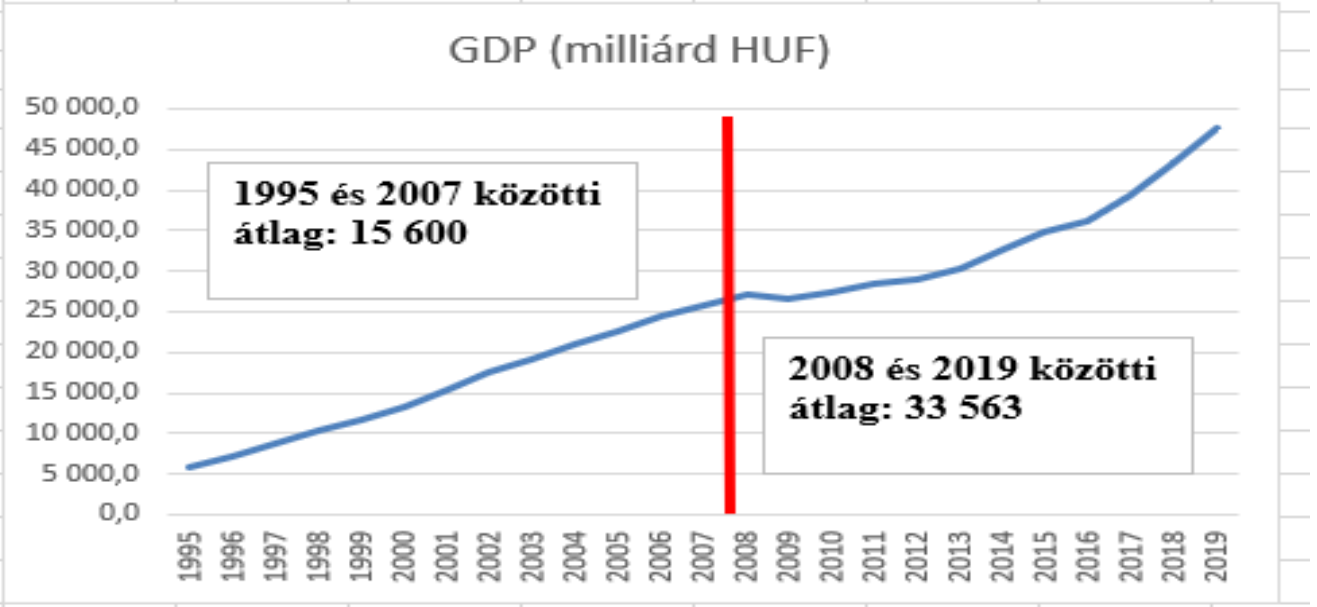
Stacionaritás

Stacionaritás:

- Olyan idősor, amelynek statisztikai tulajdonságai nem változnak az idővel.
  - Ez nem jelenti azt, hogy az idősor nem változik az idővel, csak azt, hogy ahogyan változik, az nem változik az idővel.
  - Paraméterek, mint például az átlag és a variancia nem változik az idővel.
- Jelentősége
  - Stacionárius folyamatokat könnyebb elemezni.
  - Előre tudjuk alakulásukat jelezni, hiszen ahogyan változnak az előrejelezhető.
- Stacionaritás sérülésének leggyakoribb okai
  - trend
  - szezonalitás

Egyváltozós idősor: a stacionaritás sérülése - trend

Év	GDP (milliárd HUF)
1995	5 836,5
1996	7 122,3
1997	8 834,6
1998	10 442,8
1999	11 637,5
2000	13 324,1
2001	15 398,7
2002	17 433,9
2003	19 133,8
2004	21 077,5
2005	22 549,0
2006	24 316,3
2007	25 701,4
2008	27 217,4
2009	26 458,3
2010	27 431,3
2011	28 501,5
2012	28 920,4
2013	30 290,9
2014	32 742,2
2015	34 937,3
2016	36 167,5
2017	39 233,4
2018	43 347,0
2019	47 513,9

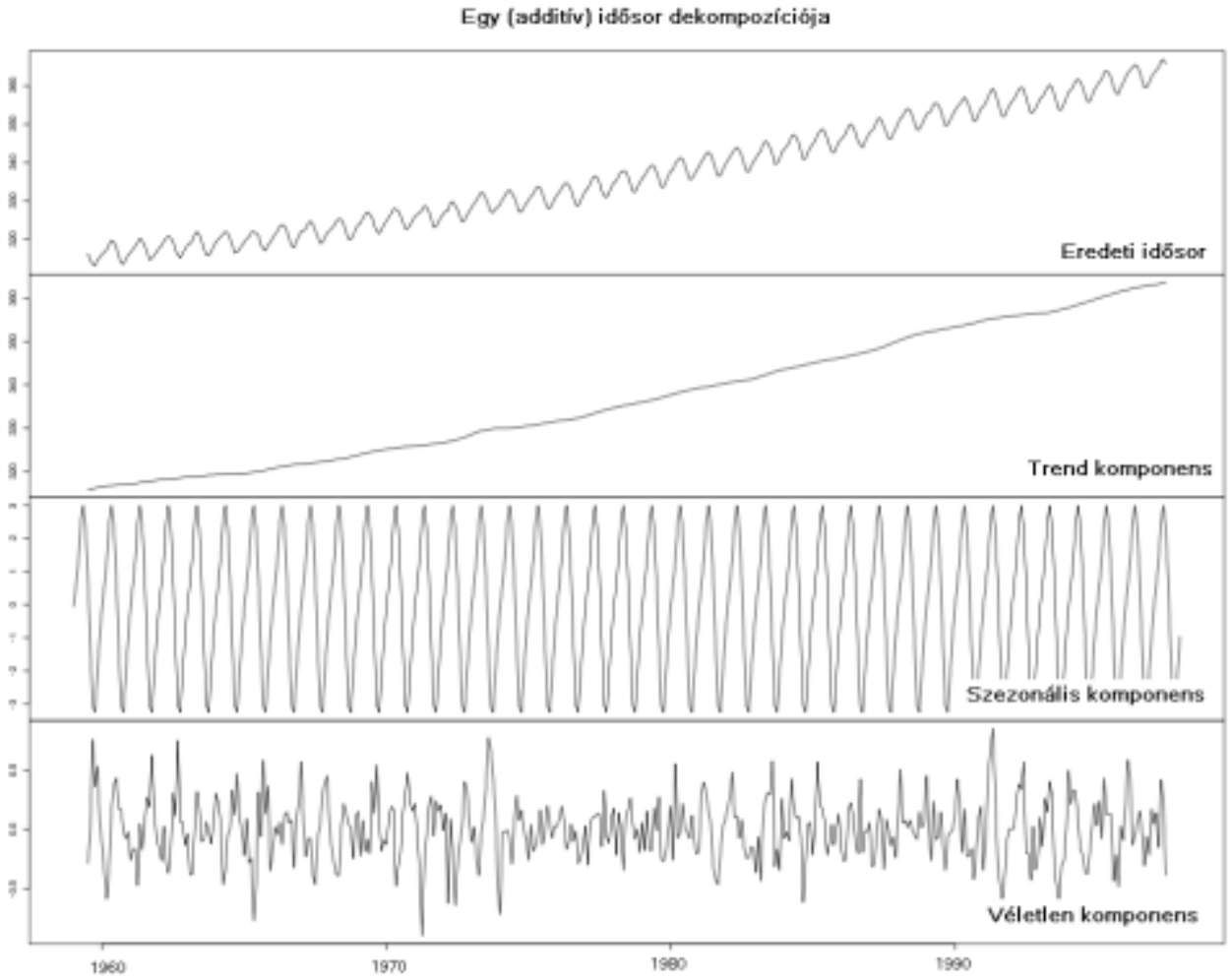


Idősorok felbontása

Bontsuk az Y_t idősort diszjunkt komponensekre:

$$Y_t = (T_t; S_t; R_t), \text{ ahol}$$

- *T* a trend-ciklus: a trend az idősor hosszú távú alapirányzata, a ciklus pedig a rövidebb távú szabálytalan ingadozás.
- *S* a szezonális komponens: az éven belüli szabályos ingadozások.
- *R* a véletlen komponens: az előre nem jelezhető véletlen hatások.



Forrás: https://www.ksh.hu/docs/hun/modsz/szezonalis_kiigazitas_ksh.pdf

Idősorok felbontása

Bontsuk az Y_t idősort diszjunkt komponensekre:

$$Y_t = (T_t;\ S_t;\ R_t)$$

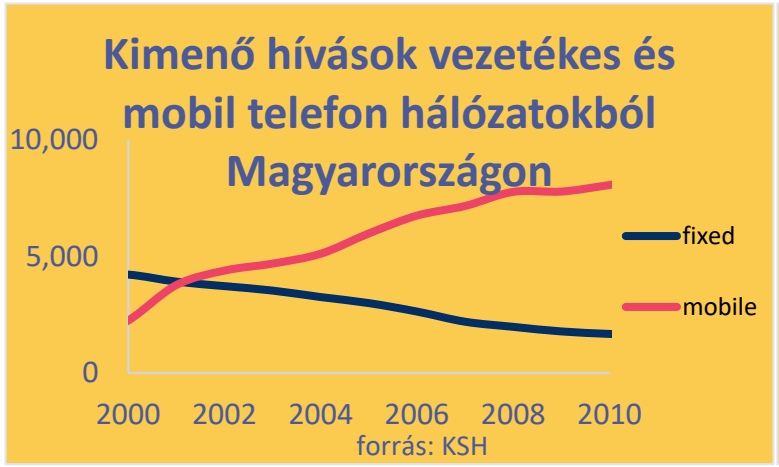
Két alapvető dekompozíciós modell:

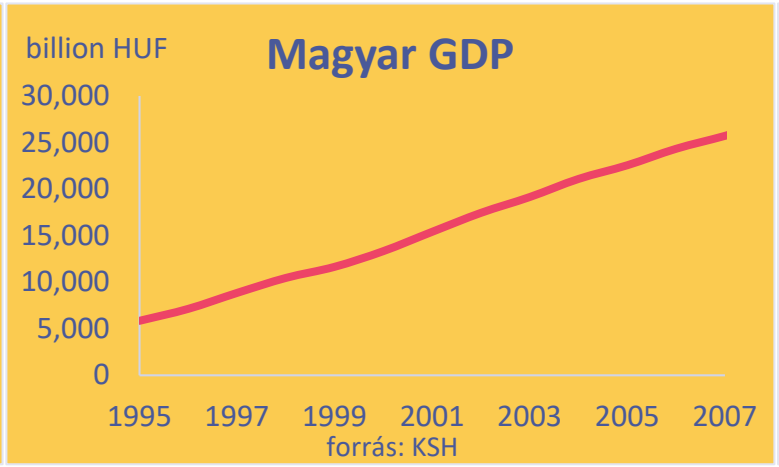
- Additív: $Y_t = T_t + S_t + R_t$
- Multiplikatív: $Y_t = T_t \cdot S_t \cdot R_t$

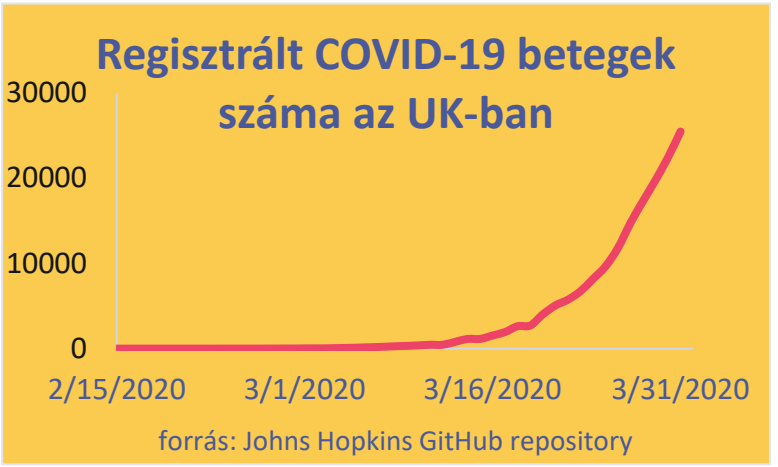
2. Trend

Trend

- Egy változó trendet követ, ha jellemzően egy irányba változik
  - pozitív trend, negatív trend
  - lineáris trend, exponenciális trend, stb.







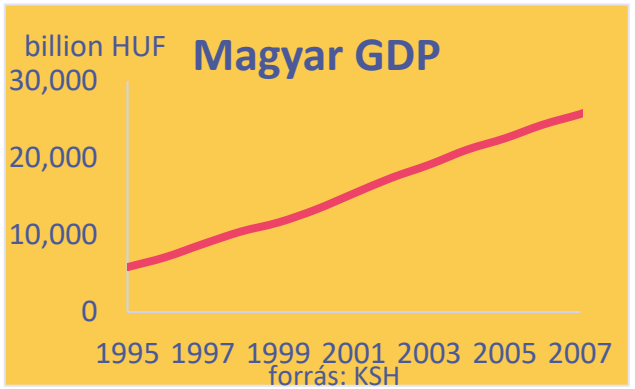
- Makroökonómiai változók többsége (fogyasztás, jövedelem stb.): jellemzően trendet követnek

Differenciaképzés

Gyakran hasznos az idősor differenciáját vizsgálni:

$$\Delta Y_t = Y_t - Y_{t-1}$$

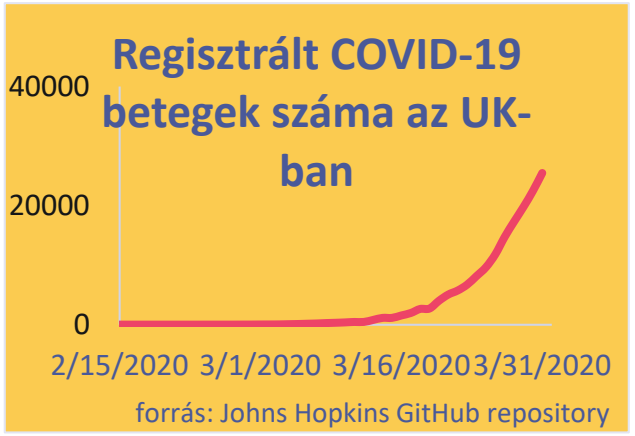
- ahol Y_{t-k} a k-adik késleltetettje
- változó időbeli változását méri



Vagy a logaritmus differenciáját:

$$\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1}) = \ln\left(\frac{Y_t}{Y_{t-1}}\right) = \ln\left(\frac{\Delta Y_t}{Y_{t-1}} + 1\right) \approx \frac{\Delta Y_t}{Y_{t-1}} = \frac{Y_t}{Y_{t-1}} - 1$$

- $(t-1)$ és t periódusok közötti százalékos változást méri

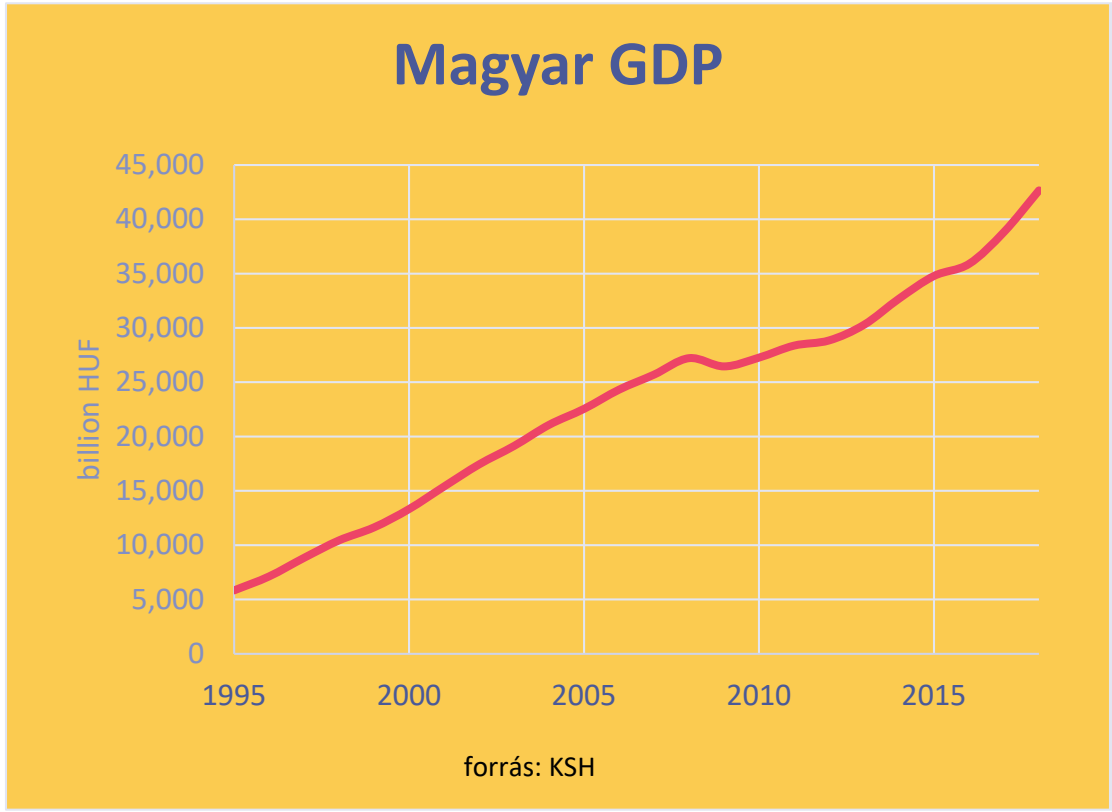


Trend és differencia

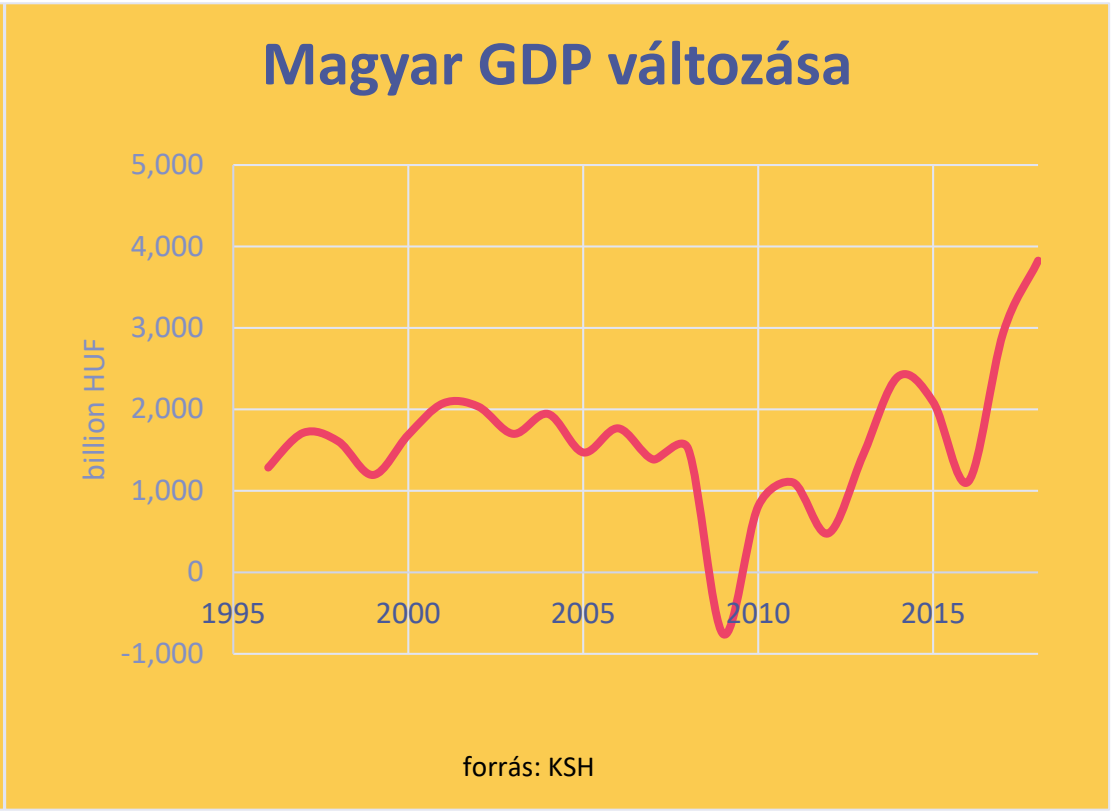
- Időben növekvő változók (pl. GDP) trendet követnek. Ebből következően az egymást követő megfigyelések korreláltak.
  - A változók erősen korreláltak a késleltetettjeikkel.
  - A jelenlegi értékből jól megbecsülhető a következő időszaki érték.
  - Vizuálisan: a változó idősora egy egyenesre illeszkedik.

- Ezzel szemben a változás idősorában (differencia vagy log differencia) jellemzően nincs trend.
  - A változások nem korrelálnak.
  - Nehéz előrejelezni.

Példa: magyar GDP és differenciája

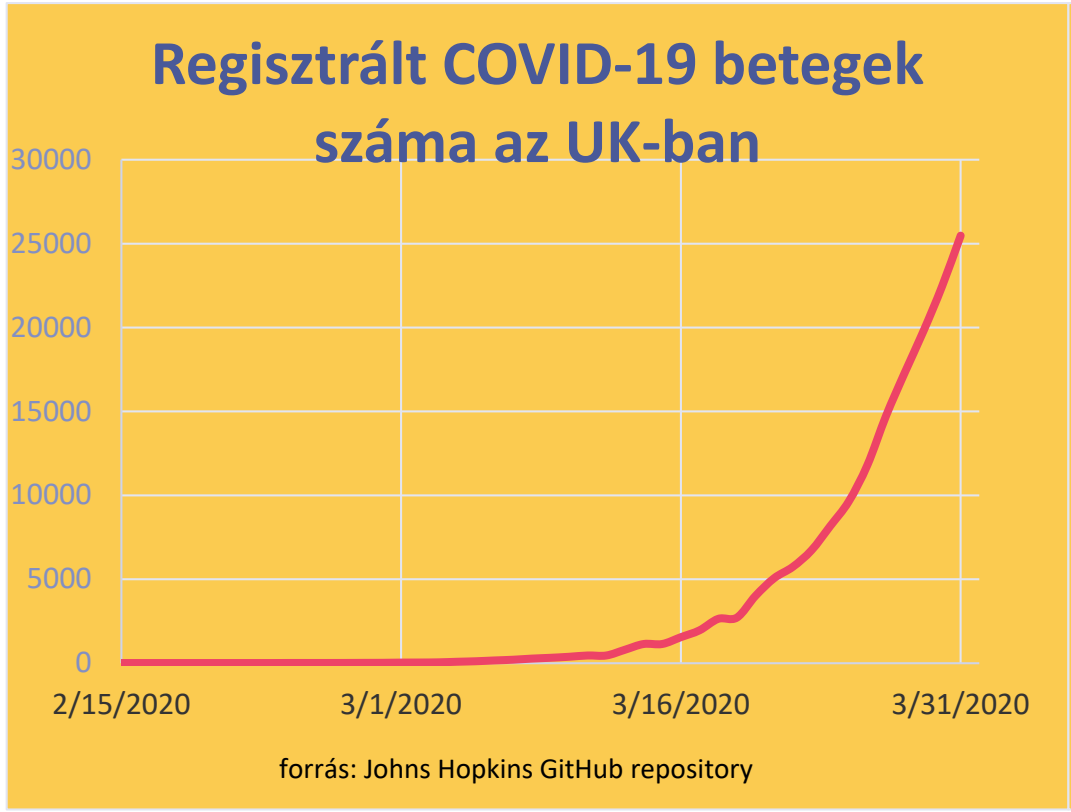


forrás: KSH

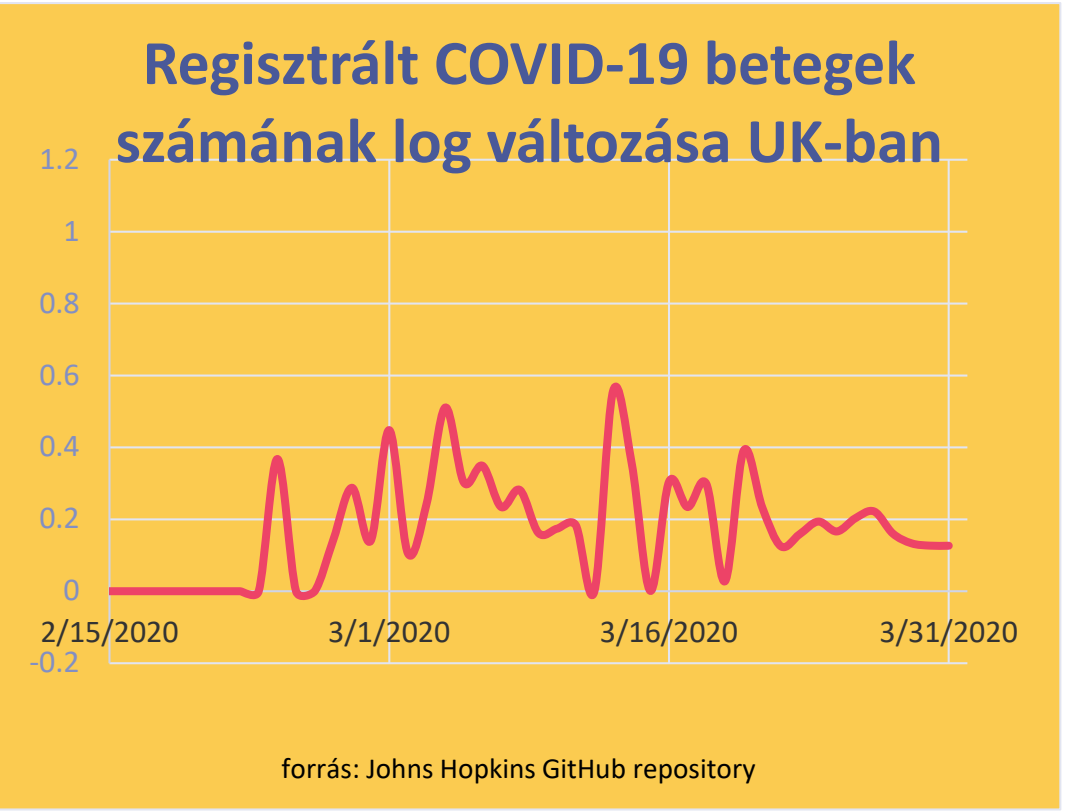


forrás: KSH

Példa: regisztrált COVID-19 betegek száma az Egyesült Királyságban és log differenciája



forrás: Johns Hopkins GitHub repository



forrás: Johns Hopkins GitHub repository

Feladat - 1

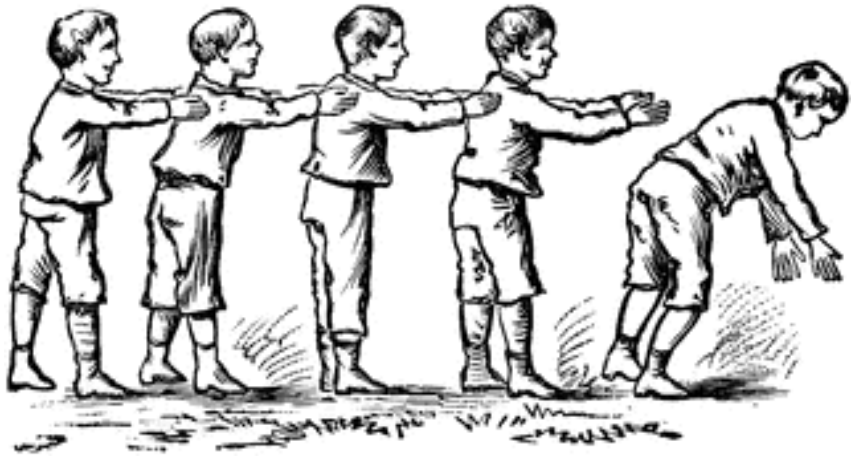
Az órai Excel fájl „GDP” és „Covid” fülein kövesse az 1a. és 1b. feladatok instrukcióit!

3. Autokorreláció

Késleltetett hatások

- Mivel az adatok rendezettek, a megfigyelések hatással lehetnek a rákövetkező megfigyelésekre.
- A sorozatok autokorreláltak: szomszédos megfigyelések egymással korreláltak.

  - https://www.youtube.com/watch?v=wJXJN4H2G4k



Forrás: http://etc.usf.edu/clipart/20000/20086/boysinline_20086.htm

Autokorreláció

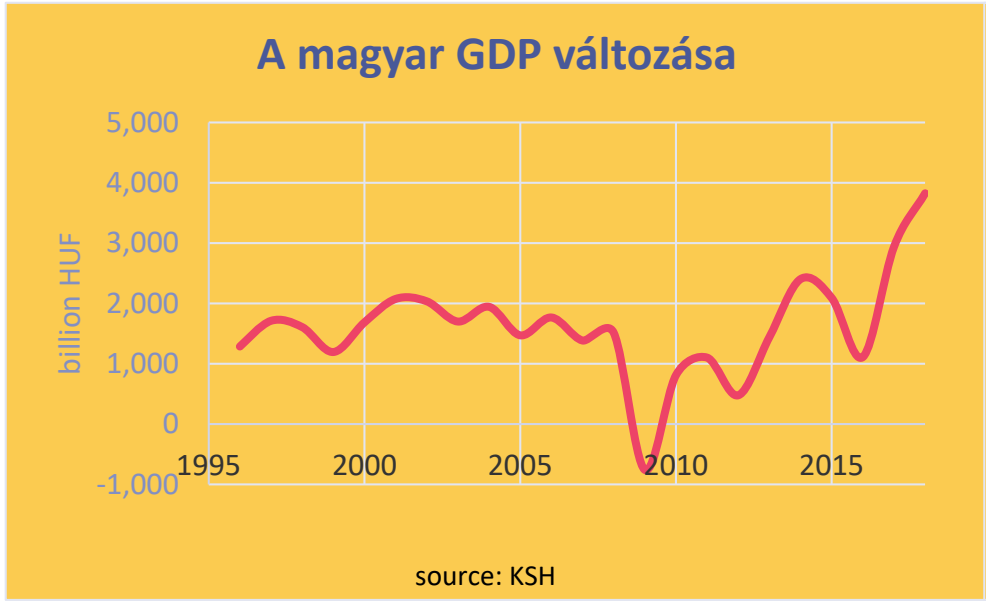
- Mivel nincs más magyarázó változónk, ezért nem más változók függvényében akarjuk megérteni a függő változót, hanem a késleltetettjei alapján.
- Fő cél: előrejelzés
- Korreláció a változó és saját késleltetettje között.

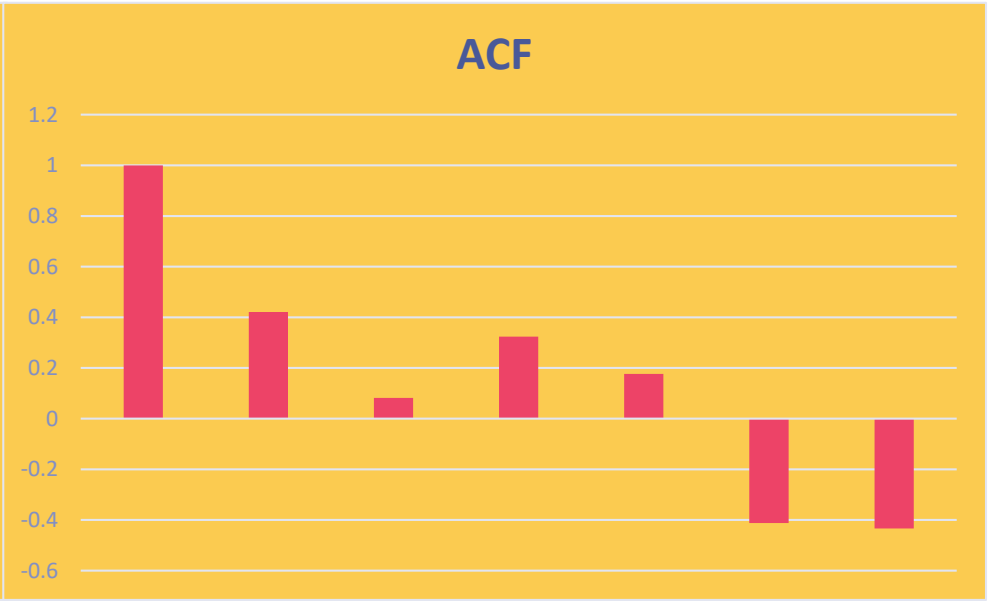
- *Y* korrelációja a p-edik késleltetettjével (Y_{t-p}):

$$r_p = corr(Y_t, Y_{t-p})$$

Autokorrelációs függvény (ACF)

- Autókorrelációk sorozata a késleltetés függvényében
  - r_p számolja ki $p=1,2,..P$-re, ahol P a leghosszabb használt késleltetés
  - Hosszabb késleltetés, kevesebb megfigyelés: ha $P=z$, akkor az első z adatot eldobjuk.
- Példa:



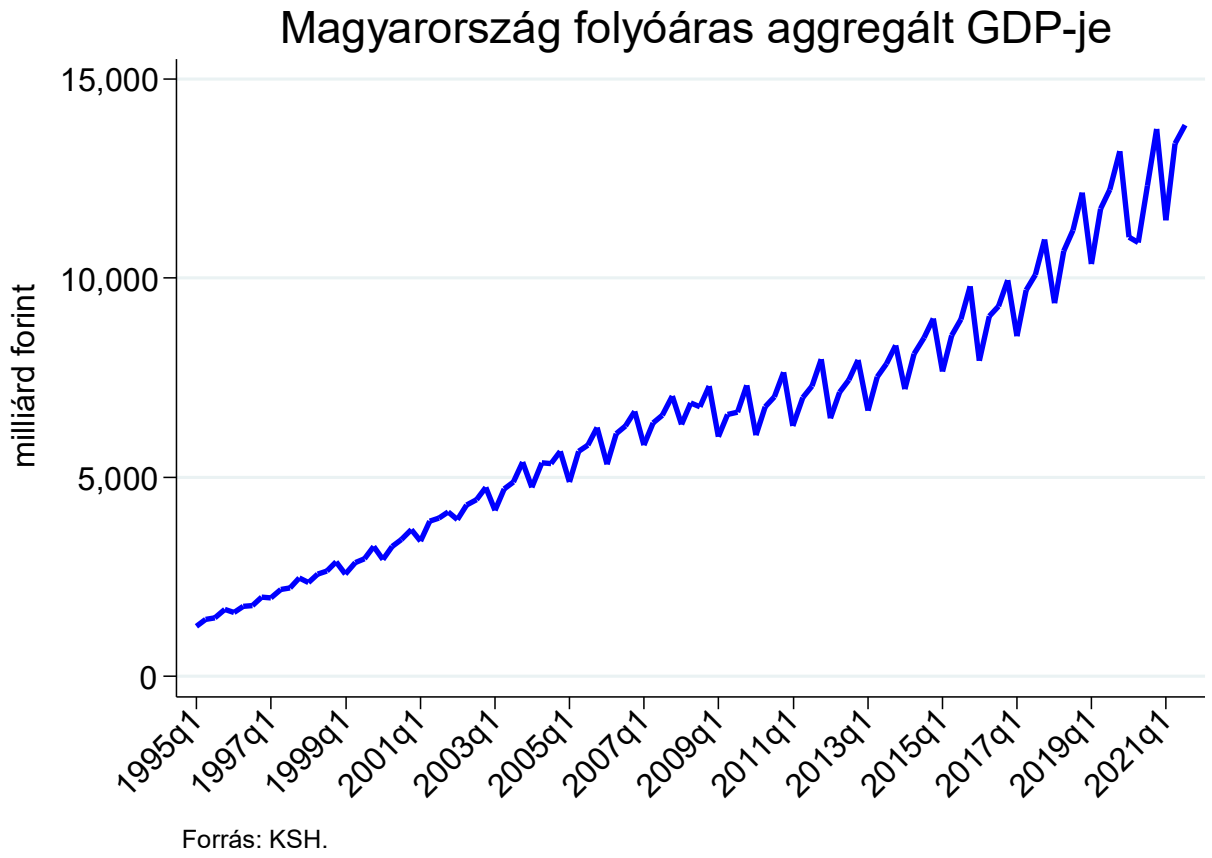


Feladat - 2

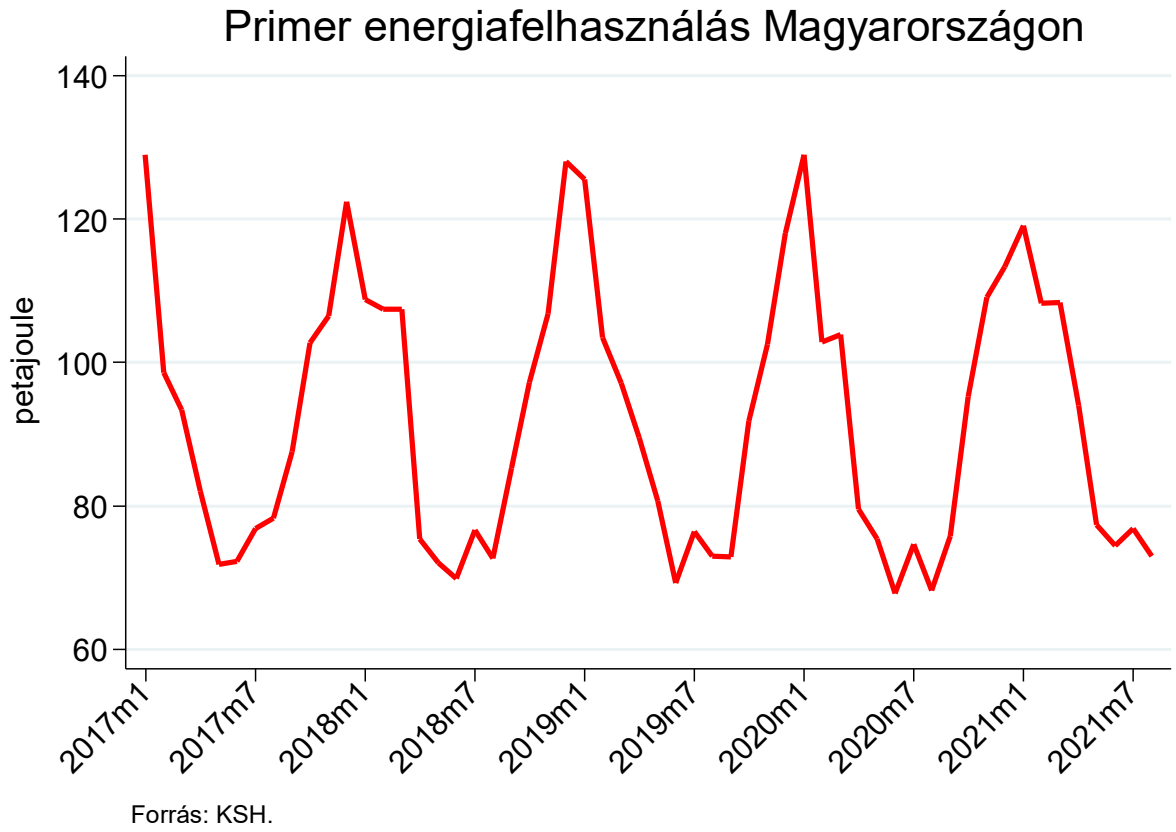
Az órai Excel fájl „GDP” és „Covid” fülein kövesse a 2. feladat instrukcióit!

4. Szezonalitás

Szezonalitás



Forrás: KSH.



Forrás: KSH.

Adott időszakban van valami

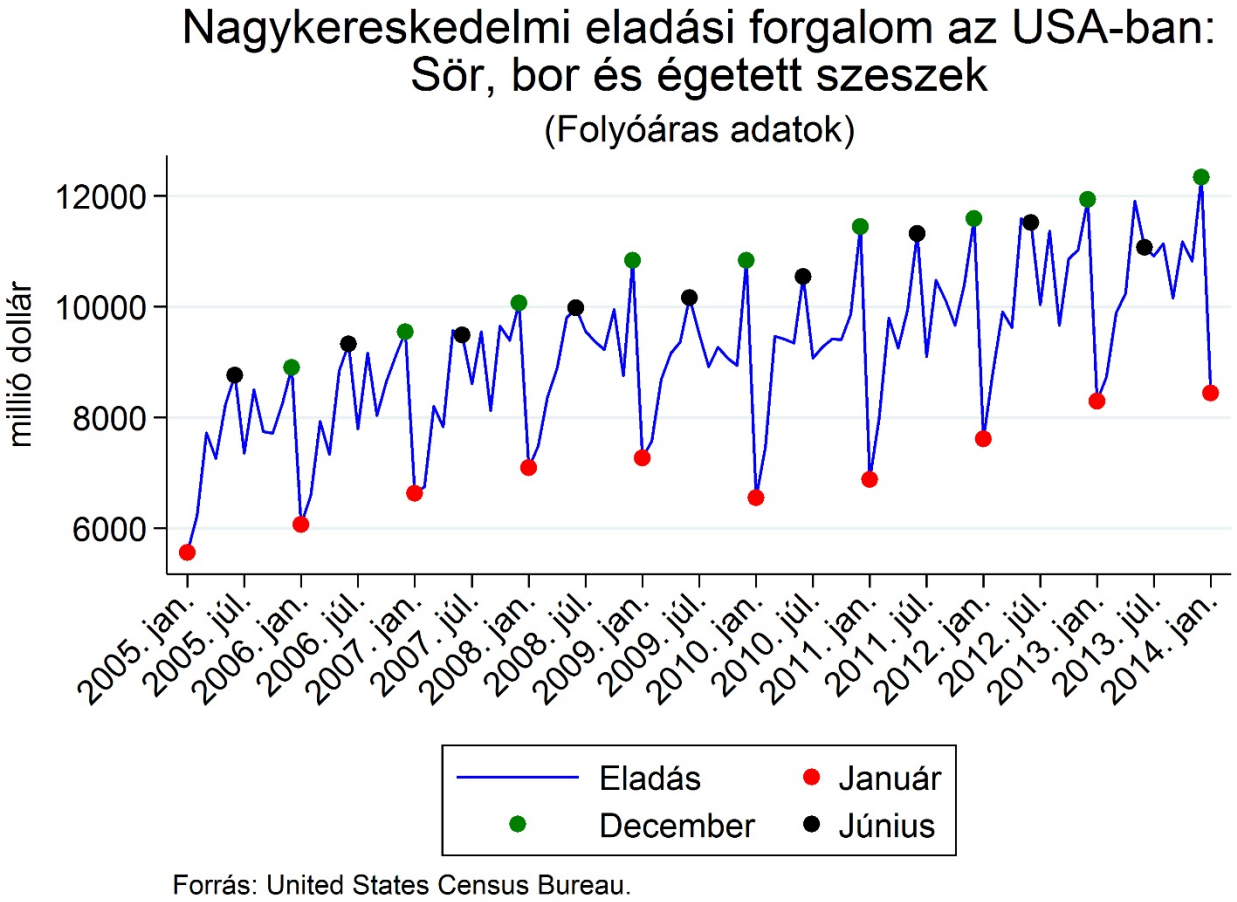
Jobb az időjárás, inkább nyílnak a virágok.

Inkább van hó a síeléshez.

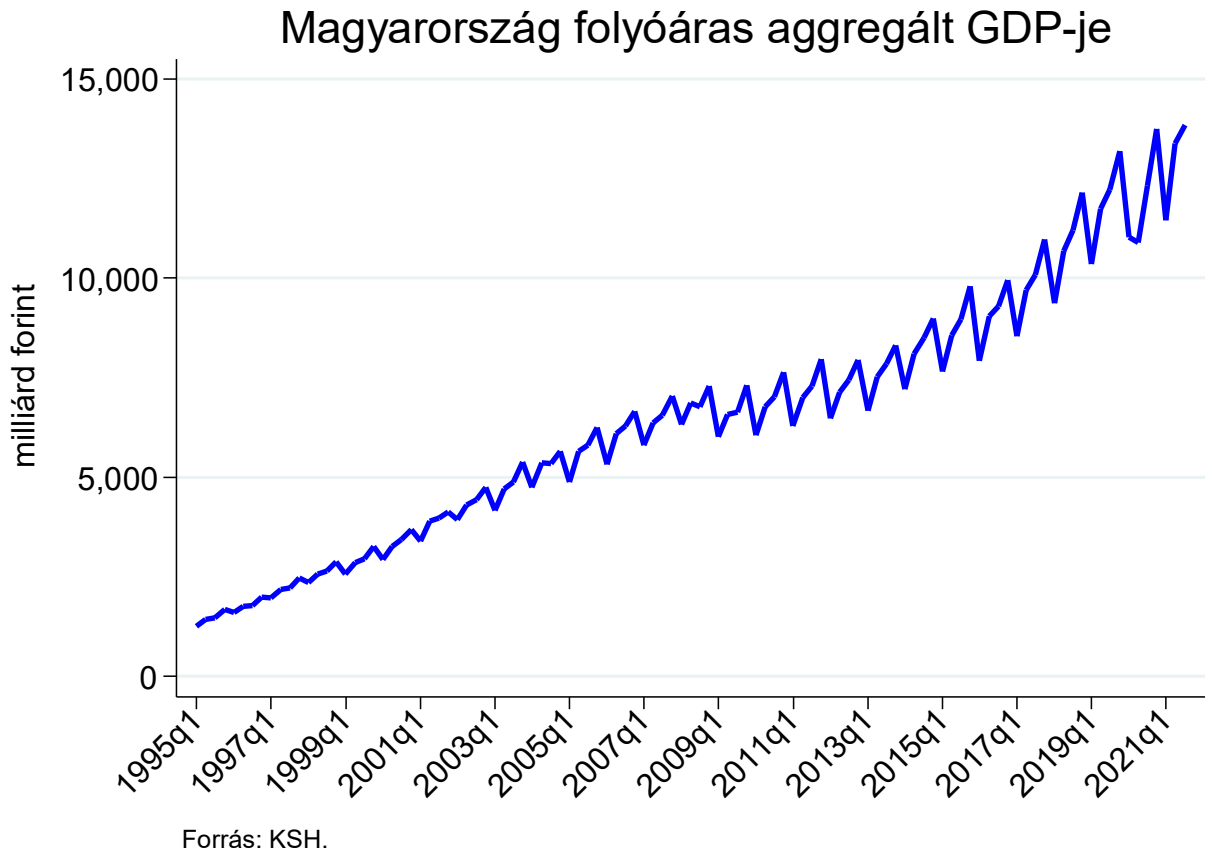
Inkább meleg a Balaton a fürdéshez.

Inkább van iskolai szünet.

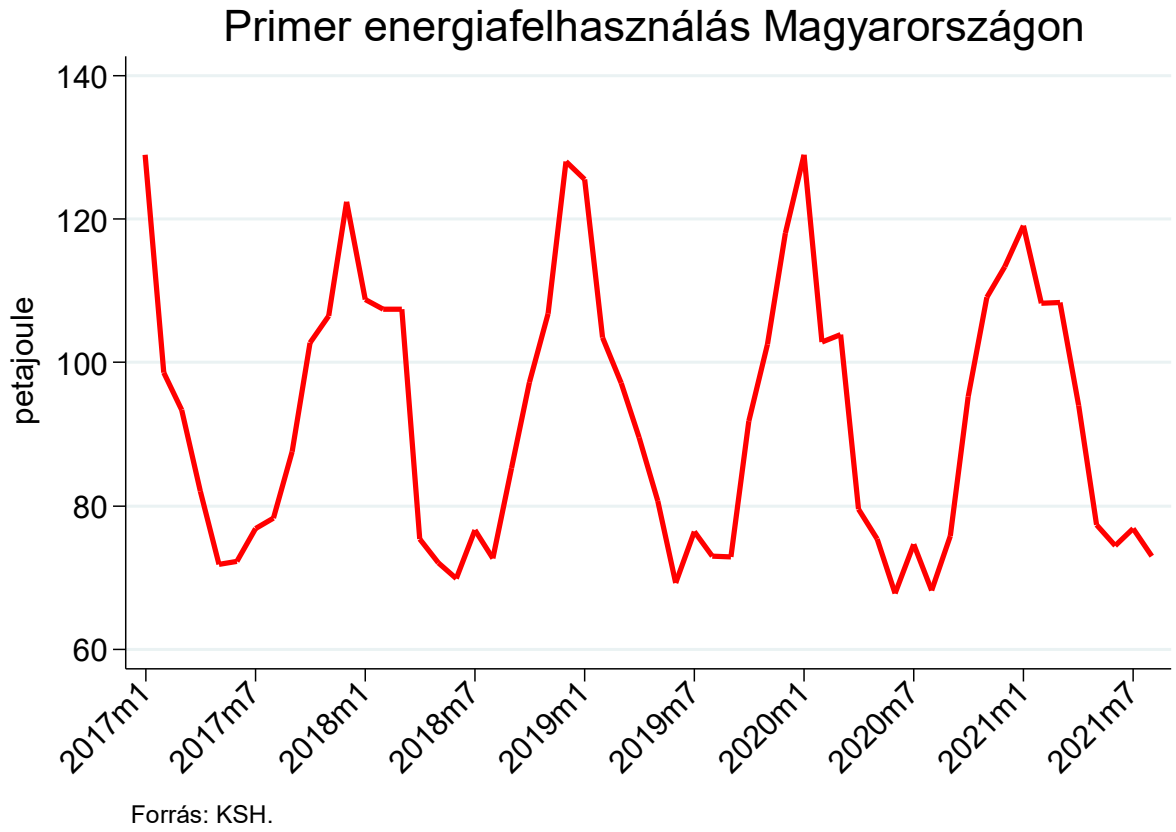
https://www.statcan.gc.ca/eng/sc/video/seasonal-adjustment



Szezonalitás



Forrás: KSH.



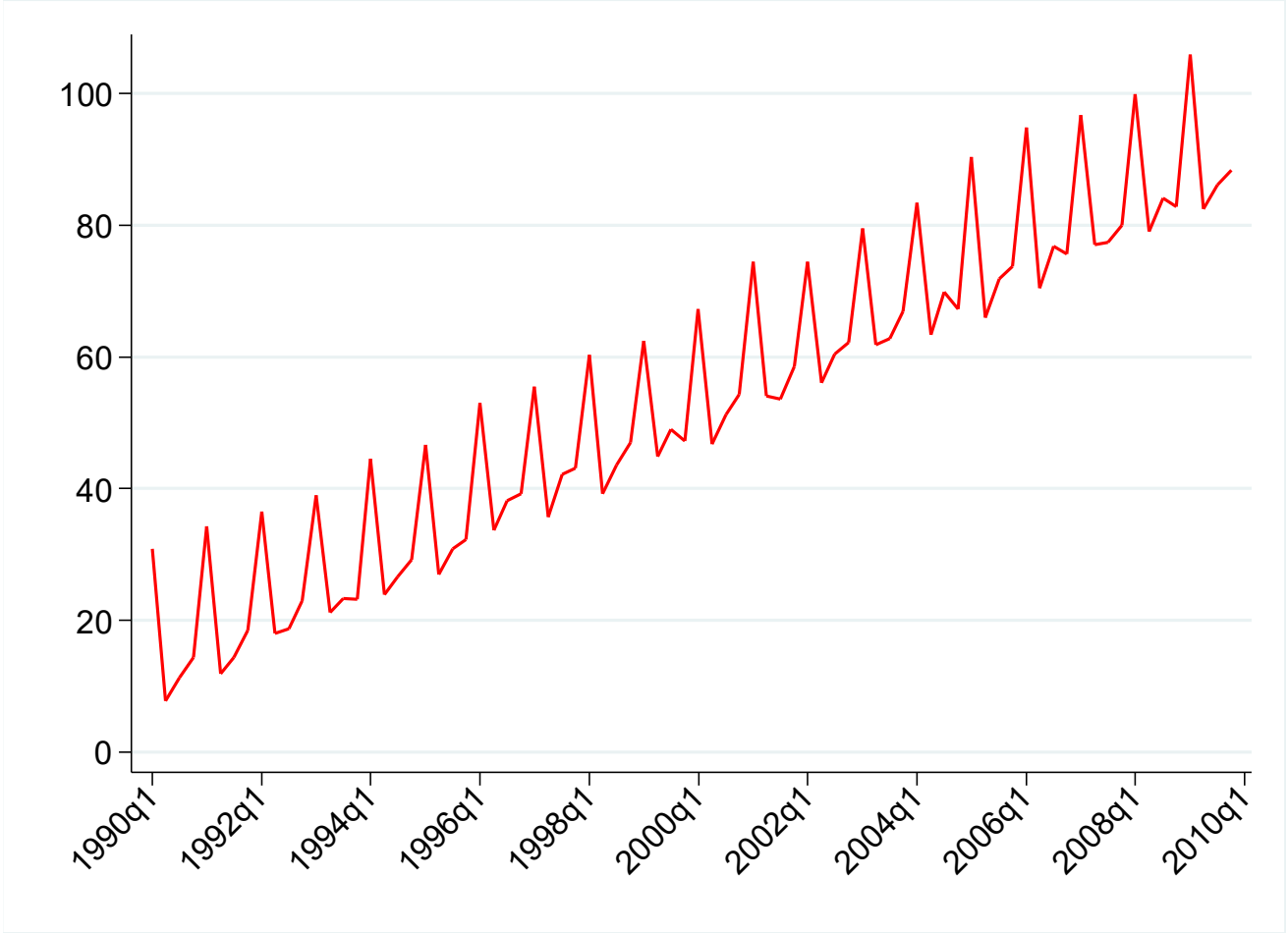
Forrás: KSH.

Szezonalitás az additív modellben: szezonális eltérés

Az additív modell: $Y_t = T_t + S_t + R_t$

4 időszak negyedévekre: 3 dummyval becslés (q1 a bázis)

$$Y_t = \alpha_0 + \alpha_{q2}D_{q2} + \alpha_{q3}D_{q3} + \alpha_{q4}D_{q4} + \varepsilon_t$$



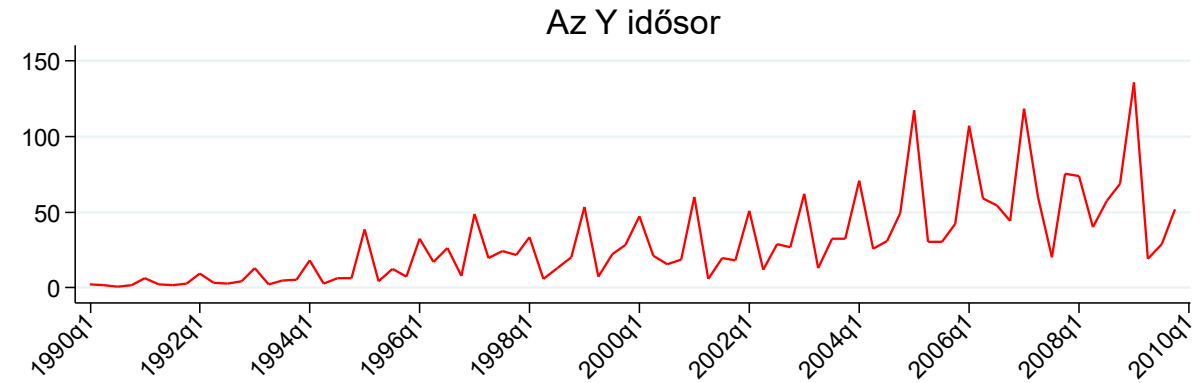
y_add	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
quarter						
2	-20.45806	7.417198	-2.76	0.007	-35.23069	-5.685427
3	-16.8564	7.417198	-2.27	0.026	-31.62903	-2.083766
4	-15.13266	7.417198	-2.04	0.045	-29.9053	-.360034
_cons	66.46352	5.244751	12.67	0.000	56.0177	76.90935

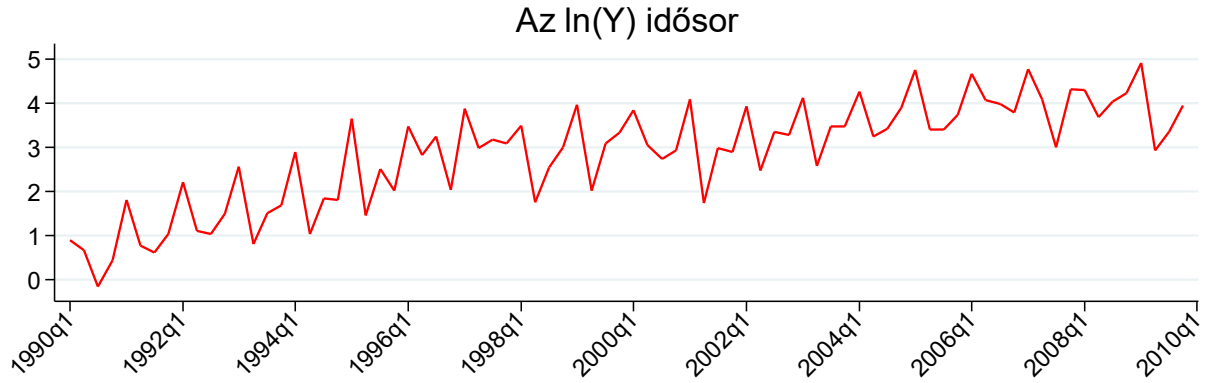
Szezonalitás a multiplikatív modellben: szezonindex

A multiplikatív modell: $Y_t = T_t \cdot S_t \cdot R_t$

4 időszak negyedévekre: 3 dummyval becslés (q1 a bázis)

$$lnY_t = \alpha_0 + \alpha_{q2}D_{q2} + \alpha_{q3}D_{q3} + \alpha_{q4}D_{q4} + \varepsilon_t$$





lny	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
quarter						
2	-1.289397	.3498729	-3.69	0.000	-1.986229	-.5925645
3	-.9639359	.3498729	-2.76	0.007	-1.660768	-.2671037
4	-.8030138	.3498729	-2.30	0.024	-1.499846	-.1061815
_cons	3.630832	.2473975	14.68	0.000	3.138097	4.123566

Kategóriák az időszakban

- 4 időszak negyedévekre: 3 dummyval becslés (q4 a bázis)

$$Y_t = \alpha_0 + \alpha_{q1} D_{q1} + \alpha_{q2} D_{q2} + \alpha_{q3} D_{q3} + \varepsilon_t$$

- Ha úgy gondoljuk, hogy a szezonalitás inkább arányaiban jelentkezik – pl. nyáron mindig kétszer annyit nyaralnak tavaszhoz képest függetlenül attól, hogy válság van-e, logaritmusban becsülünk.

$$\ln Y_t = \alpha_0 + \alpha_{q1} D_{q1} + \alpha_{q2} D_{q2} + \alpha_{q3} D_{q3} + \varepsilon_t$$

Deszezonalizálás

- Szezonális hatás eltávolítása: deszezonalizálás.
- Példa 4 negyeddel (q4 a bázis):

$$Y_t = \alpha_0 + \alpha_{q1} D_{q1} + \alpha_{q2} D_{q2} + \alpha_{q3} D_{q3} + \varepsilon_t$$

- Először eltávolítjuk a becsült szezonális hatásokat.:

$$Y\_sa\_nyers_t = Y_t - \hat{\alpha}_{q1} D_{q1} - \hat{\alpha}_{q2} D_{q2} - \hat{\alpha}_{q3} D_{q3}$$

- Majd korrigáljuk az éves átlagos szezonális hatással:

$$Y\_sa_t = Y\_sa\_nyers_t + \frac{\hat{\alpha}_{q1} + \hat{\alpha}_{q2} + \hat{\alpha}_{q3}}{4}$$

Feladat - 3

Az órai Excel fájl „energia” fülén kövesse az instrukciókat

Köszönöm a figyelmet!

Corvinus

Corvinus

www.uni-corvinus.hu