Ökonometria

1. házi feladat

Granát Marcell

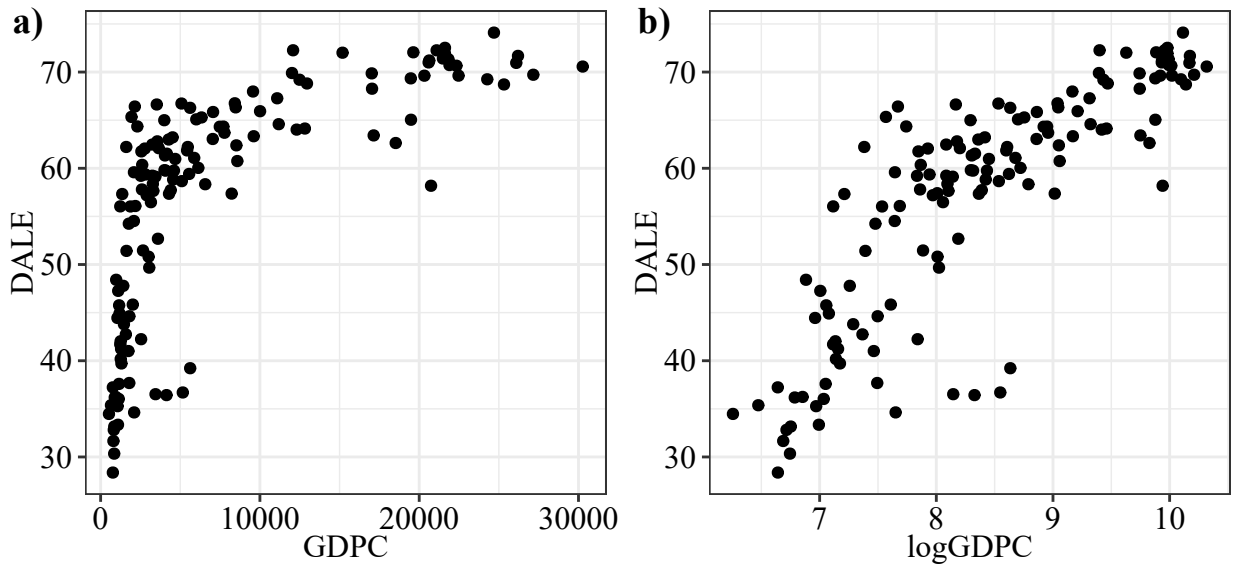
2020. december 1.

Tartalomjegyzék

1. feladat

a)

Ábrázoljuk a DALE változót a GDPC függvényében, illetve (külön ábrán) DALE-t a GDPC logaritmusa függvényében. Értelmezzük az ábrákat!



1. ábra. Az egészségkárosodással korrigált várható élettartam az egy főre eső GDP függvényében

Az 1. ábrának *a* panele ismerteti az egészségkárosodással korrigált várható élettartamot az egy főre eső bruttó kibocsátás függvényében, míg a *b* panelen ugyanezt láthatjuk, de utóbbi változó logaritmikusan átskálázott értéke szerepel a vízszintes tengelyen. Kivehető, hogy **lineáris modell jobban fog illeszkedni, amennyiben regresszorként a logaritmizált GDP/fő értéket használjuk fel (lin-log modell)**.

b)

Becsüljük meg OLS módszerrel azt a modellt, amelyben DALE-t magyarázzuk a GDPC logaritmusával és GINI szintjével!

c)

Értelmezzük a GDPC logaritmusának együtthatóját és ellenőrizzük annak statisztikai szignifikanciáját! Gyakorlati (közgazdasági) értelemben jelentős az együttható nagysága?

1. táblázat: Az első modell paraméterei

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
konstans	-2,9	5,98	-0,49	62,83%
logGDPC	8,5	0,55	15,55	0,00%
GINI	-30,0	6,59	-4,55	0,00%

Statisztikalig szignifikánsnak bizonyul a GDP/fő logaritmusa, mivel minden gyakorlatban bevett szignifikanciaszinten elutasításra kerül a H_0, mely szerint nem különbözik a becsült paraméter értéke szignifikánsan 0-tól. Amennyiben 1%-kal megnő az egy főre eső GDP értéke - minden más változatlansága mellett -, úgy

átlagosan 0,085 egységgel nő meg az egészségkárosodással korrigált várható élettartam.

Gyakorlatilag is jelentős, mivel a logaritmizált értéke az egy főre eső GDP-nek még mindig 4 egységnyi terjedelemben mozog, így a teljes intervallumon várhatóan 34,51 évnyi változást okoz, amely több mint az átlagos egészségkárosodással korrigált várható élettartam érték fele. Az előzetesen megfogalmazott elméleteli megfontolásnak - magasabb életszínvonalon nagyobb a várható élettartam - megfelelő előjelet kaptunk.

d)

Értelmezzük a GINI változó együtthatóját! Gyakorlati (közgazdasági) értelemben jelentős az együttható nagysága?

Statisztikalig szignifikánsnak bizonyul a GINI mutató, mivel minden gyakorlatban bevett szignifikanciaszinten elutasításra kerül a H_0, mely szerint nem különbözik a becsült paraméter értéke szignifikánsan 0-tól. Amennyiben 1 egységgel megnő a GINI értéke - minden más változatlansága mellett -, úgy várhatóan 30 egységgel csökken az egészségkárosodással korrigált várható élettartam.

Gyakorlatilag is jelentős, mivel a GINI mutató 0,4 egységnyi terjedelemben mozog, így a teljes intervallumon várhatóan 12,063 évnyi változást okoz az átlagos egészségkárosodással korrigált várható élettartamban. Előzetesen szakmai ismeretek hiányában nem tudok megfogalmazni feltevést a becsült paraméter előjelére, de a kapott eredmény hihetőnek tűnik.

e)

Határozzuk meg a GINI paraméterbecslésének 95%-os konfidenciaintervallumát, teszteljük szignifikanciáját 1%-os szinten, és határozzuk meg a tesztstatisztika p-értékét!

A GINI becsült paraméterének t-statisztikája -4,54, ami kisebb, mint az 1%-os szignifikanciaszinthez tartozó kritikus alsó érték (-2,61), tehát a nullhipotézist - miszerint nem szignifikáns a GINI hatása - elutasítjuk. A teszthez tartozó p-érték (empirikus szignifikanciaszint) - az 1. táblázatból kiolvasható - 0,00%, tehát az előbb említett H_0-t minden gyakorlatban bevett szignifikanciaszinten elutasítjuk.

2. táblázat: GINI paraméterbecslésének 95%-os konfidenciaintervalluma

Alsó határ	Felső határ
-43,03	-16,96

f)

Következtethetünk-e az eredmények alapján arra, hogy a magasabb Gini-együttható alacsonyabb egészségkárosodással korrigált várható élettartamot okoz? Miért vagy miért nem?

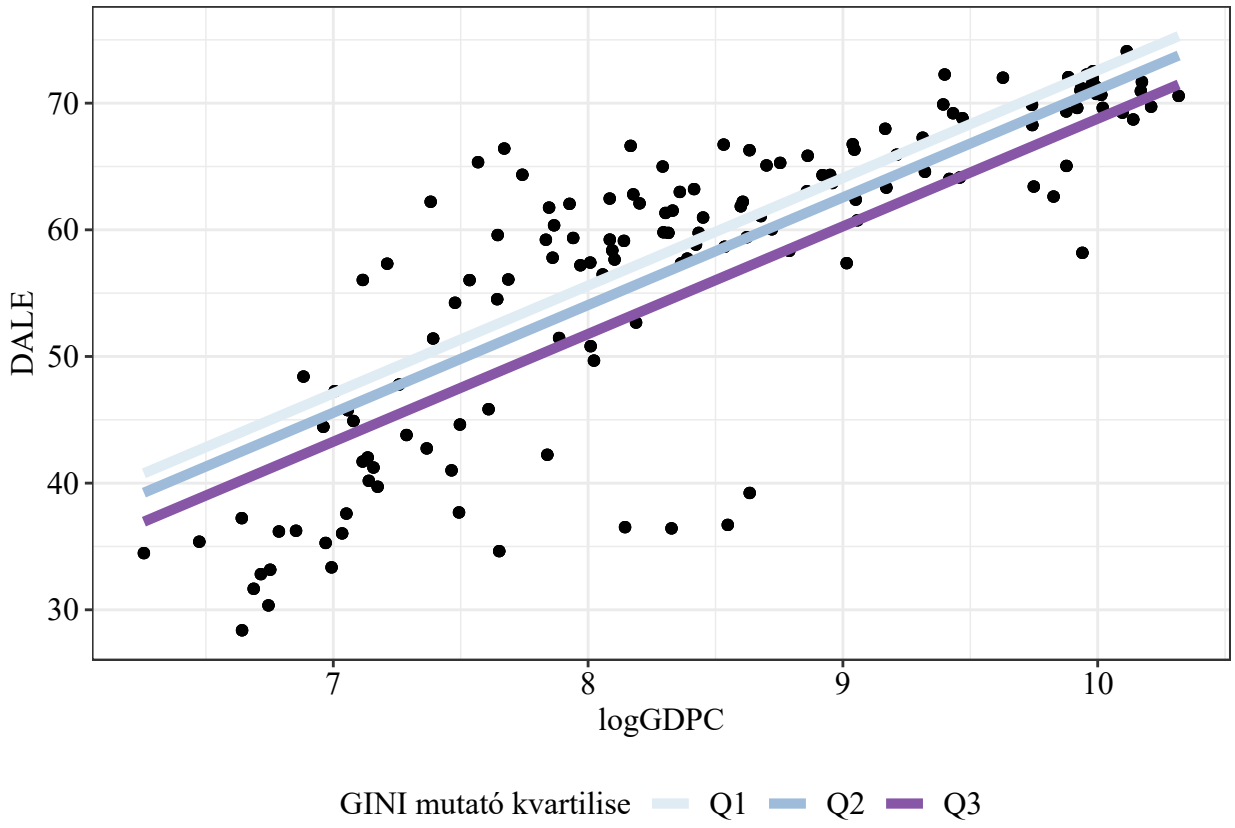
A modellben szereplő parciális hatás egyértelműen ezt az eredményt sugallja. Mindazontál szükséges lenne még megvizsgálni a teljes hatást (kiszámítani a logGDPC-n keresztüli közvetett hatást), illetve gondolni kell a kihagyott változók okozta torzításra is.

g)

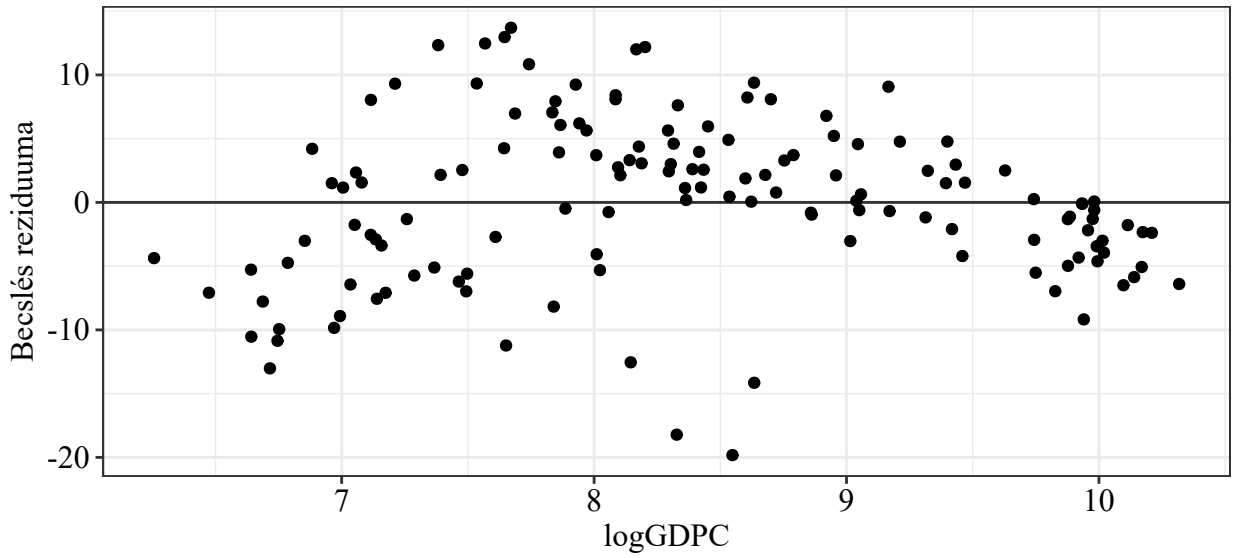
Az a) rész logaritmikus ábráján rajzoljuk be a DALE becsült regressziós függvényét a logaritmikus GDPC függvényében, ha GINI az eloszlásának also kvartilisét, mediánját illetve felső kvartilisét veszi fel!

h)

Ábrázoljuk a reziduálisokat a logaritmikus GDPC függvényében!



2. ábra. Az első modellben a GDP/fő hatása különböző GINI értékek mellett



3. ábra. Az első modell becsléséből származó reziduumok

2. feladat

a)

A legutolsó ábra sugallata alapján bővítsük ki a modellt a logaritmikus GDPC négyzetével!

b)

Az 1. vagy a 2. rész modelljét választanánk? Miért?

3. táblázat: Az 1. és 2. modell jellemzői

	R négyzet	Korrigált R négyzet	AIC	BIC
1. modell	0,72	0,72	929,49	941,28
2. modell	0,78	0,78	897,96	912,71

A 2. modellt választanánk, ugyanis mind az R^2, mind a korrigált R^2 értéke magasabb a 2. modellben, illetve a közölt információs kritériumok[1] (AIC, BIC) értékei alacsonyabbak. Így minden illeszkedés jóságát jellemző mutató alapján arra a döntésre jutunk, hogy a 2. modell jobban írja le a regressziós kapcsolatot.

c)

Értelmezzük a modellből származó előrejelzés és reziduális értékét egy tetszőlegesen választott országra!

4. táblázat: A 2. modell Magyarországra készült becslése

Valós érték	Becslés	Reziduum
63,04	67,08	-4,04

A 3. táblázatból kiolvasható, hogy a modell Magyarország 1993-as egészségkárosodással korrigált várható élettartamának 67,08 évet becsül, ami 4,04 évvel magasabb, mint a valós érték, amely 63,04 év.

[1] Hibára alapuló mutatók, így értéküket minimalizálni kell.

3. feladat

a)

Becsüljük meg DALE modelljét a GDPC (nem pedig a logaritmikus GDPC) négyzetes függvényét és a GINI szintjét használva magyarázó változóként!

b)

Teszteljük 1%-os szinten, hogy a GDPC és négyzete együttesen szignifikáns-e ebben a regresszióban!

A teszthez Wald-féle F-próbát hajtok végre, melynek nullhipotézise, hogy $\beta_{GDPC} = \beta_{GDPC^2} = 0$. Az F-próba értéke 84,341, amely minden gyakorlatban bevett szignifikanciaszinthez tartozó felső kritikus értéket meghalad (a p-érték 0,00%). Mivel a H_0-t elutasítjuk, így kijelenthetjük, hogy a két tárgyal regresszor együttesen szignifikáns a modellben.

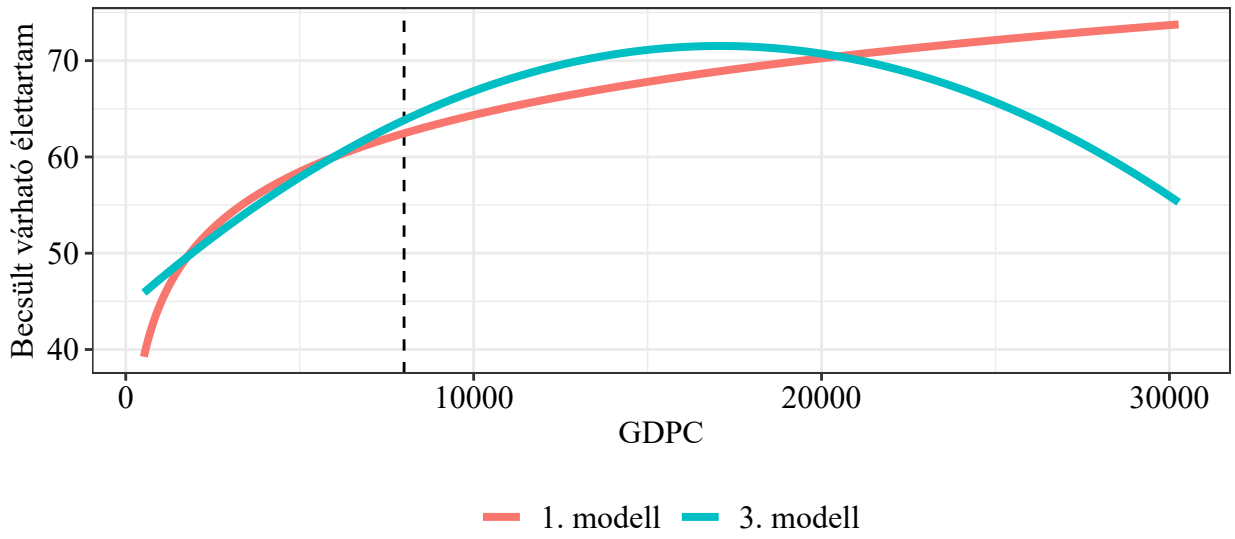
c)

A fenti modell alapján számítsuk ki GDPC parciális hatását, ha GDPC egy, a hallgató vezetéknevének kezdőbetűje által meghatározott értéket vesz fel. (GDPC=8000 USD A-F kezdőbetű esetén, GDPC=10000 USD G-P kezdőbetű esetén és GDPC=12000 USD Q-Z kezdőbetű esetén.)

Amennyiben 8000-ről 8001-re nőne a GDP/fő egy országban, úgy minden más változatlansága mellett várhatóan **19,56** évvel nőne meg ott az egészségkárosodással korrigált várható élettartam.

d)

Hasonlítsuk össze az eredményt az 1. részben kapott parciális hatással!



4. ábra. Az 1. és 3. modellel készített becslés a várható élettartamra a GDP/fő függvényében

Mivel a 3. modellben a $GDPC^2$-hez tartozó becsült paraméter előjele negatív, így az első modellel konzisztens módon ellaposodó hatása van a GDP/fő-nek a várható élettartamra. Az előbbiekben alkalmazott 8000 USD pontban - medián GINI érték mellett - a 3. modell parciális hatása nagyobb (meredekebb az egyenes).

e)

Számítsuk ki a parciális hatás standard hibáját!

A parciális hatás standard hibája **0,00031**.

f)

Az 1. vagy a 3. rész modelljét választanánk? Miért?

5. táblázat: Az 1. és a 3. modell jellemzői

	R négyzet	Korrigált R négyzet	AIC	BIC
1. modell	0,72	0,72	929,49	941,28
3. modell	0,66	0,65	961,09	975,84

Az 1. modellt választanánk, ugyanis mind az R^2, mind a korrigált R^2 értéke magasabb az 1. modellben, illetve a közölt információs kritériumok (AIC, BIC) értékei alacsonyabbak. Így minden illeszkedés jóságát jellemző mutató alapján arra a döntésre jutunk, hogy az 1. modell jobban írja le a regressziós kapcsolatot.

4. feladat

Összességében, melyik modellt választanánk azon modellek közül, amelyek a GINI szintje mellett a GDPC vagy a logaritmikus GDPC tetszőleges polinomját használják magyarázó változóként (azaz a modellhalmaz az 1., 2. és 3. rész modelljeit is tartalmazza speciális esetként)?

A legjobban illeszkedő változók körének meghatározásához kiindulási modellnek vettem azt, amelyikben csak a GINI szerepel, mint magyarázóváltozó. Ezt követően AIC információs kritérium minimalizálási céllal bővítettem a modellt. Minden egyes lépésnél a GDP/fő és a GDP/fő logaritmusának polinómjai (maximum 6 rendű) közül azt emeltem be a modellbe, amellyel az új modell AIC információs kritériuma a legalacsonyabb volt. Akkor áltam meg a bővítéssel, mikor bármely változó bevonásával csökkent volna az AIC. Így legjobb modellnek mutatkozik az, melyben a GDP/fő logaritmusa és a GDP/fő logaritmusának négyzete szerepel.

5. feladat

A log(bér)-t mint függő változót modellezzük az IQ-val mint magyarázó változóval egyváltozós regresszió segítségével, egy elég nagy mintán. A log(bér) átlaga 12 és szórása 0,5, míg IQ átlaga 100 és szórása 15. A két változó mintabeli korrelációja 0,4.

a)

Számítsuk ki a regresszió R-négyzet értékét!

$$R^2 = (r)^2 = 0,4^2 = 0,16$$

b)

Számítsuk ki a hibatag varianciáját!

$$MSE = \sigma_y^2 \times (1 - R^2) = 0,21$$

c)

Mi a meredekségi paraméter OLS becslése?

$$\hat{\beta}_1 = \hat{\rho}\frac{\hat{\sigma_y}}{\hat{\sigma_y}} = 0,4\frac{0,5}{15} = 0,013\dot{3}$$

d)

Mi a tengelymetszet OLS becslése?

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 12 - 0,4\frac{0,5}{15} \times 100 = 10,6\dot{6}$$

R kódok

```r
# setup ------------------------------------------------------------------------------------
library(tidyverse)
library(granatlib) # my personal package: https://github.com/MarcellGranat/granatlib
theme_set(theme_granat())
dat <- rio::import("health_small.xlsx") %>% filter(YEAR == 1993)
            # data import, NEPTUN: AYCOPF

# 1 ##########################################################################################

# a ----------------------------------------------------------------------------------------
dat <- dat %>% mutate(logGDPC = log(GDPC)) # add the log of GDPC

ggpubr::ggarrange(
ggplot(dat, aes(GDPC, DALE)) + geom_point(),
ggplot(dat, aes(logGDPC, DALE)) + geom_point(), labels = c("a)", "b)")
)

# b ----------------------------------------------------------------------------------------
model1 <- lm(data = dat, formula = DALE ~ logGDPC + GINI)

# c, d -------------------------------------------------------------------------------------
model1 %>% broom::tidy() %>% prtbl("Az első modell paraméterei", ufc = F)

# e ----------------------------------------------------------------------------------------
confint(model1, 'GINI', level = .95) %>% data.frame() %>%
  set_names("Alsó határ", "Felső határ") %>%
  knitr::kable(format.args = list(decimal.mark = ","), digits = 2,
               caption = "GINI paraméterbecslésének 95%-os konfidenciaintervalluma",
               row.names = F, align = c("c", "c"))

# f ----------------------------------------------------------------------------------------
data.frame(
  logGDPC = dat$logGDPC, DALE = dat$DALE,
  Q1 = dat %>% select(logGDPC, GINI) %>%
    mutate(GINI = quantile(GINI, .25)) %>% predict.lm(object = model1),
  Q2 = dat %>% select(logGDPC, GINI) %>%
    mutate(GINI = quantile(GINI, .5)) %>% predict.lm(object = model1),
  Q3 = dat %>% select(logGDPC, GINI) %>%
    mutate(GINI = quantile(GINI, .75)) %>% predict.lm(object = model1)
 ) %>% pivot_longer(3:5) %>%
  ggplot() + geom_point(aes(logGDPC, DALE)) +
  geom_line(aes(logGDPC, value, color = name), size = 1.8) +
  scale_color_brewer(palette = "BuPu") +
  labs(color = "GINI mutató kvartilise")

model1 %>% broom::augment() %>% ggplot() +
  geom_hline(yintercept = 0, color = "grey20") +
  geom_point(aes(x = logGDPC, y = .resid)) +
  labs(y = "Becslés reziduuma")

# 2 ##########################################################################################

```

```r
# a ----------------------------------------------------------------------------------------
dat <- dat %>% mutate(logGDPC2 = logGDPC^2)
model2 <- lm(data = dat, formula = DALE ~ logGDPC + logGDPC2 + GINI)

# b ----------------------------------------------------------------------------------------

rbind(broom::glance(model1), broom::glance(model2)) %>% mutate(
  model = c("1. modell", "2. modell")
) %>% column_to_rownames(var = 'model') %>% select(
  r.squared, adj.r.squared, AIC, BIC
) %>% rename(c("R négyzet" = r.squared, "Korrigált R négyzet" = adj.r.squared)) %>%
  knitr::kable(digits = 2, format.args = list(decimal.mark = ","),
               caption =  "Az 1. és 2. modell jellemzői", align = rep("c", ncol(.)))

# c ----------------------------------------------------------------------------------------
model2 %>% broom::augment() %>% cbind(dat$COUNTRYNAME) %>%
  filter(dat$COUNTRYNAME == "Hungary") %>%
  select(DALE, .fitted, .resid) %>%
  set_names("Valós érték", "Becslés", "Reziduum") %>% knitr::kable(
    caption = "A 2. modell Magyarországra készült becslése",
    align = c("c", "c", "c"), format.args = list(decimal.mark = ","), digits = 2
  )

# 3 ##########################################################################################

# a ----------------------------------------------------------------------------------------
dat <- dat %>% mutate(GDPC2 = GDPC^2)
model3 <- lm(data = dat, formula = DALE ~ GINI + GDPC + GDPC2)

# b ----------------------------------------------------------------------------------------
car::linearHypothesis(model3, c("GDPC = 0","GDPC2 = 0"), test="F")

# c ----------------------------------------------------------------------------------------
# Neptun: AYCOPF -> GDPC = 8000
GDPC.partialeffect <- format(sum(model3$coefficients[3:4]*
              c(8000, 8000^2)), digits = 4, decimal.mark = ",")

# d ----------------------------------------------------------------------------------------
data.frame(GINI = rep(median(dat$GINI), 1000),
           GDPC = seq(from = min(dat$GDPC), to = max(dat$GDPC), length.out = 1000)) %>%
  mutate(logGDPC = log(GDPC), GDPC2 = GDPC^2) %>%
  cbind(data.frame(model1 = predict.lm(object = model1, newdata = .))) %>%
  cbind(data.frame(model3 = predict.lm(object = model3, newdata = .))) %>%
  select(GDPC, model1, model3) %>% set_names("GDPC", "1. modell", "3. modell") %>%
  pivot_longer(-1) %>%
  ggplot(aes(x = GDPC, y = value, color = name)) +
  geom_vline(xintercept = 8000, linetype = "dashed") +
  geom_line(size = 1.4) +
  labs(y = "Becsült várható élettartam", color = "")
# e ----------------------------------------------------------------------------------------
answer_3e <- vcov(model3) %>%
  {.["GDPC","GDPC"] + .["GDPC2","GDPC2"] + .["GDPC","GDPC2"]} %>%
  sqrt()
```

```r

# f ----------------------------------------------------------------------------------

rbind(broom::glance(model1), broom::glance(model3)) %>% mutate(
  model = c("1. modell", "3. modell")
) %>% column_to_rownames(var = 'model') %>% select(
  r.squared, adj.r.squared, AIC, BIC
) %>% rename(c("R négyzet" = r.squared, "Korrigált R négyzet" = adj.r.squared)) %>%
  knitr::kable(digits = 2, format.args = list(decimal.mark = ","),
               caption =  "Az 1. és a 3. modell jellemzői")

# 4 ##############################################################################
dat2 <- dat %>% select(DALE, GDPC, GDPC2, GINI, logGDPC, logGDPC2) %>%
  mutate(
    GDPC3 = GDPC^3,
    logGDPC3 = logGDPC^3,
    GDPC4 = GDPC^4,
    logGDPC4 = logGDPC^4,
    GDPC5 = GDPC^5,
    logGDPC5 = logGDPC^5,
    GDPC6 = GDPC^6,
    logGDPC6 = logGDPC^6
  )

model4 <- lm(data = dat2, formula = DALE ~ GINI)

MASS::stepAIC(model4, scope = list(lower = DALE ~ GINI,
  upper = DALE ~ GINI + GDPC + GDPC2 + GDPC3 + GDPC4 + GDPC5 + GDPC6 +
  logGDPC + logGDPC2 + logGDPC3 + logGDPC4 + logGDPC5 + logGDPC6),
  direction = "forward", trace = F)
```