Ökonometria

3. házi feladat

Granát Marcell

2020. december 29.

Tartalomjegyzék

1. feladat

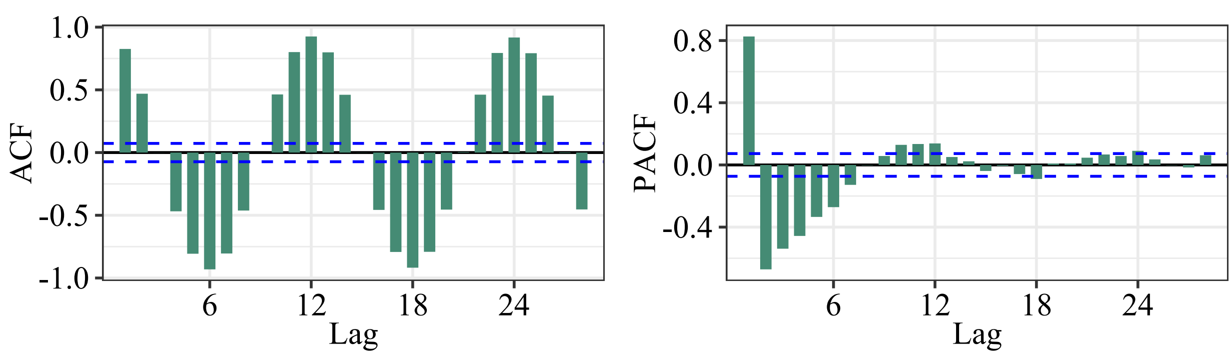
A mellékelt bptempm.csv fájl tartalmazza a budapesti havi átlaghőmérsékletet 1901 januárja és 2000 decembere között. (A read.csv függvény használható az adatok beolvasására.) A hallgatóknak különböző, 60 éves idősorokat kell elemezniük, a Neptun-kódjuk első karaktere alapján: 1901-1960 (A-C), 1911-1970 (D-F), 1921-1980 (G-I), 1931-1990 (J-P), 1940-1999 (Q-Z vagy szám).

```
library(tidyverse)
library(granatlib) # my personal package: https://github.com/MarcellGranat/granatlib
library(tseries)
library(forecast)
theme_set(theme_granat())
rio::import('bptempm.csv') %T>%
  {filter(., 190012 < ym & ym < 196101) %T>%  # data import, NEPTUN: AYCOPF
      {dat <<- .} %>%
      pull(temp) %>%
      ts(start = c(1911, 1), frequency = 12) %>%
      {x <<- .} # as ts
  } %>% mutate(
    AR = lag(temp) # auto-regressive term
  ) %>%
  filter(ym < 196201 & ym > 196012) %>% transmute(
    t = V1,
    m = as.factor(ym %% 100),
    temp, AR
  ) %>%
  {test_dat <<- .} # test data for 1.j
```

a)

Ábrázoljuk az idősor autokorreláció-függvényét! Mire utal a függvény alakja?

```
x %>% {
  ggpubr::ggarrange(
    forecast::ggAcf(., size = 2, color = "aquamarine4") + labs(title = ''),
    forecast::ggAcf(., size = 2, type = 'partial', color = "aquamarine4") +
      labs(title = '')
  )
}
```



1. ábra. A budapesti havi átlaghőmérséklet korrelogramja

A autkorreláció-függvény alakja arra utal, hogy szezonalitás van a modellben (1. ábra).

b)

Illesszünk lineáris trendet és hónap-dummykat tartalmazó modellt az idősorra!

```
mod1 <- dat %>% transmute(
  t = V1,
  m = as.factor(ym %% 100),
  temp = temp
) %>% lm(formula = temp ~ .)
```

c)

Értelmezzük a modell paramétereit!

```
f.month <- function(x) { # for nice outputs
  y <- switch(x,
              't' = 'trend',
              'm1' = 'január',
              'm2' = 'február',
              'm3' = 'március',
              'm4' = 'április',
              'm5' = 'május',
              'm6' = 'június',
              'm7' = 'július',
              'm8' = 'augusztus',
              'm9' = 'szeptermber',
              'm10' = 'október',
              'm11' = 'november',
              'm12' = 'december',
              'AR' = 'AR',
              'AR2' = 'AR(2)'
  )
  ifelse(is.null(y), x, y)
}

mod1 %>% broom::tidy() %>%
  select(1:2) %>% mutate(
    term = sapply(term, f.month)
  ) %>%
  prtbl('1. modell becsült paramétereinek értéke', ufc = T)
```

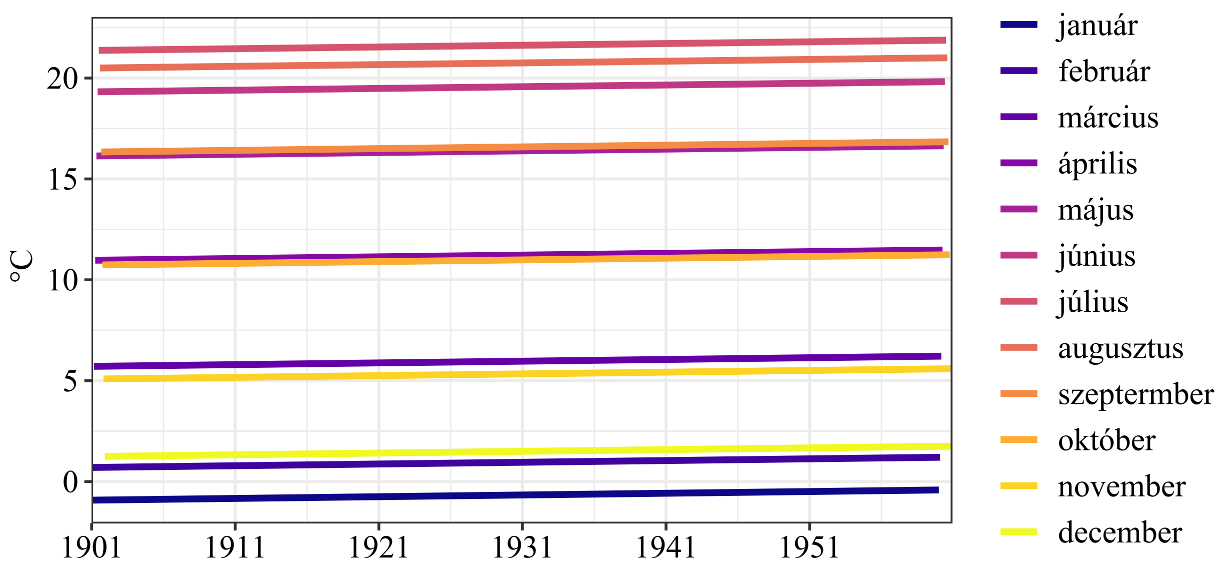
1. táblázat: 1. modell becsült paramétereinek értéke

Változó	Koefficiens
Konstans	-0,92
Trend	0,00
Február	1,62
Március	6,63
Április	11,89
Május	17,06
Június	20,23
Július	22,29

Változó	Koefficiens
Augusztus	21,41
Szeptermber	17,25
Október	11,65
November	6,00
December	2,16

A modellben szerepel 1 lineáris trend komponens, 1 konstans és 11 dummy változó. Előbbi megmutatja, hogy hónapról hónapra átlagosan miként módosul az átlaghőmérséklet kontrolálva az adott hónapok jellemző átlaghőmérsékletére, ami a konstans és a 11 dummy változó alapján határozódik meg. Az egyes dummy változók értékei megmutatják, hogy c.p. várhatóan mennyivel nagyobb az átlaghőmérséklet egy adott hónapban, mint januárban (referencia érték). Például a $\beta_{Február}$ alapján elmondható, hogy a februári hónapokban c.p. átlagosan 1,62°C-kal van melegebb, mint a januári hónapokban. Januári hónap esetén a várható átlagos hőmérsékletet a $\beta_{Konstans}$ adja meg. Ezek az értékek azonban csak 1911. januárjában igazak, mivel mindegyik hónapban nő a várható érték a β_{Trend} értékkel. Az így meghatározott becsült átlagos havi hőmérsékletet a 2. ábra mutatja be.

```
mod1 %>%
  broom::augment() %>%
  mutate(m = sapply(str_c('m', m), f.month)) %>%
  mutate(m = factor(m, levels = unique(m))) %>%
  {ggplot(.) +
  geom_line(aes(t, .fitted, color = m), size = 1.2) +
  scale_x_continuous(expand = c(0,0),
                     breaks = seq(from = 1, to = nrow(.), by = 120),
                     labels = seq(from = 1901, to = 1960, by = 10)
                     )} +
  scale_color_viridis_d(option = 'C') +
  labs(x = NULL, y = expression(''*~degree*C*''), color = NULL) +
  theme(legend.position = 'right')
```



2. ábra. Modell alapján becsült átlaghőmérséklet a megfigyelt időszakra hónap szerinti csoportosításban

d)

Teszteljük, hogy a modell reziduálisai autokorreláltak-e!

```
f.Ljung.Box.p <- function(x) { # frequently used in this assignment ---> function(...)
  capture.output( # avoid redundant printing
    checkresiduals(x, plot = F)
  ) %>% .[5] %>% # main text description
    str_remove('.*p-value = ') %>% # extract p-value of the test
    as.numeric() %>%
    scales::percent(decimal.mark = ',', accuracy = .01)
}

answer_1d <- f.Ljung.Box.p(mod1) # used as inline code as the p-value
```

A modell rezidulálisainak autokorrelációra való tesztelésére Ljung-Box tesztet alkalmaztam 16[1] késleltetéssel. Miután a p-érték 0,00%, így elvetjük a H_0-t, miszerint nem lenne autokorreláció a maradéktagban. Tehát autokorreláltak a reziduálisok.

e)

Illesszünk lineáris trendet, hónap-dummykat és elsőrendű autoregresszív tagot tartalmazó modellt az idősorra!

```
mod2 <- dat %>% transmute(
  t = V1,
  m = as.factor(ym %% 100), # shortcut for getting the month variable
  AR = lag(temp), # auto-regressive part
  temp = temp
) %>% na.omit() %>%
  lm(formula = temp ~ .)
```

f)

Teszteljük, hogy a reziduálisok autokorreláltak-e!

```
answer_1f <- f.Ljung.Box.p(mod2) # defined in this doc formerly
```

A már autoregresszív tagot is tartalmazó modell reziduálisán elvégzett Ljung-Box teszt empirikus szignifikaszintje 98,87%, így nem tudjuk elvetni a nullhipotézist, azaz a **modell maradéktagjai nem autokorreláltak.**

g)

Végül illesszünk lineáris trendet, hónap-dummykat és AR(2) tagokat tartalmazó modellt az idősorra!

```
mod3 <- dat %>% transmute(
  t = V1,
  m = as.factor(ym %% 100),
  AR_2 = lag(temp, n = 2),
  temp = temp
) %>% na.omit() %>%
  lm(formula = temp ~ .)
```

h)

A fenti három modell közül melyikkel vagyunk a legelégedettebbek?

[1] forecast::checkresiduals függvény által az idősor hossza alapján optimálisnak ítélt érték.

```
answer_1h <- f.Ljung.Box.p(mod3)
```

Mivel az AR(2) tagot tartalmazó modell maradéktagjain elvégzett Ljung-Box teszt p-értéke 0,02%, így csak az **AR(1) tagot tartalmazó modell** az, amelynek reziduumai nem autokorreláltak, tehát ezzel vagyunk a „legelégedettebbek" (paraméterek intevallumai és a tesztek csak ezen validak, előrejelezni is vele érdemes).

i)

Összességében van bizonyítékunk a klímaváltozásra ezen az időtávon Budapesten?

Amennyiben a klímaváltozás alatt azt értjük, hogy valamely irányba megváltozott volna a várható átlaghőmérséklet, úgy a tesztelendő nullhipotézisünk, hogy β_{trend} nem különbözik nullától (kétoldali próba). Ebben az esetben felhasználhatjuk a modellünk paraméterbecsléseit tartalmazó táblázatból a trend komponens parcális t-próbájához tartozó p-értéket a döntésünkhöz. Mivel annak értéke 14,01% (2. táblázat), így **nem tudjuk elutasítani a nullhipotézist és kijelenteni, hogy lenne bizonyítékunk a klímaváltozásra.** Ezzel szemben a próba megfogalmazható egyoldalúként is, így: $H_0 : \beta_{trend} \leq 0$, és csak felső kritikus értéket kell meghatároznunk a t-statisztikához. Mindazonáltal

$$c_f(jobboldali) : t_{1-\alpha}(n-k-1) > c_f(kétoldali) : t_{1-\frac{\alpha}{2}}(n-k-1),$$

így **ebben az esetben sem tudnánk a gyakorlatban bevett szignifikanciaszinteken elutasítani** „klímaváltozatlanságot" jelentő **H_0-t.**

```
mod2 %>%
  broom::tidy() %>%
  mutate(
    term = sapply(term, f.month)
  ) %>%
  prtbl('Az AR(1) tagot tartalmazó modell paraméterbecslései')
```

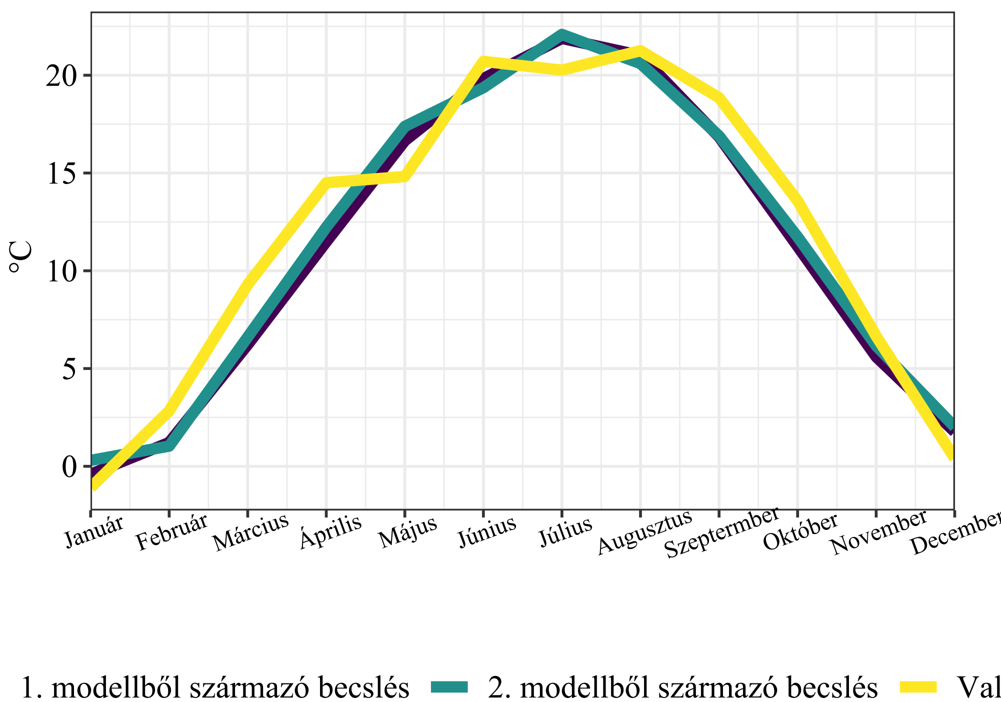
2. táblázat: Az AR(1) tagot tartalmazó modell paraméterbecslései

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
konstans	-1,14	0,28	-4,11	0,00%
trend	0,00	0,00	1,48	14,02%
február	2,08	0,35	5,88	0,00%
március	6,69	0,35	19,38	0,00%
április	10,71	0,38	28,05	0,00%
május	14,58	0,50	29,43	0,00%
június	16,49	0,64	25,59	0,00%
július	17,76	0,74	23,85	0,00%
augusztus	16,38	0,81	20,17	0,00%
szeptermber	12,43	0,78	15,87	0,00%
október	7,86	0,65	12,08	0,00%
november	3,59	0,49	7,33	0,00%
december	1,14	0,37	3,06	0,23%
AR	0,25	0,04	6,76	0,00%

j)

Jelezzük előre az első és a második modell alapján a havi átlaghőmérsékletet a mintaidőszak utáni évre, és ábrázoljuk az előrejelzéseket a ténylegesen bekövetkezett értékekkel együtt!

```
data.frame(
  t = 1:12,
  temp = test_dat$temp,
  mod1 =  predict.lm(mod1, newdata = test_dat),
  mod2 = predict.lm(mod2, newdata = test_dat)
) %>% pivot_longer(-1) %>%
  mutate(
    name = case_when(
      name == 'temp' ~ 'Valós érték',
      name == 'mod1' ~ '1. modellből származó becslés',
      name == 'mod2' ~ '2. modellből származó becslés',
      T  ~ name
    )
  ) %>%
  ggplot(aes(t, value, color = name)) +
  geom_line(size = 2) +
  scale_color_viridis_d() +
  scale_x_continuous(breaks = 1:12, expand = c(0, 0),
                     labels = str_to_title(sapply(str_c('m', 1:12), f.month))) +
  theme(axis.text.x = element_text(angle = 20, size = 9),
        plot.margin = unit(c(1,3,1,1), "cm")
  ) +
  labs(x = '', y = expression(''*~degree*C*''), color = NULL)
```



3. ábra. A budapesti havi átlaghőmérséklet előrejelzése

2. feladat

A wooldridge package-ben szereplő earns adatbázist használjuk, amely a versenyszféra mezőgazdaságon kívüli ágazataira tartalmazza az éves egy munkaórára jutó kibocsátást (azaz a termelékenységet) (outphr) és az órabért (hrwage), valamint ezek növekedési ütemét (logaritmusának éves változását) (goutphr és ghrwage). Az A-L kezdetű vezetéknévvel rendelkező hallgatóknak az 1947-1979 éveket , az M-Zs kezdetű vezetéknévvel rendelkező hallgatóknak az 1957-1987 éveket kell vizsgálniuk.

```
data(earns, package = 'wooldridge')
dat <- earns %>%
  filter(year <= 1979) # surname: Granat
```

a)

Modellezzük először az órabér növekedési ütemét (ghrwage) a termelékenység növekedési üteme (goutphr) függvényében!

```
mod4 <- lm(data = dat, formula = ghrwage ~ goutphr)
```

b)

Értelmezzük a becsült paramétereket!

```
mod4 %>% broom::tidy() %>% prtbl('Egyszerű lineáris regresszió az idősorokon') # TODO interpret
```

3. táblázat: Egyszerű lineáris regresszió az idősorokon

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
konstans	0,00	0,00	0,53	59,74%
goutphr	0,66	0,18	3,62	0,11%

Ha az idei termelékenységi ütem megegyezik a tavalyival, akkor várhatóan az idei bérnövekedés is meg fog egyezni a tavalyival (3. táblázat: $\beta_{konstans}$). Ha az idei termelékenységi ütem a tavalyihoz képest nő 1%P-al, akkor várhatóan az idei bérnövekedés is emelkedni fog a tavalyihoz képest $0,66\%P$-al.

c)

Autokorreláltak-e a modell hibatagjai?

```
answer_2c <- f.Ljung.Box.p(mod4)
```

Mivel a Ljung-Box teszt p-értéke 67,31%, így nem tudjuk elutasítani a H_0-t, tehát nem autokorreláltak a hibatagok.

d)

Teszteljük, hogy goutphr becsült paramétere különbözik-e egytől! Hogyan használtuk fel a teszt elvégzése során a c. feladatrész eredményét?

```
mod4 %>%
  car::linearHypothesis('goutphr = 1')
```

A tesztet Wald-féle F-próbával hajtom végre a megjelölt lineáris restrikcióval, így a H_0 szerint a becsült paraméter nem különbözik szignifikánsan 0-tól. Az empirikus szignifikancia szint 7,41%, így a döntés függ a választott α értéktől. Mivel az idősor igen rövid (kevesebb, mint 30 megfigyelés), így javasolt magasabb szignifikanciaszintet választani (10%), és elvetni a nullhipotézist, tehát különbözik szignifikánsan 1-től a becsült paraméter értéke. A *c)* feladatrész eredménye jelen számítás során azért fontos, mert ha autokorreláció lenne a regresszióban, úgy az intervallumbecslések és a *tesztek torzítottá válnának*, ebből következően az előbbiekben leírtak nem állnák meg a helyüket.

e)

Elemezzük ezután az órabér növekedési ütemét a termelékenység növekedési ütemének elsőrendű osztott késleltetésű modelljével!

```
mod5 <- dat %>%
  mutate(
    l.goutphr = lag(goutphr) # add lagged variable
  ) %>%
  lm(formula = ghrwage ~ goutphr + l.goutphr)
```

f)

Értelmezzük a becsült paramétereket!

```
mod5 %>%
  broom::tidy() %>%
  prtbl('Az osztott késleltetsésű modell paramétereinek becslése')
```

4. táblázat: Az osztott késleltetsésű modell paramétereinek becslése

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
konstans	0,00	0,01	-0,65	52,36%
goutphr	0,64	0,18	3,55	0,14%
l.goutphr	0,33	0,20	1,70	9,99%

Ha a termelékenység növekedési üteme nem változik, akkor hosszútávon az órabér növekedési üteme $0,37\%P$-al fog csökkenni évente ($\beta_{konstans}$). Amennyiben $1\%P$-al nagyobb a termelékenység növekedési üteme az egyik évben, mint a korábbiban, úgy várhatóan c.p. $0,64\%P$-al ($\beta_{goutphr}$) lesz magasabb az órabér növekedési üteme ugyanabb az évben a korábbihoz képest. Amennyiben permanens módon $1\%P$-al megemelkedik a termelékenység növekedési üteme, úgy várhatóan c.p. $0,97\%P$-al ($\beta_{goutphr} + \beta_{l.goutphr}$) lesz majd magasabb az órabér növekedési üteme hosszútávon.

g)

Autokorreláltak-e a modell hibatagjai?

```
answer_2g <- f.Ljung.Box.p(mod5)
```

A modell reziduumain elvégzett Ljung-Box teszt p-értéke 18,23%, ami alapján nincs autokorreláció a modell hibatagjaiban.

h)

Teszteljük, hogy az azonnali és a késleltetett hatás összege különbözik-e egytől! Értelmezzük az eredményt!

```
mod5 %>% car::linearHypothesis('goutphr + l.goutphr = 1')
```

A tesztet Wald-féle F-próbával végeztem el, az empirikus szignifikanciaszint 92,41%, így semmilyen gyakorlatban bevett α mellett nem tudjuk elvetni a H_0-t, tehát a két paraméter becsült értékének összege **nem különbözik szignifikánsan 1-től.**