Exercise 13 - Model comparison and model selection

Zoltan Kekecs

8 May 2020

Table of Contents

Abstract

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

Data management and descriptive statistics

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. The data is about accomodations in King County, USA (Seattle and sorrounding area).

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

The data can be downloaded at:

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_2/House%20price%20King%20County.sav

Check the dataset

You should always check the dataset for coding errors or data that does not make sense, by eyeballing the data through the data view tool, checking descriptive statistics and through data visualization.

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.

Hierarchical regression with two predictor blocks

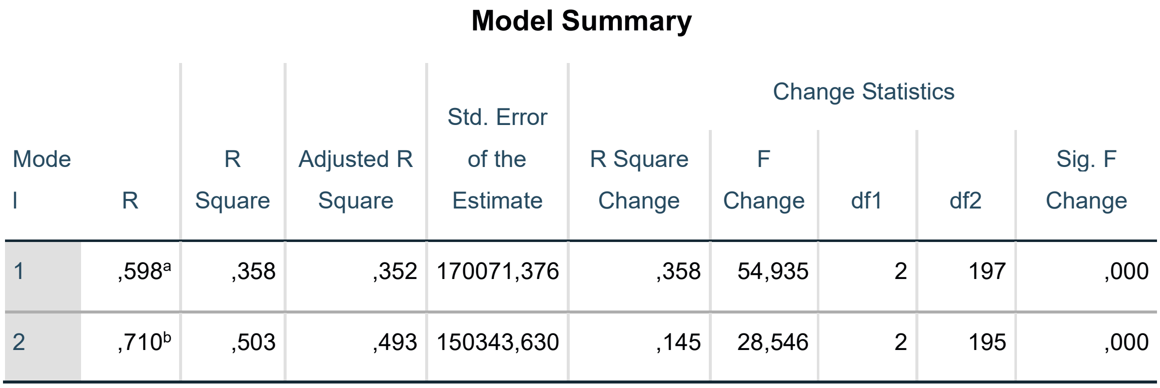
Here we first build a model to predict the price of the apartment by using only sqft_living and grade as predictors. Next, we want to see whether we can improve the effectiveness of our prediction by taking into account geographic location in our model, in addition to living space and grade.

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

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

However, we may want to see whether this increase is statistically significant. To do that we would have to fit these models in the same regression as separate blocks in the **Analyze > Regression > Linear** Independents panel by entering the predictors from the less complex model as Independents, then pressing the "next" button and entering the predictors that are only included in the more complex model. You also need to ask for the "R squared change" statistics in the Statistics menu!

This way SPSS will compare the more complex model (with the more predictors) to the less complex model. You can find this comparison in the model summary table in the output.



Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	,598[a]	,358	,352	170071,376	,358	54,935	2	197	,000
2	,710[b]	,503	,493	150343,630	,145	28,546	2	195	,000

a. Predictors: (Constant), grade, sqft_living

b. Predictors: (Constant), grade, sqft_living, lat, long

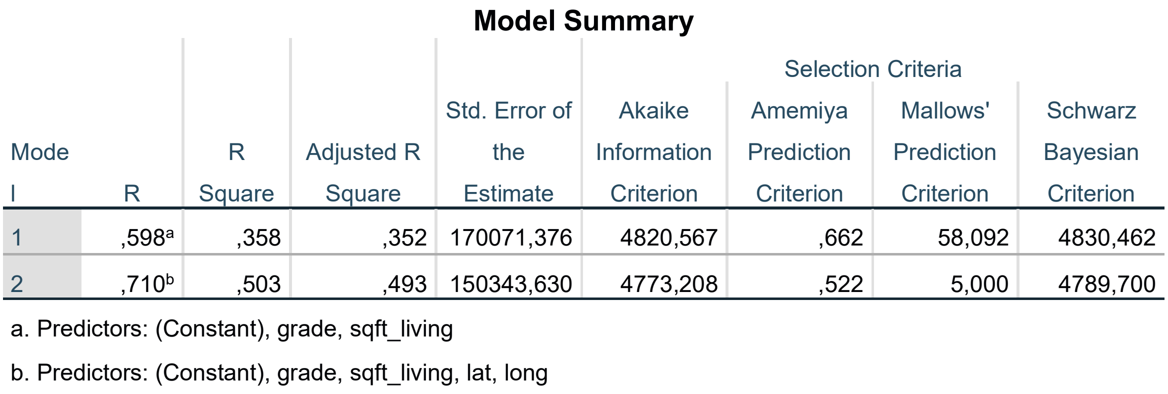
Now, we should compare residual error and model fit thought the Likelihood ratio test (F – test) listed in the Model summary table. If the anova F test is significant, it means that the models are significantly different in terms of their residual errors.

This method can only be used for comparing models when they are “nested”, that is, predictors in one of the models are a subset of predictors of the other model.

The two models can also be compared using the Akaike Information Criteria (AIC). In SPSS AIC needs to be requested through the syntax. The word SELECTION needs to be added in the end of the line starting with /STATISTICS. For the above example, the syntax would look like this (the green highlighting is just there to show where SELECTION should go, you don’t need to highlight the syntax for it to work.):

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA SELECTION
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT price
 /METHOD=ENTER sqft_living grade
 /METHOD=ENTER lat long.
```

Akaike Information Criteria will show up in the Model Summary table in the output.



Model Summary

					Selection Criteria			
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	,598a	,358	,352	170071,376	4820,567	,662	58,092	4830,462
2	,710b	,503	,493	150343,630	4773,208	,522	5,000	4789,700

a. Predictors: (Constant), grade, sqft_living

b. Predictors: (Constant), grade, sqft_living, lat, long

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. Ususally we stick with the less complicated model in this case, but theoretical considerations and previous results should also be considered when doing model selection.

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.

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 ‘condition’ as a new predictor in block 3.

Again, don’t forget to add SELECTION to the syntax and run the syntax command again!

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA SELECTION
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT price
 /METHOD=ENTER sqft_living grade
 /METHOD=ENTER lat long
 /METHOD=ENTER condition.
```

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

First rule of model selection:

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