

S08 Specialis prediktorok

Zoltan Kekecs

03 November 2020

Contents

1	Specialis prediktorok	1
1.1	Absztrakt	1
1.2	Adatmenedzsment es leiro statisztikak	1
1.3	Kategorikus valtozok mint prediktorok	4
1.4	Ket valtozo interakciojanak beillesztese a modellbe	6
1.5	Hatvany prediktorok a nem-linearis osszefugesekek modellezesehez	8

1 Specialis prediktorok

1.1 Absztrakt

Ebben a gyakorlatban megismerjuk majd hogyan használjuk es értelmezzük a különbozo tipusu prediktorokat a linearis regressziós modellekben.

1.2 Adatmenedzsment es leiro statisztikak

1.2.1 Package-ek betoltese

```
library(tidyverse)
library(psych)
library(gridExtra)
```

1.2.2 A fogyasi kutatas adatbazis betoltese

Az adatbazis egy olyan kutatas szimulalt adatait tartalmazza ahol különbozo kezelesek hatékonysagát tesztelték a súlyvesztésre tulsúlyos személyeknél.

Valtozok:

- ID - vizsgalati személy azonosítójele
- Gender - nem
- Age - életkor
- BMI_baseline - Body mass index (BMI) a kezelés előtt
- BMI_post_treatment - Body mass index (BMI) a kezelés után
- treatment_type - A kezelés amit a vizsgalati személy kapott (no treatment - nem kapott kezelést; pill - étvagycsökkentő gyógyszer; psychotherapy - kognitív behavior terápia (CBT); treatment 3 - egy harmadik fajta kezelés, lásd lentebb)
- motivation - önbevallásos motivációs szint a fogyásra (0-10-es skálán, ahol a 0 extrém alacsony motiváció a fogyásra, a 10 pedig extrém magas motiváció a fogyásra)
- body_acceptance - a személy mennyire érzi elégedettnek magát jelenleg testével (-7 - +7, ahol a -7 nagyon elégedetlen, a +7 nagyon elégedett)

```
data_weightloss = read.csv("https://tinyurl.com/weightloss-data")
```

1.2.3 Adatellenorzes

Nezzuk at eloszor az altalunk hasznalt adattablat.

```
data_weightloss %>%
  summary()
```

```
##      ID      gender      age      BMI_baseline      BMI_post_treatment
## ID_1   : 1  female:117  Min.   :21.00  Min.   :27.00  Min.   :22.00
## ID_10  : 1  male  :123  1st Qu.:33.00  1st Qu.:33.00  1st Qu.:31.00
## ID_100 : 1                Median :35.00  Median :35.00  Median :34.00
## ID_101 : 1                Mean   :34.78  Mean   :34.98  Mean   :33.78
## ID_102 : 1                3rd Qu.:38.00  3rd Qu.:37.00  3rd Qu.:37.00
## ID_103 : 1                Max.   :50.00  Max.   :43.00  Max.   :44.00
## (Other):234
##      treatment_type  motivation  body_acceptance
## no_treatment :60      Min.   : 2.000  Min.   :-6.000
## pill         :60      1st Qu.: 5.000  1st Qu.: -3.000
## psychotherapy:60      Median : 6.000  Median : -2.000
## treatment_3  :60      Mean    : 6.004  Mean    : -1.812
##              3rd Qu.: 7.000  3rd Qu.: -1.000
##              Max.    :10.000  Max.    : 3.000
##
```

```
describe(data_weightloss)
```

```
##      vars    n  mean    sd median trimmed  mad min max range
## ID*      1 240 120.50 69.43  120.5  120.50 88.96   1 240  239
## gender*   2 240   1.51  0.50    2.0    1.52  0.00   1   2    1
## age       3 240  34.78  3.99   35.0   34.85  4.45  21  50   29
## BMI_baseline  4 240  34.98  2.89   35.0   35.01  2.97  27  43   16
## BMI_post_treatment 5 240  33.78  3.82   34.0   33.86  4.45  22  44   22
## treatment_type* 6 240   2.50  1.12    2.5    2.50  1.48   1   4    3
## motivation   7 240   6.00  1.53    6.0    5.99  1.48   2  10    8
## body_acceptance 8 240  -1.81  1.60   -2.0   -1.84  1.48  -6   3    9
##      skew kurtosis  se
## ID*      0.00   -1.22 4.48
## gender*  -0.05   -2.01 0.03
## age      -0.11    0.90 0.26
## BMI_baseline -0.04    0.09 0.19
## BMI_post_treatment -0.16  -0.06 0.25
## treatment_type*  0.00   -1.37 0.07
## motivation   0.00    0.08 0.10
## body_acceptance  0.18   -0.34 0.10
```

Szeretnenk megerteni a kulonbozo kezelestipusok hatasat a BMI-re. Vegezzunk feltaro elemzest az adatokon.

```
fig_1 = data_weightloss %>%
  ggplot() +
  aes(y = BMI_baseline, x = treatment_type) +
  geom_boxplot()
ylim(c(20, 45))
```

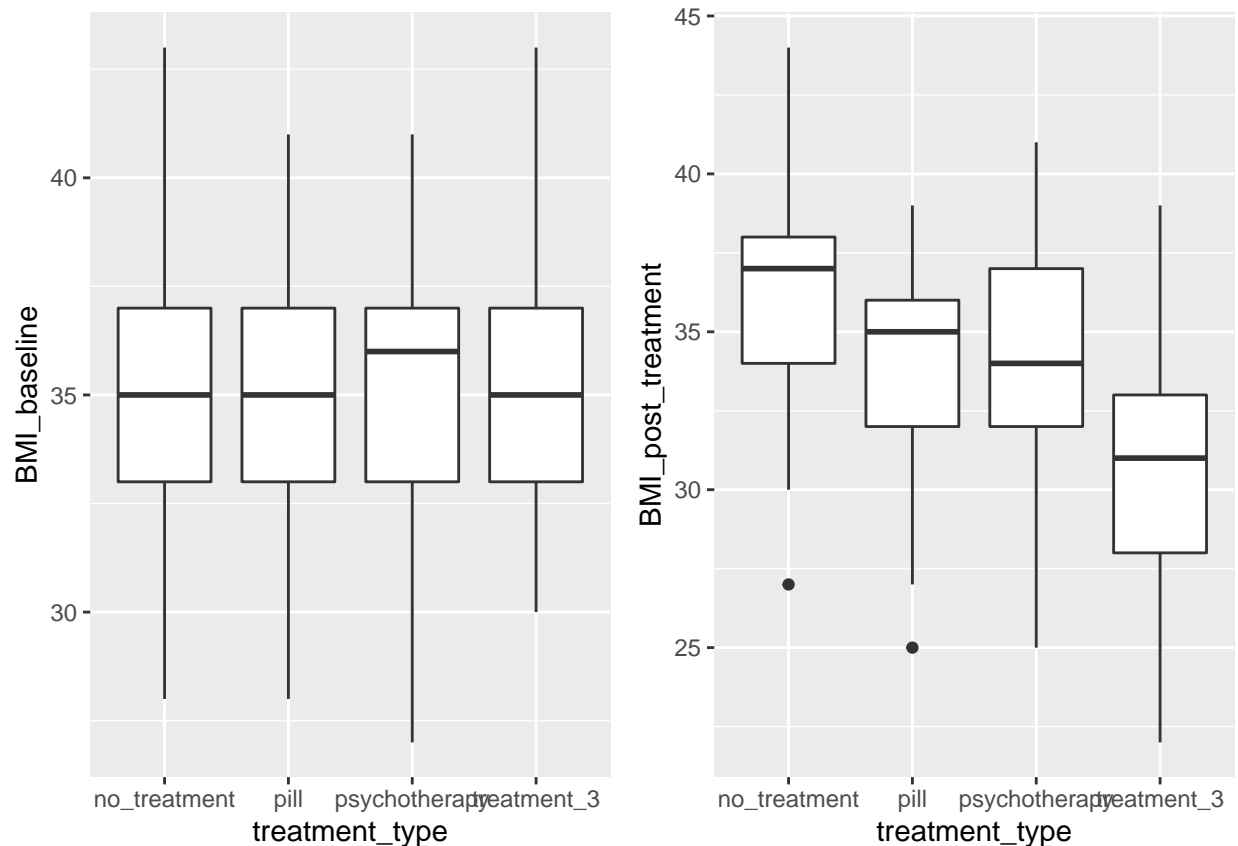
```
## <ScaleContinuousPosition>
```

```
## Range:
## Limits: 20 -- 45

fig_2 = data_weightloss %>%
  ggplot() +
  aes(y = BMI_post_treatment, x = treatment_type) +
  geom_boxplot()
  ylim(c(20, 45))

## <ScaleContinuousPosition>
## Range:
## Limits: 20 -- 45

grid.arrange(fig_1, fig_2, nrow=1)
```



```
data_weightloss %>%
  group_by(treatment_type) %>%
  summarize(mean_pre = mean(BMI_baseline),
            sd_pre = sd(BMI_baseline),
            mean_post = mean(BMI_post_treatment),
            sd_post = sd(BMI_post_treatment))

## # A tibble: 4 x 5
##   treatment_type mean_pre sd_pre mean_post sd_post
##   <fct>          <dbl> <dbl>    <dbl>    <dbl>
## 1 no_treatment    34.9  3.06     36.1     3.49
## 2 pill           35.0  2.50     34.0     2.95
## 3 psychotherapy   34.8  3.09     34.1     3.40
```

```
## 4 treatment_3          35.2    2.95    30.8    3.41
```

1.3 Kategorikus változók mint prediktorok

Mivel úgy tűnik, a csoportok összehasonlíthatóak voltak a kezelés előtt, fókuszáljunk most a kezelés utáni BMI-re (BMI_post_treatment).

A kezelés típusa (treatment_type) egy kategorikus változó, a BMI pedig egy folytonos numerikus változó. Ahogy azt korábban tanultuk, egyik módja annak, hogy kiderítsük, van-e különbség csoportok között egy adott folytonos változó átlagos szintjében, ha lefuttatunk egy **egyszempontos ANOVA**-t (aov()).

Az eredmény elárulja, hogy a kezelés utáni BMI átlaga szignifikánsan különbözik a csoportok között ($F(3, 236) = 26.51$, $p < 0.001$), (ami azt jelenti, hogy legalább két csoport szignifikánsan különbözik egymástól a BMI átlagában a négy csoport közül).

```
anova_model = aov(BMI_post_treatment ~ treatment_type, data = data_weightloss)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## treatment_type  3      877   292.33    26.51 8.17e-15 ***
## Residuals      236    2602    11.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A lineáris regresszió fontos, hogy a függő változó (a bejósolt változó) folytonos numerikus változó legyen. Viszont a modell prediktorai lehetnek akár folytonos, akár kategorikus változók (csoportosított változók mint pl. a kezelés a mi esetünkben).

Vagyis a fenti aov() modellt megépíthetjük lm() segítségével is ahogy az alábbi példa is mutatja. A **teljes modell F-tesztje** ugyan azt az eredményt adja ki, mint az aov(). Vegyük észre hogy a funkció kívül aov() vs. lm() **a két modell szintaktikailag pontosan ugyan úgy épül fel.**

```
mod_1 = lm(BMI_post_treatment ~ treatment_type, data = data_weightloss)
summary(mod_1)
```

```
##
## Call:
## lm(formula = BMI_post_treatment ~ treatment_type, data = data_weightloss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.133 -2.133 -0.050  2.200  8.200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      36.1333    0.4287   84.287 < 2e-16 ***
## treatment_typepill      -2.0833    0.6063  -3.436 0.000697 ***
## treatment_typepsychotherapy -2.0000    0.6063  -3.299 0.001121 **
## treatment_typetreatment_3  -5.3333    0.6063  -8.797 3.02e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.321 on 236 degrees of freedom
## Multiple R-squared:  0.2521, Adjusted R-squared:  0.2425
## F-statistic: 26.51 on 3 and 236 DF,  p-value: 8.173e-15
```

A regressziós együtthatók táblázata ebben az esetben maskepp néz ki a megszokotthoz képest, hiszen majdnem minden kezelési típusnak külön sora van.

1.3.1 Eredmények értelmezése

Az egyes változokhoz tartozó regressziós együtthatókat úgy értelmezzük általában, hogy **mekkora változást jelent a bejósolt változó értékeiben ha a prediktor változó értéke egy szinttel emelkedik**.

Viszont a **nominalis** változók nem sorrendezett, szóval nem tudjuk eldönteni, hogy hogyan rakjuk sorba a szinteket, hogy az egy szintnyi emelkedés hatását megbecsüljük. Ezt egy másik trükkel oldjuk meg: **dummy változokkal**.

A dummy változók gyakorlatilag azt jelentik, hogy készítünk új változókat, ami **a faktorszint megleletet (1), vagy hiányát (0) jelenti**. Vagyis lesz egy változó, ami akkor vesz fel 1-es értéket, ha valaki “pill”-t kapott, minden más esetben 0 értéket vesz fel, lesz egy másik változó ami akkor vesz fel 1-es értéket amikor valaki “psychotherapy”-t kapott, minden másik esetben 0 értéket vesz fel, és lesz egy változó ami akkor vesz fel 1-es értéket amikor valaki “treatment_3”-t kapott, minden másik esetben 0 értéket vesz fel.

Az alapszintnek nem szoktunk külön dummy változót csinálni, mert az már a többi dummy eredményéből evidens (ha minden másik dummy értéket 0, akkor az alapszint értéke 1).

```
data_weightloss = data_weightloss %>%
  mutate(
    got_pill = recode(treatment_type,
                      "no_treatment" = "0",
                      "pill" = "1",
                      "psychotherapy" = "0",
                      "treatment_3" = "0"),
    got_psychotherapy = recode(treatment_type,
                               "no_treatment" = "0",
                               "pill" = "0",
                               "psychotherapy" = "1",
                               "treatment_3" = "0"),
    got_treatment_3 = recode(treatment_type,
                             "no_treatment" = "0",
                             "pill" = "0",
                             "psychotherapy" = "0",
                             "treatment_3" = "1")
  )

mod_2 = lm(BMI_post_treatment ~ got_pill + got_psychotherapy + got_treatment_3, data = data_weightloss)
summary(mod_2)

##
## Call:
## lm(formula = BMI_post_treatment ~ got_pill + got_psychotherapy +
##     got_treatment_3, data = data_weightloss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.133  -2.133  -0.050   2.200   8.200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.1333     0.4287  84.287  < 2e-16 ***
## got_pill1      -2.0833     0.6063  -3.436  0.000697 ***
## got_psychotherapy1 -2.0000     0.6063  -3.299  0.001121 **
## got_treatment_31 -5.3333     0.6063  -8.797  3.02e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.321 on 236 degrees of freedom
## Multiple R-squared:  0.2521, Adjusted R-squared:  0.2425
## F-statistic: 26.51 on 3 and 236 DF,  p-value: 8.173e-15
```

Ez a megoldás lehetővé teszi, hogy a program **minden faktorszintet egyenként hasonlítson az alapszinthez**. Ennek az eredményt látjuk a regressziós együtthatók táblázatában.

Az **intercept**-hez tartozó regressziós együtthatót mindig úgy lehet értelmezni, hogy ez mutatja a bejosolt változó (ebben az esetben a BMI) értéket abban az esetben, **ha minden prediktor változó nulla értéket vesz fel**. Mivel itt dummy változokkal dolgozunk, ez azt jelenti, hogy az alapszinten kívül minden más szinthez tartozó dummy változó értéke 0. Vagyis mi az a BMI érték, amit akkor várhatunk ha az ember se nem “pill”-t, se nem “psychotherapy”-t, se nem “treatment_3”-t kapott (vagyis a “no_treatment” csoportban volt).

A regressziós **együtthatókat így már szokás szerint értelmezhetjük**, hogy abban az esetben ha az adott dummy változó értéke egy szinttel no (vagyis 0 helyett 1 lesz), akkor mekkora változást várhatunk a bejosolt változó értékeiben.

Az **lm()** **fuggvény mindezt elvégzi** helyettünk, nem kell manuálisan dummy változókat generalni, de az fontos, hogy megertjük, hogyan történik ez a folyamat. A kategorikus változóknak (pl. a mi esetünkben treatment type) nincs nulla értéke. Ezt az R úgy oldja meg, hogy a csoportosító változó (faktor) szintjei közül kiválaszt egyet, ami az alapszint (**default level**), és azt veszi nullának.

Fontos, hogy ahogy korábban is, az alapszint ha nem rendelkezünk maskepp alapértelmezett módon a faktor szintjei közül az **abc sorrendben legelső** lesz, a mi esetünkben ez a “no_treatment”.

Vagyis

- a “no_treatment” eseten 36.13 BMI-t várhatunk,
- ha valaki “pill”-t kap, akkor -2.08 BMI változást jósolunk a “no_treatment”-hez képest,
- ha valaki “psychotherapy”-t kap -2 BMI változást jósolunk a “no_treatment”-hez képest,
- ha valaki “treatment_3”-t kap -5.33 BMI változást jósolunk a “no_treatment”-hez képest.

Gyakorlás

Nyisd meg a data_house adattáblát amivel a korábbi gyakorlatokon foglalkoztunk, és építs egy lineáris regressziós modellt a lakás eladási árának (price) bejósolására a következő prediktorokkal: sqm_living, grade, has_basement. Értelmezd a fentiek alapján a regressziós együtthatók táblázatát. Mit jelent az intercept regressziós együtthatója? Mit jelent a has_basement prediktorhoz tartozó regressziós együttható?

```
data_house = read.csv("https://bit.ly/2DpwK0r")

data_house = data_house %>%
  mutate(price_mill_HUF = (price * 293.77)/1000000,
         sqm_living = sqft_living * 0.09290304,
         sqm_lot = sqft_lot * 0.09290304,
         sqm_above = sqft_above * 0.09290304,
         sqm_basement = sqft_basement * 0.09290304,
         sqm_living15 = sqft_living15 * 0.09290304,
         sqm_lot15 = sqft_lot15 * 0.09290304
  )
```

1.4 Ket változó interakciójának beillesztése a modellbe

A treatment_3 változóban egy olyan kondíció volt a kutatásban, ahol az emberek mind gyógyszeres, mind pszichoterapiás kezelést kaptak.

Most atalakitjuk az adattáblát, hogy ezt helyesen tukrozzek az iment generalt dummy változók.

```
data_weightloss = data_weightloss %>%  
  mutate(  
    got_pill = replace(got_pill, treatment_type == "treatment_3", "1"),  
    got_psychotherapy = replace(got_psychotherapy, treatment_type == "treatment_3", "1")  
  )
```

Most feltehetjük a kérdést, hogy **van-e interakció** a gyógyszeres kezelés és a pszichoterapiás kezelés között, vagyis **van-e valami hozzáadott értéke annak, hogy az emberek a két kezelést egyszerre kaptak** azon felül, amit a két kezelés hatása alapján varnának külön-külön.

Ezt az interakciót a modellbe úgy tudjuk beépíteni, ha ****a +** helyett ***-ot**** rakunk a két változó közé, amiknek az interakciója érdekel minket.

```
mod_3 = lm(BMI_post_treatment ~ got_pill * got_psychotherapy, data = data_weightloss)  
summary(mod_3)
```

```
##  
## Call:  
## lm(formula = BMI_post_treatment ~ got_pill * got_psychotherapy,  
##     data = data_weightloss)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -9.133 -2.133 -0.050   2.200   8.200   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)      36.1333     0.4287  84.287 < 2e-16 ***  
## got_pill1        -2.0833     0.6063  -3.436 0.000697 ***  
## got_psychotherapy1 -2.0000     0.6063  -3.299 0.001121 **   
## got_pill1:got_psychotherapy1 -1.2500     0.8574  -1.458 0.146194   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 3.321 on 236 degrees of freedom  
## Multiple R-squared:  0.2521, Adjusted R-squared:  0.2425   
## F-statistic: 26.51 on 3 and 236 DF,  p-value: 8.173e-15
```

Alternatív szintaxis: Egy másik szintaxisal is felírhatjuk ugyan ezt, ahol “got_pill * got_psychotherapy” helyett “got_pill + got_psychotherapy + got_pill:got_psychotherapy” írunk. Ez akkor lehet fontos, ha több mint két változó valamilyen komplexebb interakciós mintázatot akarjuk modellezni ahol nem akarjuk minden mindennel való interakciókat beépíteni a modellbe. Ez pontosan ugyan azt eredményezi mint a korábbi szintaxis, hiszen itt minden változó interakciókat beépítettük a modellbe (csak két prediktor változónk volt).

```
mod_3 = lm(BMI_post_treatment ~ got_pill + got_psychotherapy + got_pill:got_psychotherapy, data = data_weightloss)  
summary(mod_3)
```

```
##  
## Call:  
## lm(formula = BMI_post_treatment ~ got_pill + got_psychotherapy +  
##     got_pill:got_psychotherapy, data = data_weightloss)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -9.133 -2.133 -0.050   2.200   8.200   
##
```

```
## -9.133 -2.133 -0.050 2.200 8.200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      36.1333     0.4287  84.287 < 2e-16 ***
## got_pill1        -2.0833     0.6063  -3.436 0.000697 ***
## got_psychotherapy1 -2.0000     0.6063  -3.299 0.001121 **
## got_pill1:got_psychotherapy1 -1.2500     0.8574  -1.458 0.146194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.321 on 236 degrees of freedom
## Multiple R-squared:  0.2521, Adjusted R-squared:  0.2425
## F-statistic: 26.51 on 3 and 236 DF,  p-value: 8.173e-15
```

1.4.1 Eredmények értelmezése

Az interakcios tenyezohoz tartozo regresszios egyutthatot ugy értelmezhetjuk, hogy abban az esetben, **ha a ket valtozo szorzata egyel magasabb** erteket vesz fel (a mi esetunkben ez csak akkor lesz 1, ha mind a got_pill, mind a got_psychoterapy erteke 1), milyen valtozast varhatunk a bejosolt valtozo ertekeben **AZON FELUL, amit a ket valtozo onallo hatasam felul varnank**. Ez azert van, mert mind a got_pill, mind a got_psychotherapy valtozok erteke 1 ebben az esetben, es azok hatasa (-2.08 es -2.00) igy mar bele van kalkulalva a modellbe. Vagyis ha mind a got_pill, mind a got_psychotherapy valtozok erteke 1, akkor azon felul hogy kifejtik egyenkent hatasukat, egy extra -1.25 BMI csokkenest varhatunk az eredmenyek alapjan. Mivel ebben a kutatasban a kivanatos kimenetel a BMI csokkenes, ezert igy elmondhatjuk hogy a ket kezelesegyutt alkalmazva felerositi egymas hatasat.

Gyakorlas

Epits egy modellt a data_weightloss adatbazison ahol a **BMI_post_treatment**-t becsuljuk meg a **motivation** es a **body_acceptance** prediktorokkal, a ket prediktor interakciojat is epitsd be a modellbe.

Ertekez a regresszios egyutthatokat.

- Milyen valtozast varhatunk a BMI szintjeben ha a motivation szintje 1-el no?
- Milyen valtozast varhatunk a BMI szintjeben ha a body_acceptance szintje 1-el no?
- Van szignifikans interakcio a ket prediktor kozott?
- Hogyan ertemezhethetjuk az interakciohoz tartozo regresszios egyutthatot?

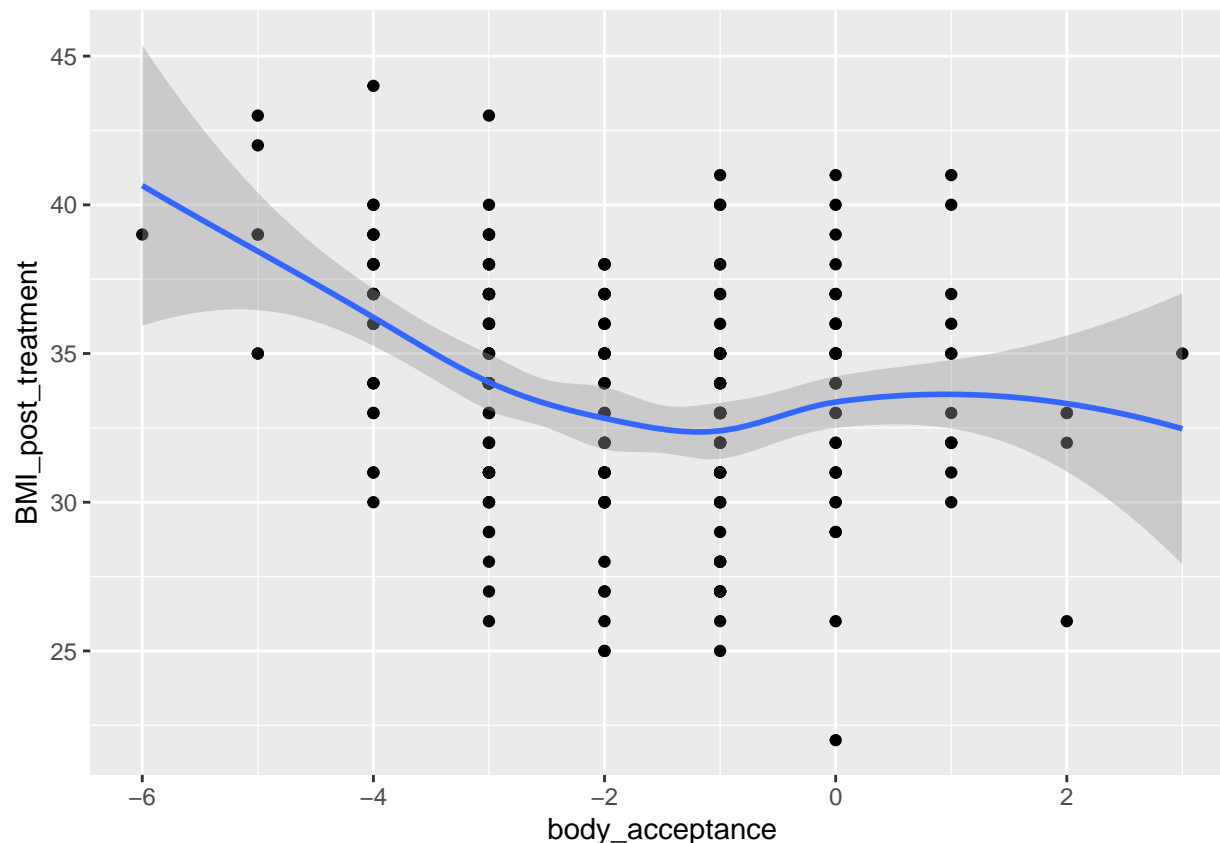
1.5 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez

A linearis regresszios modelleket eredetileg linearis osszefuggesek modellezesere talaltak ki, de egy kis matematikai trukk elerhetjuk, hogy modellezzunk **nem-linearis osszefuggesek** is.

Az alabbi abra alapjan ugy tunik, hogy BMI_post_treatment es a body_acceptance osszefuggese nem teljesen linearis, hanem egy gorbe vonal jobban leirja a ket valtozo osszefuggeset.

```
data_weightloss %>%
  ggplot() +
  aes(y = BMI_post_treatment, x = body_acceptance) +
  geom_point() +
  geom_smooth()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Ezt úgy epithetjük be a modellünkbe, hogy a prediktorok közé a `body_acceptance` mellé annak **masodik hatványát** is betesszük. Ezt a következő formula hozzáadásával tehetjük a modellben: $+ I(\text{body_acceptance}^2)$.

Ha **összehasonlítjuk** azt a modellt amiben szerepel ez a hatványtenyező azzal a modellel amiben ez nem szerepel (hierarchikus regresszio), azt találjuk hogy ez az úgynevezett kvadrátikus hatás (quadratic effect) szignifikáns hozzáadott értékkel bír a BMI bejósolásban.

```
mod_4 = lm(BMI_post_treatment ~ body_acceptance, data = data_weightloss)
summary(mod_4)
```

```
##
## Call:
## lm(formula = BMI_post_treatment ~ body_acceptance, data = data_weightloss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6960  -2.2936   0.0052   2.5112   8.9136
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32.6960     0.3613  90.490 < 2e-16 ***
## body_acceptance -0.5976     0.1495  -3.996 8.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.701 on 238 degrees of freedom
```

```
## Multiple R-squared:  0.06287,    Adjusted R-squared:  0.05894
## F-statistic: 15.97 on 1 and 238 DF,  p-value: 8.595e-05

mod_5 = lm(BMI_post_treatment ~ body_acceptance + I(body_acceptance^2), data = data_weightloss)
summary(mod_5)
```

```
##
## Call:
## lm(formula = BMI_post_treatment ~ body_acceptance + I(body_acceptance^2),
##     data = data_weightloss)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-10.7602	-2.2547	0.1633	2.3218	8.7453

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	32.76024	0.35018	93.552	< 2e-16 ***
body_acceptance	0.37209	0.27684	1.344	0.18
I(body_acceptance^2)	0.29008	0.07059	4.110	5.46e-05 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.584 on 237 degrees of freedom
## Multiple R-squared:  0.1252, Adjusted R-squared:  0.1178
## F-statistic: 16.96 on 2 and 237 DF,  p-value: 1.305e-07
```

```
AIC(mod_4)
```

```
## [1] 1313.253
```

```
AIC(mod_5)
```

```
## [1] 1298.732
```

Fontos, hogy amikor hatvagy-prediktorokat használunk mindenkeppen tegyük be a modellbe a prediktor **minden alacsonyabb hatványát is** egészen az első hatványig (ami maga az eredeti prediktor).

```
mod_6 = lm(BMI_post_treatment ~ body_acceptance + I(body_acceptance^2) + I(body_acceptance^3), data = data_weightloss)
summary(mod_6)
```

```
##
## Call:
## lm(formula = BMI_post_treatment ~ body_acceptance + I(body_acceptance^2) +
##     I(body_acceptance^3), data = data_weightloss)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-11.0437	-2.0752	0.1402	2.1689	8.9924

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.04373	0.40618	81.352	<2e-16 ***
body_acceptance	0.36393	0.27639	1.317	0.189
I(body_acceptance^2)	0.11268	0.14740	0.764	0.445
I(body_acceptance^3)	-0.03858	0.02815	-1.370	0.172

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.577 on 236 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1211
## F-statistic: 11.98 on 3 and 236 DF,  p-value: 2.513e-07
```

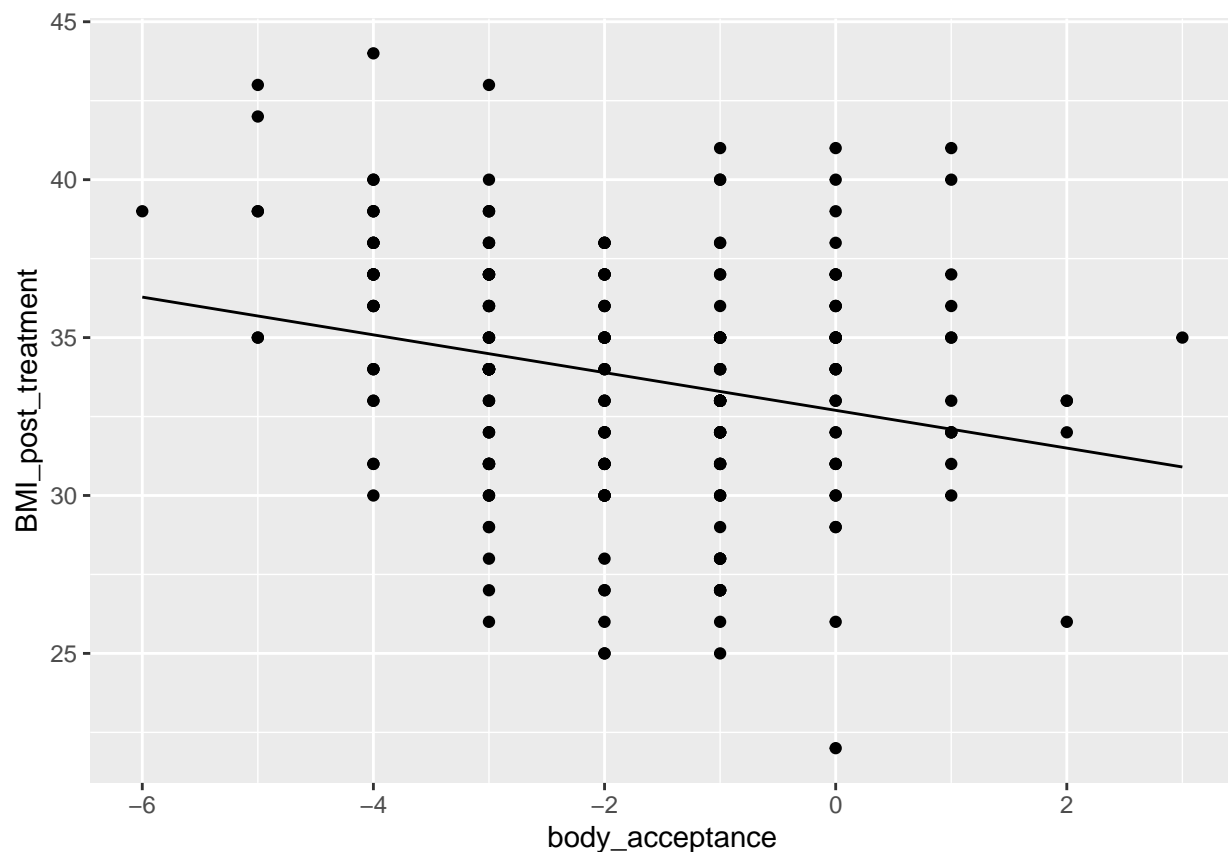
```
AIC(mod_6)
```

```
## [1] 1298.83
```

A regressziós vonal így néz ki ha csak az első hatvány szerepel a modellben:

```
data_weightloss = data_weightloss %>%
  mutate(pred_mod_4 = predict(mod_4),
         pred_mod_5 = predict(mod_5),
         pred_mod_6 = predict(mod_6))

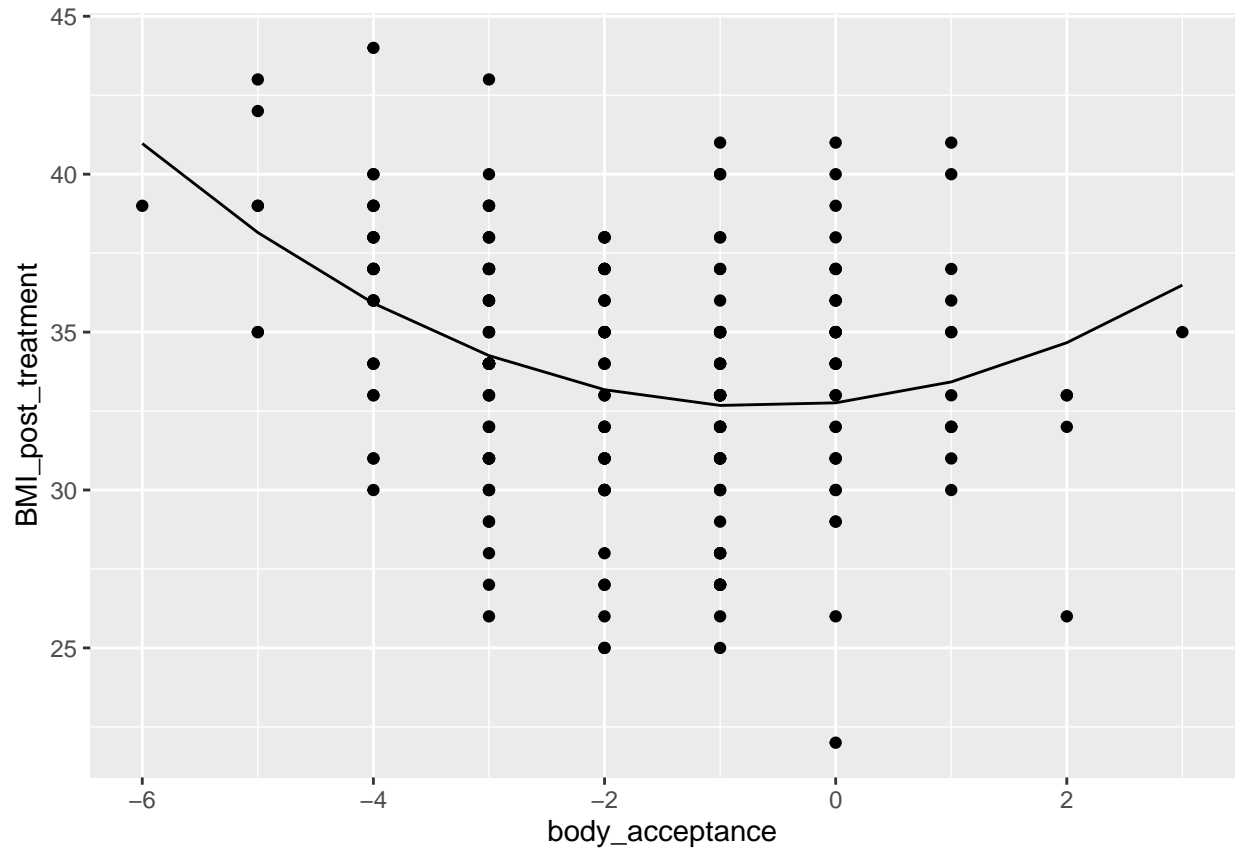
data_weightloss %>%
  ggplot() +
  aes(y = BMI_post_treatment, x = body_acceptance) +
  geom_point() +
  geom_line(aes(y = pred_mod_4))
```



Igy amikor a második hatvány szerepel a modellben:

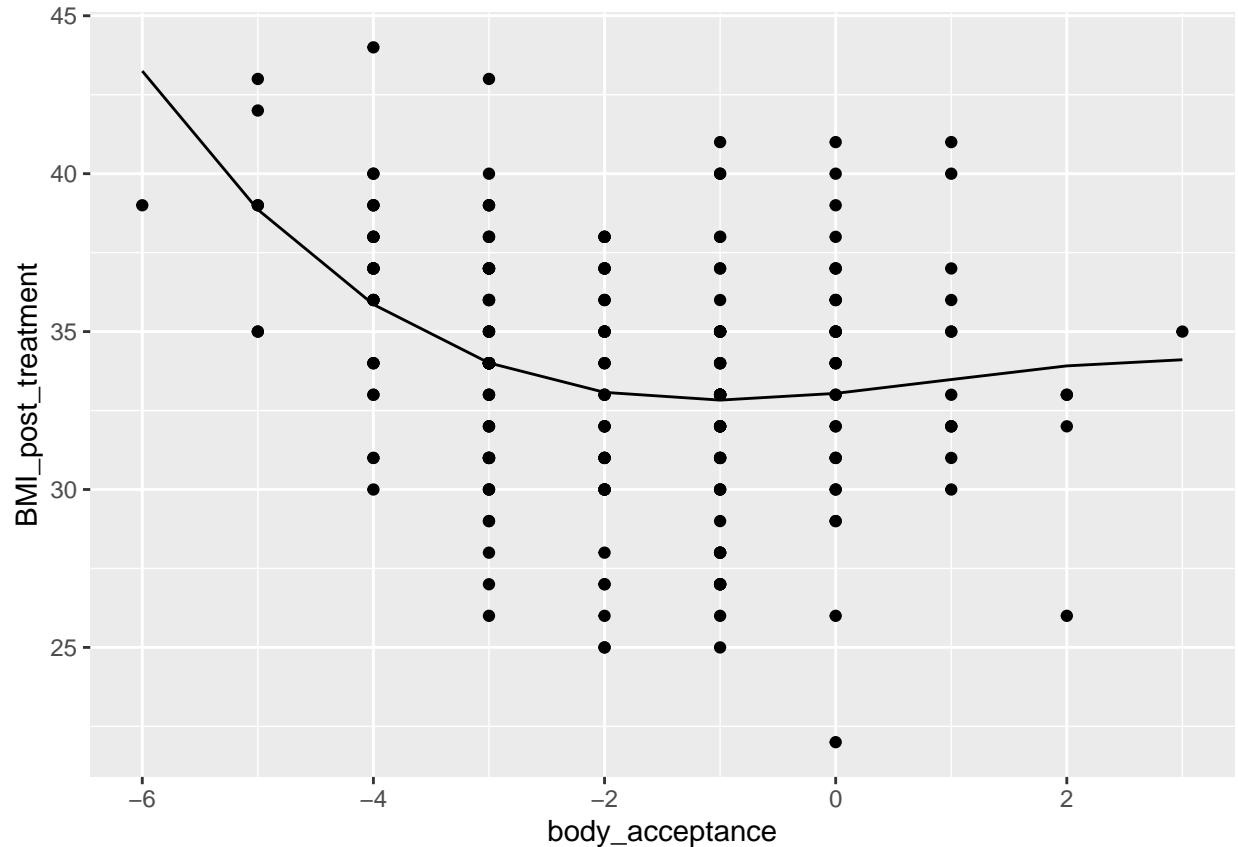
```
data_weightloss %>%
  ggplot() +
  aes(y = BMI_post_treatment, x = body_acceptance) +
  geom_point() +
```

```
geom_line(aes(y = pred_mod_5))
```



Es így amikor a harmadik hatvány szerepel a modellben:

```
data_weightloss %>%  
  ggplot() +  
  aes(y = BMI_post_treatment, x = body_acceptance) +  
  geom_point() +  
  geom_line(aes(y = pred_mod_6))
```



Lathato hogy **minel nagyobb hatvanyt illesztunk a modellbe, annal tob “gorbuletet” engedunk** a regressziós egyenesnek. (A fenti ábrák alapján lathato hogy mindig egyel kevesebb gorbületi (inflexios) pontot engedunk mint ahanyadik hatvanyt beletettuk a modellbe prediktorkent.)

Azonban a túl nagy felxibilitas nem célravezető, mert minel felxibilisebb a modell, annal inkább hajlamos arra, hogy a saját mintához illeszkedjen, és nem a populacioban megtalálható összefüggéseket ragadja meg.

Ezt túlillesztesnek (**overfitting**) nevezzük. Ezért legtöbbször nem teszünk a modellekbe haramdik hatvanynál nagyobb hatvanyprediktort, és csak akkor használunk hatvanyprediktorokat, amikor az elméletileg megalapozottnak tunik.

Gyakorlas

Experiment with different models in the house price dataset based on your theories about what could influence housing prices.

Try to increase the adjusted R^2 above 52%. If you want to get access to the whole dataset or get ideas on which model works best, go to Kaggle, check out the top kernels, and download the data. <https://www.kaggle.com/harlfoxem/housesalesprediction/activity>
