Lab of ML&DA

Model selection

Matilde Trevisani

YA 2017/2018

Analysis on crime data in USA

The dataset in `UScrime.dat` contains some demographic and social statistics for 47 american states in 1960. Among statistics, the criminality rate is given by # crimes reported per million inhabitants.

Let's upload the data and look at the first records:

Crime rate (y) is the response variable, the covariates are:

young	# males age 14-24 per 1000
sud	Southern or Northern state (1 = Yes, 0 = No)
edu	(10×average) # of years adults (age $\geq$ 25) have been educated
expp0	1960 per capita expenditure on police (by state and local government)
expp1	1959 per capita expenditure on police (by state and local government)
LFy	Labor force participation rate per 1000 civilian urban males age 14-24
male	# males per 1000 females
pop	State population (× 100000)
nonW	# non-whites per 1000
unempl1	Unemployment rate of urban males of age 14-24 per 1000
unempl2	Unemployment rate of urban males of age 35-39 per 1000
assets	Median value of transferable goods and assets or family income (×10 $)
poverty	# families earning below 1/2 the median income (of state)

Let's have a summary of the variables.
Check whether there is any categorical variable.

After excluded any categorical variable, let's represent graphically the remaining variables one by one. (Try different graphical methods.)

Let's now look at the joint distribution of all variables.
In particular, inspect any presence of multicollinearity.

What are the most relevant covariates for y?
What covariates are more troubling as far as collinearity is concerned?

Multicollinearity

Let's try different linear models which include either just `expp0` or just `expp1` or both.
What do you note?

Now let's try the same experiment with the couple `(assets,poverty)`.
What do you note? (look also at the standard error estimates.)

It can be convenient to esclude some collinear covariates before starting with model building.

In any event, in the current study it is sensible to exclude both `expp0` e `expp1`. Why?

Remove them from the dataset.

Criteria for comparing models

Let us recall the most known model selection criteria: R_a^2, AIC, BIC.

R command: AIC and BIC

See `AIC(x)` and `BIC(x)` (they use function `logLik`)
`extractAIC` uses a different normalization const.

Variable Selection

Let's select a model for USC dataset.
We start by estimating a model with all variables.

Several covariates appear important.
Let's check the model by some diagnostic tools. (Such analysis is useful to examine whether any variable transformation is needed.)

Do you see any misfitting?

Given that sample size is small as well as the number of regressors is limited, we can estimate all the possible models.

How many are they?

Best subset selection

R command: Best subset selection

Function `leaps` (homonymous pkg) estimates all possible models.

Let's load the pkg and run the best subset selection by `leaps`.

What does the matrix `which` of the output describe?

Then, what is the best simple model (with 1 regressor) on the basis of R^2?

And what is the best model with 4 regressors on the basis of R^2?

Plot R^2 values for all the best models.
Now, can you decide what is the overall best model by looking at this plot?

Choose a proper criterion for model selection and then find the best subset.

NOTE: if you find that the best model includes collinear variables you should repeat the analysis by properly removing some of them.

0.0.1 Stepwise selection

Let's try now a stepwise selection method which will possibly lead to a suboptimal solution.

R command: **Stepwise selection**

Function `stepAIC` in `MASS` pkg carries out a stepwise subset selection.