Ökonometria

2. házi feladat

Granát Marcell

2020. december 15.

Tartalomjegyzék

```
library(tidyverse)
library(granatlib) # my personal package: https://github.com/MarcellGranat/granatlib
theme_set(theme_granat())
```

1. feladat

Az alvással és a munkával töltött idő közötti átváltást, valamint az alvásidőt befolyásoló egyéb tényezőket vizsgáljuk a sleep75 adatbázis alapján (amely a „wooldridge" csomagban található). A függő változó az éjszakai alvással töltött összes idő percben (sleep), a magyarázó változók pedig a teljes heti munkaidő (totwrk), az iskolai évek száma és az életkor (educ, age), nem (male) és egy dummy változó, ami a kisgyerek jelenlétét mutatja a családban (yngkid).

```
data(sleep75, package = "wooldridge")
dat <- sleep75
```

a)

Definiálja a foglalkoztatás három kategóriáját a heti ledolgozott órák alapján (hozzávetőlegesen: nem dolgozik [<4 óra], részmunkaidős [4-35 óra], teljes munkaidős [> 35 óra])! Vizsgálja meg az alvással töltött idő eloszlását boxplot-tal a foglalkoztatási kategóriák szerint! Adjon az ábrának címet, informatív tengelyfeliratokat stb.! Megjegyzés: itt használhatja a cut() függvényt.

```
dat %>% mutate(
  totwrk = totwrk/60,
  wrktype = factor(case_when(
    totwrk < 4 ~ 'Nem dolgozik',
    totwrk >= 4 & totwrk < 35 ~ 'Részmunkaidős',
    T ~ 'Teljes munkaidős'))) %T>%
  {dat <<- .} %>% # rewrite totwrk + wrktype
  ggplot(aes(x = wrktype, y = sleep)) + geom_boxplot() + # plot
  labs(x = 'Foglalkozatási típus', y = 'Alvási idő')
```

b)

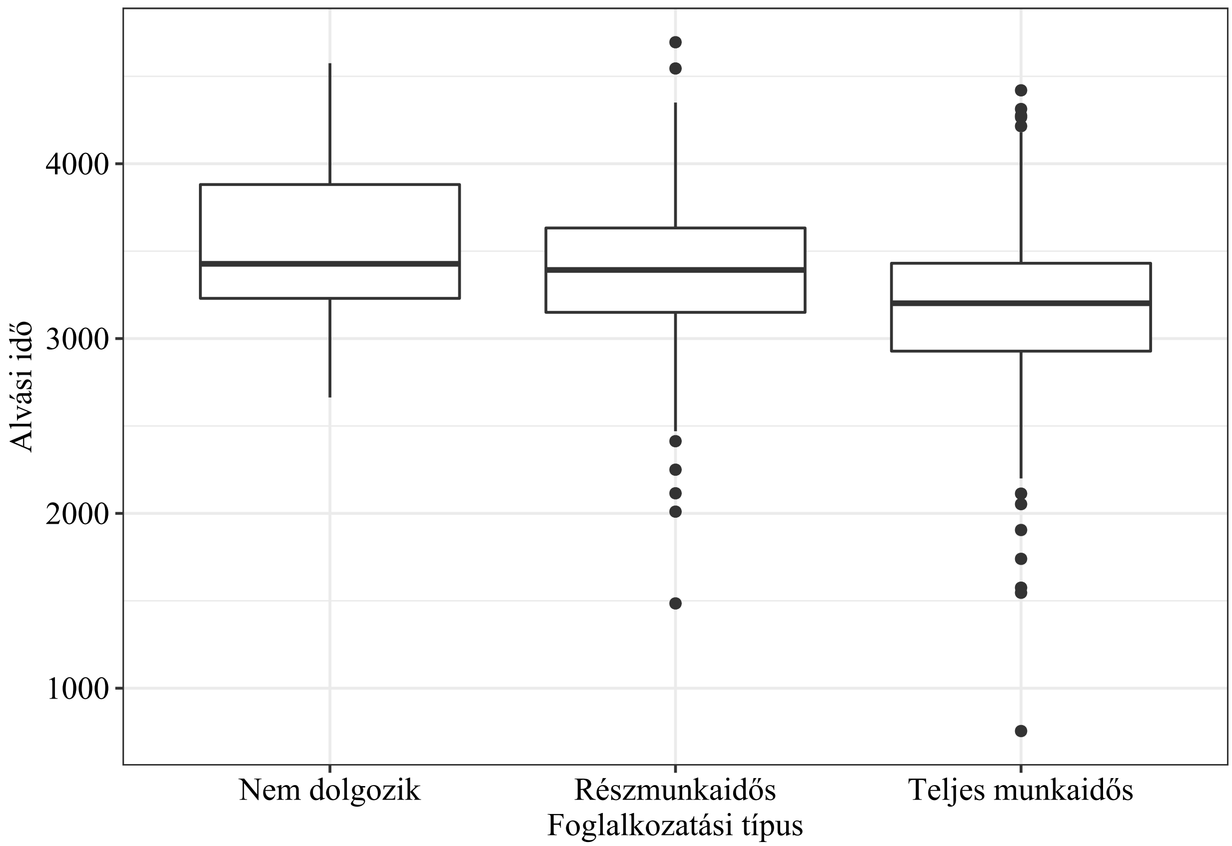
Becsüljön meg három olyan regressziós modellt, amelyek függő változója a sleep, és az első modell egyetlen magyarázó változója totwrk, a második modell magyarázó változói totwrk és négyzete, a harmadik modell pedig a foglalkoztatási kategóriákat tartalmazza magyarázó változóként! (Természetesen minden modellben szerepeljen a konstans is.) Megjegyzés: a factor() függvény hasznos lehet ebben a részben.

```
dat <- dat %>% mutate(totwrk2 = totwrk^2)
model1 <- lm(data = dat, formula = sleep ~ totwrk)
model2 <- lm(data = dat, formula = sleep ~ totwrk + totwrk2)
model3 <- lm(data = dat, formula = sleep ~ wrktype)
```

c)

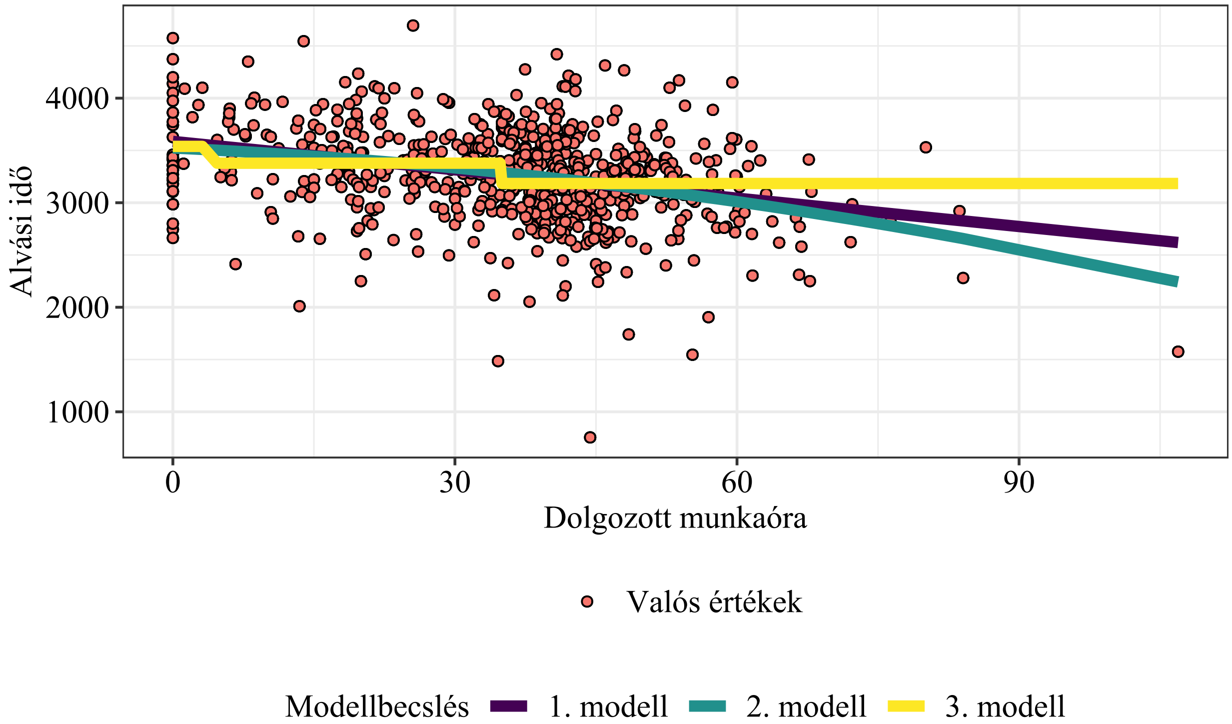
Ábrázolja a sleep becsült függését a totwrk változótól egyetlen ábrában a három modell alapján kiszámítva!

```
data.frame(sleep = dat$sleep, totwrk = dat$totwrk, m1 = model1$fitted.values,
           m2 = model2$fitted.values, m3 = model3$fitted.values) %>%
  pivot_longer(-c(1, 2)) %>% mutate(
    name = paste0(str_remove(name, "m"), ". modell")
  ) %>%
  ggplot(data = .) +
```



1. ábra. Alvási idő dobozábrája foglalkozatatási típusonként

```
  geom_point(aes(x = totwrk, y = sleep, fill = "Valós értékek"),
             color = 'black', shape = 21) +
  geom_line(aes(totwrk, value, color = name), size = 2) +
  scale_color_viridis_d() +
  labs(x = "Dolgozott munkaóra", y = "Alvási idő", color = "Modellbecslés", fill = "")
```



2. ábra. Becsült alvási idő különböző modellekből

d)

Melyik modellt választaná a modellszelekciós kritériumok és az értelmezhetőség alapján?

```
rbind(broom::glance(model1), broom::glance(model2)) %>%
  rbind(broom::glance(model3)) %>%
  mutate(model = c("1. modell", "2. 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 = 4, format.args = list(decimal.mark = ","),
      caption =  "A 3 modell jellemzői", align = rep("c", ncol(.)))
```

1. táblázat: A 3 modell jellemzői

	R négyzet	Korrigált R négyzet	AIC	BIC
1. modell	0,1033	0,1020	10540,19	10553,87
2. modell	0,1075	0,1049	10538,90	10557,14
3. modell	0,0614	0,0588	10574,40	10592,64

A korrigált R^2 az Akaike-féle információs mutató alapján alapján a 2. modellt, míg a BIC alapján az első modellt választanám. Mivel az értelmezhetőség az első modell mellett szól (nincsen kvadratikus hatás, így a β-kat egyszerűen lehet a parciális hatásként leolvasni), így azt választanám.

2. feladat

a)

Becsüljön meg egy többváltozós regressziós modellt úgy, hogy az alvással töltött idő a függő változó, és a munkával töltött idő, az életkor, az életkor négyzete, az iskolázottság, a nem és a kisgyermek jelenléte a magyarázó változó!

```
model4 <- dat %>%
  mutate(age2 = age^2) %T>%
  {dat <<- .} %>% # refresh dat
  lm(formula = sleep ~ totwrk + age + age2 + educ + male + yngkid)
```

b)

Értelmezze a nem és a kisgyermek jelenlétének paraméterbecslését!

Ceteris paribus egy férfi várhatóan 8,7 perccel alszik többet, mint egy nő. Amennyiben van kisgyermek, amely 3 évnél fiatalabb (yngkid = 1), úgy az alvásidő várhatóan ceteris paribus 0,0228 perccel kevesebb.

c)

Tesztelje 5%-os szinten, hogy a hibatag varianciája nem függ-e a magyarázó változóktól!

- *Adja meg a tesztstatisztikát és a hozzá tartozó p-értéket!*
- *Értékelje a teszteredményt!*
- *Kell-e heteroszkedaszticitás-robusztus standard hibákat használni?*

```
lmtest::bgtest(model4)
```

A Breusch-Godfrey teszt-statisztikájának értéke (1) **0,6660**, amelyhez (1) **41,44%-os** p-érték tartozik. Mivel jelen esetben a (2) **nullhipotézist - mely szerint nincs heteroszkedaszticitás a modellben - elfogadjuk minden gyakorlatban bevett szignifikanciaszinten**, így a stanard hibákat tekinthetjük torzítatlannak, és (3) **nem kell heteroszkedaszticitás-robusztus standard hibákat használni**.

d)

Becsülje meg az alvással töltött időt egy 40 éves, teljes munkaidőben dolgozó, kisgyerekes és középfokú végzettséggel (azaz 12 éves oktatással) rendelkező férfi munkavállaló számára!

```
answer_2d <-
  data.frame(totwrk = 35, age = 40, age2 = 40^2, educ = 12, male = 1, yngkid = 1) %>%
  predict.lm(object = model4)
```

Egy 40 éves, teljes munkaidőben dolgozó, kisgyeremekes és középfokú végzettséggel (azaz 12 éves oktatással) rendelkező férfi munkavállaló várhatóan **3302,445** percet tölt hetente alvással.

e)

Adja meg a várható érték és a konkrét érték előrejelzésének standard hibáját!

```
data.frame(totwrk = 35, age = 40, age2 = 40^2, educ = 12, male = 1, yngkid = 1) %>%
  predict.lm(object = model4, se.fit = T, interval = "confidence") %>% .$se.fit %T>%
```

```
  {answer_3e_a <<- .} %>%
  {answer_3e_b <<- . + sd(dat$sleep)}
```

A várható érték előrejelzésének standard hibája 52,69 perc, míg a konkrét érték előrejelzésének standard hibája 497,1 perc.

f)

Adjon 95%-os konfidencia-intervallumot a fenti két mennyiségre!

```
data.frame(totwrk = 35, age = 40, age2 = 40^2, educ = 12, male = 1, yngkid = 1) %>%
{rbind(predict.lm(object = model4, newdata = ., se.fit = T, interval = "confidence")$fit,
predict.lm(object = model4, newdata = ., se.fit = T, interval = "prediction")$fit)} %>%
  data.frame() %>%
  transmute("type" = c('Várható érték', 'Konkrét érték'), lwr, upr) %>%
  mutate_at(-1, function(x) format(x, decimal.mark = ',', digits = 6)) %>%
  column_to_rownames('type') %>%
  set_names('Alsó határ', 'Felső határ') %>%
  knitr::kable(caption = 'Várható és konkrét érték konfidencia intervalluma',
               align = c('c', 'c'))
```

2. táblázat: Várható és konkrét érték konfidencia intervalluma

	Alsó határ	Felső határ
Várható érték	3198,99	3405,90
Konkrét érték	2475,21	4129,68

g)

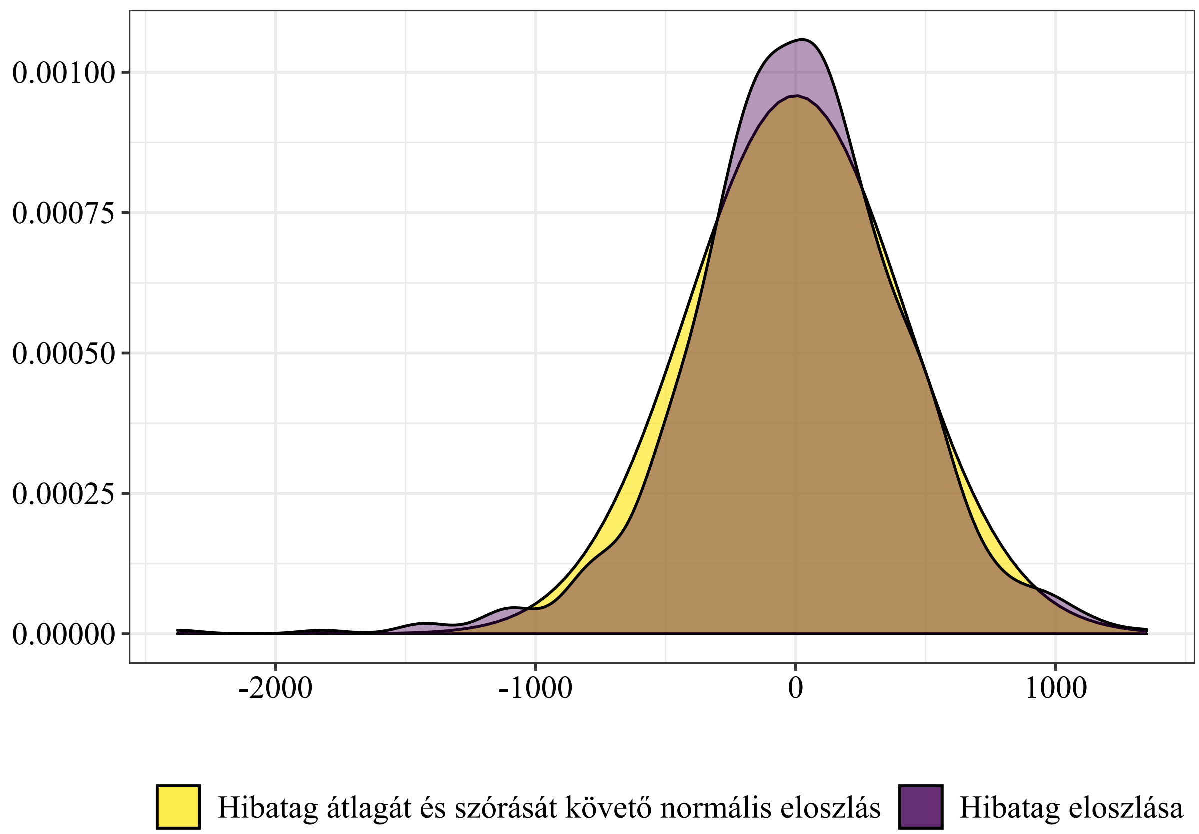
A (d) - (f) során melyik ponton feltételeztük a hibatag normális eloszlását a számítások során?

Az **f feladatban**.

h)

Vizsgálja meg egy megfelelő ábra segítségével, hogy a normalitás feltételezése (hozzávetőlegesen) igaz-e!

```
model4 %>%
  broom::augment() %>% select(.resid) %>%
  {ggplot(data = .)  +
  stat_function(fun = dnorm, args = list(mean = mean(.$.resid), sd = sd(.$.resid)),
  aes(fill = 'Hibatag átlagát és szórását követő normális eloszlás'),
                  geom = 'area', alpha = .7, color = "black") +
    geom_density(mapping = aes(.resid, fill = 'Hibatag eloszlása'),
                 position = 'stack', alpha = .4) +
    scale_fill_viridis_d(direction = -1) +
    labs(x = "", y = "", fill = "")}
```



i)

A mintát ossza fel véletlenszerűen két azonos méretű részre! Becsülje meg a fenti modellt az egyik részmintán! Számítsa ki az átlagos négyzetes eltérést (MSE) a becslési mintán és a másik (teszt-) mintán is! Hasonlítsa össze az MSE értékeket és értelmezze az esetleges különbséget!

```
set.seed(3)
dat %>%
  mutate(sample = runif(n = nrow(.)),
  sample = ifelse(sample < median(sample), 'train', 'test')) %T>%
  {dat <<- .} %>%
  filter(sample == 'train') %>%
  lm(formula = sleep ~ totwrk + age + age2 + educ + male + yngkid) %>%
  {model5 <<- .}

data.frame('Becslési minta' = model5$residuals,
           'Teszt minta' = pull(filter(dat, sample == 'test'), sleep) -
            predict.lm(object = model5,
            newdata = filter(dat, sample == 'test'))) %>%
  apply(2, function(x) mean(x^2)) %>% matrix() %>% t() %>%
  data.frame() %>% set_names(c('Becslési minta', 'Teszt minta')) %>%
  prtbl("MSE a becslési és a tesztmintán", un = F)
```

3. táblázat: MSE a becslési és a tesztmintán

Becslési minta	Teszt minta
144697,7	208418,7

A teszt mintán számított MSE nagyobb, mint a becslési mintán, azonban ez futtatásonként eltérő. Ennek oka a mintavételi ingadozás. Amennyiben ez az eredmény gyakorta ismétlődik, úgy levonható következtetés lenne, hogy a modell túlilleszkedik, de jelenleg nem levonható ez a következtetés, mert ismételt mintavételek esetén előrfodul gyakran, hogy a teszt mintán kisebb az MSE.

3. feladat

a)

Mennyivel különbözik a kisgyermekes és nem kisgyermekes szülők alvással töltött átlagos ideje a férfiak illetve a nők esetében?

```
dat %>% group_by(male, yngkid) %>% summarise(y = mean(sleep)) %>%
  pivot_wider(names_from = 'yngkid', values_from = 'y') %>%
  {data.frame(c('nő', 'férfi'),.[,2] - .[,3])} %>%
  set_names("Nem", "Különbség") %>% prtbl('Különbség a gyermek nemekre való hatásában')
```

4. táblázat: Különbség a gyermek nemekre való hatásában

Nem	Különbség
nő	66,74
férfi	-12,09

b)

Egészítse ki a yngkid és a male interakciójával a 2. feladat modelljét! Értelmezze a paraméterbecsléseket (és azok statisztikai szignifikanciáját), majd hasonlítsa össze azokat a 2. feladat megfelelő becslésével!

```
lm(data = dat, formula = sleep ~ totwrk + age + age2 + educ + male + yngkid + yngkid:male) %>%
  broom::tidy() %>% prtbl("A bővített modell paramétereinek becslése")
```

5. táblázat: A bővített modell paramétereinek becslése

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
konstans	3861,00	239,85	16,10	0,00%
totwrk	-9,88	1,09	-9,06	0,00%
age	-9,43	11,34	-0,83	40,57%
age2	0,14	0,13	1,02	30,96%
educ	-11,38	5,88	-1,94	5,31%
male	74,56	36,22	2,06	3,99%
yngkid	-88,77	86,81	-1,02	30,68%
male:yngkid	128,04	102,12	1,25	21,03%

Statisztikailag szignifikáns magyarázóváltozónka bizonyult a dolgozott munkaóra és a nem 5%-os szignifikanciaszinten. Az alvási időt csökkenti a dolgozott heti munkaóra, ha az illető nő, ha van 3 évnél fiatalabb gyermeke és az életkor növekedése kvadratikusan hat, kezdetben csökkenti az alvási időt. A fő változás a 2 feladatban becsült modellhez képest, hogy a fiatal gyermek jelenlétének paramétere jelentőset nőtt abszolút értékben és a p-értéke is csökkent, bár a bevett szignifikanciaszinteken még mindig nem szignifikáns. Ezzel szemben a férfi nem és fiatal gyermek jelenlétének interakciójának paramétere pozitív előjelet kapott a becsült modellben, amely arra utal, hogy a kisgyermek eltérő módon hat a férfi és női szülő alvás idejére.

4. feladat

A titanic_small.xls fájl tartalmazza a Titanic utasainak egyéni jellemzőit: nem (sex=1 férfiak esetében, sex=2 nők esetében); az osztály, amelyen utaztak (pclass), életkor (age) és hogy túlélték-e a katasztrófát (survived).

```
dat <- rio::import('titanic_small.xls')
```

a)

Számítsa ki a katasztrófát túlélő utazók százalékos arányát! Mennyire különbözik ez osztályonként?

```
dat %>% group_by(pclass) %>% summarise(r = mean(survived)) %>% mutate(
  r = scales::percent(r, decimal.mark = ',', accuracy = .01)) %>%
  set_names("Osztály", "Túlélési arány") %>% prtbl("Túlélési arány utazási osztályonként")
```

6. táblázat: Túlélési arány utazási osztályonként

Osztály	Túlélési Arány
1	61,92%
2	42,96%
3	25,53%

b)

Becsüljön meg egy lineáris valószínűségi modellt (LPM), egy logit és egy probit modellt, függő változóként a túlélés valószínűségét, magyarázó változóként pedig az osztályt (mint kategorikus változót), a nemet és az életkort használva!

```
lpm <- dat %>% lm(formula = survived~factor(pclass)+factor(sex)+age)
logit <- dat %>% glm(formula = survived~factor(pclass)+factor(sex)+age,
            family = binomial(link = "logit"))
probit <- dat %>% glm(formula = survived~factor(pclass)+factor(sex)+age,
            family = binomial(link = "probit"))
```

c)

Számítsa ki az LPM, logit és probit modellek alapján a harmadosztályon és a másodosztályon utazók átlagos kontrollált túlélésvalószínűség-különbségét!

```
merge(lpm %>% broom::tidy() %>%
  transmute(term, LPM = estimate),
mfx::logitmfx(atmean = F, data = dat,
  formula = survived ~ factor(pclass) + factor(sex) + age) %>%
  .$mfxest %>% data.frame() %>% rownames_to_column() %>%
  select(1:2) %>% set_names('term', 'Logit')
) %>% merge(
  mfx::probitmfx(atmean = F, data = dat,
  formula = survived ~ factor(pclass) + factor(sex) + age) %>%
  .$mfxest %>% data.frame() %>% rownames_to_column() %>%
```

```
  select(1:2) %>% set_names('term', 'Probit')
) %>% filter(
  term == 'factor(pclass)2' | term == 'factor(pclass)3'
) %>% mutate(term = str_remove(term, 'factor\\(pclass\\)')) %>%
  mutate_at(-1, function(x) scales::percent(x, accuracy = .01, decimal.mark = ',')) %>%
  rename('Osztály' = term) %>%
  prtbl(align = c('c', 'c', 'c', 'c'))
```

Osztály	Lpm	Logit	Probit
2	-21,14%	-18,14%	-18,64%
3	-37,04%	-36,68%	-35,97%

d)

Hasonlítsa össze ezt a három számot egymással és az a) rész eredményeivel!

A 3 modell alapján készült kontrolállt valószínűség-különbség jól közelíti a sokkaságban megfigyelhető arányokat. Az eltérés fő oka a magyarázóváltozók közötti multikollinearitás, de itt ez most nem számottevő.

e)

A klasszifikációhoz használja a 0,5 értéket küszöbként. Számítsa ki a logit modell alapján a kétfajta klasszifikációs hibát és a helyesen besorolt megfigyelések arányát!

```
regclass::confusion_matrix(M = logit, DATA = dat) %>%
  {.[-3,-3]} %>%
  {(./sum(.))} %>%
  data.frame(row.names = c('Valós 0', 'Valós 1')) %>%
  rownames_to_column() %>%
  mutate_at(-1, function(x) scales::percent(x, accuracy = .01, decimal.mark = ',')) %>%
  column_to_rownames() %>%
  set_names(c('Becsült 0', 'Becsült 1')) %>%
  knitr::kable(caption = 'A logit modellel készített kalsszifikáció konfúziós mátrixa', align = c('c',
```

8. táblázat: A logit modellel készített kalsszifikáció konfúziós mátrixa

	Becsült 0	Becsült 1
Valós 0	49,71%	9,46%
Valós 1	12,05%	28,78%

A táblázatból kiolvasható, hogy a helyesen besoroltak aránya 78%, a hibásan klasszifikált valóságban 1-esek aránya 12%, míg a hibásan klasszifikált 0-sok aránya 10%.