Exercise 12 - Multiple regression

Zoltan Kekecs

20 november 2018

Contents

1 Tobbszoros regresszio

1.1 Abstract

Ennek a gyakorlatnak az a celja hogy az egyszeru regresszorol szerzett tudast altalanositsuk olyan esetekre, ahol tobb prediktor (bejoslo valtozo) is szerepel a modellben.

Ennek a dokumentumnak a legfrissebb valtozatat megtalalod itt: https://osf.io/e23by/

1.2 Package-ek betoltese

A kovetkezo package-ek betoltesere lesz szukseg:

```
library(psych)  # for describe
library(lm.beta)  # for lm.beta
library(car)  # for scatter3d
library(ggplot2)  # for ggplot
library(rgl)  # for scatter3d
library(tidyverse)  # for tidy format
library(gridExtra)  # for grid.arrange
```

1.3 Az adatfajl betoltese: Lakasarak adattabla

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")
```

1.4 Adatellenoryes

Mindig ellenorizd az adatok strukturajat es integritasat.

Eloszor at valtjuk a USA dollar-t millio forint mertekegysegre, es a negyzetlab adatokat negyzetmeterre.

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

```
##        id                 date                          price
##  Min.   :1.600e+07   Min.   :2014-05-06 00:00:00   Min.   : 153503
##  1st Qu.:1.885e+09   1st Qu.:2014-07-22 18:00:00   1st Qu.: 299250
##  Median :3.521e+09   Median :2014-10-29 12:00:00   Median : 425000
##  Mean   :4.113e+09   Mean   :2014-11-08 10:19:12   Mean   : 453611
##  3rd Qu.:6.424e+09   3rd Qu.:2015-02-28 00:00:00   3rd Qu.: 550000
##  Max.   :9.819e+09   Max.   :2015-05-12 00:00:00   Max.   :1770000
##     bedrooms      bathrooms     sqft_living      sqft_lot
##  Min.   :1.00   Min.   :0.75   Min.   : 590   Min.   :   914
##  1st Qu.:3.00   1st Qu.:1.00   1st Qu.:1240   1st Qu.:  4709
##  Median :3.00   Median :1.75   Median :1620   Median :  7270
##  Mean   :2.76   Mean   :1.85   Mean   :1728   Mean   : 12985
##  3rd Qu.:3.00   3rd Qu.:2.50   3rd Qu.:1985   3rd Qu.: 10187
##  Max.   :3.00   Max.   :3.50   Max.   :4380   Max.   :217800
##      floors        waterfront         view         condition
```

```
##  Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :3.00
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:3.00
##  Median :1.000   Median :0.000   Median :0.000   Median :3.00
##  Mean   :1.472   Mean   :0.005   Mean   :0.145   Mean   :3.42
##  3rd Qu.:2.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:4.00
##  Max.   :3.000   Max.   :1.000   Max.   :4.000   Max.   :5.00
##      grade         sqft_above   sqft_basement       yr_built
##  Min.   : 5.00   Min.   : 590   Min.   :   0.0   Min.   :1900
##  1st Qu.: 7.00   1st Qu.:1090   1st Qu.:   0.0   1st Qu.:1946
##  Median : 7.00   Median :1375   Median :   0.0   Median :1968
##  Mean   : 7.36   Mean   :1544   Mean   : 184.1   Mean   :1968
##  3rd Qu.: 8.00   3rd Qu.:1862   3rd Qu.: 315.0   3rd Qu.:1993
##  Max.   :11.00   Max.   :4190   Max.   :1600.0   Max.   :2015
##   yr_renovated        zipcode           lat             long
##  Min.   :   0.00   Min.   :98001   Min.   :47.18   Min.   :-122.5
##  1st Qu.:   0.00   1st Qu.:98033   1st Qu.:47.49   1st Qu.:-122.3
##  Median :   0.00   Median :98065   Median :47.58   Median :-122.2
##  Mean   :  79.98   Mean   :98078   Mean   :47.57   Mean   :-122.2
##  3rd Qu.:   0.00   3rd Qu.:98117   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :2014.00   Max.   :98199   Max.   :47.78   Max.   :-121.7
##  sqft_living15    sqft_lot15     has_basement
##  Min.   : 740   Min.   :   914   Length:200
##  1st Qu.:1438   1st Qu.:  5000   Class :character
##  Median :1715   Median :  7222   Mode  :character
##  Mean   :1793   Mean   : 11225
##  3rd Qu.:2072   3rd Qu.: 10028
##  Max.   :3650   Max.   :208652
```

```
data_house = data_house %>% mutate(price_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, )
```

Egyszeru leiro statisztikak es abrak.

Kezdetben a lakasok arat a **sqm_living** (a lakas lakoreszenek alapterulete negyzetmeterben), es a **grade** (a lakas altalanos minositese a King County grading system szerint, ami a lakas minoseget, poziciojat, a haz minoseget stb. is tartalmazza) prediktorok felhasznalasaval josoljuk majd be. Kesobb a **has_basement** (tartozik-e a lakashoz pince) valtozot is hasznaljuk majd. Szoval fokuszaljuk ezekre az adatellenorzes soran.

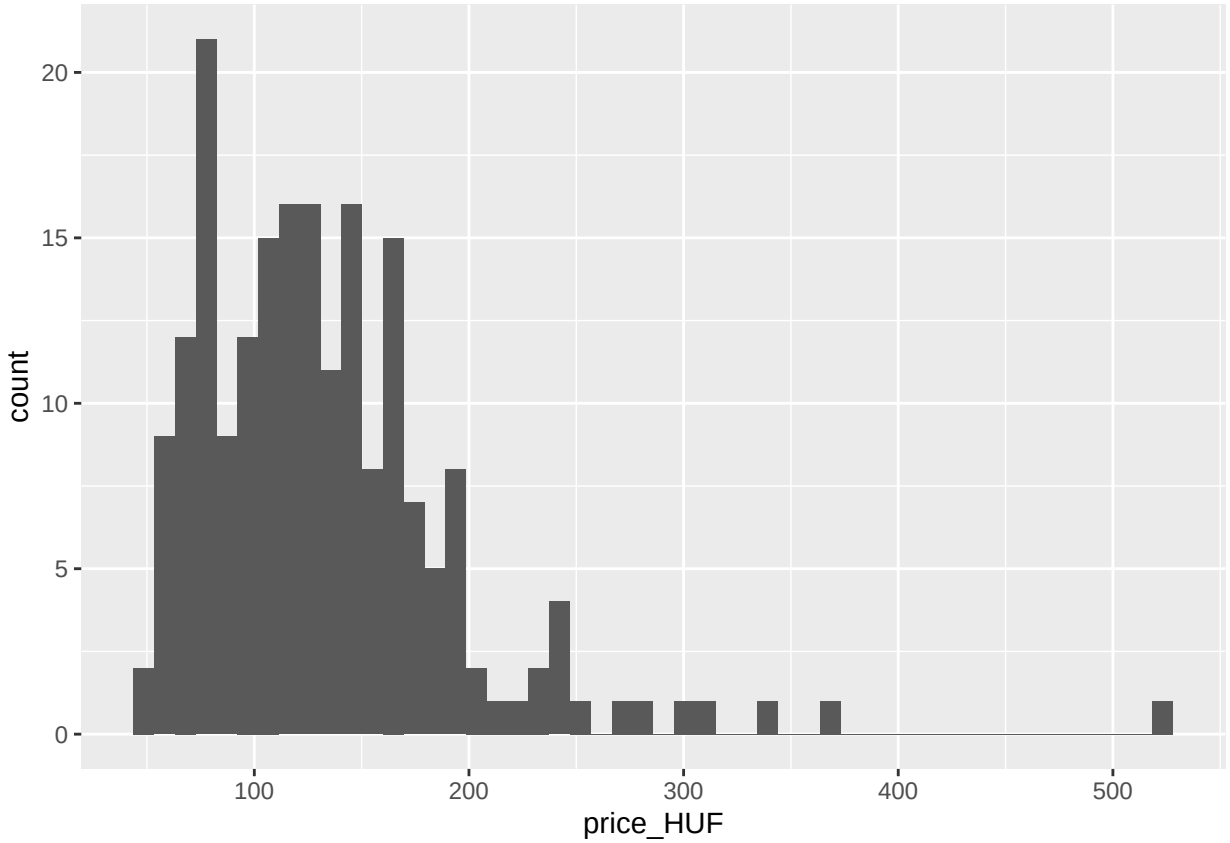
```
# leiro statiszikaka
describe(data_house)
```

```
## Warning in describe(data_house): NAs introduced by coercion

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning
## Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning
## Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning
## -Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning
## -Inf
```

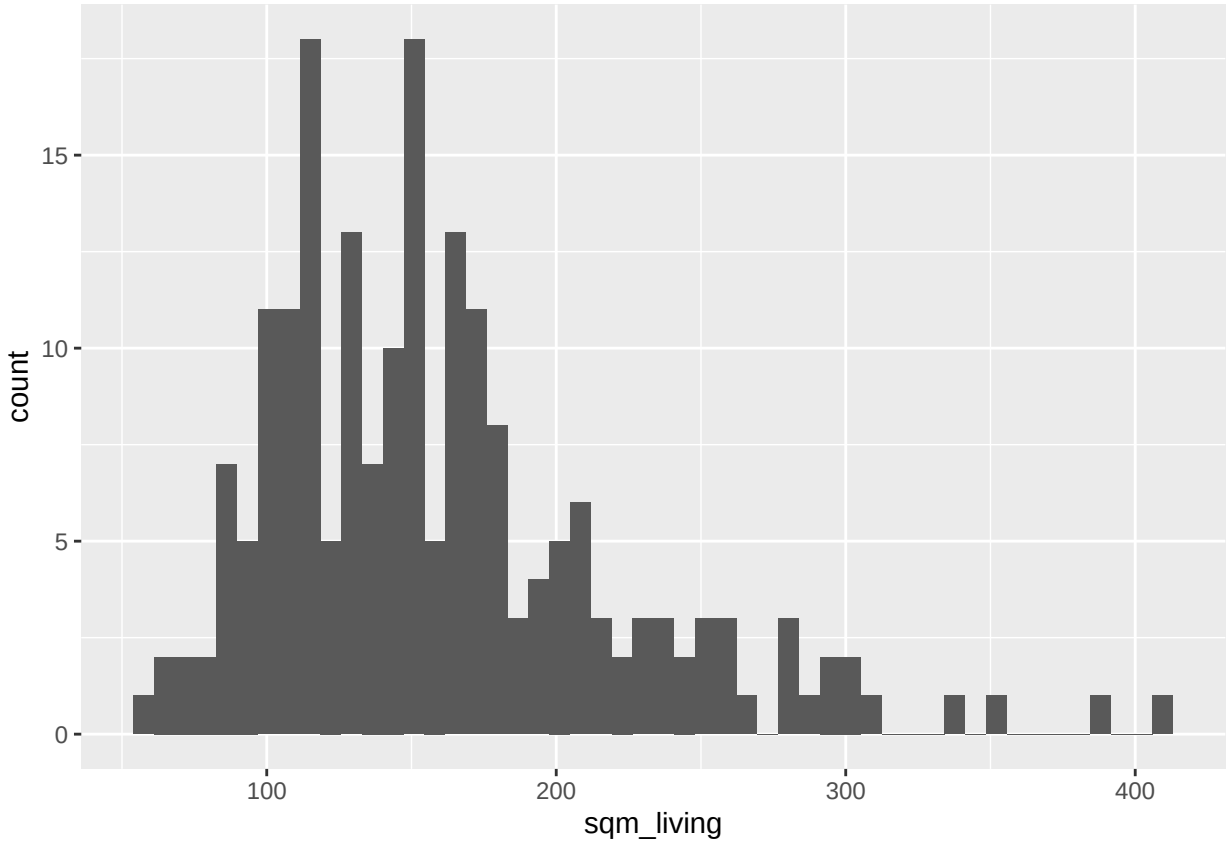
```
##                vars   n          mean           sd        median
## id                1 200 4112747619.38 2.746825e+09 3520875095.00
## date              2 200           NaN           NA           NA
## price             3 200     453610.89 2.111943e+05     425000.00
## bedrooms          4 200          2.76 4.500000e-01          3.00
## bathrooms         5 200          1.85 6.600000e-01          1.75
## sqft_living       6 200       1727.61 6.629200e+02       1620.00
## sqft_lot          7 200      12985.36 2.773609e+04       7270.00
## floors            8 200          1.47 5.500000e-01          1.00
## waterfront        9 200          0.00 7.000000e-02          0.00
## view             10 200          0.14 6.000000e-01          0.00
## condition        11 200          3.42 6.200000e-01          3.00
## grade            12 200          7.36 1.020000e+00          7.00
## sqft_above       13 200       1543.51 6.298700e+02       1375.00
## sqft_basement    14 200        184.10 3.250700e+02          0.00
## yr_built         15 200       1967.64 2.956000e+01       1968.50
## yr_renovated     16 200         79.98 3.928100e+02          0.00
## zipcode          17 200      98077.98 5.407000e+01      98065.00
## lat              18 200         47.57 1.400000e-01         47.58
## long             19 200       -122.20 1.700000e-01       -122.25
## sqft_living15    20 200       1793.34 5.127800e+02       1715.00
## sqft_lot15       21 200      11225.47 1.966363e+04       7222.00
## has_basement*    22 200           NaN           NA           NA
## price_HUF        23 200        133.26 6.204000e+01        124.85
## sqm_living       24 200        160.50 6.159000e+01        150.50
## sqm_lot          25 200       1206.38 2.576770e+03        675.41
## sqm_above        26 200        143.40 5.852000e+01        127.74
## sqm_basement     27 200         17.10 3.020000e+01          0.00
## sqm_living15     28 200        166.61 4.764000e+01        159.33
## sqm_lot15        29 200       1042.88 1.826810e+03        670.95
##                    trimmed          mad         min           max
## id           3956631056.34 2.981805e+09 16000200.00 9818700320.00
## date                  NaN           NA         Inf          -Inf
## price          427743.09 1.853250e+05   153503.00    1770000.00
## bedrooms            2.84 0.000000e+00        1.00          3.00
## bathrooms           1.83 1.110000e+00        0.75          3.50
## sqft_living      1650.86 5.633900e+02      590.00       4380.00
## sqft_lot         7728.61 3.977820e+03      914.00     217800.00
## floors              1.42 0.000000e+00        1.00          3.00
## waterfront          0.00 0.000000e+00        0.00          1.00
## view                0.00 0.000000e+00        0.00          4.00
## condition           3.31 0.000000e+00        3.00          5.00
## grade               7.29 1.480000e+00        5.00         11.00
## sqft_above       1464.11 5.115000e+02      590.00       4190.00
## sqft_basement     110.75 0.000000e+00        0.00       1600.00
## yr_built         1969.17 3.484000e+01     1900.00       2015.00
## yr_renovated        0.00 0.000000e+00        0.00       2014.00
## zipcode         98074.58 6.227000e+01    98001.00      98199.00
## lat                47.58 1.500000e-01       47.18         47.78
## long             -122.22 1.600000e-01     -122.45       -121.73
## sqft_living15    1742.61 4.596100e+02      740.00       3650.00
## sqft_lot15       7559.91 3.624960e+03      914.00     208652.00
## has_basement*        NaN           NA         Inf          -Inf
## price_HUF         125.66 5.444000e+01       45.09        519.97
```

```
## sqm_living           153.37 5.234000e+01       54.81        406.92
## sqm_lot              718.01 3.695500e+02       84.91      20234.28
## sqm_above            136.02 4.752000e+01       54.81        389.26
## sqm_basement          10.29 0.000000e+00        0.00        148.64
## sqm_living15         161.89 4.270000e+01       68.75        339.10
## sqm_lot15            702.34 3.367700e+02       84.91      19384.41
##                      range  skew kurtosis           se
## id            9.802700e+09  0.45    -1.04 194229829.92
## date                  -Inf    NA       NA           NA
## price         1.616497e+06  2.02     7.84     14933.69
## bedrooms      2.000000e+00 -1.53     1.15         0.03
## bathrooms     2.750000e+00  0.12    -0.88         0.05
## sqft_living   3.790000e+03  1.20     1.78        46.88
## sqft_lot      2.168860e+05  6.16    40.75      1961.24
## floors        2.000000e+00  0.74    -0.31         0.04
## waterfront    1.000000e+00 13.93   193.03         0.00
## view          4.000000e+00  4.27    17.86         0.04
## condition     2.000000e+00  1.18     0.28         0.04
## grade         6.000000e+00  0.62     1.00         0.07
## sqft_above    3.600000e+03  1.29     1.90        44.54
## sqft_basement 1.600000e+03  1.91     3.43        22.99
## yr_built      1.150000e+02 -0.32    -0.96         2.09
## yr_renovated  2.014000e+03  4.66    19.81        27.78
## zipcode       1.980000e+02  0.42    -0.85         3.82
## lat           6.000000e-01 -0.56    -0.66         0.01
## long          7.200000e-01  0.79    -0.22         0.01
## sqft_living15 2.910000e+03  0.94     0.88        36.26
## sqft_lot15    2.077380e+05  6.61    54.45      1390.43
## has_basement*         -Inf    NA       NA           NA
## price_HUF     4.748800e+02  2.02     7.84         4.39
## sqm_living    3.521000e+02  1.20     1.78         4.35
## sqm_lot       2.014937e+04  6.16    40.75       182.20
## sqm_above     3.344500e+02  1.29     1.90         4.14
## sqm_basement  1.486400e+02  1.91     3.43         2.14
## sqm_living15  2.703500e+02  0.94     0.88         3.37
## sqm_lot15     1.929949e+04  6.61    54.45       129.18
```

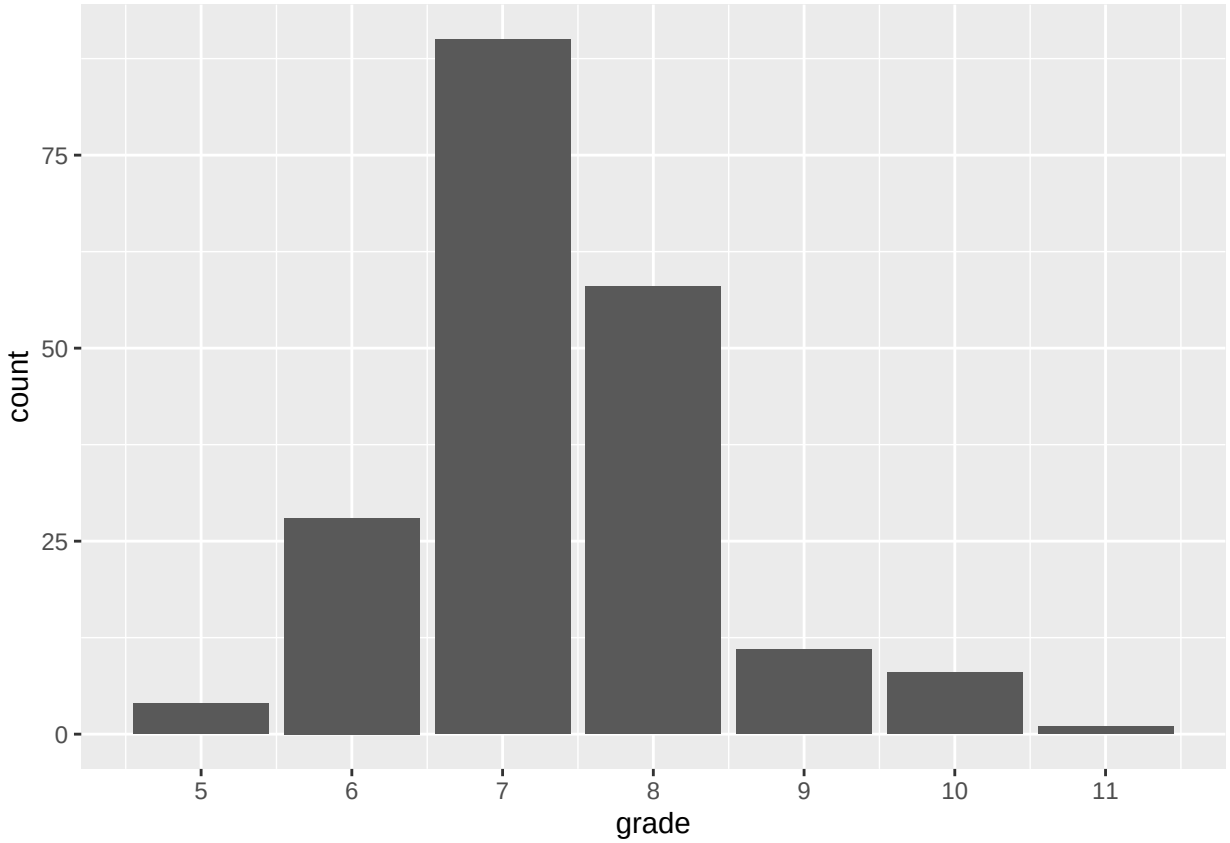
```
# hisztogramok
data_house %>% ggplot() + aes(x = price_HUF) + geom_histogram(bins = 50)
```



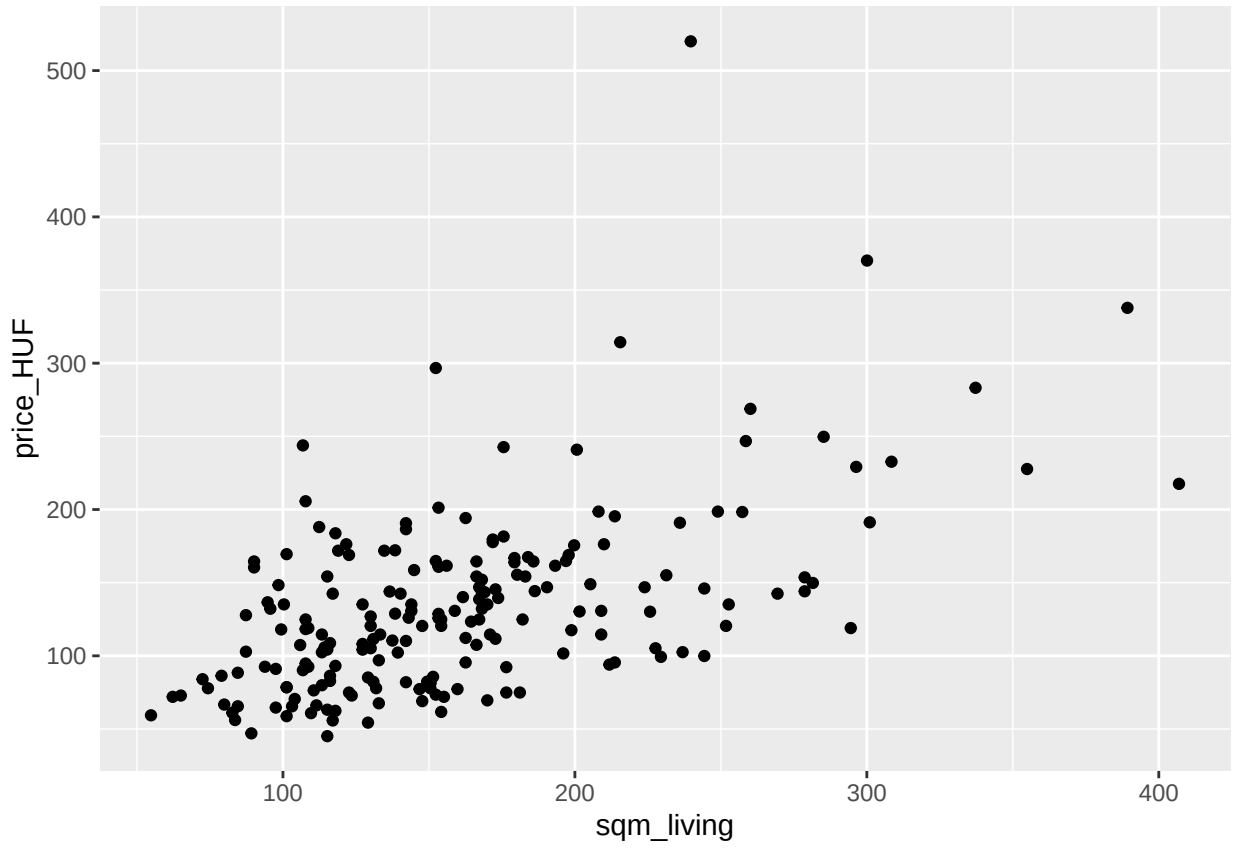

```
data_house %>% ggplot() + aes(x = sqm_living) + geom_histogram(bins = 50)
```



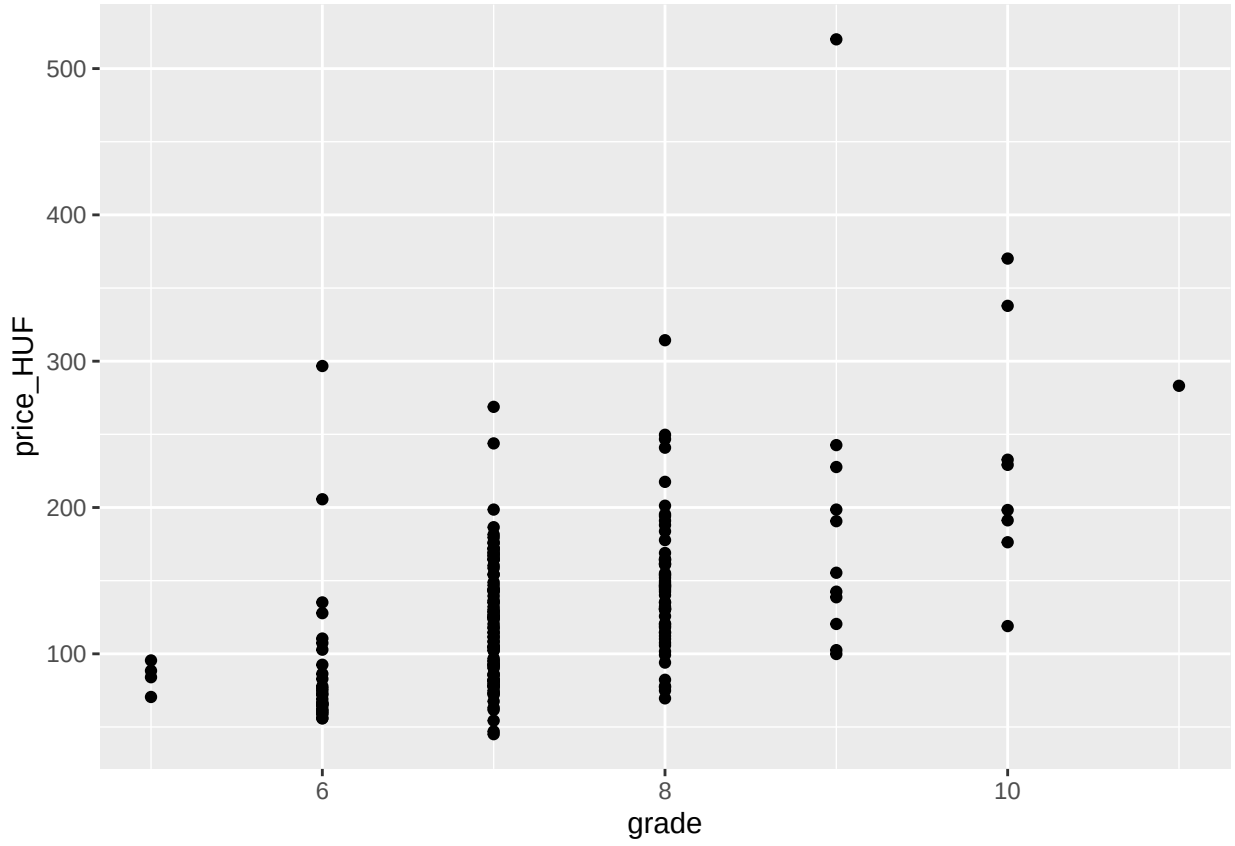

```
data_house %>% ggplot() + aes(x = grade) + geom_bar() + scale_x_continuous(breaks = 4:12)
```




```
# scatterplot
data_house %>% ggplot() + aes(x = sqm_living, y = price_HUF) +
    geom_point()
```



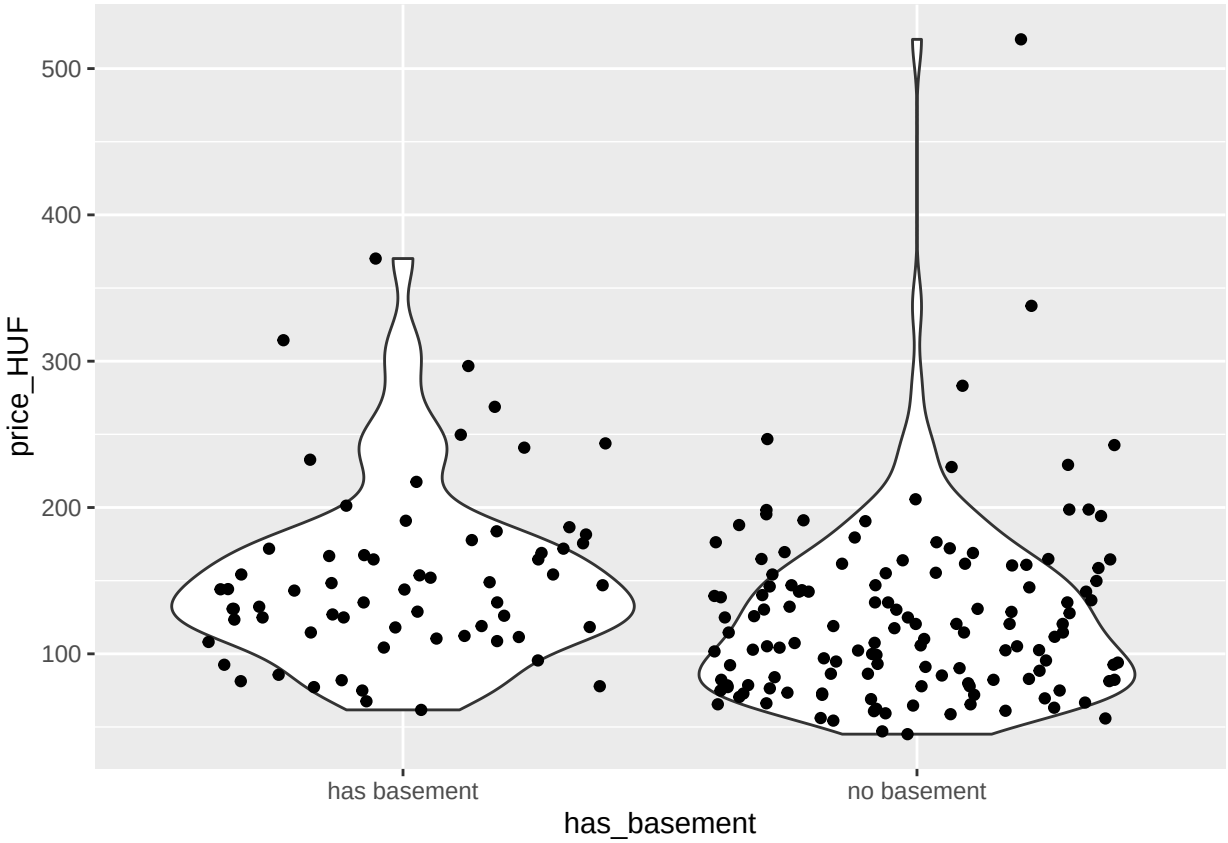

```
data_house %>% ggplot() + aes(x = grade, y = price_HUF) + geom_point()
```




```
# leiro statisztika
table(data_house$has_basement)
```

```
##
## has basement  no basement
##           65           135
```

```
# violin plot
data_house %>% ggplot() + aes(x = has_basement, y = price_HUF) +
    geom_violin() + geom_jitter()
```



1.5 Tobbszoros regresszio

1.5.1 A regresszios modell felepitese (fitting a regression model)

A tobbszoros regresszios modellt ugyan ugy epeitjuk mint az egyszeru regresszios modellt, csak csak tobb prediktort is betehetunk a modellbe. Ezeket a prediktorvaltozokat + jellen valasztjuk el egymastol a regresszios formulaban.

Alabb **price_HUF** a bejosolt valtozo, es a **sqm_living** es a **grade** a prediktorok.

```
mod_house1 = lm(price_HUF ~ sqm_living + grade, data = data_house)
```

A regresszios egyenletet a modell objektumon keresztul erhetjuk el:

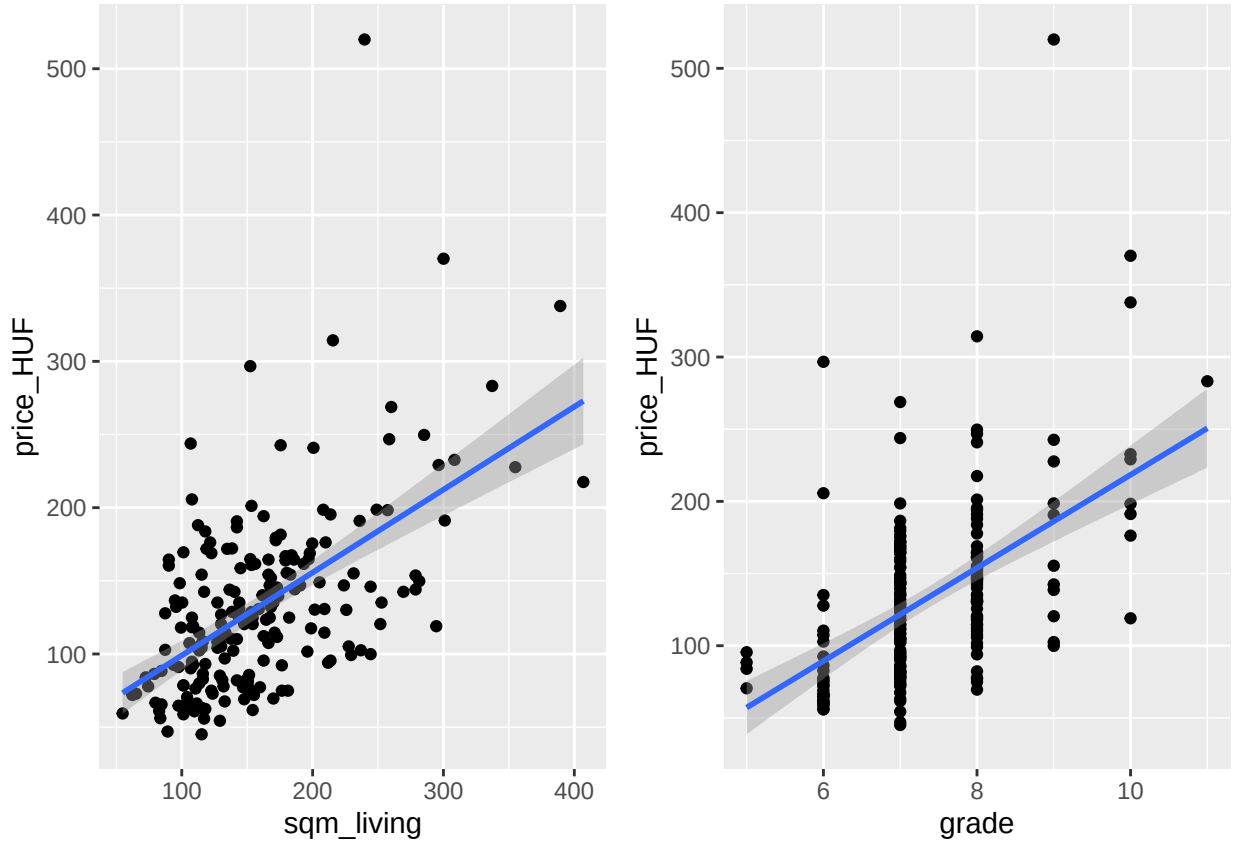
```
mod_house1
```

```
##
## Call:
## lm(formula = price_HUF ~ sqm_living + grade, data = data_house)
##
## Coefficients:
## (Intercept)   sqm_living        grade
##    -51.2305       0.3768      16.8485
```

A tobbszoros regresszios modellek vizualizacioja nem olyan egyertelmu mint az egyszeru regresszios modelleke.

Az egyik megoldas hogy a paronkenti osszefuggeseket vizualizaljuk egyenkent, de ez nem ragadja meg a modell tobbvaltozos jelleget.

```
# scatterplot
plot1 = data_house %>% ggplot() + aes(x = sqm_living, y = price_HUF) +
    geom_point() + geom_smooth(method = "lm")

plot2 = data_house %>% ggplot() + aes(x = grade, y = price_HUF) +
    geom_point() + geom_smooth(method = "lm")

grid.arrange(plot1, plot2, nrow = 1)
```



Egy alternativa hogy egy haromdimenzios abran abrazoljuk a regresszios sikot.Bar ez szepen nez ki, de nem tul hasznos, es ez is csak ket prediktorvaltozoig mukodik, harom es tobb prediktor eseten mar egy tobbdimenzios terben kepzelheto csak el a regresszios felulet, ezert a vizualizaciora altalaban megis az paronkenti scatterplot-ot szoktuk hasznalni.

```
# plot the regression plane (3D scatterplot with regression
# plane)
scatter3d(price_HUF ~ sqm_living + grade, data = data_house)
```

1.5.2 Becsles (prediction)

Ugyan ugy ahogy az egyszeru regresszional, itt is kerhetjuk a prediktorok bizonyos uj ertekekeire a kimeneti valtozo ertekenek megbecsleset a predict() fuggveny segitsegevel.

Fontos, hogy a prediktorok ertekeit egy data.frame vagy tibble formatumban kell megadnunk, es a prediktor-valtozok valtozoneveinek meg kell egyeznie a regresszios modellben hasznalt valtozonevekkel.

```r
sqm_living = c(60, 60, 100, 100)
grade = c(6, 9, 6, 9)
newdata_to_predict = as.data.frame(cbind(sqm_living, grade))
predicted_price_HUF = predict(mod_house1, newdata = newdata_to_predict)

cbind(newdata_to_predict, predicted_price_HUF)
```

```
##   sqm_living grade predicted_price_HUF
## 1         60     6            72.47102
## 2         60     9           123.01660
## 3        100     6            87.54459
## 4        100     9           138.09017
```

1.5.3 Hogyan kozoljuk az eredmenyeinket egy kutatasi jelentesben

Egy kutatsi jelentesben (pl. cikk, muhelymunka, ZH) a kovetkezo informaciokat kell leirni a regresszios modellrol:

Eloszor is le kell irni a regresszios modell tulajdonsagait (altalaban a "Modszerek" reszben):

"Egy linearis regresszios modellt illesztettem, melyben a lakas arat (millio HUF-ban) a lakas lakoreszenek teruletevel (m^2-ben) es a lakas King County lakas-minosites ertekevel becsultem meg."

"I built a linar regression model in which I predicted housing price (in million HUF) with the size of the living area (in m^2) and King County housing grade as predictors."

Ezutan a teljes modell bejoslasi hatekonysagat kell jellemezni. Ezt a modellhez tartozo adjusted R^2 ertek (modositott R^2), es a modell-t a null-modellel osszehasonlito anova F-tesztjenek statiszikainak megadasaval szoktuk tenni (F-ertek, df, p-ertek). Mindezen informaciot a summary() funkcioval tudjuk lekerdezni. A modell illeszkedeset az AIC (Akaike information criterion) ertekkel is szoktuk jellemezni, amit az AIC() funcio ad meg.

Az APA publikacios kezikonyv alapjan minden szamot ket tizedesjegy pontossaggal kell megadni, kiveve a p erteket, amit harom tizedesjegy pontossaggal.

```r
sm = summary(mod_house1)
sm
```

```
##
## Call:
## lm(formula = price_HUF ~ sqm_living + grade, data = data_house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.26  -29.55   -6.79   19.65  329.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -51.2305    27.9831  -1.831 0.068646 .
## sqm_living    0.3768     0.0783   4.813 2.96e-06 ***
## grade        16.8485     4.7158   3.573 0.000444 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.96 on 197 degrees of freedom
## Multiple R-squared:  0.358,  Adjusted R-squared:  0.3515
## F-statistic: 54.94 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
AIC(mod_house1)
```

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

Vagyis az "Eredmenyek" reszben igy irnank a fenti pelda eredmenyeirol:

"A tobbszoros regresszios modell mely tartalmazta a lakoterulet es a lakas minosites prediktorokat hatekonyabban tudta bejosolni a lakas arat mint a null modell. A modell a lakasar varianciajanak 35.15%-at magyarazta (F (2, 197) = 54.94, p < .001, Adj. R^2 = 0.35, AIC = 2137.06)."

Ezen felul meg kell adnunk a regresszios egyenletre es az egyes prediktorok becsleshez valo hozzajarulasara vontkozo adatokat. Ezt altalaban egy osszefoglalo tablazatban szoktuk megadni, melyben a kovetkezo adatok szerepelnek prediktoronkent:

- regresszios egyutthato (regression coefficients, estimates) - summary()
- az egyutthatokhoz tartozo konfidencia intervallum (coefficient confidence intervals) - confint()
- standard beta ertekek (standardized beta values) - lm.beta() az lm.beta pakcage-ben
- a t-teszthez tartozo p-ertek (p-values of the t-test) -summary()

```
confint(mod_house1)
lm.beta(mod_house1)
```

A vegso tablazat valahogy igy nez majd ki:

```
##                   b 95%CI lb 95%CI ub Std.Beta p-value
## (Intercept) -51.23  -106.42     3.95        0    .069
## sqm_living    0.38     0.22     0.53     0.37   <.001
## grade        16.85     7.55    26.15     0.28   <.001
```

1.5.4 regresszios eghyutthato ertelmezese

A regresszios egyutthatot ugy lehet ertelmezni, hogy a prediktor ertekenek egy ponttal valo novekedese eseten a kimeneti valtozo erteke ennyivel valtozik. Pl. ha a sqm_living-hez tartozo regresszios egyutthato 0.38, az azt jelenti hogy minden egyes ujabb negyzetmeter teruletnovekedes 0.38 millio forint arvaltozassal jar.

1.5.5 standard beta ertelmezese

A regresszios egyutthato elonye, hogy a kimeneti valtozo mertekegysegeben van, es nagyon egyszeru ertelemzni. Ezert ez egy "nyers" hatasmeret mutato. Viszont a hatranya hogy az erteke a hozza tartozo prediktor valtozo skalajan mozog. Ez azt jelenti, hogy az egyes egyutthato ertekek nem konnyen osszehasonlithatoak, mert a prediktorok mas skalan mozognak. Pl. az sqm_living egyutthatoja alacsonyabb mint az grade egyutthatoja, de ez onmagaban nem mond arrol semmit, hogy melyik prediktornak van nagyobb szerepe a kimeneti valtozo bejoslasaban, mert a sqm_living skalaja sokkal kiterjedtebb (50-400 m^2) mint a grade skalaja (5-11).

Ahhoz hogy ossze tudjuk hasonlitani az egyes prediktorok becsleshez hozzaadott erteket, a ket egyutthatot ugyan arra a skalara kell helyeznunk, amit standardizalassal erhetunk el. A standard Beta egy ilyen standardizalt mutato. Ez mar direkt modon osszehasonlithato a prediktorok kozott. Ebbol mar latszik hogy a sqm_living hozzaadott erteke a price_HUF bejoslasahoz nagyobb mint a grade hozzaadott erteke.

Amikor tobb prediktor van, ez nem feltetlenul jelenti azt, hogy ha egyenkent megneznenk a prediktorok korrelaciojat a kimeneti valtozoval, akkor ugyan ilyen osszefuggest kapnank. Ez az egyutthato es a std.Beta ertek a prediktor egesz modellben betoltott szerepet jeloli, a tobbi prediktor bejoslo erejenek leszamitasaval. Vagyis elkepzelheto, hogy egy prediktor onmagaban jobban korrelal a kimeneti valtozoval mint barmelyik masik prediktor, viszont a modellben kisebb szerepet jatszik, mert a tobbi prediktor ugyan azt a reszet magyarazza a kimeneti valtozo varianciajanak, mint ez a prediktor.

_____Gyakorlas_____

1. Epits egy tobbszoros linearis regresszio modellt az lm() fugvennyel amiben az **price_HUF** a kimeneti valtozot becsuljuk meg. Hasznalhatod a **data_house** adatbazisban szereplo barmelyik valtozot felhasznalhatod a modellben, ami szerinted realisan hozzajarulhat a lakas aranak meghatarozasahoz.
2. Hatarozd meg, hogy szignifikansan jobb-e a modelled mint a null modell (a teljese modell F-teszthez tartozo p-ertek alapjan)?
3. Mekkora a teljes modell altal bejosolt varianciaarany (adj.R^2)?
4. Melyik az a prediktor, mely a legnagyobb hozzadaott ertekkel bir a becslesben?
