S8 Specialis prediktorok

Zoltan Kekecs

13 November 2019

Contents

1 Specialis prediktorok **1**
1.1 Absztrakt 1
1.2 Adatmenedzsment es leiro statisztikak 1
1.3 Kategorikus valtozok mint prediktorok 3
1.4 Ket valtozo interakciojanak beillesztese a modellbe 6
1.5 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez 7

1 Specialis prediktorok

1.1 Absztrakt

Ebben a gyakorlatban megismerjuk majd hogyan hasznaljuk es ertelmezzuk a kulonbozo tipusu prediktorokat a linearis regresszios 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 tartalmazzam ahol kulonbozo kezelesek hatekonysagat teszteltek a sulyvesztesre tulsolyos szemelyeknel.

Valtozok:

- ID - vizsgalati szemlely azonositojele
- Gender - nem
- Age - eletkor
- BMI_baseline - Body mass index (BMI) a kezeles elott
- BMI_post_treatment - Body mass index (BMI) a kezeles utan treatment_type - A kezeles amit a vizsgalati szemely kapott (no treatment - nem kapott kezelest; pill - etvagycsokkento gyogyszer; psychotherapy - kognitiv behavior terapia (CBT); treatment 3 - egy harmadik fajta kezeles, lasd lentebb)
- motivation - onbevallasos motivacioszint a fogyasra (0-10-es skalan, ahol a 0 extremen alacsony motivacio a fogyasra, a 10 pedig extremen magas motivacio a fogyasra)
- body_acceptance - a szemely mennyire erzi elegedettnek magat jelenleg testevel (-7 - +7, ahol a - 7 nagyon elegedetlen, a +7 nagyon elegedett)

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

```
describe(data_weightloss)
```

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

Ebben a gyakorlatban 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))
```

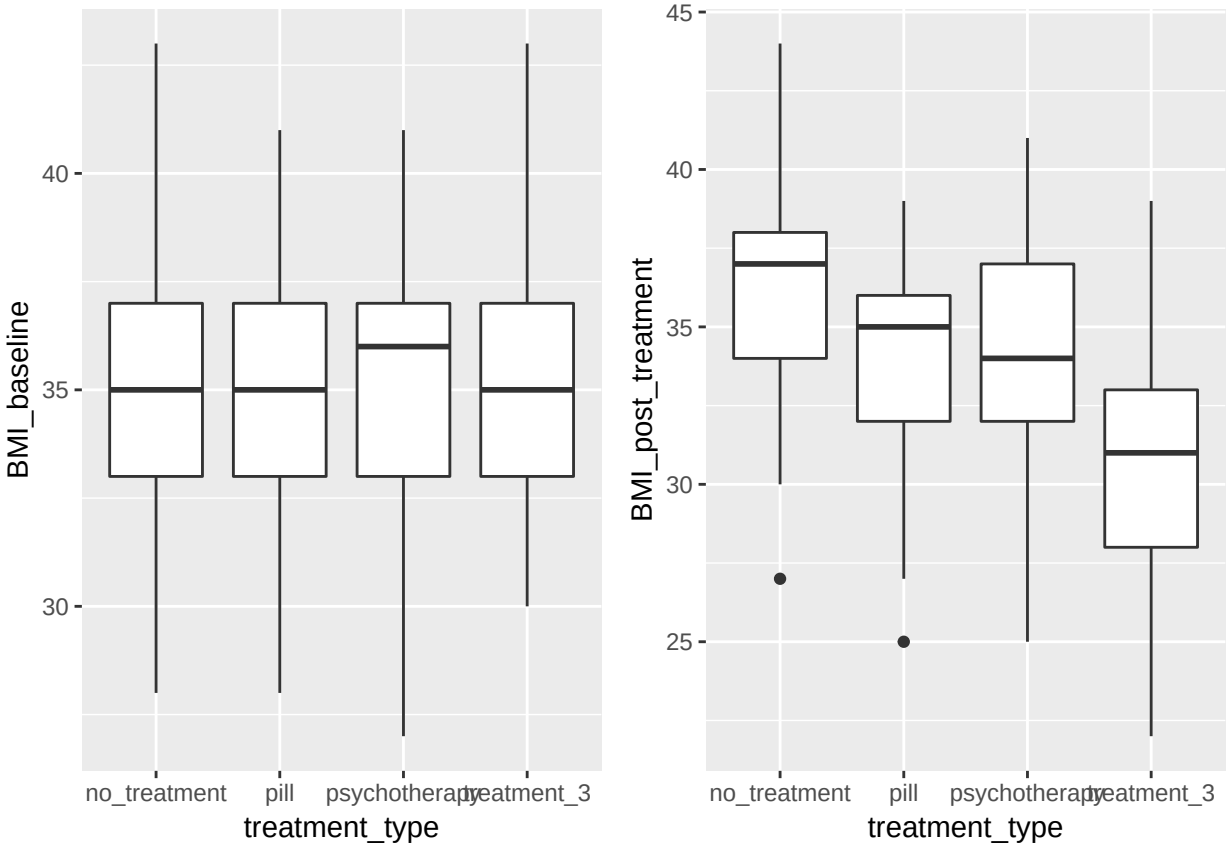
```
## <ScaleContinuousPosition>
##  Range:
##  Limits:   20 --   45
```

```
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 valtozok mint prediktorok

Mivel ugy tunik, a csoportok osszehasonlithatoak voltak a kezeles elott, fokuszaljunk most a kezeles utani BMI-re (BMI_post_treatment).

A kezeles tipus (treatment_type) egy kategorikus valtozo, a BMI pedig egy folytonos numerikus valtozo. Ahogy azt korabban tanultuk, egyik modja annak hogy kideritsuk, van-e kulonbseg csoportok kozott egy adott folytonos valtozo atlagos szintjeben, ha lefuttatunk egy egyszempontos ANOVA-t (aov()).

Az eredmeny elarulja, hogy a kezeles utani BMI atlaga szignifikansan kulonbozik a csoportok kozott ($F(3, 236) = 26.51$, $p < 0.001$), (ami azt jelenti, hogy legalabb ket csoport szignifikansan kulonbozik egymastol a

BMI atlagaban a negy csoport kozul).

```
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 linearis regresszional fontos, hogy a fuggo valtozo (a bejosolt valtozo) folytonos numerikus valtozo legyen. Viszont a modell prediktorai lehetnek akar folytonos, akar kategorikus valtozok (csoportosito valtozok mint pl. a kezeles a mi esetunkben).

Vagyis a fenti aov() modellt megepithetjuk lm() segitsegevel is ahogy az alabbi pelda is mutatja. A teljes modell F-tesztje ugyan azt az eredmenyt adja ki, mint az aov().

```
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 regresszios egyutthatok tablazata ebben az esetben maskepp nez ki a megszokotthoz kepest, hiszen majdnem minden kezelesi tipusnak kulon sora van.

Az egyes valtozokhoz tartozo regresszios egyutthatokat pedig ugy ertelmezzuk altalaban, hogy mekkora valtozast jelent a bejosolt valtozo ertekeben ha a prediktor valtozo erteke egy szinttel emelkedik. Viszont a nominalis valtozok nem sorrendezettek, szoval nem tudjuk eldonteni, hogy hogyan rakjuk sorba a szinteket, hogy az egy szintnyi emelkedes hatasat megbecsuljuk. Ezt egy masik trukkel oldjuk meg: dummy valtozokkal. A dummy valtozok gyakorlatilag azt jelentik, hogy keszitunk uj valtozokat, ami a faktorszint megletet (1), vagy hianyat (0) jelenti. Vagyis lesz egy valtozo, ami akkor vesz fel 1-es erteket, ha valaki “pill”-t kapott, minden mas esetben 0 erteket vesz fel, lesz egy masik valtozo ami akkor vesz fel 1-es erteket amikor valaki “psychotherapy”-t kapott, minden masik esetben 0 erteket vesz fel, es lesz egy valtozo ami akkor vesz fel 1-es erteket amikor valaki “treatment_3”-t kapott, minden masik esetben 0 erteket vesz fel. Az alapszintnek nem szoktunk kulon dummy valtozot csinalni, mert az mar a tobbi dummy eredmenyebol evidens (ha minden masik dummy erteket 0, akkor az alapszint erteke 1).

```r
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 megoldas lehetove teszi, hogy a program minden valtozoszintet egyenkent hasonlitson az alapszinthez. Ennek az eredmenyet latjuk a regresszios egyutthatok tablazataban.

Az intercept-hez tartozo regresszios egyutthatot mindig ugy lehet ertelmezni, hogy ez mutatja a bejosolt valtozo (ebben az esetben a BMI) erteket abban az esetben, ha minden prediktor valtozo nulla erteket vesz fel. Mivel itt dummy valtozokkal dolgozunk, ez azt jelenti, hogy az alapszinten kivul minden mas szinthez tartozo dummy valtozo erteke 0. Vagyis mi az a BMI ertek, amit akkor varhatunk ha az ember se nem “pill”-t, se nem “psychotherapy”-t, se nem “treatment_3”-t nem kapott.

A regresszios egyutthatokat igy mar szokas szerint ertelmezhetjuk, hogy abban az esetben ha az adott dummy valtozo erteke egy szinttel no (vagyis 0 helyett 1 lesz), akkor mekkora valtozast varhatunk a bejosolt valtozo ertekeben.

Az R mindezt elvegzi helyettunk, nem kell nekunk manualisan dummy valtozokat generalni, de az fontos, hogy megertsuk, hogyan tortenik ez a folyamat. De a kategorikus valtozoknak (pl. a mi esetunkben treatment type) onmagaban nincs nulla erteke. Ezt az R ugy oldja meg, hogy a csoportosito valtozo szintjei kozul kivalaszt egyet, ami az alapszint (default level), es azt veszi nullanak. Ahogy korabban is, az alapszint ha nem rendelkezunk maskepp alapertelmezett modon a faktor szintjei kozul az abc sorrendben legeleso lesz, a mi esetunkben ez a “no_treatment”.

Vagyis

- a “no_treatment” eseten 36.13 BMI-t varhatunk,
- ha valaki “pill”-t kap, akkor -2.08 BMI valtozast josolunk,

- ha valaki “psychotherapy”-t kap -2 BMI valtozast josolunk,
- ha valaki “treatment_3”-t kap -5.33 BMI valtozast josolunk.

Gyakorlas

Nyisd meg a data_house adattablat amivel a korabbi gyakorlatokon foglalkoztunk, es epits egy linearis regresszios modellt a lakas eladasi aranak (price) bejoslasara a kovetkezo prediktorokkal: sqm_living, grade, has_basement. Ertelmezd a fentiek alapjan a regresszios egyutthatok tablazatat. Mit jelent az intercept regresszios egyutthatoja? Mit jelent a has_basement prediktorhoz tartozo regresszios egyutthato?

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

data_house = data_house %>% mutate(price_mill_HUF = (price *
    293.77)/1e+06, 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 valtozo interakciojanak beillesztese a modellbe

A treatment_3 valojaban egy olyan kondicio volt a kutatasban, ahol az emberek mind gyogyszeres, mind pszichoterapias kezelest kaptak.

Most atalakitjuk az adattablat, hogy ezt helyesen tukrozzek az iment generalt dummy valtozok.

```
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 feltehetjuk a kerdest, hogy van-e interakcio a gyogyszeres kezeles es a pszichoterapias kezeles kozott, vagyis van-e valami hozzaadott erteke annak, hogy az emberek a ket kezelest egyszerre kaptak azon felul, amit a ket kezeles hatasa alapjan varnanak kulon-kulon.

Ezt az interakciot a modellbe ugy tudjuk beepiteni, ha a + helyett *-ot rakunk a ket valtozo koze, amiknek az interakcioja erdekel mindket.

```
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
```

Itt az interakcios tenyezohoz tartozo regresszios egyutthatot ugy ertelmezhetjuk, 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 (-2.0833 es -2.0000) igy mar bele van kalkulalva a modellbe. Vagyis ha mind a pill, mind a got_pill, mind a got_psychotherapy valtozok erteke 1, akkor azon felul hogy kifejtik egyenkent hatasukat, egy extra -1.2500 BMI csokkenest varhatunk az eredmenyek alapjan.

_____*Gyakorlas*_____

A data_house adatokon epits egy linearis regresszios modellt a lakas eladasi aranak (price) bejoslasara a kovetkezo prediktorokkal: sqm_living, grade, lat, long. De ahelyett hogy csak a fohatasokat nezned, kalkulald be a lat es long interakciojanak hatasat is! Ertelmezd az interakciohoz tartozo regresszios egyutthatot annak tudataban, hogy a latitude ertek minel magasabb, annal inkabb eszakra van a hely, es a longitude ertek minel magasabb annal inkabb keletre van a hely.

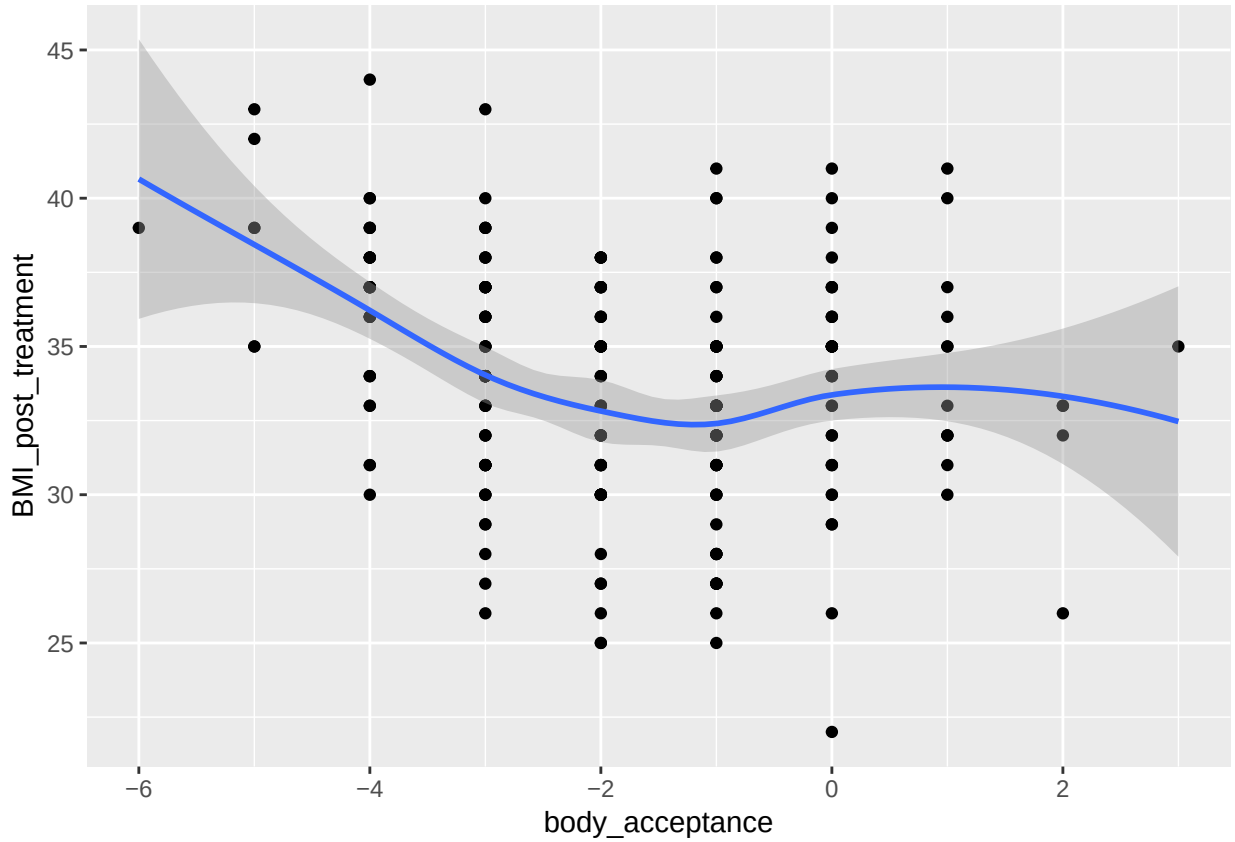
1.5 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez

A linearis regresszios modelleket eredetileg linearis osszefuggesek modellezesere talaltak ki, de egy kis matematikai trukkel 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 ugy epithetjuk be a modellunkbe, hogy a prediktorok koze a body_acceptance melle annak masodik hatvanyat is betesszuk. Ezt a kovetkezo formula hozzadasaval tehetjuk a modellben: + I(body_acceptance^2).

A modell summary es a modell illeszkedesi mutato alapjan ugy tunik, hogy ez az ugynevezett qyadratikus hatas szignifikans hozzaadott ertekkel bir a BMI bejoslasaban.

```
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
```

```r
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
```

```r
AIC(mod_4)
```

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

```r
AIC(mod_5)
```

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

Fontos, hogy amikor hatvagy-prediktorokat hasznalunk mindenkeppen tegyuk be a modellbe a prediktor minden alacsonyabb hatvanyat is egeszen az elso hatvanyig (ami maga az eredeti prediktor).

```r
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
```

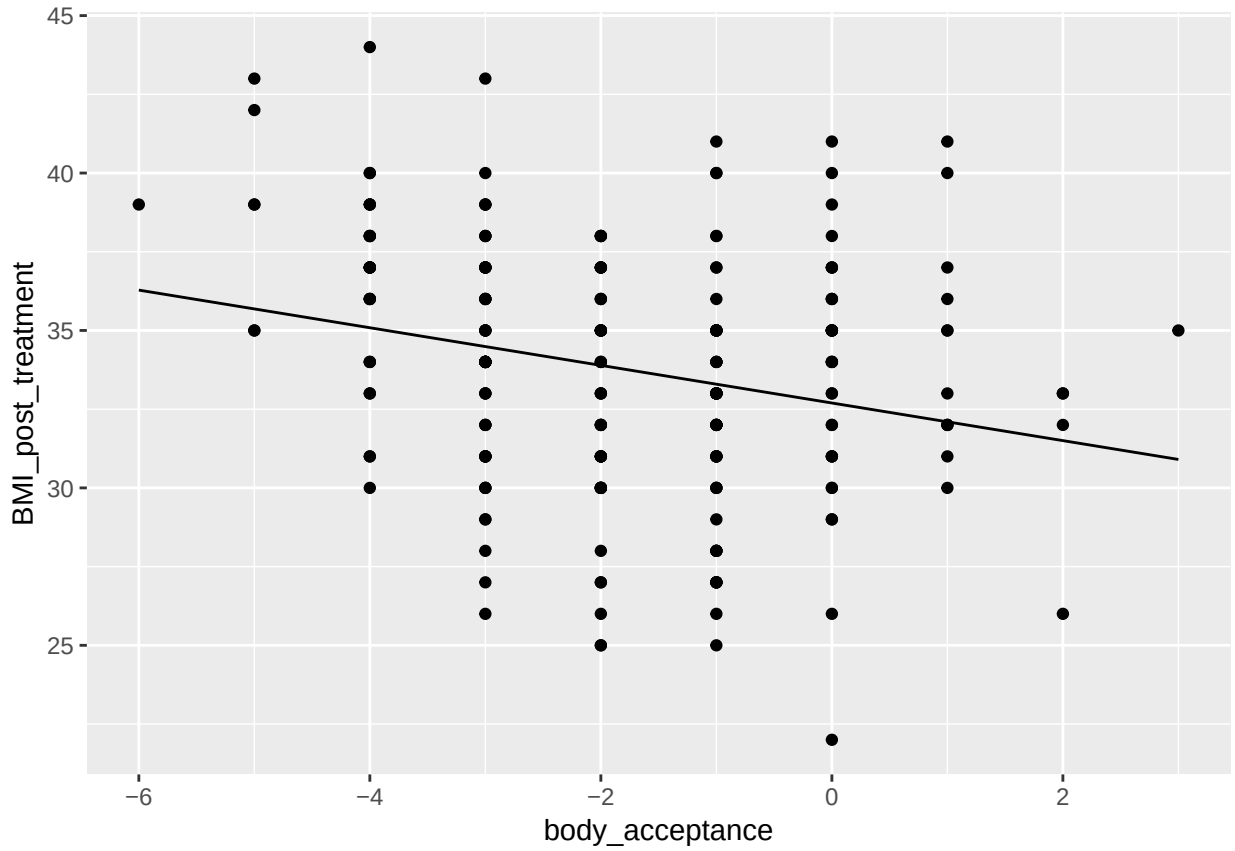
```
AIC(mod_6)
```

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

A regresszios "egyenes" igy nez ki ha csak az elso hatvany szerepel a modellben:

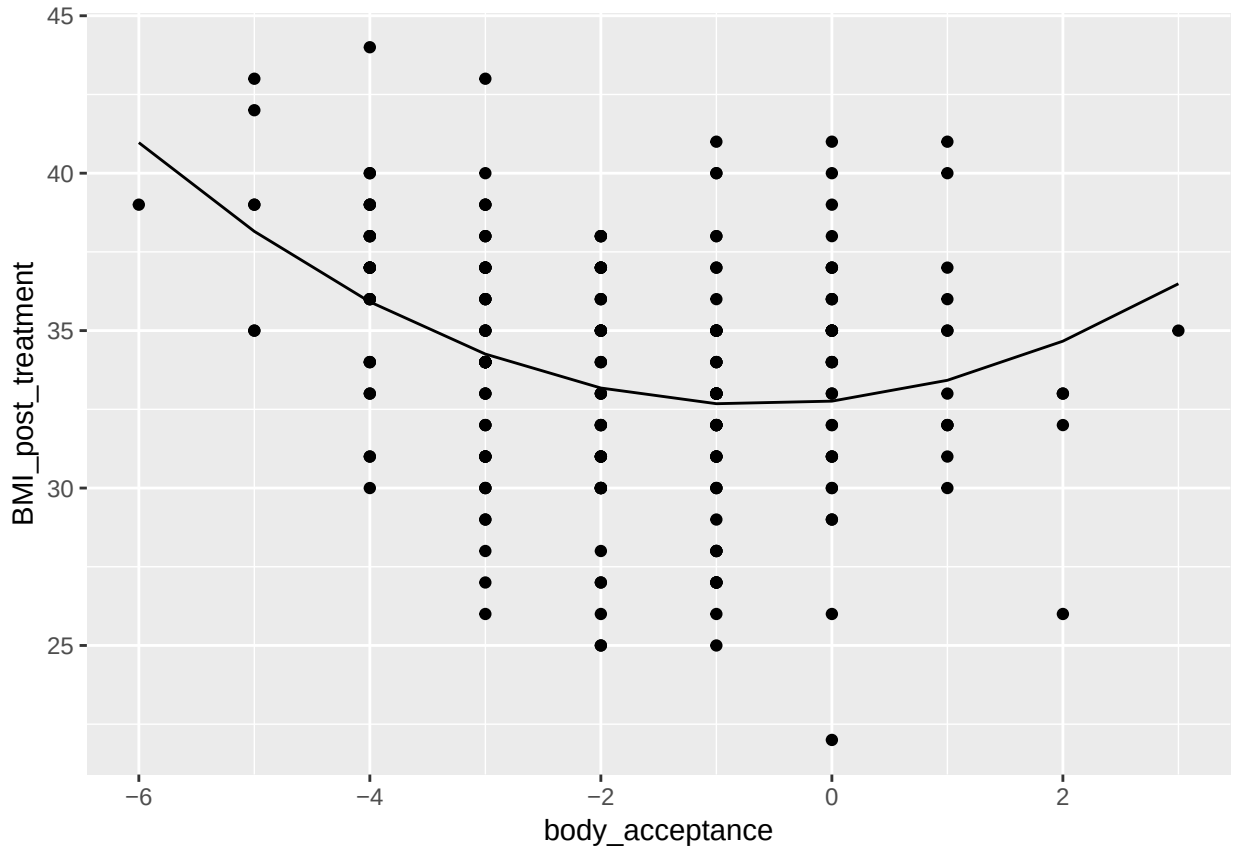
```
data_weightloss = data_weightloss %>% mutate(pred_mod_4 = predict(mod_4),
    pred_mod_5 = predict(mod_5), pred_mod_6 = predict(mod_6))

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



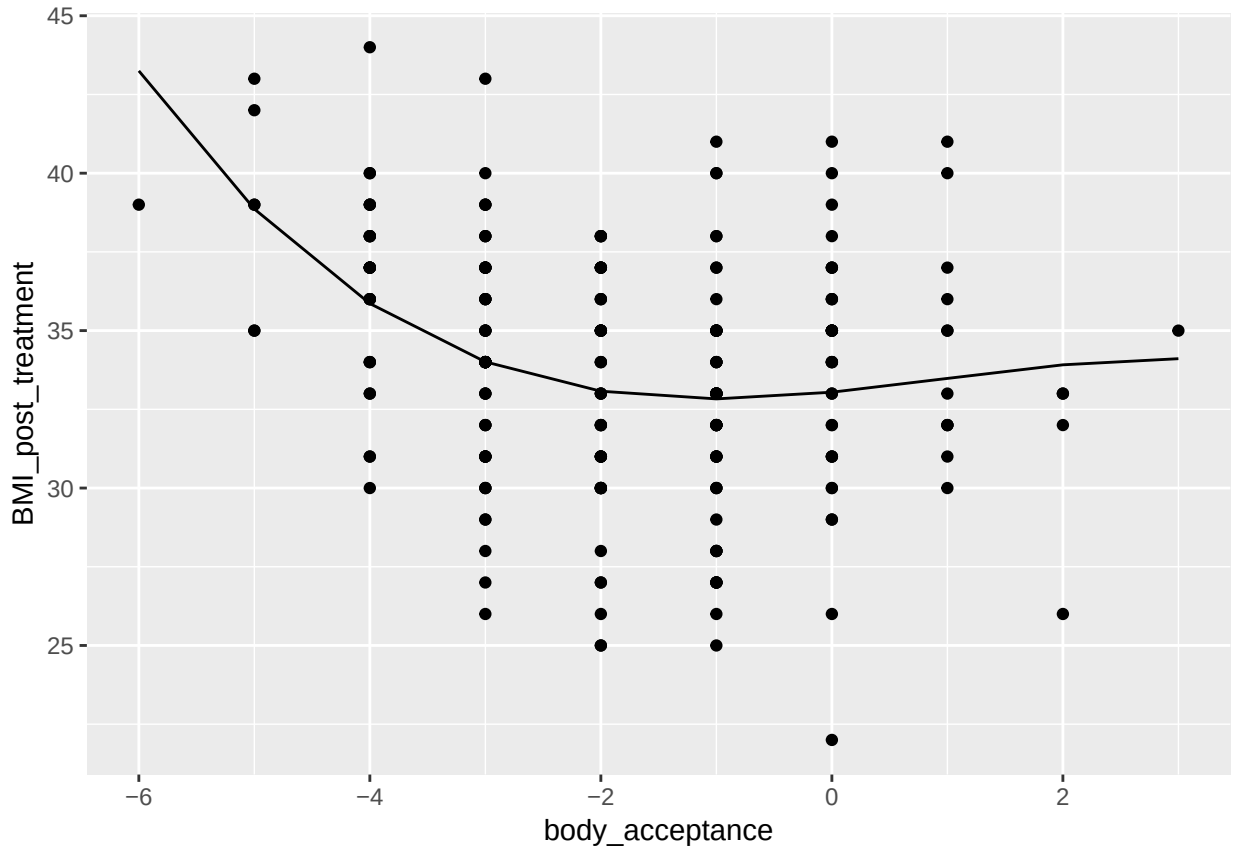
Igy amikor a masodik hatvany szerepel a modellben:

```
data_weightloss %>% ggplot() + aes(y = BMI_post_treatment, x = body_acceptance) +
    geom_point() + geom_line(aes(y = pred_mod_5))
```



Es igy amikor a harmadik hatvany 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 regresszios egyenesnek. (Mindig egyel kevesebb gorbuleti (inflexios) pontot engedunk mint ahanyadik hatvanyt beletettuk a modellbe prediktorkent.)

Azonban a tul nagy felxibilitas nem celravazeto, mert minel felxibilisebb a modell, annal inkabb hajlamos arra, hogy a sajat mintankhoz illeszkedjen, es nem a populacioban megtalalhato osszefuggeseket ragadja meg. Ezt tulillesztesnek (overfitting) nevezzuk. Ezert legtobbszor nem teszunk a modellekbe haramdik hatvanynal nagyobb hatvanypediktort, es csak akkor hasznalunk hatvanyprediktorokat, amikor az elmeletileg megalapozottnak tunik.

_____*Gyakorlas*_____

Experiment with different models 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