

S8 Model comparison and model selection

Zoltan Kekecs

11 November 2019

Contents

1	Modell összehasonlítás és Modellválasztás	1
1.1	Absztrakt	1
1.2	Adatmenedzsment és leíró statisztikák	1
1.3	Hierarchikus regresszió	3
1.4	Hierarchikus regresszió két prediktor-blokkal	3
1.5	Hierarchikus regresszió több mint két blokkal	4

1 Modell összehasonlítás és Modellválasztás

1.1 Absztrakt

Ez a gyakorlat megmutatja majd, hogyan lehet különböző prediktorokat tartalmazó modelleket összehasonlítani egymással. Demonstráljuk majd a hierarchikus regressziót. Néhány modell szelekciós módszerre is kiterünk majd, és megemlítjük a “tulullesztés” (overfitting) fogalmát.

1.2 Adatmenedzsment és leíró statisztikák

1.2.1 Package-ek betöltése

```
library(tidyverse)
```

1.2.2 A King County lakaseladás adattábla betöltése

Ebben a gyakorlatban lakások és házak árát fogjuk megbecsülni.

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

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

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

1.2.3 Adatellenőrzés

Mindig nézd át az általad használt adattáblát. Ezt már megtettük az előző gyakorlatban, így ezt most itt mellozzuk, de a korábbi tapasztalatok alapján átalakítjuk az árát (price) millió forintra, és a negyzetlabban szereplő terület értékeit negyzetmeterre.

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

```
##           id           date           price  
## Min.      :1.600e+07  20140623T000000: 5   Min.      : 153503  
## 1st Qu.:1.885e+09   20141107T000000: 5   1st Qu.: 299250  
## Median :3.521e+09   20150317T000000: 4   Median : 425000
```

```

## Mean :4.113e+09 20140627T000000: 3 Mean : 453611
## 3rd Qu.:6.424e+09 20140717T000000: 3 3rd Qu.: 550000
## Max. :9.819e+09 20140902T000000: 3 Max. :1770000
## (Other) :177
## 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 has basement: 65
## 1st Qu.:1438 1st Qu.: 5000 no basement :135
## Median :1715 Median : 7222
## Mean :1793 Mean : 11225
## 3rd Qu.:2072 3rd Qu.: 10028
## 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)

```

1.3 Hierarchikus regresszio

A hierarchikus regresszióval (Hierarchical regression) meghatározhatjuk, mennyivel javul a bejóslo ero egy bonyolultabb (több prediktort tartalmazó) modell használatával ahhoz képest ha egy egyszerűbb (kevesebb prediktort tartalmazó) modellt használunk.

Ehhez két regressziós modellt fogunk építeni. Az egyszerűbb modellben szereplő prediktorok egy részhalmazát alkotják a bonyolultabb modell prediktorainak. (vagyis a bonyolultabb modell minden prediktort tartalmaz az egyszerűbb modellből, plusz meg néhány extra prediktort.)

1.4 Hierarchikus regresszio két prediktor-blokkal

Eloszor 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)
```

Az `adj. R Squared` mutató segítségével meghatározhatjuk a két modell által megmagyarázott varianciaarányt. Ezt a `model summary` kizárásával 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
```

Ugy tunik, 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 hiba és jobb modell illeszkedést jelent. Ha a különbség nem éri el a 2-t, akkor a két modell közül bármelyiket megtehetjük. 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
```

Másrészt pedig az `anova()` funkció segítségével összehasonlíthatjuk a két modell residuális hibáját.

Ha az `anova()` F-tesztje szignifikáns, az azt jelenti, 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 jobban elfogadott a szakirodalomban, ezért ha az AIC és az anova összehasonlítás különbozó 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 beagyazott modellek összehasonlítására is, (bar ezzel kapcsolatban nem teljes az egyetemes a szakirodalomban, a dolgozatokban elfogadott AIC-ot használni nem beagyazott modellek összehasonlítására).

1.5 Hierarchikus regresszio több mint két blokkal

A fenti folyamat ugyan így 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 következőképpen hasonlíthatjuk össze:

```
# R2
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) figyelembevetelével?

Gyakorlas

Tedd hozzá a modellhez az iment épített modellhez (mod_house_geolocation_cond) a ház építésének évét (yr_built) és a fürdőszobák számát (bathrooms) mint prediktorokat. Ez az új modell szignifikánsan jobban illeszkedik az adatokhoz mint a korábbi modellek?

A modellválasztás legfontosabb szabálya:

Mindig azt a modellt választjuk, ami elméletileg alátámasztott és/vagy korábbi kutatási eredmények támogatják, mert az automatikus modellválasztás rossz modellekhez vezet a túlillesztés (overfitting) miatt.