MPC Course - Part 2

Regression : predicting a variable using other variable(s)

Simon Malinowski

M1 Miage, Univ. Rennes 1

1. Variable selection
   - Problem
   - Criteria for model comparisons
   - Variable selection procedures
2. Non-linear regression

Variable selection : problem

Amongst the p predictive variables $(X_1, \ldots X_p)$, we aim at selecting a subset of variables that lead to the best model.

Difficulties :

1. number of subset of a set of p variables ?
   - 20 variables : 1 million
   - 30 variables : 1 billion
2. How to compare models with different number of variables ?

Criteria that we have already seen

1. coefficient of determination $R^2 \rightarrow$ not adapted
2. $SCE_r \rightarrow$ not adapted
3. generalization error $\rightarrow$ adapted
4. critical probability of the Student's test $\rightarrow$ adapted

We can use the following criteria :

1. Adjusted R-squared : $R^2_{adj} = \frac{R^2(n-1)-p}{n-p-1}$
2. generalization error
3. critical probability of the Student's test

Variable selection procedures

Hypothesis : we know how to compare models with diferent number of variables (with an adapted criterion, cf. above)

Several procedures :

1. Exhaustive search : we try every possible subset. Never if $p > 15$
2. Forward search
3. Backward search
4. Stepwise search
5. Stagewise search

Forward search

We have p predictive variables $x_1, \dots, x_p$, and a selection criteria C to compare models (ex : generalization error)

- Search for the best model with 1 variable (according to C)
  - $\rightarrow$ Model $M_1 = \{x_b^1\}$, its performance is $C(M_1)$
  - $\rightarrow$ Best model found up to now is $M_b = M_1$
  - $\rightarrow$ The performance C_b of the best model is $C_b = C(M_1)$
- We then search for the best variable **to go with** x_b^1
  - $\rightarrow$ Model $M_2 = \{x_b^1, x_b^2\}$, its performance is $C(M_2)$
  - $\rightarrow$ If $C(M_2)$ is *better* than $C(M_1)$, then $M_b = M_2$ and $C_b = C(M_2)$
- Iterate this procedure until a stopping criterion is met (explained after)

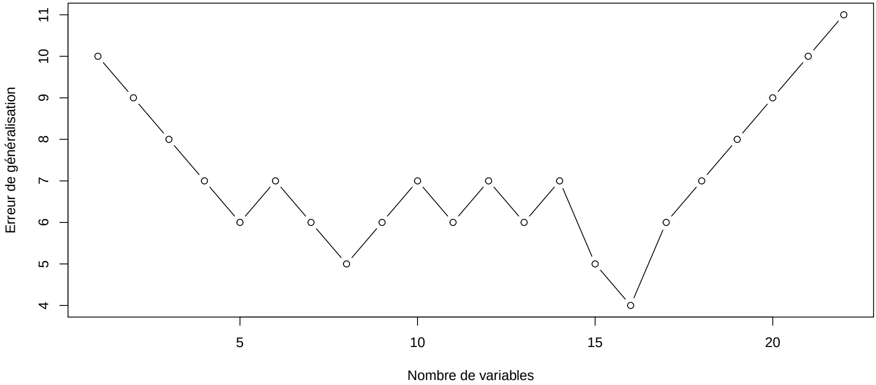
Stopping criteria

The most used stopping criterion is : stop as soon as one iteration does not improve the perforance (according to C).
It is a bit strict, sometimes better models can be found by waiting a little bit more

Other stopping criteria :

- stop when no better model has been found since δ iterations
- don't stop (go until the end) and keep the best model found

Stopping criteria



Forward search : pseudo-code

We choose the generalization error as performance criterion
Stopping criterion : as soon as the performance decreases

Input : Dataset with p predictive variables $X_1, \ldots, X_p$ and a target one Y

Output : $V_s = \{X_{\sigma_1}, \ldots, X_{\sigma_k}\}$, subset of selected variables

Forward search : pseudo-code

Initialization :

$V_s = []$: selected variables
$V_{nu} = [X_1, \ldots, X_p]$: variables not (yet) used
$C_f = \infty$: performance of the best model found
$stop = F$: variable to handle the stopping criterion

Forward search : pseudo-code

WHILE $stop = F$

$\forall x \in V_{nu}$, Compute the performance of the models with variables $[V_s, x]$

Let x_b be the best x (above), and C_b its performance

IF $C_b < C_f$,

$V_s = [V_s, x_b]$; $C_f = C_b$; $V_{nu} = V_{nu} \setminus \{x_b\}$

ELSE $stop = T$

END WHILE

Backward search

Same principle but the other way around :

- We start with the full model (with p variables)
- We search for the best variable to remove (the one that leads to the best model) $\rightarrow p-1$ variable
- Amongst these $p-1$, we search for the best one to remove
- repeat until a stopping criterion is met

Variable selection in practice

Forward and backward search are approximations of the exhaustive search

They don't always lead to the same selected model (and hence not always the best one)

Different performance criteria (R^2_{adj}, generalization error, ...) may lead to different models

The different selected models have then to be compared (on a new set)

Variable selection in practice

1. Split the dataset into a training set and a test set (with about 20% for the test set
2. Apply the variable selection algorithm using the training set
   $\rightarrow$ if the generalization error is the performance criterion, you will need to split the training set again (into training and validation)
3. Estimate the generalization error of the selected model on the test set.

1. Variable selection

2. Non-linear regression
   - Polynomial regression

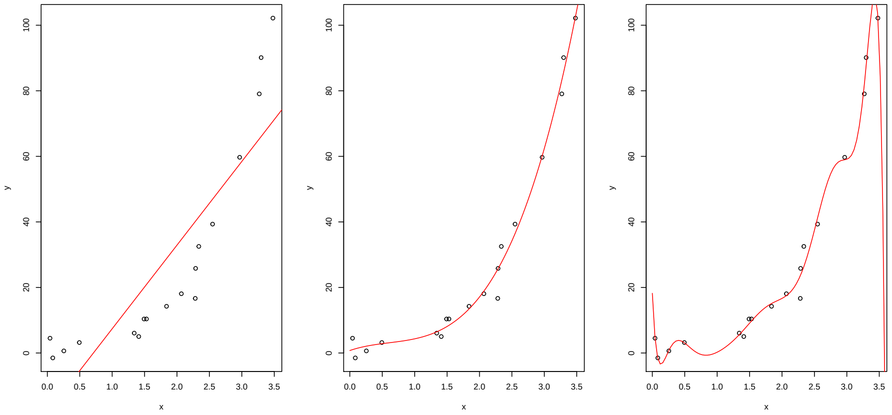
Polynomial regression

Classical linear regression : $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Ploynomial regression : same structure but including powers to the variables (by adding new columns in rhe dataset)

Problem : How to find which power needs to be included for which variable(s) ?

Example



In practice

Regression with non-numerical variables

Size	District	Price
80	Center	120
92	Bourg-Lesveque	180
75	Center	145
110	Longchamps	220
85	Longchamps	140
150	Bourg-Lesveque	225
55	Center	100
105	Longchamps	? ?
95	Center	? ?