Exercise 13 - Model comparison and model selection

Zoltan Kekecs

14 november 2019

Contents

1 Abstract

This exercise will show you how different models can be compared to each other. It will denonstrate hierarchical regression.

The latest version of this document and the code the document refers to can be found in the GitHub repository of the class at: https://github.com/kekecsz/PSYP13_Data_analysis_class-2019

2 Data management and descriptive statistics

2.1 Loading packages

You will need to load the following packages for this exercise:

```
library(tidyverse)  # for tidy format
```

2.2 Load data about housing prices in King County, USA

In this exercise we will predict the price of apartments and houses.

We use a dataset from Kaggle containing data about housing prices and variables that may be used to predict housing prices. This dataset contains house sale prices for King County, USA (Seattle and sorrounding area) which includes Seattle. It includes homes sold between May 2014 and May 2015. More info about the dataset here: https://www.kaggle.com/harlfoxem/housesalesprediction

We only use a portion of the full dataset now containing information about N = 200 accomodations.

You can load the data with the following code

```
data_house = read_csv("https://bit.ly/2DpwKOr")
```

2.3 Check the dataset

You should always get familiar with the dataset you are usung, and check for any inconsistencies that need to be corrected.

In the code below we convert the area metrics that are in square feet in the original dataset to square meters. We also specify that the variable has_basement is a factor.

```
data_house = data_house %>% mutate(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, has_basement = factor(has_basement))
```

3 Hierarchical regression

Using hierarchical regression, you can quantify the amount of information gained by adding a new predictor or a set of predictors to a previous model. To do this, you will build two models, the predictors in one is the subset of the predictors in the other model.

3.1 Hierarchical regression with two predictor blocks

Here we first build a model to predict the price of the apartment by using only sqm_living and grade as predictors.

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

Next, we want to see whether we can improve the effectiveness of our prediction by taking into account geographic location in our model, in addition to living space and grade

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

We can look at the adj. R squared statistic to see how much variance is explained by the new and the old model.

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

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

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

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

It seems that the variance explained has increased substantially by adding information about geographic location to the model.

Now, we can compare model fit using the AIC() function and residual error throught the anova() function.

```
AIC(mod_house2)
```

```
## [1] 5390.142
```

```
AIC(mod_house_geolocation)
```

```
## [1] 5342.783
```

If the difference in AIC of the two models is larger than 2, the two models are significantly different in their model fit. Smaller AIC means less error and better model fit, so in this case we accept the model with the smaller AIC. However, if the difference in AIC does not reach 2, we can retain either of the two models. In this case, theoretical considerations and previous results should should be considered when doing model selection. If both models seem plausible theoretically, we can retain the model containing less predictors.

The anova() function compares the models based on their residual error and degrees of freedom.

Importantly, the anova for model comparison is only appropriate if the two models are “nested”, that is, predictors in one of the models are a subset of predictors of the other model. By comparison, the AIC can be used to compare non-nested models as well (although there is some controversy about this in the literature).

If the anova F test is significant, it means that the models are significantly different in terms of their residual errors.

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

```
## Analysis of Variance Table
##
## Model 1: price ~ sqm_living + grade
## Model 2: price ~ sqm_living + grade + long + lat
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    197 5.6981e+12
## 2    195 4.4076e+12  2 1.2905e+12 28.546 1.338e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AIC is a more established model comparison tool, so if the anova and AIC methods return discrepant results, the AIC should be used for decision making.

3.2 Hierarchical regression with more than two blocks

The same procedure can be repeated if we have more than two steps/blocks in the hierarchical regression.

Here we build a third model, which adds even more predictors to the formula. This time, we add information about the condition of the apartment.

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

We can compare the three models now.

```
# 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 ~ sqm_living + grade
## Model 2: price ~ sqm_living + grade + long + lat
## Model 3: price ~ sqm_living + grade + long + lat + condition
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    197 5.6981e+12
## 2    195 4.4076e+12  2 1.2905e+12 29.318 7.493e-12 ***
## 3    194 4.2695e+12  1 1.3812e+11  6.276   0.01306 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## [1] 5390.142
```

```
AIC(mod_house_geolocation)
```

```
## [1] 5342.783
```

```
AIC(mod_house_geolocation_cond)
```

```
## [1] 5338.416
```

Did we gain substantial information about housing price by adding information about the condition of the apartment to the model?

_______________*Practice*_______________

Add the year the house was built (yr_built) as a new predictor to the previously built model (mod_house_geolocation_cond) and the number of bathrooms in the apartment (bathrooms). Does this increase model fit significantly?

____

3.3 First rule of model selection:

Always go with the model that is grounded in theory and prior research, result-driven model selection can lead to bad predictions on new datasets due to overfitting!