Budapesti lakáspiac túlfűtöttsége

Szakdolgozat

Morvai Zsolt

Matematika szak, elemző szakirány

Témavezető

Prőhle Tamás, Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem

Természettudományi Kar, Matematika Intézet

2020.

Absztrakt

Több befolyásoló tényező volt azzal kapcsolatban, hogy miért a budapesti lakáspiacról írom a szakdolgozatomat. Felsőfokú tanulmányaimat Budapesten folytattam és a diákmunkám során életem szerves része lett az ingatlanpiac. A korona vírus terjedése előtt nagyon népszerű téma volt a budapesti lakáspiac túlfűtöttsége, nem volt olyan hét, amikor ne jelent volna meg ezzel kapcsolatban valamilyen statisztika. Ezekből kifolyólag jutottam arra a döntésre, hogy ebből írom a szakdolgozatomat. Csak Budapesten vagy Magyarországon volt nagymértékű változás vagy esetleg egész Európát is érintette? Nem is túlfűtöttségről beszélünk, hanem egy tendenciáról, ami mindenhol megtalálható volt? Három statisztikai módszer segítségével kerestem a válaszokat az előző felvetéseimre. A három statisztikai módszerem a többfázisú regressziós modell, a modell alapú szórás analízis és a többdimenziós idősor modell.

Többfázisú regressziós modellel töréspontokat figyeltem meg, hogy esetlegesen statisztikailag is kimutatható eltéréseket találjak. Illetve, hogy meghatározzam az idősoraimban azokat az időpontokat, amikor valamilyen külső befolyásoló tényező megváltoztatta a tendenciát. Budapesti alkudozás visszaesésében, a lakásvásárlók megoszlásában és az ingatlan értékesítésében mind-mind találhattunk töréspontokat a lakásárindex emelkedése előtt, ami arra enged következtetni, hogy ezek mind fontos szerepet játszanak a lakásárindex alakulásában. Ennek a modell segítségével kiválasztottam azokat az értékeket, amelyeket későbbiekben a többdimenziós idősor modellben felhasználok.

Szórás alapú analízis segítségével hasonlítottam össze a különböző adataimat. Itt vizsgáltam meg azt a felvetésemet, hogy csak Budapesten vagy Magyaroroszágon volt-e ilyen nagymértékű változás, vagy esetleg egész Európában is. Budapesti lakásárindex viselkedése láthatóan eltér a városokétól, községekétől. Nemcsak itt vehető észre számottevő eltérés, hanem Európával szemben is.

Többdimenziós idősor modellemmel megnéztem, hogy bizonyos változók mekkora mértékben befolyásolhatták az adataimat. Legutoljára pedig az utolsó negyedévet elhagyva figyeltem meg, milyen pontosan prediktálta a változóimat. Előrejelzésem során a modell az elégtelen keresletet leszámítva jól közelítette az értékeket. Elégtelen keresletnél számottevő eltérést figyelhettünk meg. Ezek után két negyedévvel rövidebb adatsorral vizsgáltam meg az előrejelzést. Viszont itt is nagy volt az eltérés, ami arra enged következtetni, hogy egy változó tendencia alakult ki az elégtelen keresletnél.

Tartalomjegyzék

Absztrakt **2**

1. Bevezetés **4**

1.1. A téma jelentősége 4

1.2. Lakáspiaci túlfűtöttség 5

2. Alkalmazandó statisztikai modellek **9**

2.1. Modell alapú szórás analízis 9

2.1.1. Wild-bootstrap 10

2.2. Többfázisú regresszió 14

2.2.1. Többfázisú regresszió illesztési algoritmusa 20

2.3. Többdimenziós idősor modellek 20

3. Statisztikai elemzés **21**

3.1. Modell alapú szórás analízis 21

3.2. Többfázisú regressziós modell 23

3.3. Többdimenziós idősor modell 26

4. Összegzés **26**

Hivatkozások **27**

1. Bevezetés

1.1. A téma jelentősége

Nemcsak népszerű téma volt az ingatlanpiac, hanem az ingatlanszektor a nemzetgazdaságok egyik igen fontos része is, amit több fellelhető külföldi és hazai publikáció is bizonyít. Így többek között DiPasquale és Wheaton (1992) [1] az Amerikai Egyesült Államok ingatlanpiacának fontosságát az építőipari termelés és az ingatlanállomány értéke alapján vizsgálva azt találta, hogy 1990-ben az újonnan átadott lakóépületek előállítási értéke a GDP 3,3%-át tette ki, mely igen jelentős, hiszen az összes új épület előállítási értéke a GDP 5,5%-át adta. Gudell (2018) [2] szerint az amerikai lakóingatlan-állomány jelenlegi becsült értéke 31.800 billió USD, ami kicsivel több, mint a GDP 1,5-szerese és Kína GDP-jének a háromszorosát adja ki. Az elmúlt évben a lakóingatlan-állomány 1,95 billió USD-ral bővült, ami több, mint Kanada GDP-je.

Megnevezés	2019. I.	2019. II.	2019. III.	2019. IV.
Mezőgazdaság, erdőgazdálkodás, halászat	102.2	99.4	98.0	101.0
Ipar	105.8	103.8	107.6	103.2
Ebből: feldolgozóipar	106.0	103.3	108.1	103.0
Építőipar	144.5	127.2	118.4	111.3
Szolgáltatások összesen, ebből:	103.8	104.3	104.0	104.6
Kereskedelem, szálláshely-szolgáltatás és vendéglátás	106.5	107.7	109.4	110.7
Szállítás, raktározás	103.3	105.5	106.1	106.1
Információ, kommunikáció	106.5	105.1	106.5	106.9
Pénzügyi, biztosítási tevékenység	103.1	103.9	104.1	104.5
Ingatlanügyletek	104.6	104.6	103.9	104.3
Szakmai, tudományos, műszaki és adminisztratív tevékenység	107.8	107.3	106.1	105.8
Közigazgatás, oktatás, egészségügyi, szociális ellátás	99.2	99.6	97.6	98.0
Művészet, szórakoztatás, egyéb szolgáltatás	102.8	102.1	101.9	103.1
GDP összesen (piaci beszerzési áron)	105.3	104.9	105.0	104.5

Nemcsak külföldön, hanem Magyarországon is fontos szerepet játszik az építőipar, ahogy a táblázatban is látható. Az építőipar bruttó hozzáadott értéke 2019 I. negyedévében 44,5%-kal bővült, ami a negyedévek során csökkent, de még így is a legmagasabb 11,3%-os hozzájárulásával a legmagasabb maradt, megelőzve a kereskedelmet, szálláshely-szolgáltatást és a vendéglátást, amivel a gazdaság húzóágazatává vált.

”A statisztika egy eszköz, amivel válaszokat lehet keresni, de válaszokat nem ad. A válaszokat, az adatok interpretációját mindig nekünk kell adnunk.” (Kun Ádám)[1]

[1] Kutató, aki biológusként, vegyészként, valamint programozóként végezte felsőfokú tanulmányait. ELTE-n tanít evolúcióbiológiát, ökológiát, biostatisztikát, programozást és modellezést.

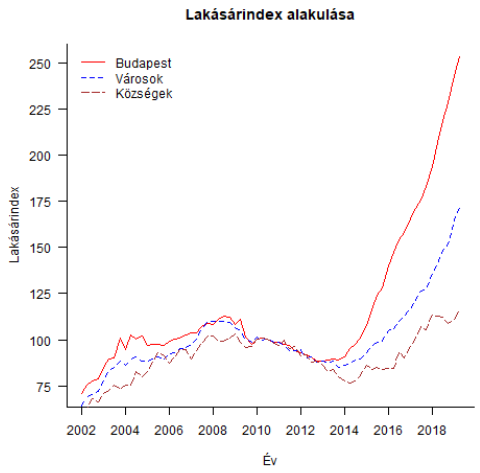
1.2. Lakáspiaci túlfűtöttség

Bevezetésben szeretnék két dolgot tisztázni, mit jelent a túlfűtöttség és a töréspont.

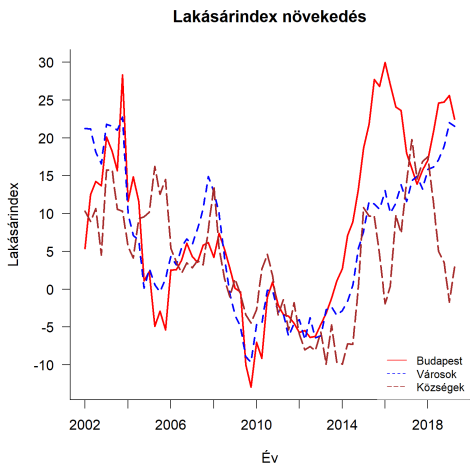
Az idősorban töréspontként definiálható az a pont, amely mentén az adatsort kettébontva a két részintervallum adatai között statisztikailag is igazolható eltérés mutatható ki. A töréspontok felfedésére több statisztikai eljárás is ismert. Az ilyen idősorok közötti kapcsolatok, pontosan a függő vagy független változóban talált töréspont(ok) miatt nem tekinthetőek alapvetően a teljes vizsgálati időszak vonatkozásában stabilnak.

Mit is nevezünk túlfűtöttségnek?
Ezzel a szóval találkozhatunk a konjuktúra-ciklusban, jelentése csúcs. A gazdaság fellendülése és lassulása/visszaesése közötti időszak. Mivel nem találunk egzakt definiciót és gyakran a túlfűtöttséget a buborék kialakulásával kötik össze, ezért Joseph Eugene Stiglitz amerikai Nobel-díjas közgazdászt szeretném idézni, aki így határozta meg az eszközár-buborékot:

> ”hogyha csakis azért magas ma az ár, mert a befektetők holnap is magas eladási árra számítanak – amikor az „alapvető” tényezők látszólag nem igazolnak egy árat – akkor beszélhetünk buborékról”



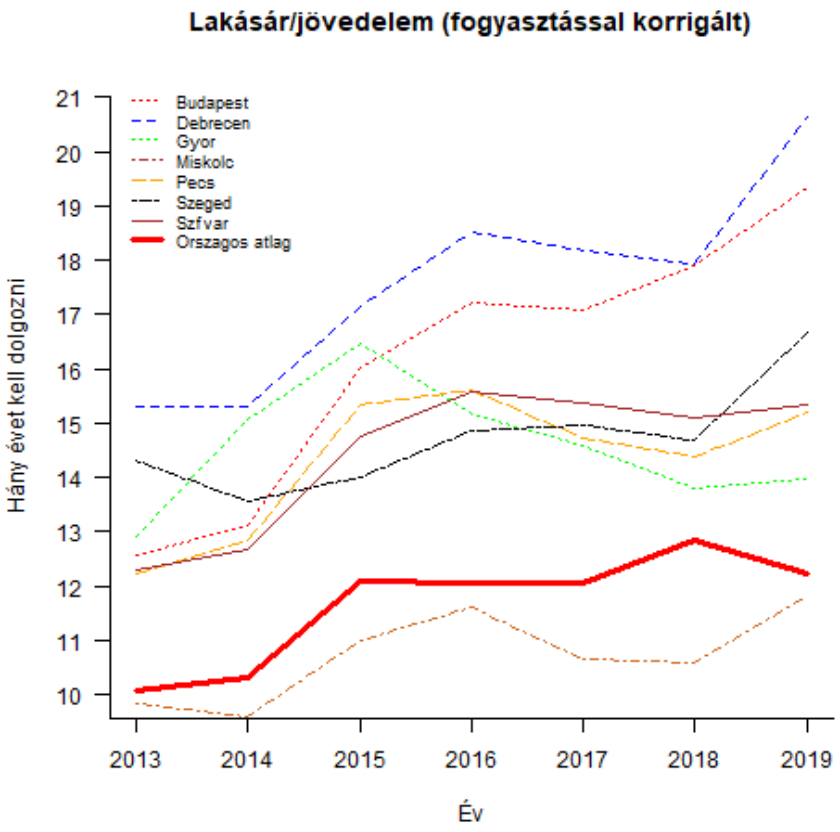
(a) Lakásárindex



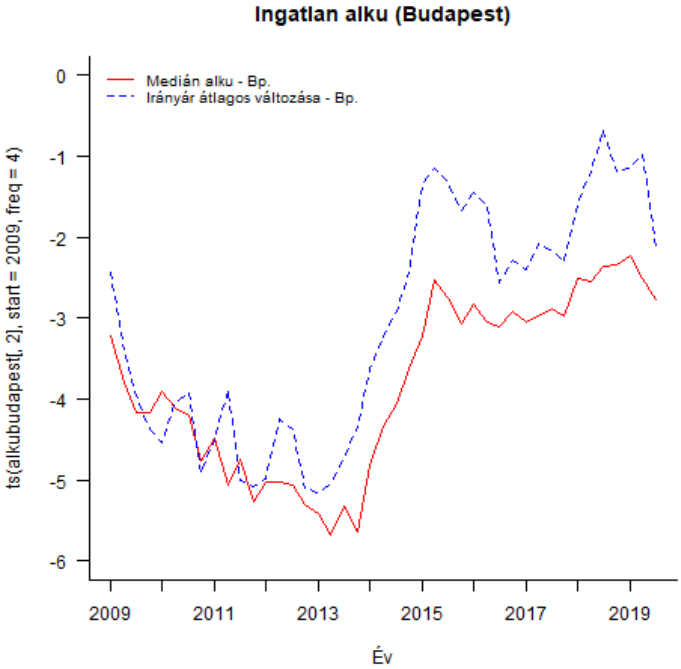
(b) Lakásárindex növekedés

A lakásár nagy mértékű növekedése a következő diagramokon jól látható. Nemcsak növekedés figyelhető meg 2014-től Budapesten, hanem egy gyorsuló növekédés, ami 2016-ban eléri a 30%-ot, amiből már következtethetünk a lakáspiaci buborék kialakulására.

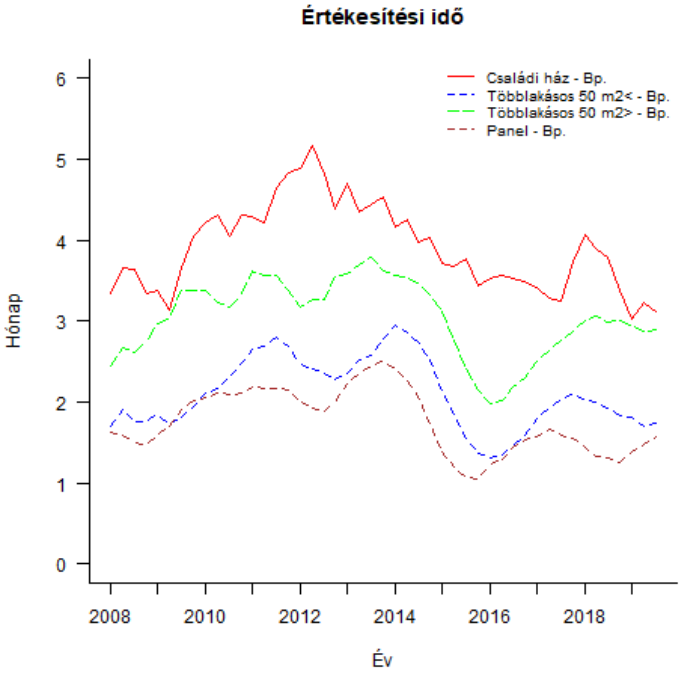
Lakáspiac alakulását jelentősen befolyásolja a gazdasági stabilitás/serkentés, a lakások iránti kereslet/kínálat és a szociális támogatások együttes hatásai. Ezekből kifolyólag olyan idősorokat gyűjtöttem a KSH [3] és az MNB [4] adatbázisából, amik segíthetnek az elemzéseim során, mint például a lakásárjövedelem alakulása különböző nagyvárosban, elégtelen kereslet, értékesítési idő és még sorohatnám. Ezenkívül összevetem Magyarország helyzetét külföldi országokéval, hogy ez a tendencia nemcsak Magyarországon található meg, hanem egész Európában is.



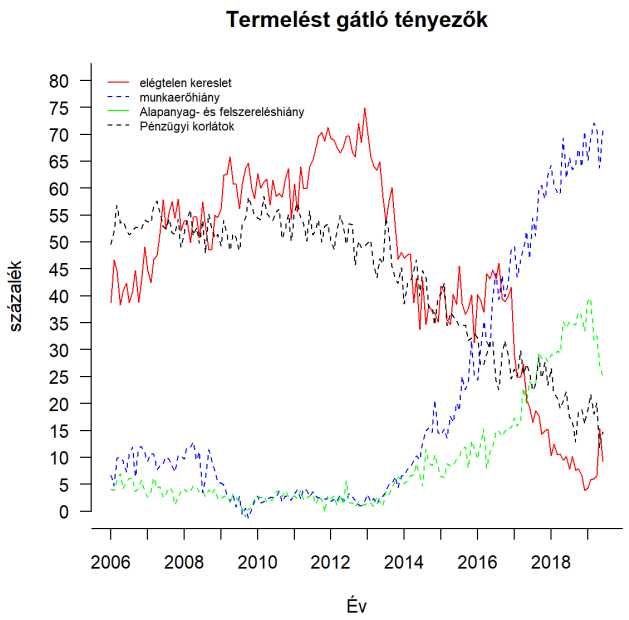
Ezen az ábrán hét nagyváros és az országos átlag lakásárjövedelem alakulását láthatjuk, hogy 2013-tól 2019-ig egy átlag lakosnak mennyit kell dolgoznia egy ingatlan megvételéhez. Itt figyelembe vesszük az adott településen kapott jövedelmet és a fogyasztással kapcsolatos költségeket. Ha a jövedelmeket, lakásárakat és a költségeket vesszük figyelembe, akkor nem Budapest áll a legrosszabb helyzetben, hanem Debrecen, ami arra enged következtetni, hogy Debrecenben sokkal rosszabb a lakáspiaci helyzet, mint Budapesten.



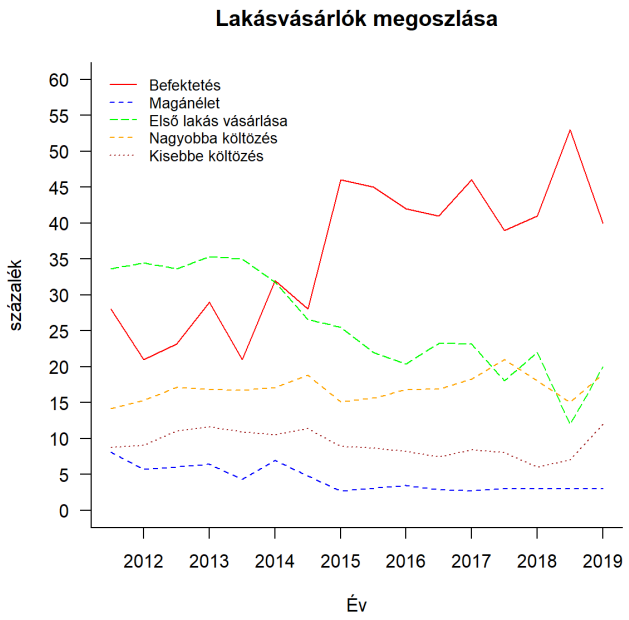
Itt a lakásvásárlók helyzetét látjuk, hogy milyen könnyen tudnak egy-egy ingatlannál alkudni a piaci szereplők. Ezen a diagrammon már szabadszemmel láthatunk egy töréspontot 2014-ben, amikor drasztikusan -5,5%-ról -2,5%-ra változott a medián alku, ezzel az eladók piaci előnyt szereztek a vásárlókkal szemben.



Itt is a lakásvásárlók helyzetét láthatjuk, hogy hány hónapig tart nekik eladni egy-egy ingatlant. Külön van szedve a családi ház, többlakásos házak $50m^2$ alatt/felett és a panel házak. Itt is látható, hogy 2014-ben átlagosan egyre hamarabb keltek el az ingatlanok. Viszont itt már egy növekedést figyelhetünk meg, a családi házakat leszámítva.



Ábrán a termelést gátló tényezőket láthatjuk, mint például az elégtelen keresletet, munkaerőhiányt, alapanyag- és felszereléshiányt és a pénzügyi korlátokat. 2014 és 2017 között egy stagnálást vehetünk észre az elégtelen keresletre, amit egy szociális támogatás, a CSOK egyre kedvezőbb feltételei mellett egy exponenciális csökkenést láthatunk.



Ezen a diagrammon a vásárlók szándékát figyelhetjük meg. Itt is észre vehetünk egy változó tendenciát 2014-től. Korábban a túlfűtöttségre használt definició, amikor csakis azért magas ma az ár, mert a befektetők holnap is magas eladási árra számítanak – amikor az „alapvető” tényezők látszólag nem igazolnak egy árat – akkor beszélhetünk buborékról. 2015-től láthatóan megemelkedtek azok a vásárlók, akik befektetés szempontjából vettek ingatlanokat. Tehát a befektetők magasabb eladási árra számítanak, de az alapvető tényezők igazolják ezeket az árakat? Statisztikai szempontból nem tudjuk az idősort felhasználni a rövidségéből adódó probléma miatt. Az idősor 2011 második felétől 2019 első feléig tartalmaz adatok, viszont ezek az adatok féléves adatok. Többdimenziós idősor modell használatakor csak kevés adatsorral tudtuk volna futtatni, amik nem megfelelő eredményhez vezettek volna.

2. Alkalmazandó statisztikai modellek

2.1. Modell alapú szórás analízis

Tegyük fel, hogy a folytonos célváltozó adataihoz két magyarázó változó áll rendelkezésre. Az egyik csoportosító jellegű, a másik folytonos. Kérdésként merül fel, hogy a célváltozónak a folytonos változó szerinti modelljei a csoportosító változó által meghatározott részhalmazokon azonos-e?

A fenti problámát abban a legegyszerűbb speciális esetben vizsgáljuk, amikor a célváltozó modellje egy egyváltozós lokális lineáris (`loess`) regresszió.

A vizsgált modell a következő:

$$y_{i,j} = m_i(x_{i,j}) + e_{i,j},$$

itt $m_i(.)$ egy esetlegesen nem-paraméteres sima függvény, az $e_{i,j}$ független, azonos eloszlású hiba. A hibáknak sem a normális eloszlása, sem pedig az azonos szórása nem feltételezett. Megengedjük, hogy a hibák az

$$e_{i,j} = \sigma_i(x_{i,j}) \cdot d_{i,j}$$

képlet szerintiek legyenek, ahol az $d_{i,j}$ független, 0 várható értékű, 1 szórású azonos eloszlású mennyiségek.
A kérdés, amit a módszer meg tud válaszolni az, hogy azok az egydimenziós regressziós görbék, vagy esetlegesen az x változók kétdimenziós léte esetén azok a regressziós felszínek, amelyek a célváltozó várható értékét a magyarázó változók függvényében leírják, statisztikailag azonosnak tekinthetőek-e vagy sem? Pontosabban:
A vizsgált hipotézis pár:

$$\begin{aligned} \mathtt{H_0} &: m_1(.) = m_2(.) = ... \\ \mathtt{H_1} &: \text{van köztük különböző} \end{aligned}$$

Az általunk alkalmazott módszer illesztett görbék `L2` távolsága alapján dönt arról, hogy az egyes csoportokban azonos függvények érvényesülnek-e. Az alkalmazott teszt statisztika az úgynevezett wild-bootstrap (vad-bootstrap).

2.1.1. Wild-bootstrap

A boostrap eljárás klasszikus referenciája D. Efron [7]. Mi ennek a módszernek egy a lényegében C.F.J. Wu szerint módosított [8] változatát alkalmazzuk. A bootstrap és speciálisan a wild-bootstrap módszer összefoglaló leírása a következő.

Általános probléma annak megadása, hogy az analitikusan nem vagy csak nehezen kezelhető statisztikák becsült paramétereinek meghatározni az eloszlását, adott szintű konfidencia tartományát és a szórását. Egyszerű eset, ha olyan nagytömegű homogén adat áll rendelkezésre, hogy annak részein is kielégítő becslését tudjunk adni a becsült paramétereknek. Ekkor ugyanis, véve a rendelkezésre álló adatok elégségesen nagy elemszámú diszjunkt részeit és kiszámítva a kérdéses paraméterek becsléseit, egy független mintát kapunk a becsült paraméterek tapasztalati eloszlásának kellő pontosságú megállapításához, a származtatható konfidencia tartományok és szórás levezetéséhez, akár közvetlenül is.

Tipikusan viszont nem áll rendelkezésre elég adat a fent vázolt eljárás követésére, ezért más módszer után kell néznünk. Ennek a problémának a kiküszöbölésére, még mindig lehetséges technika, hogy feltételezve a megfigyelt adatokat, vagy a modellhibákat egy adott eloszlás szerinti voltát, szimulációval megfigyeléseket generálunk. A generált adatokra modelleket illesszünk és az így nyert paraméter becslések alapján következtessünk a paraméterek viselkedésére.

Viszont általában nem ismert, vagy csak komoly megszorításokkal feltételezhető, hogy a megfigyelt adatok, vagy modell hibák valamely adott eloszlás szerintiek. Ebben az esetben adódik a bootstrap, mint a minta generálás egy lehetséges módszere.

A bootstrap módszer alapfeltételezése, hogy a rendelkezésre álló adatok elégségesek arra, hogy az adatok tapasztalati eloszlása jól közelítse a megfigyelések tényleges eloszlását. Márpedig ha ez fennáll, akkor pontosabb eredményt kaphatunk, ha az adatgenerálást nem a hibák feltételezett eloszlása, hanem a hibák tapasztalati eloszlása szerint végezzük. Ha elfogadjuk, hogy a statisztika eloszlásának becsléséhez szükséges mintát a megfigyelések vagy azok modell szerinti hibáinak tapasztalati eloszlása szerint generáljuk, akkor már csak azt kell észrevennünk, hogy a tapasztalati eloszlás szerinti mintavétel a rendelkezésre álló adatokból való visszatevéses mintavételnek felel meg.

A fenti megfontolások alapján, regressziós feladatok esetén kétféle bootstrap eljárás adódik. Az első a párok bootstrap-je. Ennek esetében a rendelkezésre álló célváltozó-magyarázó változó párokból veszünk visszatevéses mintát. Vagyis a minta az

$$(y_{i_j}, x_{i_j}), \quad \text{ahol } i_1, ..., i_n \text{ egy visszatevéses minta az } (1, ..., n)\text{-ig terjedő}$$

adatokból fog állni, és az így nyert adatokra illesztjük a vizsgált modellt. A második pedig a hibák bootstap-ja. Ekkor a bootstrap mintát úgy alakítjuk ki, hogy a magyarázó változó értékekhez nem azt a (hozzá tartozó) célváltozó értéket rendeljük hozzá, amelyet úgy is megkaphatunk, hogy a célváltozó illesztett modell szerinti becsléséhez hozzáadjuk az adott pontbeli hibát, hanem azt a célváltozó értéket, amelyre a becsléshez adott hibatag egy visszatevéses minta az összes (tapasztalt) hibatag közül. Vagyis a bootstrap minta ebben az esetben a

$$(\widehat{y}_i + \widehat{e}_{i_j}, x_i), \quad \text{ahol } i = 1, ..., n \text{ és az } i_1, ..., i_n \text{ egy visszatevéses minta } (1, ..., n)\text{-ből}$$

n elemű halmaz lesz, ahol $\widehat{y}_i$ az y_i célérték becslése az (y_i, x_i), $i = 1, ..., n$ adathalmaz felhasználása mellett. Az $\widehat{e}_i$ pedig ennek a becslésnek a hibája.
Tehát $y_i = \widehat{y}_i + \widehat{e}_i$ minden $i = 1, .., n$-re.

Mindkét fent vázolt módszer korrekt alkalmazhatóságának komoly feltétele, hogy az illesztett modellek homoskedasztikusak legyenek. Márpedig azokban az esetekben, amikor bootstrap eljárásra van szükség, – és a mi elemzéseink is ilyenek – homoskedaszcitás nem áll fenn. Erre az esetre megfelelő megoldás a wild bootstrap vagy magyarul vad-bootstrap.

A wild-bootstrap esetén a célváltozó bootstrap szerinti értékét a célváltozó becsült értékének az adott pontbeli hiba véletlen permutálásával nyerjük olymódon, hogy az eredeti hiba egy megfelelően standardizált értékét egy független, – tipikusan bináris – véletlen mennyiséggel szorozzuk meg. Ez a hibát szorzó véletlen mennyiség olyan, hogy a várhatóértéke 0, a szórása és a további centrális momentumai pedig lehetőleg 1-gyel egyenlőek. Ez utóbbi feltételekre azért van szükség, hogy a hiba eloszlását, vagy annak legalábbis a momentumok szerinti leírását ne változtassuk meg. Ugyanakkor ezek a szigorú kritériumok nem kielégíthetőek. Ugyanis például a 4. momentuma legalább 2 egy olyan véletlen mennyiségnek, amely 0 várható értékű, 1 szórású és amelynek a 3. momentuma is 1.

A megfelelő b_i, $i = 1, ..., n$ bootstrap szorzótényezőként két tipikus megoldás létezik. A Mammen-féle [9]:

$$b_i = \begin{cases} \frac{1-\sqrt{5}}{2} & p = \frac{5+\sqrt{5}}{10} \text{ valószínűséggel} \\ \frac{1+\sqrt{5}}{2} & 1-p \quad \text{ valószínűséggel} \end{cases}$$

Ekkor $\mathbb{E}(b_i) = 0$, $\mathbb{E}(b_i^2) = 1$, $\mathbb{E}(b_i^3) = 1$, $\mathbb{E}(b_i^4) = 2$.

A Davidson és Flachaire által [10]-ben ajánlott Rademacher-féle:

$$b_i = \begin{cases} 1 & 1/2 \text{ valószínűséggel} \\ -1 & 1/2 \text{ valószínűséggel} \end{cases}$$

Ekkor $\mathbb{E}(b_i) = 0$, $\mathbb{E}(b_i^2) = 1$, $\mathbb{E}(b_i^3) = 0$, $\mathbb{E}(b_i^4) = 1$.

Mint látható, a Rademaher-féle bootstrap tényező jó, mert nem növeli a lapultságot. Ugyanakkor rossz is, mert szükségszerűen szimmetrizál. A Mammen-féle viszont növeli ugyan a lapultságot, de nem rontja el az esetleges aszimmetriát. Fontos jellemzője mindkét megoldásnak, hogy lényegében csak annyit tesz, hogy véletlenszerűen előjelezi a hibákat. A fentiekből az is következik, hogy a wild-bootstrap módszert választva a lehetséges, különböző bootstrap minták száma n megfigyelés mellett 2^n.

A p-érték bootstrap alapján vett becslése megegyezik azzal az aránnyal, amilyen arányban valamelyik fenti módon generált minta esetén olyan becslést kapunk, amelyre a vizsgált görbék egymástól vett távolsága nagyobb, mint az eredeti minta esetén.

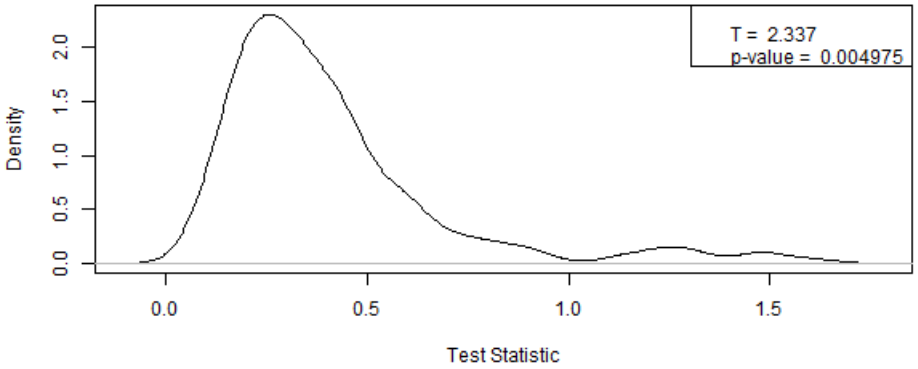
A wild-bootstrap [8] módszert abban az esetben fogjuk használni, amikor annak az eldöntése a feladat, hogy néhány trend azonosnak tekinthető-e. Számolásaink menete a következő lesz. Előbb egy lokális regressziós modellt illesztünk a rendelkezésre álló adatokra. Majd a trend görbéket adott számú pontban mintavételezve, integrálással megállapítjuk a trend görbék L_2 távolságát. Végül a wild-bootstrap alkalmazásával megadjuk, hogy az aktuális L_2 távolságnak mennyi a p-értéke.

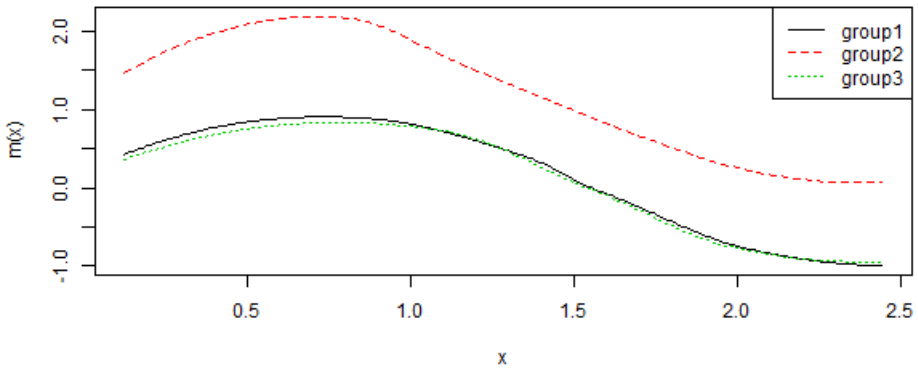
Az utóbbi lépés vázlata a következő. Vesszük a célváltozónak az eredeti adatokra illesztett alap-modell szerint becsült értékét a magyarázó változó rendelkezésre álló mérési pontjaiban. Módosítjuk ezeket a célváltozó értékeket a wild-bootstrap módszerrel nyert hibákkal. Lokális regressziós modellt illesztünk az módosított megfigyelésekre és kiszámítjuk az így kapott görbék L_2 távolságát. Többször megismételjük a célváltozó értékek alap-modell szerint becsült értékének wild-bootstrap módosítását, a regressziós modell illesztését és a módosított adatokra illesztett görbék távolságának kiszámítását. A boot-

strap módszerrel nyert p-érték az a hányad lesz, ahányszor bootstrap szimulációk során nagyobb L_2 távolságú görbéket nyertünk.

Az előzőekben leírt módszer szerint működik az R-project '`fANCOVA`' csomagjának `T.L2()` parancsa. A továbbiakban röviden ennek a módszernek az alkalmazás technikáját ismertetjük.

```
rm(list=ls())
library(fANCOVA)
set.seed(123)
# a módszer tesztje 3 csoport eseten
# generalt adatokkal mutatjuk be a parancs mukodeset

# egy dimenzios magyarazo adatok
data.gen<- function(g,n,m,s)
 {
  x <- runif(n,min=0, max=3);
  y <- sin(2*x) + rnorm(n,m,s)
  return(cbind(x,y,g))
 }

adat <- data.frame(rbind(data.gen(1,20,0,.20),
                         data.gen(2,20,1,.25),
                         data.gen(3,25,0,.25)))

M <- T.L2(adat$x, adat$y, adat$g, degree=2)
M

# Test the equality of curves based on L2 distance
#
# Comparing 3 nonparametric regression curves
# Local polynomial regression with automatic smoothing
# parameter selection via AICC is used for curve fitting.
# Wild-bootstrap algorithm is applied to obtain the null
# distribution.
#
# Null hypothesis:
#   there is no difference between the 3 curves.
# T =  2.337     p-value =  0.004975



plot(M) # a teszt statisztika eloszlásának a képe
plot(M, test.statistic=FALSE) # a becsült görbék
```





Ebben a részben leírt modellt a túlfűtöttség problémájával kapcsolatban vizsgált `ár/jövedelem` adathalmaz elemzésekor fogjuk felhasználni.

2.2. Többfázisú regresszió

A többfázisú regresszió egy igen általános regressziós módszer. Egy fontos speciális esete a spline-regresszió. Mi ezen regressziós módszereknek is csak azzal a speciális esetével foglalkozunk, amikor a regressziós függvény ismeretlen csomópontokban törő, szakaszonként lineáris függvény.

A törött vonal szerinti regresszió problémájának megoldásához a regressziós függvényt előbb zárt, paraméteres formában írjuk fel.

<u>Példa 1.</u>
Álljon a közelítő függvény mindössze két szakaszból. Legyen a töréspont ψ-ben. Legyen a függvény értéke 0-ban B_0. Legyen a függvény meredeksége ψ-től balra B_1 jobbra

pedig B_2. Ekkor a keresett $f(z) = f_{B_0,B_1,B_2}(z)$ függvény a

$$\begin{aligned}\beta_0 &= B_0\\ \beta_1 &= B_1\\ \beta_2 &= B_2 - B_1\end{aligned}$$

paraméterekkel a

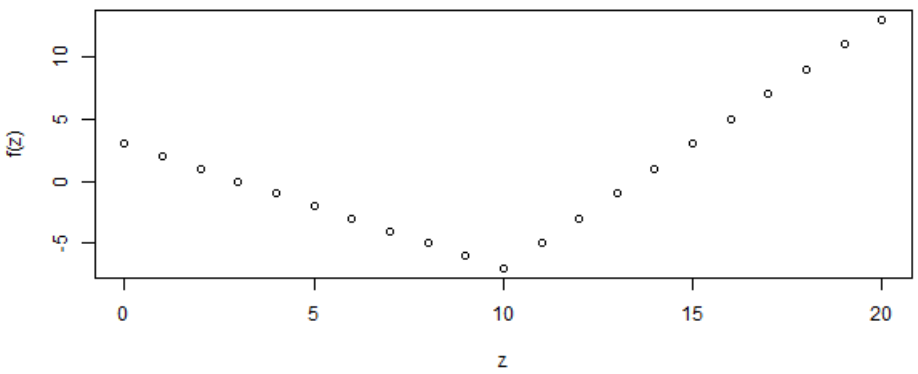
$$f(z) = \beta_0 + \beta_1 \cdot z + \beta_2 \cdot (z - \psi) \cdot \chi_\psi(z)$$

formába írható, ahol a $\chi_c(x)$ a $\chi_c(x) = ind(x > c)$ függvényt jelöli, aminek 1 az értéke, ha $x > c$ és 0, ha nem.
A függvény egy alternatív formulája

$$f(z) = \beta_0 + \big(\beta_1 + \beta_2 \cdot \chi(z)\big) \cdot z - \chi_\psi(z) \cdot \psi \cdot \beta_2$$

A függvény demonstrációja **R**-ben:

```
B0  <-  3 # a konstans es a ket meredekseg
B1  <- -1
B2  <-  2

psi <- 10 # a toresi pont
chi <- function(x) ifelse(x>psi,1,0)

b0 <- B0
b1 <- B1
b2 <- B2-B1

f<-function(x) b0+b1*x+b2*(x-psi)*chi(x)

z <- 0:20
y_alt <- b0+(b1+b2*chi(z))*z-chi(z)*psi*b2
sum((f(z)-y_alt)^2) # ugyanaz

f(0) # B0 az atmetszes jo
       B0 # 3

diff(f(z)) # B1,...B1,B2,...,B2 a meredeksegek jok
c(B1,B2) # -1, 2

plot(z,f(z))
```



Példa 2.

Álljon a közelítő függvény három szakaszból. Legyenek a töréspontok ψ_1-ben és ψ_2-ben. Legyen a függvény értéke 0-ban B_0. Legyen a függvény meredeksége ψ_1-től balra B_1 ψ_1 és ψ_2 közt B_2 és ψ_2-től jobbra B_3. Ekkor a keresett $f(z) = f_{B_0,B_1,B_2,B_3}(z)$ függvény a

$$\begin{aligned}
\beta_0 &= B_0 \\
\beta_1 &= B_1 \\
\beta_2 &= B_2 - B_1 \\
\beta_3 &= B_3 - B_2
\end{aligned}$$

paraméterekkel a

$$f(z) = \beta_0 + \beta_1 \cdot z + \beta_2 \cdot (z - \psi_1) \cdot \chi_{\psi_1}(z) + \beta_3 \cdot (z - \psi_2) \cdot \chi_{\psi_2}(z)$$

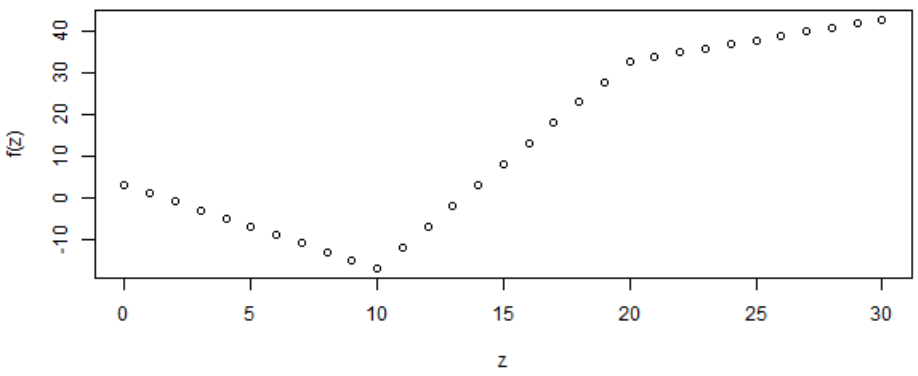
formába írható. A függvény egy alternatív formulája

$$\begin{aligned}
f(z) &= B_0 + \\
&\quad B_1 \cdot (\chi_0(z) \cdot z - \chi_{\psi_1}(z) \cdot (z - \psi_1)) + \\
&\quad B_2 \cdot \big((\chi_{\psi_1}(z) - \chi_{\psi_2}(z)) \cdot z - \chi_{\psi_1}(z)\psi_1 + \chi_{\psi_2}(z) \cdot \psi_2\big) + \\
&\quad B_3 \cdot \chi_{\psi_2}(z) \cdot (z - \psi_2)
\end{aligned}$$

A függvény demonstrációja **R**-ben

```
rm(list=ls())

B0  <-  3 # a konstans es a harom meredekseg
B1  <- -2
B2  <-  5
B3  <-  1
```

```
ps1 <- 10 # a toresi pont
ps2 <- 20 # a toresi pont
ch0 <- function(x) 1
ch1 <- function(x) ifelse(x>ps1,1,0)
ch2 <- function(x) ifelse(x>ps2,1,0)
chT <- function(x) 0

b0 <- B0
b1 <- B1
b2 <- B2-B1
b3 <- B3-B2

f<-function(x) b0+b1*x+b2*(x-ps1)*ch1(x)+b3*(x-ps2)*ch2(x)

z <- 0:30
y_alt <- B0 - ch1(z)*(B2-B1)*ps1 -ch2(z)*(B3-B2)*ps2 +
         B1*z +ch1(z)*(B2-B1)*z +ch2(z)*(B3-B2)*z
sum((f(z)-y_alt)^2) # ugyanaz
yx_alt <- B0 +
          B1*(ch0(z)*z-ch1(z)*(z-ps1)) +
          B2*((ch1(z)-ch2(z))*z -ch1(z)*ps1+ch2(z)*ps2) +
          B3*ch2(z)*(z-ps2)
sum((f(z)-yx_alt)^2) # ugyanaz


f(0) # B0 az atmetszes jo
       B0 # 3

diff(f(z)) # B1,...B1,B2,...,B2,B3,...,B3 a meredekseg jo
c(B1,B2,B3) # -2, 5, 1

plot(z,f(z))
```



Példa 3.

Álljon a közelítő függvény négy szakaszból. Legyenek a töréspontok (ψ_1, ψ_2, ψ_2). Legyen a függvény értéke 0-ban B_0. Legyen a függvény a négy szakaszban rendre B_1, B_2, B_3, B_4. Ekkor a keresett $f(z) = f_{B_0,B_1,B_2,B_3,B_4}(z)$ függvény a

$$\begin{aligned}
\beta_0 &= B_0 \\
\beta_1 &= B_1 \\
\beta_2 &= B_2 - B_1 \\
\beta_3 &= B_3 - B_2 \\
\beta_4 &= B_4 - B_3
\end{aligned}$$

paraméterekkel a

$$f(z) = \beta_0 + \beta_1 \cdot z + \beta_2 \cdot (z - \psi_1) \cdot \chi_{\psi_1}(x) + \beta_3 \cdot (z - \psi_2) \cdot \chi_{\psi_2}(x) + \beta_4 \cdot (z - \psi_3) \cdot \chi_{\psi_3}(x)$$

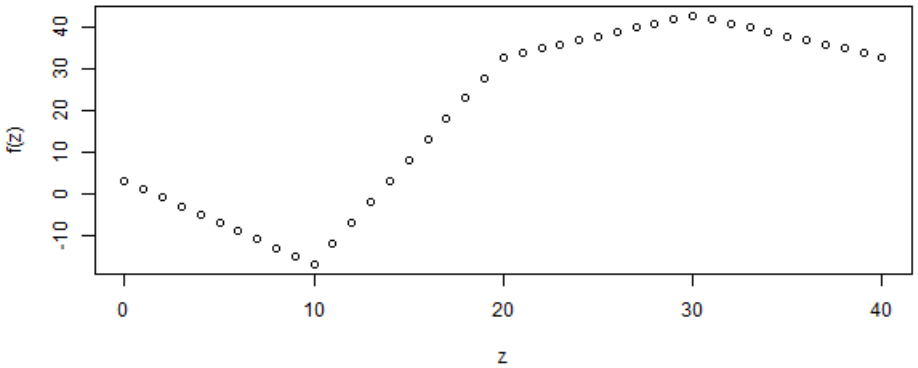
formába írható. A függvény egy alternatív formulája

$$\begin{aligned}
f(z) \;=\; & B_0 + \\
& B_1 \cdot \big(z \cdot (\chi_{\psi_0}(z) - \chi_{\psi_1}(z)) + \psi_1 \cdot \chi_{\psi_1}(z)\big) + \\
& B_2 \cdot \big(z \cdot (\chi_{\psi_1}(z) - \chi_{\psi_2}(z)) + \psi_2 \cdot \chi_{\psi_2}(z) - \psi_1 \cdot \chi_{\psi_1}(z)\big) + \\
& B_3 \cdot \big(z \cdot (\chi_{\psi_2}(z) - \chi_{\psi_3}(z)) + \psi_3 \cdot \chi_{\psi_3}(z) - \psi_2 \cdot \chi_{\psi_2}(z)\big) + \\
& B_4 \cdot \big(z \cdot (\chi_{\psi_3}(z) - \chi_{\psi_T}(z)) + -\psi_3 \cdot \chi_{\psi_3}(z)\big)
\end{aligned}$$

A függvény demonstrációja

```
rm(list=ls())

B0  <-  3 # a konstans es a negy meredekseg
B1  <- -2
B2  <-  5
B3  <-  1
B4  <- -1

ps1 <- 10 # a toresi pontok
ps2 <- 20 #
ps3 <- 30 #
ch0 <- function(x) 1
ch1 <- function(x) ifelse(x>ps1,1,0)
ch2 <- function(x) ifelse(x>ps2,1,0)
ch3 <- function(x) ifelse(x>ps3,1,0)
chT <- function(x) 0

b0 <- B0
b1 <- B1
```

```
b2 <- B2-B1
b3 <- B3-B2
b4 <- B4-B3

f <- function(x) b0 +
                 b1*x +
                 b2*(x - ps1)*ch1(x) +
                 b3*(x - ps2)*ch2(x) +
                 b4*(x - ps3)*ch3(x)

T<-40
z <- 0:T

y_alt <- B0 +
       B1*(z*(ch0(z)-ch1(z)) + ps1*ch1(z) - 0          ) +
       B2*(z*(ch1(z)-ch2(z)) + ps2*ch2(z) - ps1*ch1(z)) +
       B3*(z*(ch2(z)-ch3(z)) + ps3*ch3(z) - ps2*ch2(z)) +
       B4*(z*(ch3(z)-chT(z)) + 0          - ps3*ch3(z))

sum((f(z)-y_alt)^2) # ugyanaz

f(0) # B0 az atmetszes jo
       B0 # 3

diff(f(z)) # B1,...B1,B2,...,B2,B3,...,B3,B4,...,B4 OK
c(B1,B2,B3,B4) #  -2  5  1 -1

plot(z,f(z))
```



2.2.1. Többfázisú regresszió illesztési algoritmusa

Az eljárást az egy töréspontú esetben mutatjuk be. [5] Az illesztés módja iteratív. Megengedjük, hogy a

$$f(z) = \beta_1 + \beta_2 \cdot (z - \psi) \cdot \chi(z > \psi)$$

modell szerinti adatok esetén ψ_k még ne az 'igazi', hanem az optimális töréspont legyen. Ezt a 'hibát' egy a ψ_k ponttól jobbra, egy nem nulla konstans hozzáadásával korrigáljuk:

$$f(z) = \beta_1 + \beta_2 \cdot (z - \psi_k) \cdot \chi(z > \psi_k) + \gamma \cdot \chi(z > \psi_k)$$

Az aktuális ψ_k és a β_1, β_2, γ paraméterek becsült értéke alapján javíthatjuk a ψ_k értékét:

$$\psi_{k+1} = \psi_k - \widehat{\gamma} \Big/ \widehat{\beta}_2$$

Ha az eljárás konvergál, akkor a $\widehat{\gamma} \approx 0$. [6]

2.3. Többdimenziós idősor modellek

A legegyszerűbb idősor modellek a lineáris folyamat modellek. Ezek sokfele alkalmazott speciális esetei a könnyen és jól illeszthető egyváltozós `ARMA` folyamatok. Az egydimenziós `ARMA` folyamatok többdimenziós változatának illesztése, alkalmazása számos, az egydimenziós illesztési feladatnál lényegesen összetettebb feladat megoldását igényli. Ugyanakkor többdimenziós megfigyelések esetén a megfigyelés egyes koordinátáinak az értékét tipikusan jobban meg lehet jósolni, ha ehhez nemcsak az adott koordinátának a múltbeli értékeit, hanem a többi koordináta múltbeli értékeit is felhasználjuk. A következőkben a többdimenziós `ARMA` folyamatok egy részmodelljével, az `AR` folyamatokkal foglalkozunk. Ezeket a modelleket fogjuk alkalmazni az adathalmaz elemzésekor.

Az alkalmazott modell a következő $\mathtt{AR}(1)$ modell:

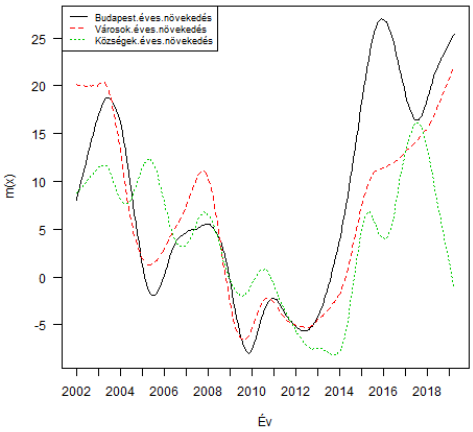
$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \varepsilon_t$$

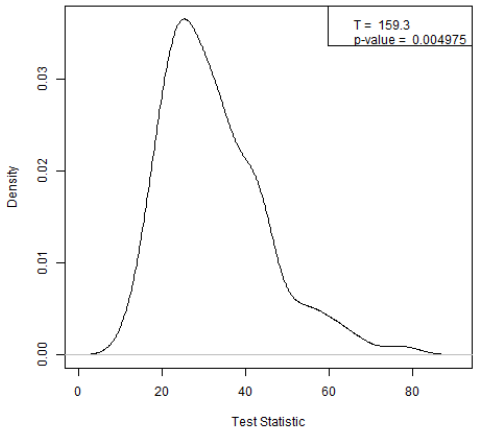
Mi esetünkben az Y_t egy 11 dimenziós folyamat, amelyre vonatkozóan egy 42 negyedévnyi megfigyeléssor áll rendelkezésre. Ennek megfelelően, a Φ_0 egy 11 dimenziós konstans vektor, a Φ_1 pedig egy 11×11 méretű mátrix, amely azt adja meg, hogy a t időpontbeli értékek hogyan függnek a $t-1$ időpontbeli értékektől. Az ε_t egy 11 dimenziós zaj sorozat, amelyről feltételezzük, hogy 0 várható értékű, korrelálatlan fehérzaj.

3. Statisztikai elemzés

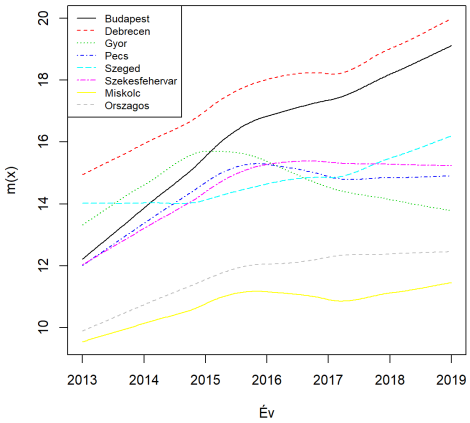
3.1. Modell alapú szórás analízis

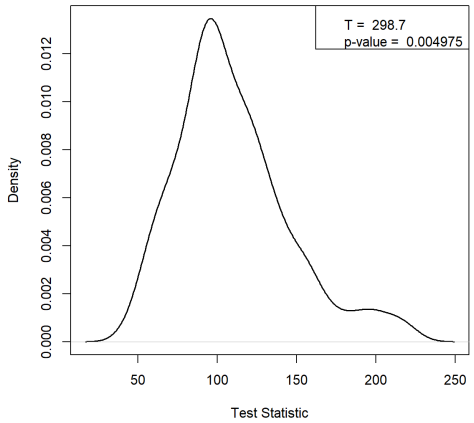
Első lépésben a lakásárindex változás ütemét vizsgáltam meg, hogy eltérően viselkedett-e az adott változó Budapest, városok és községek esetén. Ennek regresszió görbéjét a bal oldali diagrammon figyelhetjük meg. A jobb oldali diagrammon pedig az L_2 távolság tapasztalati eloszlását, a T-vel jelölt aktuális L_2 távolságot és a hozzá tartozó p értéket láthatjuk. A magas T érték és az alacsony p-érték miatt *elutasítjuk* azt a feltételezést, hogy statisztikailag azonosnak tekinthetőek.



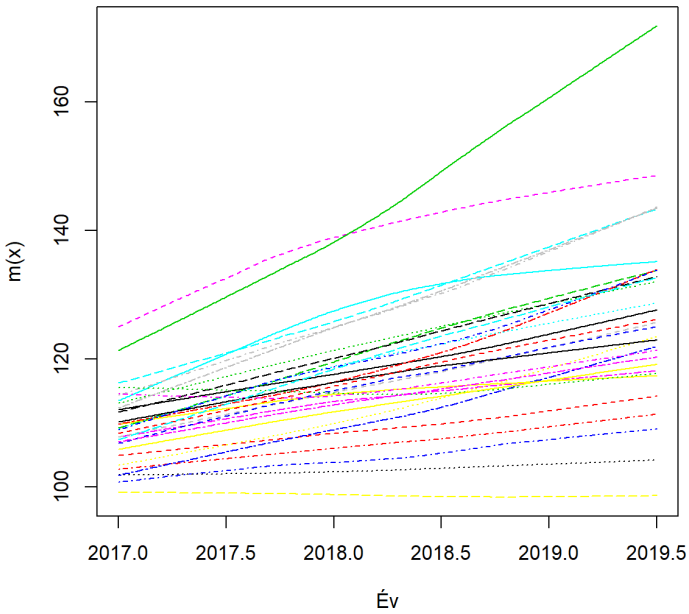


Következő esetben a lakásár/jövedelem alakulását figyeltem meg. A magas T érték és az alacsony p-érték miatt elutasítjuk azt a feltételezést, hogy az adatsoraink azonosnak tekinthetőek-e. Viszont Budapest és Debrecen esetén a regressziós görbék hasonlóan viselkednek, egy párhuzamosságot mutatnak 2017-től.

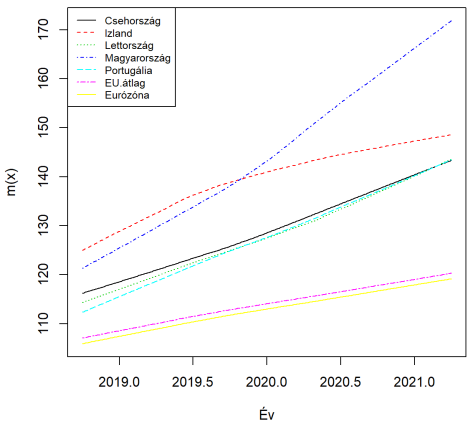


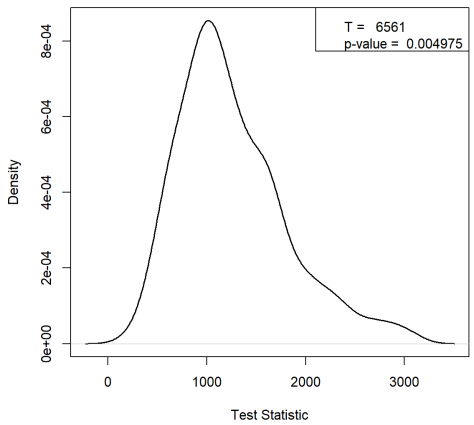


A következő diagrammon 29 ország adatai, az EU-átlag és az Eurózóna adatai tekinthetőek meg, hogy hogyan alkultak a lakásárindexeik. Mivel nagyon sok adatokat láthatunk és elvesznénk a vonalak sokaságában, ezért lecsökkentettem a dimenzió számokat. Megkerestem azokat az adatokat, amik eltérően viselkedtek az idősorban és hozzávettem az EU-átlagot és az Eurózóna átlagát. Itt már könnyebben tudjuk követni az adatokat.



Magyarország mellett még kiemelkedően teljesített országok között találhatjuk Csehországot, Izlandot, Lettországot és Portugáliát. Modellezés során viszont ezek az országok mind-mind alul maradtak Magyarországhoz képest, nem volt olyan magas lakásárindex növekedés ezekben az országokban, mint nálunk. Statisztikai értékünk is arra enged következtetni, hogy elutasíthatjuk azt a feltételezést, hogy azonosnak tekinthetőek a regressziós görbék.

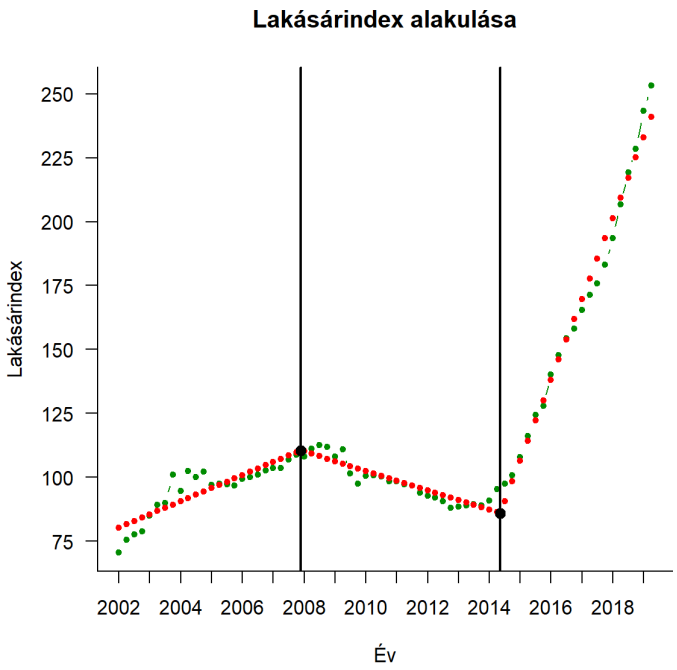




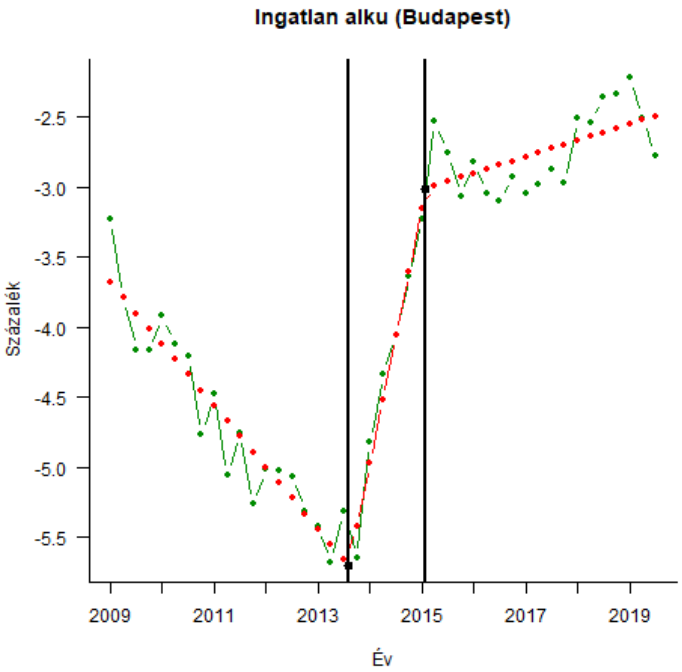
3.2. Többfázisú regressziós modell

Többfázisú regressziós modellel vizsgáltuk meg, hogy a rendelkezésre álló idősorokban mely időpontokban következett be strukturális változás. Az alábbi ábrákon ennek az elemzésnek az eredményei láthatóak.

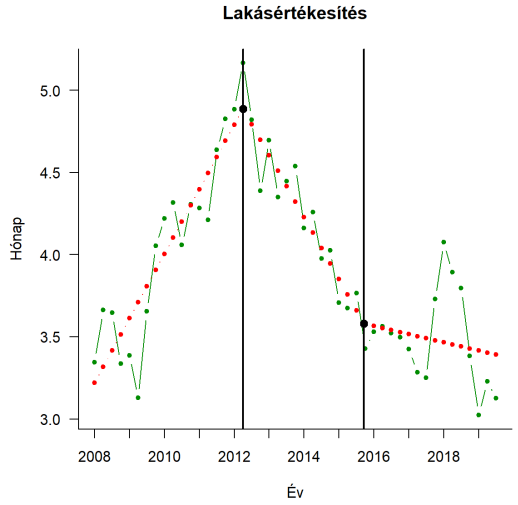
Budapesten 2014-től elindul egy növekedés, ami statiszikai módszerrel bizonyítható. A későbbiekben ezt az értéket szeretnénk megjósolni, ezért megnéztem mely idősorok esetén figyelhetőek meg hasonló változások 2014 előtt.



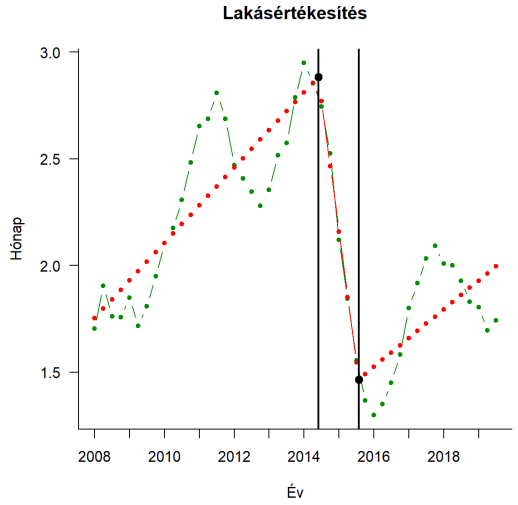
Medián alku esetén 2013 harmadik negyedévétől figyelhető meg ez a töréspont, ami 2015-ig tartott és ettől az évtől egy lassabb növekedést figyelhetünk meg. Mivel a töréspont korábbra tehető, mint a lakásárindex töréspontja, ezért feltételezhetjük azt az állítást, hogy ez az idősor egy befolyásoló tényezője lehet a lakásárindexnek.



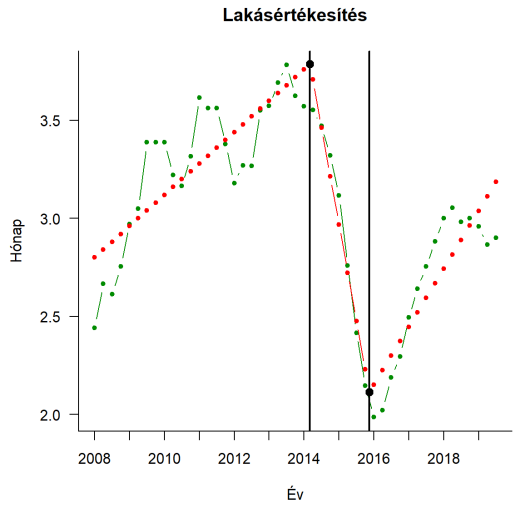
Következő diagrammokon lakástípusonként láthatjuk a hónapokban számolt átlagos értékesítési idő többfázisú regressziós modelljét.



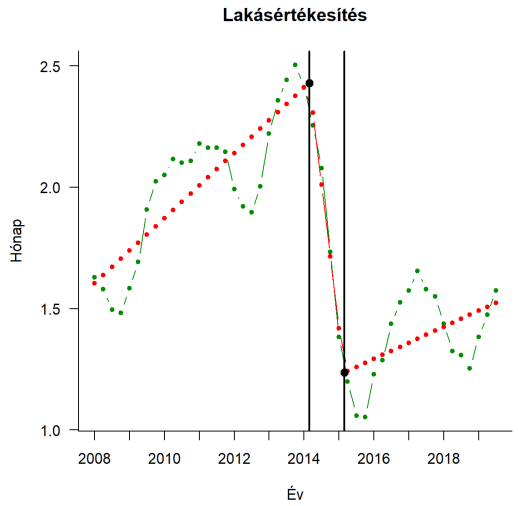
(a) Lakás



(b) Többlakásos 50 m^2<



(c) Többlakásos 50 m^2>



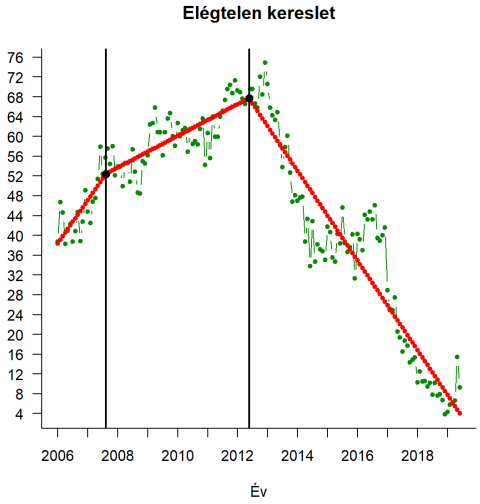
(d) Panel

Az illesztett modellek azt mutatják, hogy az értékesítési időben a lakások esetén – a 2014 körüli lakásárindex váltásnál sokkal korábban, – 2012-ben markáns csökkenést figyelhető meg. Míg a többlakásos és panel kínálat esetén a váltási időpont a 2014-essel nagyjából megegyezik.

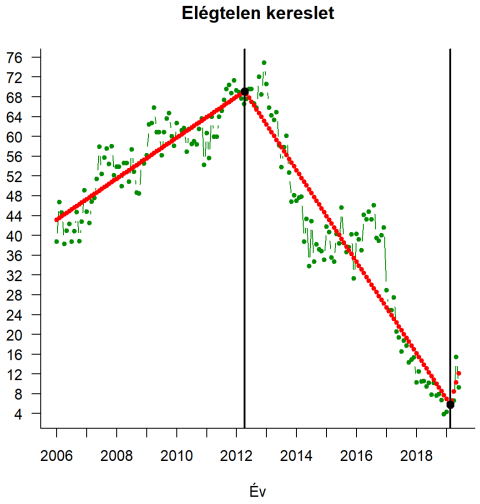
Az így nyert struktura váltási időpontokat az idősormodellezés során használjuk fel.

Későbbiekben láthatjuk, hogy a többdimenziós idősor modell nem tudja jól kezelni az elégtelen keresletet, ezért alaposabban megvizsgáltam ennek az adatsornak a többfázisú regressziós modelljét.

`Segmented()` függvény többszöri lefutása után észrevehetjük, hogy több helyre rakja le a töréspontokat. 2019-től megfigyelhetünk egy változó tendenciát.



(a) példa 1



(b) példa 2

E különböző eredményeknek az a magyarázata, hogy az illesztés során felhasznált likelihood függvény a töréspontok környékén nem feltétlenül konkáv. Az ezt vizsgáló teszt futtatások során azt tapasztaltuk, hogy a töréspont leggyakrabban a fenti két, példa 1 és példa 2 alábrának megfelelően a 2012-es évre került. A bizonytalanság a másik töréspont körül adódott. A két alábra a másodlagos törés két leggyakoribb esetét mutatja, amikor a másodlagos törés a 2007-es illetve 2019-es évre kerül.

3.3. Többdimenziós idősor modell

Korábban a többfázisú regressziós modellben megvizsgált adatokat használtam fel a többdimenziós idősor modellben. Az alábbi diagrammon azt tekinthetjük meg, hogy különböző változóinkat milyen mértékben folyásolják be az adott változók. Budapest három változó folyásolta be leginkább: város, munkaerő (mint termelést gátló tényező) és a kereslet (mint termelést gátló tényező).

	Budapest	Város	Község	alku	IrányárDiff	Családi	Nagy	Kicsi	Panel	kereslet	munkaerő
Budapest	0.72	0.17	-0.04	0	0.02	0.03	-0.02	0.01	-0.03	-0.07	0.16
Város	0.95	0.37	0.04	0.07	-0.16	0.01	0	0.05	-0.07	-0.17	-0.35
Község	0.03	0.16	0.68	0.56	-0.56	-0.02	0.07	-0.04	0.01	-0.08	0.06
alku	0.59	-0.77	0.18	0.53	0.04	-0.05	-0.04	-0.03	-0.07	-0.63	-0.33
IrányárDiff	-0.26	-0.14	-0.15	0.42	0.16	0.11	0.01	0.08	-0.21	-0.69	0.1
Családi	-0.71	-0.14	0.26	-0.19	-0.5	0.43	0.22	0	-0.39	0.03	0.6
Nagy	0.99	-0.71	0.24	-0.23	0	-0.02	0.54	-0.1	0.57	-0.19	-0.37
Kicsi	-2.13	2.05	-0.25	-0.47	0.06	-0.15	-0.01	0.66	0.14	0.19	0.83
Panel	1.11	-0.37	0.03	-0.35	-0.16	-0.13	-0.28	0.06	0.9	0.06	-0.53
kereslet	0.05	0.07	0.12	-0.14	-0.03	0.03	0.11	-0.15	-0.2	0.66	-0.55
munkaerő	0.63	-0.45	0.03	0.01	-0.01	0.01	-0.02	-0.04	0.06	-0.1	0.73

A táblázat Φ_1 együttható értékeit mutatja standardizált változók esetén. A diagonálisban az látható, hogy az egyes folyamat koordináták milyen mértékben függnek a saját korábbi értéküktől. Érdemes észrevenni, hogy a "kereslet", "munkaerő" valamint a "Budapest" és "Község" változókat kivéve egyik változó jelenértéke sem a saját multjától függ a legjobban. Ez bizonyítja, hogy a lakáspiaci helyzetet feltétlen mint többdimenziós folyamatot kell kezelni.

4. Összegzés

Az elemzés során Budapesten felfigyelhetőek a túlfűtöttség jelei. Definicióban szereplő befektetők aránya megemelkedett Budapesten, ami a növekvő kereslet miatt a lakásárak emelkedését segítette elő. Modell alapú szórás analízis során nem tudtuk egyik hipotézist sem elfogadni, miszerint hasonló változások voltak Magyarországon belül. Európán belül is csak nálunk találhattunk ilyen mértékű növekedést. Többfázisú regressziós modell esetén a keresletnél mint termelést gátló tényező egy strukturális változást figyelhetünk meg, ami arra enged következtetni, hogy változásokra számíthatunk az elkövetkező időkben.

Hivatkozások

[1] D. DiPasquale, WC. Wheaton: Housing market Dynamics and the Future od Housing Prices; J. Urban Economics 35, pp1-27, 1994.

[2] Svenja Gudell, Zillow:Total Value of All U.S. Homes:$31.8 Trillion Forbes https://www.forbes.com/sites/zillow/2018/01/03/total-value-of-all-u-s-homes-31-8-trillion/#59dfc78d3ca8

[3] https://www.ksh.hu/stadat; (elérés időpontja: 2020 január).

[4] www.mnb.hu; https://www.mnb.hu/letoltes/lakaspiaci-jelentes-abrakeszlet-2019 - november.xlsx; (elérés időpontja: 2020 január).

[5] Vito M. R. Muggeo: 'segmented': An R Package to Fit Regression Models with Broken-Line Relationships; R-News – The Newsletter of the R Project; Volume 8/1, pp20-25, May 2008.

[6] V. Muggeo: Estimating regression models with unknown break-points; Statistics in Medicine; Vol 22: pp 3055-3071, 2003.

[7] B. Efron: Bootstrap methods: Another look at the jackknife; Ann. Statist. 7 pp 1-26, 1979.

[8] C. F. J. Wu: Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis; Ann. Statist. 7 pp 1261-1350, 1986.

[9] Mammen, E.: Bootstrap and wild bootstrap for high dimensional linear models; Ann. Statist. 21, pp 255–285, 1993.

[10] Davidson, R., E. Flachaire: The wild bootstrap, tamed at last; J. of Econometrics, 146, pp 162–169, 2008.

[11] R Core Team; R Foundation for Statistical Computing; R: A Language and Environment for Statistical Computing; https://www.R-project.org; Vienna, Austria; 2020.

[12] Xiao-Feng Wang: 'fANCOVA': Nonparametric Analysis of Covariance; https://CRAN.R-project.org/package=fANCOVA; R package version 0.5-1; 2010.