

S08 Modell összehasonlítás, és speciális prediktorok

Zoltan Kekecs

1 November 2021

Contents

| | | |
|----------|---|----------|
| 0.1 | Absztrakt | 1 |
| 0.2 | Package-ek betöltése | 1 |
| 1 | Modell összehasonlítás és Modellválasztás | 1 |
| 1.1 | Adatmenedzsment és leíró statisztikák | 1 |
| 1.2 | Hierarchikus regresszió | 3 |
| 1.3 | Hierarchikus regresszió két prediktor-blokkal | 3 |
| 1.4 | Hierarchikus regresszió több mint két blokkal | 4 |
| 2 | Speciális prediktorok | 5 |
| 2.1 | Adatmenedzsment és leíró statisztikák | 5 |
| 2.2 | Kategorikus változók mint prediktorok | 8 |
| 2.3 | Két változó interakciójának beillesztése a modellbe | 10 |
| 2.4 | Hatvány prediktorok a nem-lineáris összefüggések modellezéséhez | 13 |

0.1 Absztrakt

Ez a gyakorlat megmutatja majd, hogyan lehet különböző prediktorokat tartalmazó modelleket összehasonlítani egymással. Demonstráljuk majd a hierarchikus regressziót. Nehány modell szelekciós módszerre is kiterünk majd, és megemlítjük a “tulullesztés” (overfitting) fogalmát. Ezen felül megismerjük majd hogyan használjuk és értelmezzük a különböző típusú speciális prediktorokat a lineáris regressziós modellekben.

0.2 Package-ek betöltése

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

1 Modell összehasonlítás és Modellválasztás

1.1 Adatmenedzsment és leíró statisztikák

1.1.1 A King County lakaslétszáma adatbázis betöltése

Ebben a gyakorlatban lakások és házak adatait fogjuk megvizsgálni.

Egy Kaggle-ról származó adatbázist használunk, melyben olyan adatok szerepelnek, melyeket valószínűsíthetően alkalmasak lakások adatainak bejósolására. Az adatbázisban az USA King County-ból származnak az adatok (Seattle és környéke).

Az adatbázisnak csak egy kis részét használjuk ($N = 200$).

```
# data from
# github/kekecs/PSYP13_Data_analysis_class-2018/master/data_house_small_sub.csv.
data_house = read.csv("https://bit.ly/2DpwK0r")
```

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 terület ertekeket negyzetmeterre.

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

```
##           id           date           price           bedrooms
## Min.      :1.600e+07   20140623T000000: 5   Min.      : 153503   Min.      :1.00
## 1st Qu.:1.885e+09   20141107T000000: 5   1st Qu.: 299250   1st Qu.:3.00
## Median :3.521e+09   20150317T000000: 4   Median : 425000   Median :3.00
## Mean    :4.113e+09   20140627T000000: 3   Mean    : 453611   Mean     :2.76
## 3rd Qu.:6.424e+09   20140717T000000: 3   3rd Qu.: 550000   3rd Qu.:3.00
## Max.    :9.819e+09   20140902T000000: 3   Max.    :1770000   Max.     :3.00
##           (Other)           :177
##   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   has basement: 65
## 1st Qu.: -122.3   1st Qu.:1438   1st Qu.: 5000   no basement :135
## Median : -122.2   Median :1715   Median : 7222
## 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)/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)
```

1.2 Hierarchikus regresszio

A hierarchikus regresszióval (Hierarchical regression) meghatározhatjuk, hogy különbözik-e két modell egymástól a bejóslo hatékonyságukat tekintve, és ha igen, **ennyivel javul a bejóslo ero** egy bonyolultabb (több prediktort tartalmazó) modell használatával, ahhoz képest ha egy egyszerűbb (kevesebb prediktort tartalmazó) modellt használunk.

Mivel a hierarchikus regresszió gyakorlatilag két regressziós modell (egy egyszerűbb és egy összetettebb) összehasonlítása, ezért most mi is két regressziós modellt fogunk építeni.

1.3 Hierarchikus regresszio két prediktor-blokkal

1.3.1 Modellelépítés

Elsőször építünk egy egyszerű modellt, amiben a ház vételárát csak a *sqm_living* és a *grade* változók alapján jósoljuk be.

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

Majd építünk egy bonyolultabb modellt, amiben a *sqm_living* és a *grade* prediktorokon kívül szerepelnek még a lakás földrajzi hosszúság és szélesség adatai is (*long* és *lat*).

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

Vegyünk észre, hogy az egyszerűbb modellben szereplő prediktorok egy **reszhalmazat** alkotják a bonyolultabb modell prediktorainak. vagyis **a bonyolultabb modell minden prediktort tartalmaz az egyszerűbb modellből**, plusz még néhány extra prediktort. Ezt úgy nevezzük hogy **“nested models”** vagyis **“egymásba ágyazott modellek”**, hiszen a modellek úgy épülnek fel mint a matrjoska babák.

1.3.2 Modellosszehasonlítás

Az **adj. R Squared** mutató segítségével meghatározhatjuk a két modell által megmagyarázott varianciaarányt. Ezt a model summary kilitázásával is megtehetjük, de a model summary-ból csak ez az információ is kinyerhető a \$adj.r.squared hozzáadásával az alábbi módon:

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

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

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

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

Úgy tűnik, hogy a megmagyarázott varianciaarány magasabb lett azzal, hogy a modellhez hozzátettük a geolokációval kapcsolatos információt.

Most meghatározhatjuk, hogy ez a bejósloeroben bekövetkezett **javulás szignifikáns-e**. Ezt egyrészt a két modell AIC modell-illeszkedési mutatójának összehasonlításával tehetjük meg.

Ha a két **AIC** érték közötti különbség nagyobb mint 2, a két modell illeszkedése szignifikánsan különbözik egymástól. Az alacsonyabb AIC kevesebb hibat és jobb modell illeszkedést jelent. Ha a különbség nem éri el a 2-t, akkor a két modell közül bármelyiket megtarthatjuk. Ilyenkor általában azt a modellt tartjuk meg amelyik elméletileg megalapozottabb, de ha nincs erős elméletünk, akkor az egyszerűbb 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** segítségével összehasonlíthatjuk a két modell residuális hibáját.

Ha az anova() F-tesztje szignifikáns, az azt jeletni, hogy a két modell reziduális hibája szignifikánsan különbözik egymástól.

```
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** mutató alapján való modell-összehasonlítás **jobbán elfogadott** a szakirodalomban, ezért ha az AIC és az anova összehasonlítás különbözõ eredményre vezet, akkor az AIC eredményt érdemes használni.

Fontos, hogy az **anova összehasonlításnak az eredménye csak akkor valid, ha egymásba ágyazott (nested)** modellek összehasonlítására használjuk őket, vagyis az egyik modell prediktorai a másik modell prediktorainak részhalmaza alkotják.

Az AIC legtöbbször alkalmas nem beágyazott modellek összehasonlítására is, (bar ezzel kapcsolatban nem teljes az egyetemes a szakirodalomban, a dolgozatokban elfogadott AIC-ot használni nem beágyazott modellek összehasonlítására).

1.4 Hierarchikus regresszio több mint két blokkal

A fenti folyamat ugyan úgy megismételhető ha több mint két blokkban adjuk hozzá a prediktorokat a modellhez.

Itt egy harmadik modellt építünk, a “condition” prediktor hozzáadásával.

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

A három modellt következõképpen hasonlíthatjuk össze:

```
# 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 eredmények alapján javult a bejoslo ereje a modellunknek a lakas allapotanak (condition) figyelembevetelevel?

Gyakorlas

Tedd hozza az iment epített 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 alatasztott** es/vagy korabbi kutatasi eredmények 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 kutatias adatbazis betoltese

Az adatbazis egy olyan kutatias szimulalt adatait tartalmazzam ahol kulonbozo kezelesek hatekonysagat tesztelték a sulyvesztesre tulsulyos személyekkel.

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 személy kapott (no treatment - nem kapott kezelest; pill - etvagycsokkentó gyógyszer; psychotherapy - kognitiv behavior terapia (CBT); treatment 3 - egy harmadik fajta kezeles, lasd lentebb)

- motivation - onbevallásos motivációs szint 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 személy mennyire erzi elegetettnek magat jelenleg testevel (-7 - +7, ahol a -7 nagyon elegetetlen, a +7 nagyon elegetett)

```
data_weightloss = read.csv("https://raw.githubusercontent.com/kekecsz/PSZB17-210-Data-analysis-seminar/1")
```

2.1.2 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
```

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

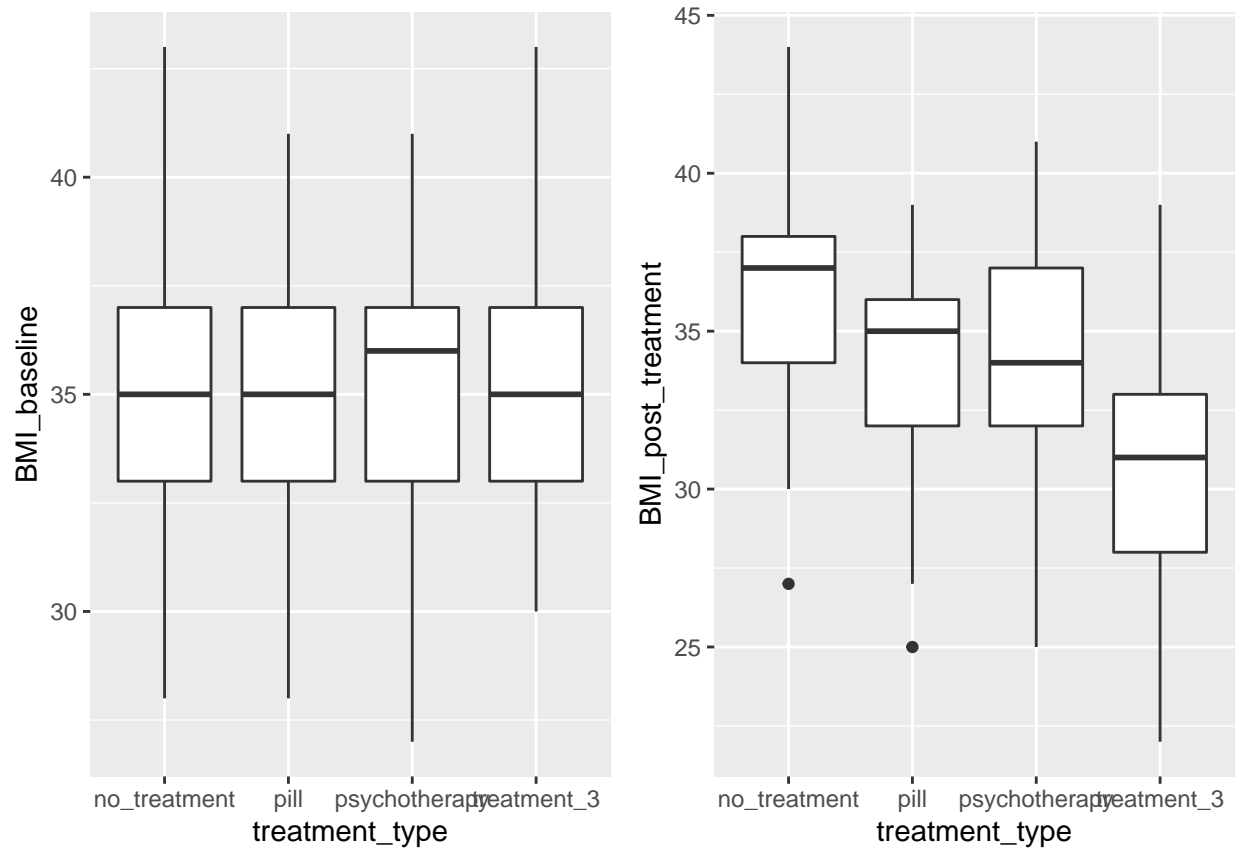
```
fig_1 = data_weightloss %>% ggplot() + aes(y = BMI_baseline,
      x = treatment_type) + geom_boxplot()
ylim(c(20, 45))
```

```
## <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
```

2.2 Kategorikus változók mint prediktorok

Mivel úgy tűnik, a csoportok összehasonlíthatók 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 megepíthetjük 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 funkcion 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_typedtreatment_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.

2.2.1 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.**

Visszont 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 eredményet látjuk a regressziós együtthatók táblázatában.

Az **intercept**-hez tartozó regressziós együtthatót mindig úgy lehet értelmezni, hogy ez mutatja a bejósolt 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 **egyutthatokat így mar szokas szerint értelmezhetjük**, hogy abban az esetben ha az adott dummy változó értéke egy szinttel no (vagyis 0 helyett 1 lesz), akkor mekkora változást varhatunk a bejosolt változó értékeiben.

Az **lm()** **fuggvény mindezt elvégzi** helyettünk, nem kell manualisan dummy változókat generalni, de az fontos, hogy megertsek, 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 varhatunk,
- 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 adatbázist 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 egyutthatok táblázatát. - Mit jelent az intercept regressziós egyutthatója? - Mit jelent a has_basement prediktorhoz tartozó regressziós egyuttható?

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

```
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)
```

2.3 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 átalakítjuk az adatbázist, hogy ezt helyesen tükrözzék az imént generált dummy változók.

Ugy alakítom át a got_pill változót, hogy akkor is 1-es értéket vegyen fel, amikor “treatment_3” volt a treatment_type értéke, és a got_psychotherapy változót, hogy akkor is 1-es értéket vegyen fel, amikor “treatment_3” volt a treatment_type értéke. Így a got_pill változó azt jelenti, hogy az illető kapott-e gyógyszert a kezelése során, és a got_psychotherapy változó azt jelenti, az illető kapott-e pszichoterápiát a kezelése során.

```
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 feltehetjük a kérdést, hogy **van-e interakció** a gyógyszeres kezelés és a pszichoterapiás kezelés között.

2.3.1 Mit jelent az, hogy interakció van változók között?

A klasszikus lineáris regressziós modellekben azt feltételezzük, hogy az egyes prediktorok “hatása” a bejósolt változóra független a többi prediktor értékeitől. pl. a regressziós modellben amit így írunk le:

$\text{price} \sim \text{sqm_living} + \text{grade}$

azt feltételezzük, hogy bármilyen értéket is vesz fel a lakás minősége (grade), a lakás méretének a hatása (sqm_living) változatlan marad.

Ha jó okunk van feltételezni, hogy ez nem így van, vagyis hogy **az egyes prediktorok hatása függ egy (vagy több) másik prediktor értékeitől**, akkor **interakcióról** beszélünk, vagyis a prediktorváltozók interakcióban vannak egymással, és ez az interakció (is) befolyásolja a kimeneti változóra gyakorolt hatást, nem csak a változók egymástól független hatása.

A példánkban ez úgy jelenik meg, hogy feltételezhető, hogy van hozzáadott értéke annak, hogy az emberek a két kezelést egyszerre kaptak azon felül, amit a két kezelés hatása alapján várnának külön-külön. Masszóval, úgy feltételezzük, hogy az, hogy a pszichoterápiának mekkora a hatása a BMI-re **attól függ**, hogy az ember kap-e mellette gyógyszeres kezelést is, vagy sem.

2.3.2 Az interakció beépítése a lineáris regressziós modellbe

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

```
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)      38.9667     2.1435  18.179  <2e-16 ***
## got_pill         -0.8333     1.3557  -0.615    0.539
## got_psychotherapy -0.7500     1.3557  -0.553    0.581
## got_pill:got_psychotherapy -1.2500     0.8574  -1.458    0.146
## ---
## 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 szintaxissal 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_b = lm(BMI_post_treatment ~ got_pill + got_psychotherapy +
  got_pill:got_psychotherapy, data = data_weightloss)
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)      38.9667     2.1435  18.179  <2e-16 ***
## got_pill         -0.8333     1.3557   -0.615    0.539
## got_psychotherapy -0.7500     1.3557   -0.553    0.581
## got_pill:got_psychotherapy -1.2500     0.8574   -1.458    0.146
## ---
## 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 Eredmények értelmezése

Ahhoz hogy megertsek, hogyan kell értelmezni az eredményeket, érdemes egy marmadik modjat megnezni annak, hogy hogyan tudjuk ugyan azt a modellt leirni: Az alabbi kodban a ket prediktor valtozo ertekeket osszeszorozom, es ezt az ertekeket elmentem egy uj valtozoba. Ezt a szorzat ertekeket 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 = data_weightloss)
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)      38.9667     2.1435  18.179  <2e-16 ***
## got_pill         -0.8333     1.3557   -0.615    0.539
## got_psychotherapy -0.7500     1.3557   -0.553    0.581
## got_pill_times_got_psychotherapy -1.2500     0.8574   -1.458    0.146

```

```
## ---
## 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 értelmezhetjuk, hogy abban az esetben, **ha a ket valtozo szorzata egyel magasabb** ertekeket 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 alapján 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 alapján. 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 ertelmezhetjuk az interakciohoz tartozo regresszios egyutthatot?
-

2.4 Hatvany prediktorok a nem-linearis osszefuggesek modellezesehez

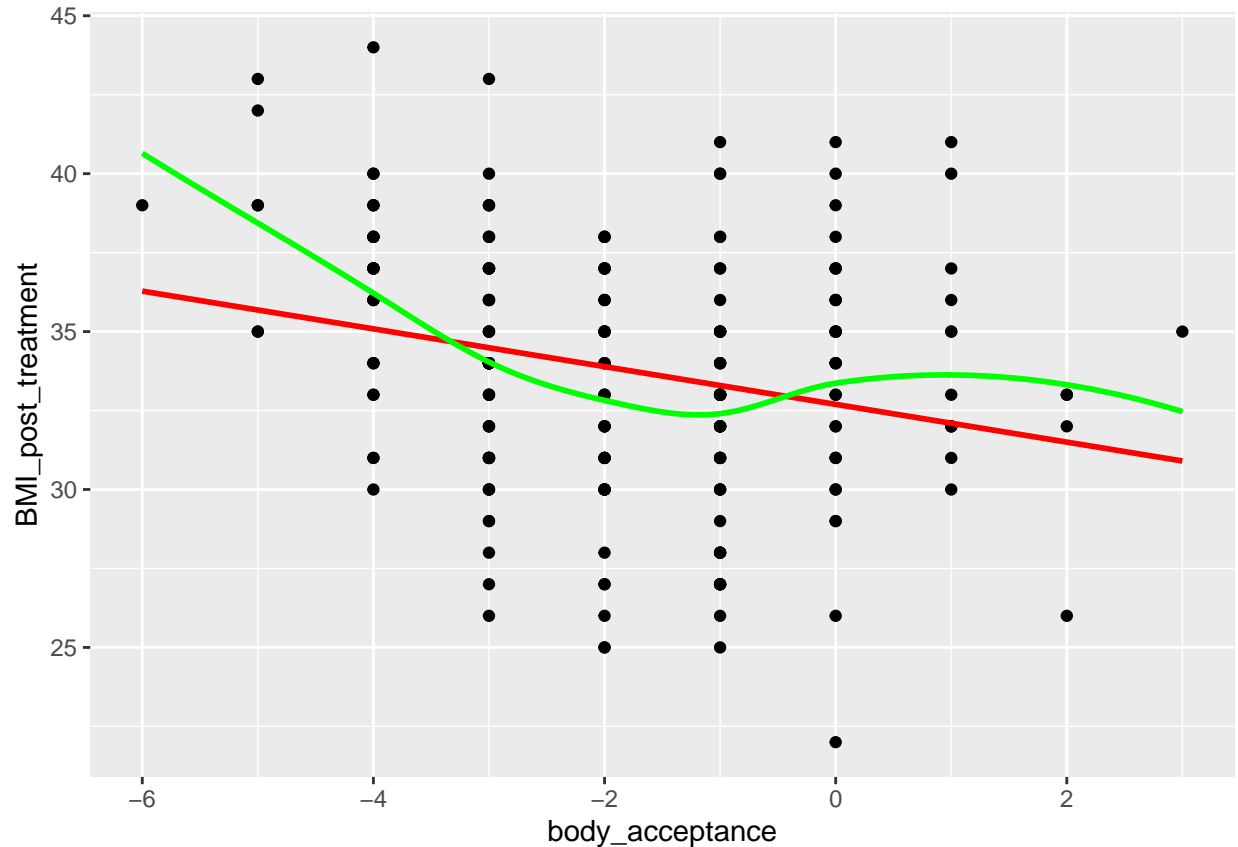
A linearis regresszios modelleket eredetileg linearis osszefuggesek modellezesere talaltak ki, de egy kis matematikai trukkel elerhetjuk, hogy modellezzunk **nem-linearis osszefuggesek** is.

Az alabbi abra alapján 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 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 hasznalunk 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
```

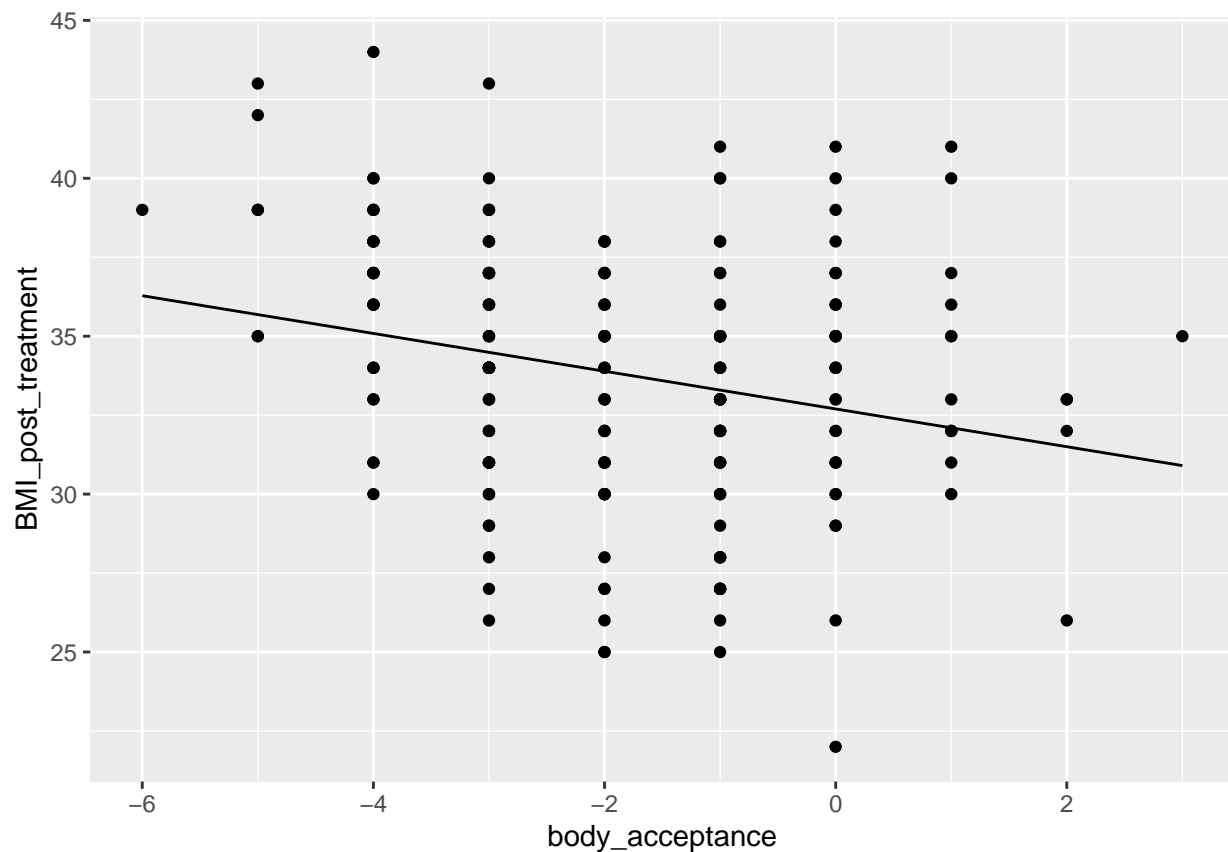
```
AIC(mod_6)
```

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

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

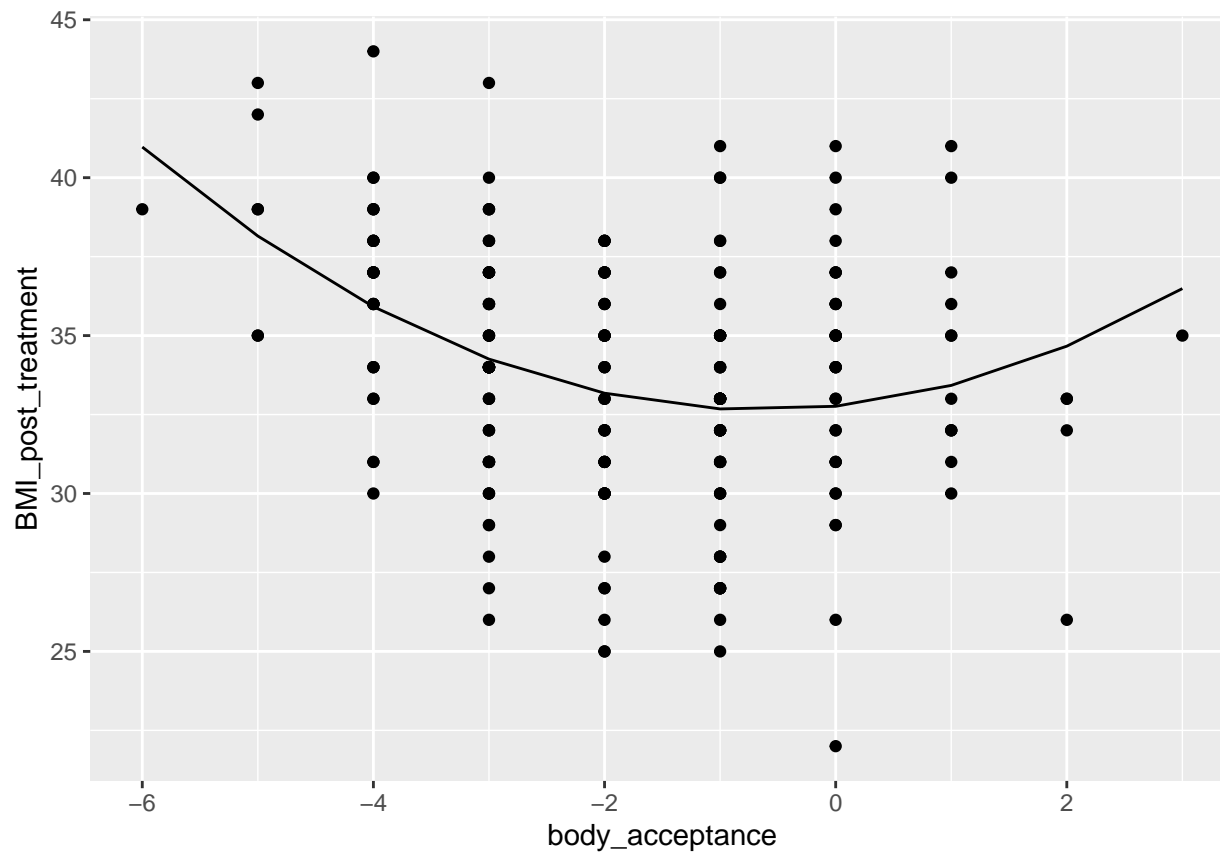
```
data_weightloss = data_weightloss %>% mutate(pred_mod_4 = predict(mod_4),
  pred_mod_5 = predict(mod_5), pred_mod_6 = predict(mod_6))

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



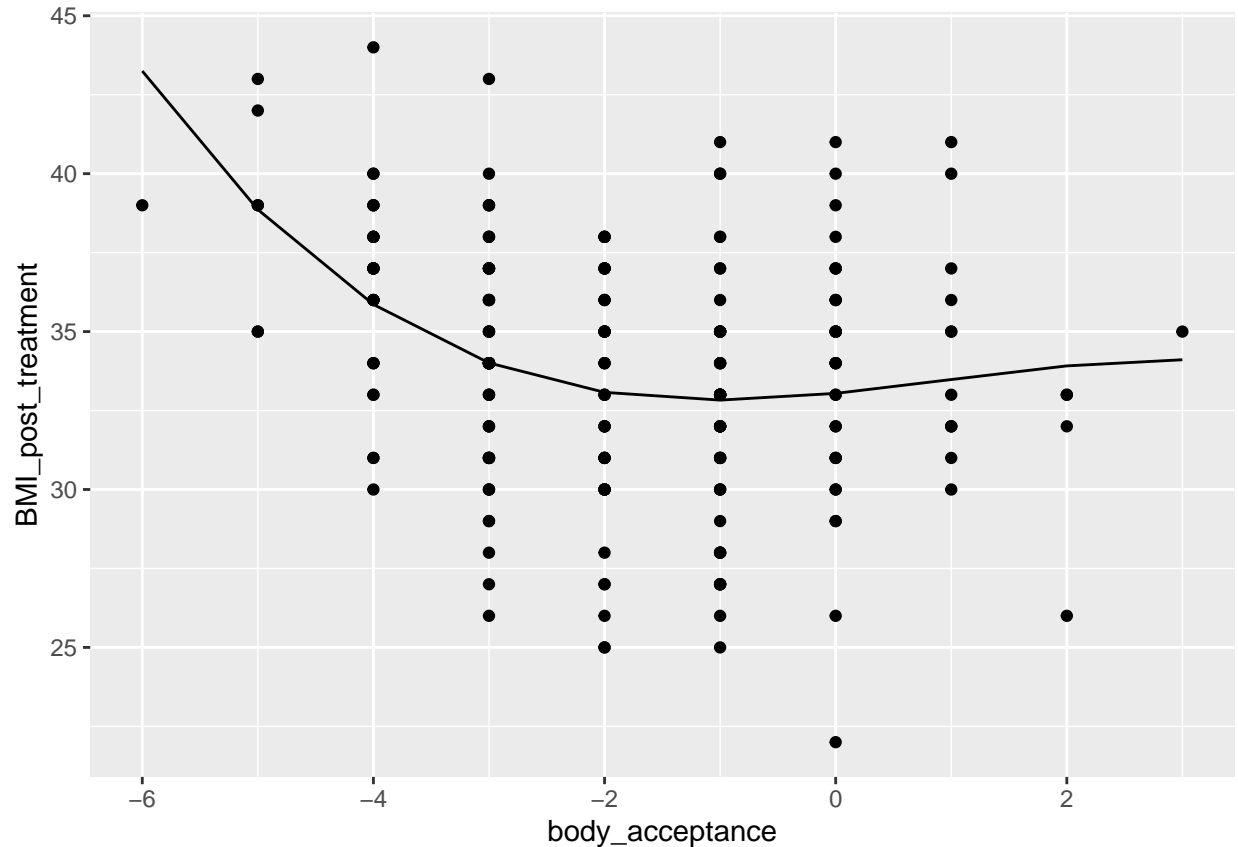
Igy amikor a második hatvány szerepel a modellben:

```
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 gorbuleti pontot engedunk** a regressziós egyenesnek. (A fenti ábrák alapján lathato hogy mindig egyel kevesebb gorbuleti (inflexios) pontot engedunk mint ahanyadik hatvanyt beletettuk a modellbe prediktorkent.)

Azonban a túl nagy flexibilitás nem célravezető, mert minél flexibilisebb a modell, annál inkább hajlamos arra, hogy a saját mintához illeszkedjen, és nem a populációban megtalálható összefüggéseket ragadja meg.

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

Gyakorlas

Kísérletezz a lakasarakat tartalmazó adatbázissal. Gondold át, milyen változók játszhatnak szerepet a lakasar meghatározásában, és hogy van-e értelme interakciókat, vagy nem-lineáris összefüggéseket feltételezni.

Próbálj elérni a modellel 52%-nál magasabb adjusted R^2 értéket.

Ha szeretnéd a teljes adatbázist megszerezni, és szeretnéd látni, mások milyen modellekkel kísérleteztek és milyen eredményesek voltak, a Kaggle-on megtalálod az ezzel az adatbázissal foglalkozó modelleket ezen a linken:

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