S08 Modell osszehasonlitas, es specialis prediktorok

Zoltan Kekecs

29 March 2021

Contents

0.1 Absztrakt 1
0.2 Package-ek betoltese 1

1 Modell osszehasonlitas es Modellvalasztas **1**
1.1 Adatmenedzsment es leiro statisztikak 1
1.2 Hierarchikus regresszio 3
1.3 Hieararchikus regresszio ket prediktor-blokkal 3
1.4 Hierarchikus regresszio tobb mint ket blokkal 4

2 Specialis prediktorok **5**
2.1 Adatmenedzsment es leiro statisztikak 5
2.2 Kategorikus valtozok mint prediktorok 8
2.3 Ket valtozo interakciojanak beillesztese a modellbe 11
2.4 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez 13

0.1 Absztrakt

Ez a gyakorlat megmutatja majd, hogyan lehet kulonbozo prediktorokat tartalmazo modelleket osszehasonlitani egymassal. Demonstraljuk majd a hierarchikus regressziot. Nehany modell szelekcios modszerre is kiterunk majd, es megemlitjuk a “tulillesztes” (overfitting) fogalmat. Ezen felul megismerjuk majd hogyan hasznaljuk es ertelmezzuk a kulonbozo tipusu specialis prediktorokat a linearis regresszios modellekben.

0.2 Package-ek betoltese

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

1 Modell osszehasonlitas es Modellvalasztas

1.1 Adatmenedzsment es leiro statisztikak

1.1.1 A King County lakaseladas adattabla betoltese

Ebben a gyakorlatban lakasok es hazak arait fogjuk megbecsulni.

Egy Kaggle-rol szarmazo adatbazist hasznalunk, melyben olyan adatok szerepelnek, melyeket valoszinusithetoen alkalmasak lakasok aranak bejoslasara. Az adatbazisban az USA Kings County-bol szarmaznak az adatok (Seattle es kornyeke).

Az adatbazisnak csak egy kis reszet hasznaljuk (N = 200).

```
# data from github/kekecsz/PSYP13_Data_analysis_class-2018/master/data_house_small_sub.csv.
data_house = read.csv("https://bit.ly/2DpwKOr")
```

1.1.2 Adatellenorzes

Mindig nezd at az altalad hasznalt adattablat. Ezt mar megtettuk az elozo gyakorlatban, igy ezt most itt mellozzuk, de a korabbi tapasztalatok alapjan atalakitjuk az arat (price) millio forintra, es a negyzetlabban szereplo terulet ertekeket negyzetmeterre.

```
data_house %>%
  summary()
```

```
##        id                 date               price            bedrooms
##  Min.   :1.600e+07   Length:200         Min.   : 153503   Min.   :1.00
##  1st Qu.:1.885e+09   Class :character   1st Qu.: 299250   1st Qu.:3.00
##  Median :3.521e+09   Mode  :character   Median : 425000   Median :3.00
##  Mean   :4.113e+09                      Mean   : 453611   Mean   :2.76
##  3rd Qu.:6.424e+09                      3rd Qu.: 550000   3rd Qu.:3.00
##  Max.   :9.819e+09                      Max.   :1770000   Max.   :3.00
##    bathrooms      sqft_living      sqft_lot          floors        waterfront
##  Min.   :0.75   Min.   : 590   Min.   :   914   Min.   :1.000   Min.   :0.000
##  1st Qu.:1.00   1st Qu.:1240   1st Qu.:  4709   1st Qu.:1.000   1st Qu.:0.000
##  Median :1.75   Median :1620   Median :  7270   Median :1.000   Median :0.000
##  Mean   :1.85   Mean   :1728   Mean   : 12985   Mean   :1.472   Mean   :0.005
##  3rd Qu.:2.50   3rd Qu.:1985   3rd Qu.: 10187   3rd Qu.:2.000   3rd Qu.:0.000
##  Max.   :3.50   Max.   :4380   Max.   :217800   Max.   :3.000   Max.   :1.000
##       view          condition        grade          sqft_above   sqft_basement
##  Min.   :0.000   Min.   :3.00   Min.   : 5.00   Min.   : 590   Min.   :   0.0
##  1st Qu.:0.000   1st Qu.:3.00   1st Qu.: 7.00   1st Qu.:1090   1st Qu.:   0.0
##  Median :0.000   Median :3.00   Median : 7.00   Median :1375   Median :   0.0
##  Mean   :0.145   Mean   :3.42   Mean   : 7.36   Mean   :1544   Mean   : 184.1
##  3rd Qu.:0.000   3rd Qu.:4.00   3rd Qu.: 8.00   3rd Qu.:1862   3rd Qu.: 315.0
##  Max.   :4.000   Max.   :5.00   Max.   :11.00   Max.   :4190   Max.   :1600.0
##     yr_built     yr_renovated        zipcode           lat
##  Min.   :1900   Min.   :   0.00   Min.   :98001   Min.   :47.18
##  1st Qu.:1946   1st Qu.:   0.00   1st Qu.:98033   1st Qu.:47.49
##  Median :1968   Median :   0.00   Median :98065   Median :47.58
##  Mean   :1968   Mean   :  79.98   Mean   :98078   Mean   :47.57
##  3rd Qu.:1993   3rd Qu.:   0.00   3rd Qu.:98117   3rd Qu.:47.68
##  Max.   :2015   Max.   :2014.00   Max.   :98199   Max.   :47.78
##       long        sqft_living15    sqft_lot15     has_basement
##  Min.   :-122.5   Min.   : 740   Min.   :   914   Length:200
##  1st Qu.:-122.3   1st Qu.:1438   1st Qu.:  5000   Class :character
##  Median :-122.2   Median :1715   Median :  7222   Mode  :character
##  Mean   :-122.2   Mean   :1793   Mean   : 11225
##  3rd Qu.:-122.1   3rd Qu.:2072   3rd Qu.: 10028
##  Max.   :-121.7   Max.   :3650   Max.   :208652
```

```
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
        )
```

1.2 Hierarchikus regresszio

A hierarchikus regresszioval (Hierarchical regression) meghatarozhatjuk, hogy kulonbozik-e ket modell egymastol a bejoslo hatekonysagukat tekintve, es ha igen, **mennyivel javul a bejoslo ero** egy bonyolultabb (tobb prediktort tartalmazo) modell hasznalataval, ahhoz kepest ha egy egyszerubb (kevesebb prediktort tartalmazo) modellt hasznalnank.

Mivel a hierarchikus regresszió gyakorlatilag ket regresszios modell (egy egyszerubb es egy osszetettebb) osszehasonlitasa, ezert most mi is ket regresszios modellt fogunk epiteni.

1.3 Hieararchikus regresszio ket prediktor-blokkal

1.3.1 Modellepites

Eloszor epitunk egy egyszeru modellt amiben a haz vetelarat csak a *sqm_living* es a *grade* valtozok alapjan josoljuk be.

```
mod_house2 <- lm(price_mill_HUF ~ sqm_living + grade, data = data_house)
```

Majd epitunk egy bonyolultabb modellt, amiben a sqm_living es a grade prediktorokon kivul szerepelnek meg a lakas foldrajzi hosszusag es szelesseg adatai is (*long* es *lat*).

```
mod_house_geolocation = lm(price_mill_HUF ~ sqm_living + grade + long + lat, data = data_house)
```

Vegyuk eszre, hogy az egyszerubb modellben szereplo prediktorok egy **reszhalmazat** alkotjak a bonyolultabb modell prediktorainak. vagyis **a bonyolultabb modell minden prediktort tartalmaz az egyszerubb modellbol**, plusz meg nehany extra prediktort. Ezt ugy nevezzuk hogy "**nested models**" vagyis "**egymasba agyazott modellek**", hiszen a modellek ugy epulnek fel mint a matrjoska babak.

1.3.2 Modellosszehasonlitas

Az **adj. R Squared** mutato segitsegevel meghatarozhatjuk a ket modell altal megmagyarazott varianciaaranyt. Ezt a model summary kilistazasaval is megtehetjuk, de a model summary-bol csak ez az informacio is kinyerheto a $adj.r.squared hozzaadasaval az alabbi modon:

```
summary(mod_house2)$adj.r.squared
```

```
## [1] 0.3515175
```

```
summary(mod_house_geolocation)$adj.r.squared
```

```
## [1] 0.4932359
```

Ugy tunik, hogy a megmagyarazott varianciaarany magasabb lett azzal, hogy a modellhez hozzatettuk a geolokacioval kapcsolatos informaciot.

Most meghatarozhatjuk, hogy ez a bejosloeroben bekovetkezett **javulas szignifikans-e**. Ezt egyreszt a ket modell AIC modell-illeszkedesi mutatojanak osszehasonlitasaval tehetjuk meg.

Ha a ket **AIC** ertek kozotti kulonbseg nagyobb mint 2, a ket modell illeszkedese szignifikansan kulonbozik egymastol. Az alacsonyabb AIC kevesebb hibat es jobb modell illeszkedest jelent. Ha a kulonbseg nem eri el a 2-t, akkor a ket modell kozul barmelyiket megtarthatjuk. Ilyenkor altalaban azt a modellt tartjuk meg amelyik elmeletileg megalapozottabb, de ha nincs eros elmeletunk, akkor az egyszerubb modellt szoktuk megtartani (amelyikben kevesebb prediktor van).

```
AIC(mod_house2)
```

```
## [1] 2137.057
```

```
AIC(mod_house_geolocation)
```

```
## [1] 2089.698
```

Masreszt pedig az **anova() funkcio** segitsegevel osszehasonlithatjuk a ket modell residualis hibajat.

Ha az anova() F-tesztje szignifikans, az azt jeletni, hogy a ket modell rezidualis hibaja szignifikansan kulonbozik egymastol.

```
anova(mod_house2, mod_house_geolocation)
```

```
## Analysis of Variance Table
##
## Model 1: price_mill_HUF ~ sqm_living + grade
## Model 2: price_mill_HUF ~ sqm_living + grade + long + lat
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 491749
## 2    195 380382  2    111367 28.546 1.338e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Az **AIC** mutato alapjan valo modell-osszehasonlitas **jobban elfogadott** a szakirodalomban, ezert ha az AIC es az anova osszehasonlitas kulonbozo eredmenyre vezet, akkor az AIC eredmenyet erdemes hasznalni.

Fontos, hogy az **anova osszehasonlitasnak az eredmenye csak akkor valid, ha egymasba agyazott (nested)** modellek osszehasonlitasara hasznaljuk oket, vagyis az egyik modell prediktorai a masik modell prediktorainak reszhalmazat alkotjak.

Az AIC legtobbszor alkalmas nem beagyazott modellek osszehasonlitasara is, (bar ezzel kapcsolatban nem teljes az egyetertes a szakirodalomban, a dolgozatokban elfogadott AIC-ot hasznalni nem beagyazott modellek osszehasonlitasara).

1.4 Hierarchikus regresszio tobb mint ket blokkal

A fenti folyamat ugyan ugy megismetelheto ha tobb mint ket blokkban adjuk hozza a prediktorokat a modellhez.

Itt egy harmadik modellt epitunk, a “condition” prediktor hozzaadasaval.

```
mod_house_geolocation_cond = lm(price_mill_HUF ~ sqm_living + grade + long + lat + condition, data = da
```

A harom modellt kovetkezokeppen hasonlithatjuk ossze:

```
# R^2
summary(mod_house2)$adj.r.squared
```

```
## [1] 0.3515175
```

```
summary(mod_house_geolocation)$adj.r.squared
```

```
## [1] 0.4932359
```

```
summary(mod_house_geolocation_cond)$adj.r.squared
```

```
## [1] 0.5065859
```

```
# anova
anova(mod_house2, mod_house_geolocation, mod_house_geolocation_cond)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: price_mill_HUF ~ sqm_living + grade
## Model 2: price_mill_HUF ~ sqm_living + grade + long + lat
## Model 3: price_mill_HUF ~ sqm_living + grade + long + lat + condition
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 491749
## 2    195 380382  2    111367 29.318 7.493e-12 ***
## 3    194 368462  1     11920  6.276   0.01306 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC
AIC(mod_house2)
```

```
## [1] 2137.057
```

```
AIC(mod_house_geolocation)
```

```
## [1] 2089.698
```

```
AIC(mod_house_geolocation_cond)
```

```
## [1] 2085.33
```

A fenti eredmenyek alapjan javult a bejoslo ereje a modellunknek a lakas allapotanak (condition) figyelembevetelevel?

_____*Gyakorlas*_____

Tedd hozza a modellhez az iment epitett modellhez (mod_house_geolocation_cond) a haz epitesenek evet (yr_built) es a furdoszobak szamat (bathrooms) mint prediktorokat. Ez az uj modell szignifikansan jobban illeszkedik az adatokhoz mint a korabbi modellek?

A modellvalasztas legfontosabb szabalya:

Mindig azt a modellt valasztjuk, ami **elmeletileg alatamasztott** es/vagy korabbi kutatasi eredmenyek tamogatjak, mert az automatikus modellvalasztas rossz modellekhez vezet a tulillesztes (overfitting) miatt.

2 Specialis prediktorok

2.1 Adatmenedzsment es leiro statisztikak

2.1.1 A fogyasi kutatas adatbazis betoltese

Az adatbazis egy olyan kutatas szimulalt adatait tartalmazzam ahol kulonbozo kezelesek hatekonysagat teszteltek a sulyvesztesre tulsulyos 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")
```

2.1.2 Adatellenorzes

Nezzuk at eloszor az altalunk hasznalt adattablat.

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

```
##       ID                gender               age         BMI_baseline
##  Length:240         Length:240         Min.   :21.00   Min.   :27.00
##  Class :character   Class :character   1st Qu.:33.00   1st Qu.:33.00
##  Mode  :character   Mode  :character   Median :35.00   Median :35.00
##                                        Mean   :34.78   Mean   :34.98
##                                        3rd Qu.:38.00   3rd Qu.:37.00
##                                        Max.   :50.00   Max.   :43.00
##  BMI_post_treatment treatment_type       motivation     body_acceptance
##  Min.   :22.00      Length:240         Min.   : 2.000   Min.   :-6.000
##  1st Qu.:31.00      Class :character   1st Qu.: 5.000   1st Qu.:-3.000
##  Median :34.00      Mode  :character   Median : 6.000   Median :-2.000
##  Mean   :33.78                         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 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.

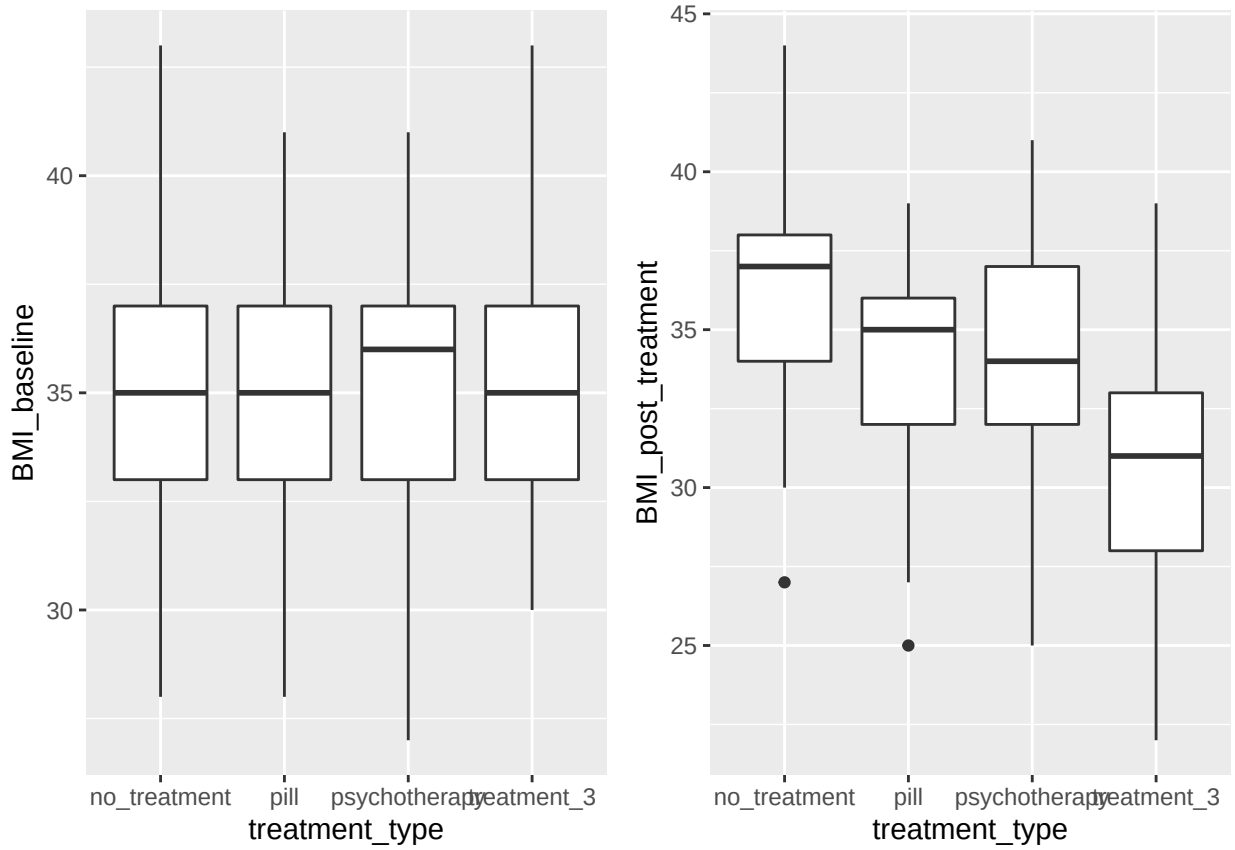
```
fig_1 = data_weightloss %>%
  ggplot() +
  aes(y = BMI_baseline, x = treatment_type) +
  geom_boxplot()
  ylim(c(20, 45))
```

```
## <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))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 5
##   treatment_type mean_pre sd_pre mean_post sd_post
##   <chr>             <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
```

2.2 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(). Vegyuk eszre hogy a funkcion kivul aov() vs. lm() **a ket modell szintaktikailag pontosan ugyan ugy epul 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 regresszios egyutthatok tablazata ebben az esetben maskepp nez ki a megszokotthoz kepest, hiszen majdnem minden kezelesi tipusnak kulon sora van.

2.2.1 Eredmenyek ertelmezese

Az egyes valtozokhoz tartozo regresszios egyutthatokat 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).

```
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 faktorszintet 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 kapott (vagyis a "no_treatment" csoportban volt).

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 **lm() fuggveny mindezt elvegzi** helyettunk, nem kell manualisan dummy valtozokat generalni, de az fontos, hogy megertsuk, hogyan tortenik ez a folyamat. A kategorikus valtozoknak (pl. a mi esetunkben treatment type) nincs nulla erteke. Ezt az R ugy oldja meg, hogy a csoportosito valtozo (faktor) szintjei kozul kivalaszt egyet, ami az alapszint (**default level**), es azt veszi nullanak.

Fontos, hogy ahogy korabban is, az alapszint ha nem rendelkezunk maskepp alapertelmezett modon a faktor szintjei kozul az **abc sorrendben legelso** 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 a "no_treatment"-hez kepest,
- ha valaki "psychotherapy"-t kap -2 BMI valtozast josolunk a "no_treatment"-hez kepest,
- ha valaki "treatment_3"-t kap -5.33 BMI valtozast josolunk a "no_treatment"-hez kepest.

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)/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
         )
```

2.3 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.

Ugy alakitom at a got_pill valtozot, hogy akkor is 1-es erteket vegyen fel, amikor "treatment_3" volt a treatment_type erteke, es a got_psychotherapy valtozot, hogy akkor is 1-es erteket vegyen fel, amikor "treatment_3" volt a treatment_type erteke. Igy a got_pill vatlozo azt jelenti, hogy az illeto kapott-e gyogyszert a kezelese soran, es a got_psychotherapy valtozo azt jelenti, az illeto kapott-e pszichoterapiat a kezelese soran.

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

Most feltehetjuk a kerdest, hogy **van-e interakcio** a gyogyszeres kezeles es a pszichoterapias kezeles kozott.

2.3.1 Mit jelent az, hogy interakcio van valtozok kozott?

A klasszikus linearis regresszios modellekben azt feltetelezzuk, hogy az egyes prediktorok "hatasa" a bejosolt valtozora fuggetlen a tobbi prediktor erteketol. pl. a regresszios modellben amit igy irunk le:

price ~ sqm_living + grade

azt feltetelezzuk, hogy barmilyen erteket is vesz fel a lakas minosege (grade), a lakos meretenek a hatasa (sqm_living) valtozatlan marad.

Ha jo okunk van feltetelezni, hogy ez nem igy van, vagyis hogy **az egyes prediktorok hatasa fugg egy (vagy tobb) masik prediktor erteketol**, akkor **interakciorol** beszelunk, vagyis a prediktorvaltozok interakcioban vannak egymassal, es ez az interakcio (is) befolyasolja a kimeneti valtozora gyakorolt hatast, nem csak a valtozok egymastol fuggetlen hatasa.

A peldankban ez ugy jelenik meg, hogy feltetelezheto, hogy van hozzaadott erteke annak, hogy az emberek a ket kezelest egyszerre kaptak azon felul, amit a ket kezeles hatasa alapjan varnanak kulon-kulon. Masszoval, ugy feltetelezzuk, hogy az, hogy a pszichoterapianak mekkora a hatasa a BMI-re **attol fugg**, hogy az ember kap-e mellette gyogyszeres kezelest is, vagy sem.

2.3.2 Az interakcio beepitese a linearis regresszios modellbe

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

```
mod_3_a = lm(BMI_post_treatment ~ got_pill * got_psychotherapy, data = data_weightloss)
summary(mod_3_a)
```

```
##
## 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_pill                    -2.0833     0.6063  -3.436 0.000697 ***
## got_psychotherapy           -2.0000     0.6063  -3.299 0.001121 **
## got_pill:got_psychotherapy  -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
```

Alternativ szintaxis: Egy masik szintaxissal is felirhatjuk ugyan ezt, ahol "got_pill * got_psychotherapy" helyett **"got_pill + got_psychotherapy + got_pill:got_psychotherapy"** irunk. Ez akkor lehet fontos, ha tobb mint ket valtozo valamilyen komplexebb interakcios mintazatat akarjuk modellezni ahol nem akarjuk minden mindennel valo interakciojat beepiteni a modellbe. Ez pontosan ugyan azt eredmenyezi mint a korabbi szintaxis, hiszen itt minden valtozo interakciojat beepitettuk a modellbe (csak ket prediktor valtozonk volt).

```
mod_3_b = lm(BMI_post_treatment ~ got_pill + got_psychotherapy + got_pill:got_psychotherapy, data = data
summary(mod_3_b)
```

```
##
## 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
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  36.1333     0.4287  84.287  < 2e-16 ***
## got_pill                     -2.0833     0.6063  -3.436 0.000697 ***
## got_psychotherapy            -2.0000     0.6063  -3.299 0.001121 **
## got_pill:got_psychotherapy   -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
```

2.3.3 Eredmenyek ertelmezese

Ahhoz hogy megertsuk, hogyan kell ertelmezni az eredmenyeket, erdemes egy marmadik modjat megnezni annak, hogy hogyan tudjuk ugyan azt a modellt leirni: Az alabbi kodban a ket prediktor valtozo erteket osszeszorzom, es ezt az erteket elmentem egy uj valtozoba. Ezt a szorzat erteket beepitjuk a modellunkbe mint egy uj prediktort.

```
data_weightloss = data_weightloss %>%
  mutate(got_pill_times_got_psychotherapy = got_pill * got_psychotherapy)

mod_3_c = lm(BMI_post_treatment ~ got_pill + got_psychotherapy + got_pill_times_got_psychotherapy, data
summary(mod_3_c)
```

```
##
## Call:
## lm(formula = BMI_post_treatment ~ got_pill + got_psychotherapy +
##     got_pill_times_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_pill                         -2.0833     0.6063  -3.436 0.000697 ***
## got_psychotherapy                -2.0000     0.6063  -3.299 0.001121 **
## got_pill_times_got_psychotherapy -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
```

Vegyuk eszre, hogy ez ugyan azt az eredmenyt adja, mint a korabbi modellek (mod_3_a es mod_3_b). Vagyis 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 hatasa alapjan 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.

Ertelmezd 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 ertemezhetjuk az interakciohoz tartozo regresszios egyutthatot?

2.4 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez

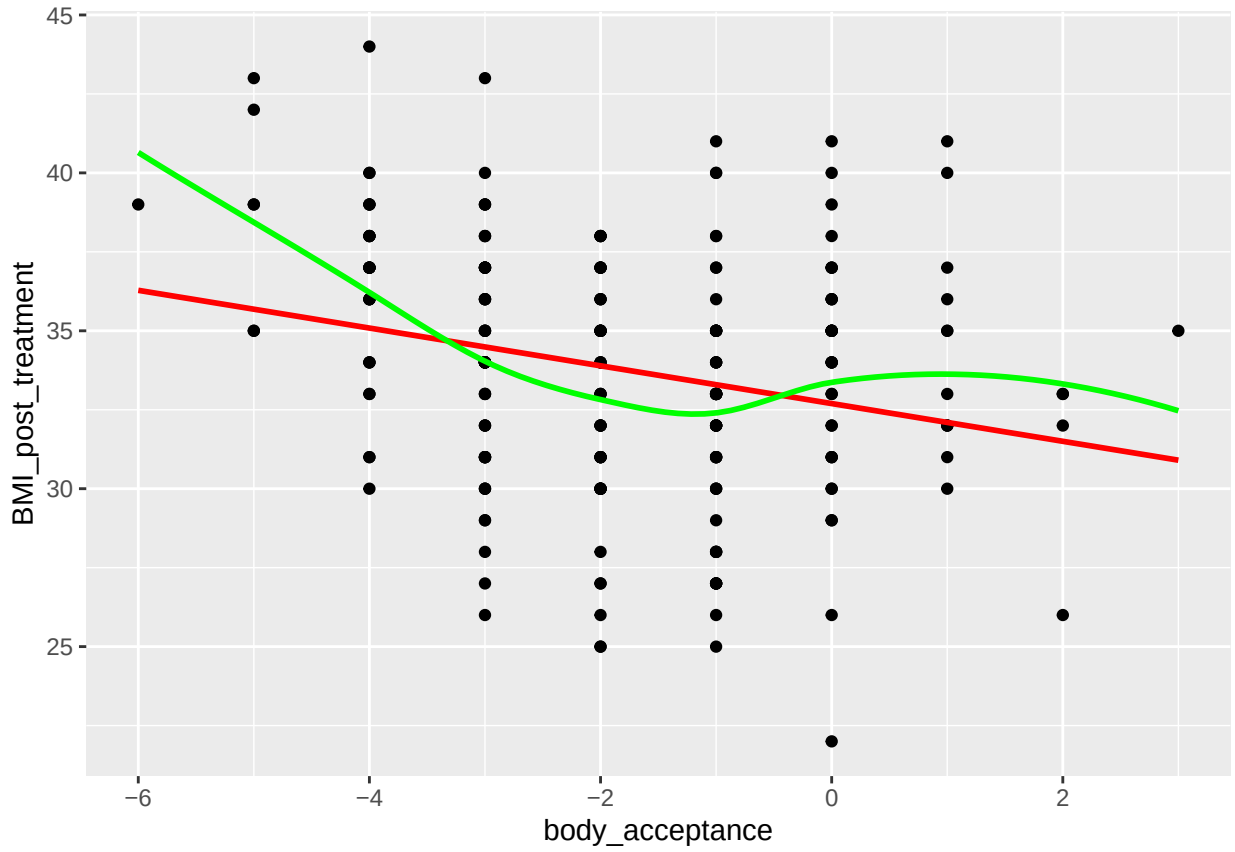
A linearis regresszios modelleket eredetileg linearis osszefuggesek modellezesere talaltak ki, de egy kis matematikai trukkel elerhetjuk, hogy modellezzunk **nem-linearis osszefuggesek** is.

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(method = "lm", se = F, color = "red") +
  geom_smooth(se = F, color = "green")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `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)**.

Ha **osszehasonlitjuk** azt a modellt amiben szerepel ez a hatvanytenyezo azzal a modellel amiben ez nem szerepel (hiearchikus regresszio), azt talaljuk hogy ez az ugynevezett kvadratikus hatas (quadratic effect) 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
```

```
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 hasznalunk mindenkeppen tegyuk be a modellbe a prediktor **minden alacsonyabb hatvanyat is** egeszen az elso hatvanyig (ami maga az eredeti prediktor).

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

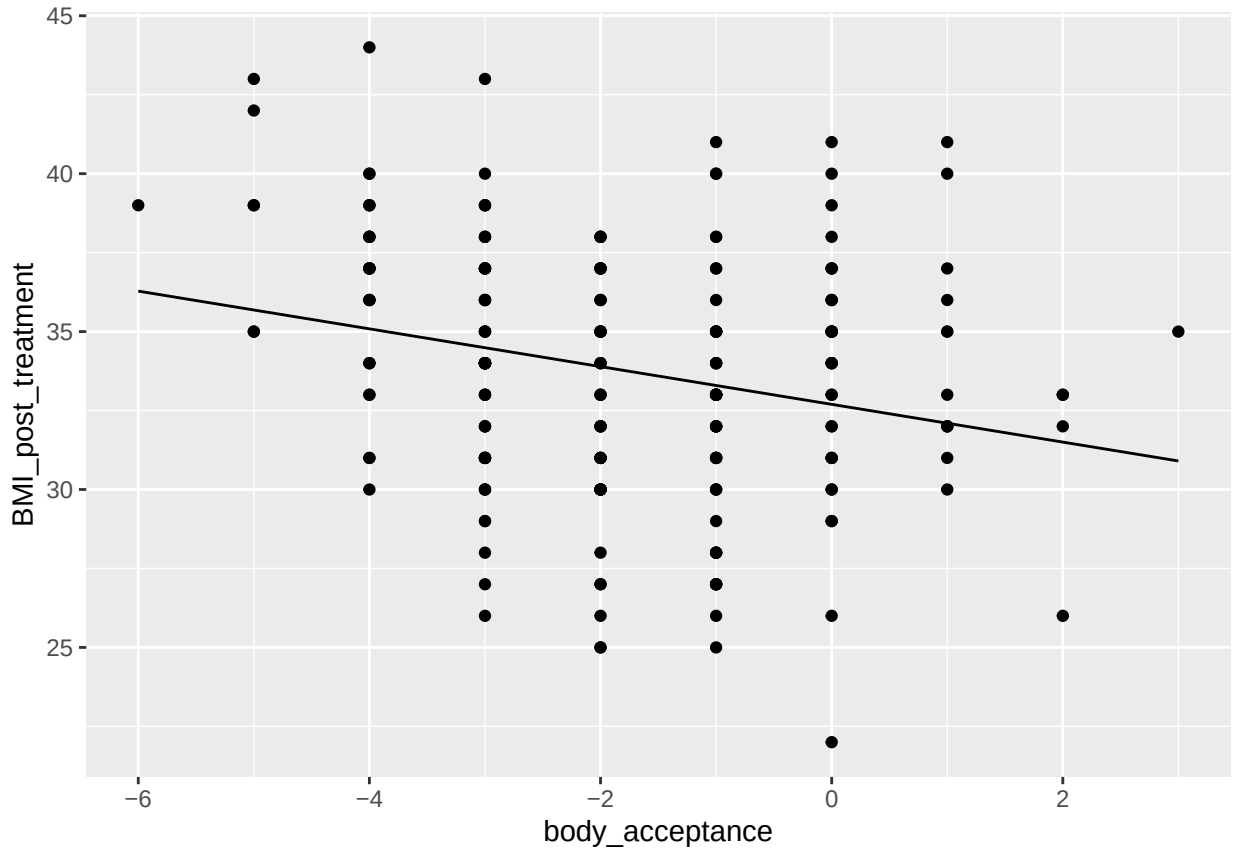
```
AIC(mod_6)
```

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

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

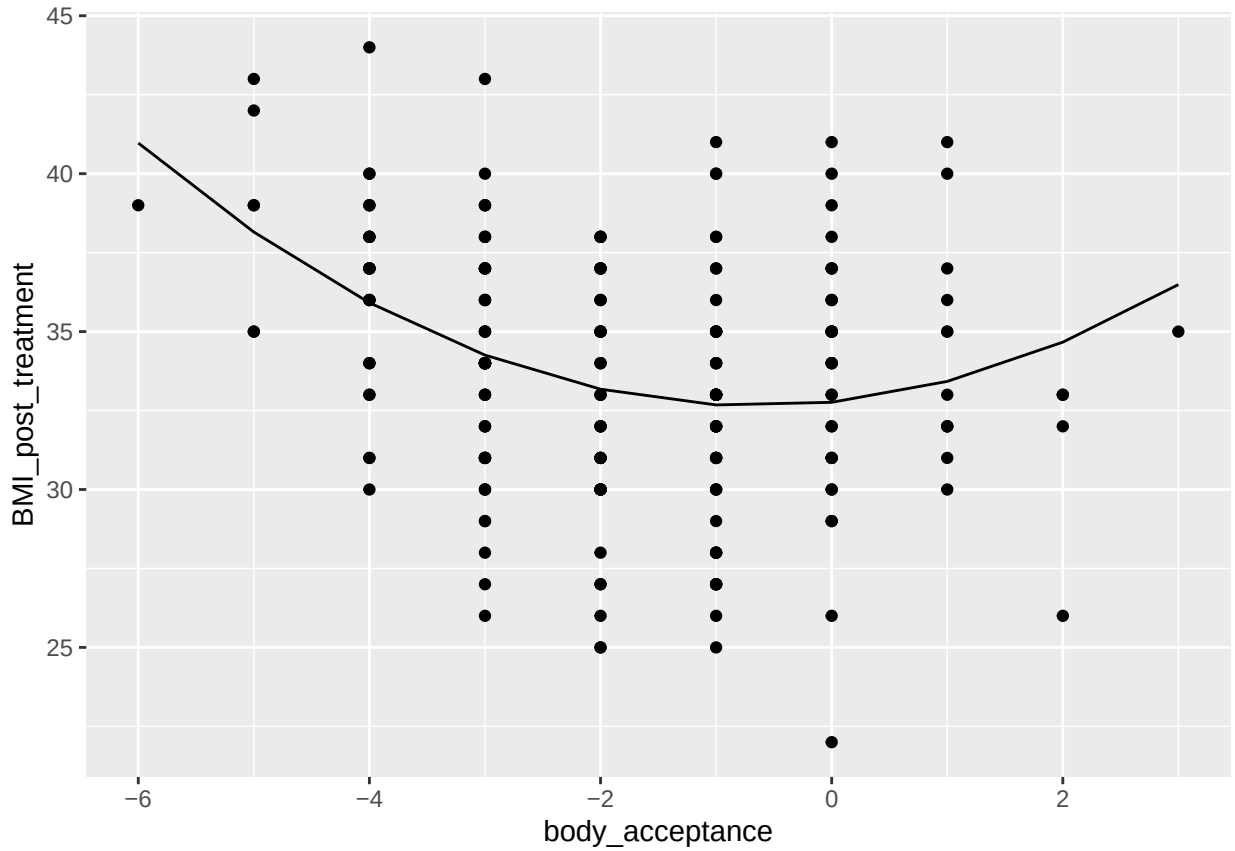
```
data_weightloss = data_weightloss %>%
  mutate(pred_mod_4 = predict(mod_4),
         pred_mod_5 = predict(mod_5),
         pred_mod_6 = predict(mod_6))

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



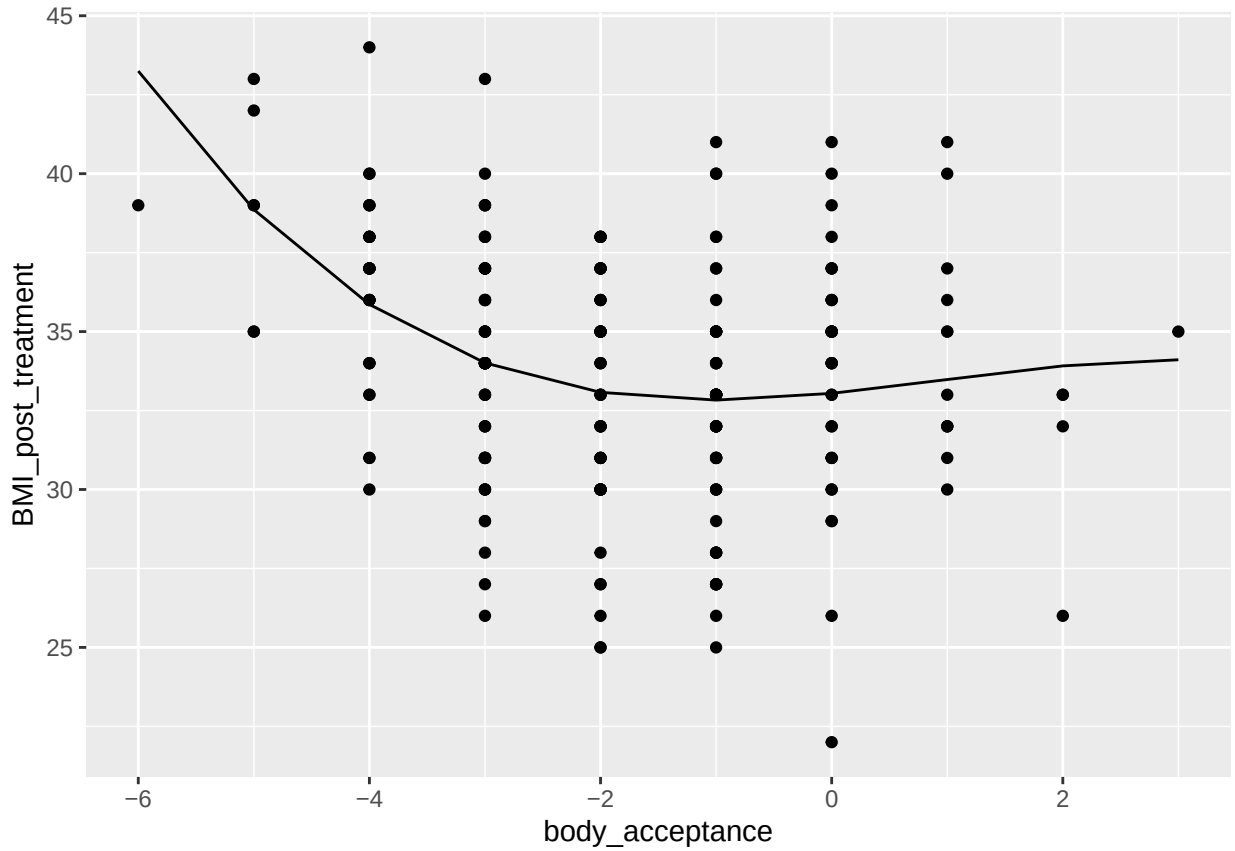
Igy amikor a masodik hatvany szerepel a modellben:

```
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 gorbuleti pontot engedunk** a regresszios egyenesnek. (A fenti abrak alapjan lathato hogy 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*_____

Kiserletezz a lakasarakat tartalmazo adatbazissal. Gondold at, milyen valtozok jatszhatnak szerepet a lakasar meghatarozasaban, es hogy van-e ertelme interakciokat, vagy nem-linearis osszefuggeseket feltetelezni.

Probalj elerni a modellel 52%-nal magasabb adjusted R^2 erteket.

Ha szeretned a teljes adatbazist megszerezni, es szeretned latni, masok milyen modellekkel kiserletezte es milyen eredmenyesek voltak, a Kaggle-on megtalalod az ezzel az adatbazissal foglalkozo modelleket ezen a linken:

https://www.kaggle.com/harlfoxem/housesalesprediction/activity