# S08 Modell osszehasonlitas es Modellvalasztas

Zoltan Kekecs

03 November 2020

## Contents



# 1 Modell osszehasonlitas es Modellvalasztas

## 1.1 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.

## 1.2 Adatmenedzsment es leiro statisztikak

### 1.2.1 Package-ek betoltese

```
library(tidyverse)
```

### 1.2.2 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")
```

### 1.2.3 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   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)/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
         )
```

## 1.3 Hierarchikus regresszio

A hierarchikus regresszioval (Hierarchical regression) meghatarozhatjuk, **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.

Ugy fogjuk a ket modellt felepiteni, 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.

## 1.4 Hieararchikus regresszio ket prediktor-blokkal

### 1.4.1 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)
```

### 1.4.2 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).

## 1.5 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.