Lecture 3 : Model selection

M2-Modèles pour la régression

K. Meziani

Dauphine Université Paris | PSL

Packages

```
library(MASS)
library(car)
library(carData)
library(knitr)
library(ggplot2)
library(caret)
library(cowplot)
library(reshape2)
library(mlbench)
library(GGally)
library(corrplot)
library(questionr)
library(multcomp)
library(dplyr)
```

Section 1

1. Introduction

Introduction

- The purpose of the regression is twofold: **Explain and predict** using estimation tools.
- In previous chapters, it has been assumed that the model

$$Y = X\beta + \varepsilon$$

is the "*good model*" where $X = (X_1, \ldots, X_p)$. **In practice, nothing assures us that we have not forgotten variables.**

It is also possible that too many variables are used.

- If the goal is to explain, it seems justified to take the model having the largest R^2.
- If the goal is to estimate or predict, we will see that this is not necessarily the case. To do this, we use the mean squared error (MSE).

The mean squared error (MSE)

Definition Let $\theta \in \mathbb{R}^k$ be the parameter to be estimated and $\hat{\theta}$ an estimator of θ. **The mean squared error (MSE)** of $\hat{\theta}$ is given by:

$$\mathbb{E}[\|\hat{\theta} - \theta\|^2] = \sum_{j=1}^{p} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2].$$

Comment:

- The use of $\|\cdot\|^2$ is consistent with the idea of ordinary least squares estimation.

Proposition For all $\theta \in \mathbb{R}^p$:

$$\mathbb{E}[\|\hat{\theta} - \theta\|^2] = \sum_{j=1}^{p} (\mathbb{V}\mathrm{ar}(\hat{\theta}_j) + (\mathbb{E}[\hat{\theta}_j] - \theta_j)^2).$$

Why section model ?

To illustrate it, let us consider the following example. We assume the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon = X\beta + \varepsilon, \quad \text{where} \tag{1}$$

- $X = [X_1 \; X_2]$ is a $n \times 2$ matrix of rank 2
- $\beta = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$ *s.t.* $\beta_1 \neq 0$.

Question : Is the variable X_2 useful ?

Study the case $\beta_2 = 0$ (even if it is false), and look for when to omit an explanatory variable can be advantageous in terms of risk .

Why section model ?

- Let define the following model

$$Y = X_1\beta_1 + \varepsilon, \quad \widetilde{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top Y$$

  and the associated OLSE estimator of β_1 where Y is defined by the model (1),

- Denote by $\widehat{\beta}$, the OLSE calculate from the model (1).
- Thus, we have 2 estimators, one biased and the other one unbiased

$$\widetilde{\beta} = (\widetilde{\beta}_1, 0)^\top \text{ and } \widehat{\beta} = (X^\top X)^{-1} X^\top Y$$

Proposition In the previous context, $\forall \beta \in \mathbb{R}^2$

$$\mathbb{E}[\|\hat{\beta} - \beta\|^2] - \mathbb{E}[\|\widetilde{\beta} - \beta\|^2] \geq \sigma^2 \frac{\|X_1\|^2}{D} - \beta_2^2 \left(1 + \frac{(X_1^\top X_2)^2}{\|X_1\|^4}\right),$$

where D denote the determinant of the matrix $(X^\top X)$.

rajouter du biais peut diminuer le risque de généralisation

mais pas accès, β est inconnu.

Comment

This result does not contradict the Gauss-Markov theorem, because $\widetilde{\beta}$ is biased. By introducing (for $\beta_2 \neq 0$ and small enough) a slightly biased estimator with a lower variance, the quadratic risk is improved. For the estimation (and therefore the prediction), we must be wary of too rich models.

Proof

We easily prove that

$$(X^\top X)^{-1} = \frac{1}{D}\begin{pmatrix} \|X_2\|^2 & -X_1^\top X_2 \\ -X_1^\top X_2 & \|X_1\|^2 \end{pmatrix}, \quad \text{where } : D := \|X_1\|^2\|X_2\|^2 - (X_1^\top X_2)^2 > 0.$$

Moreover, the estimator $\widehat{\beta}$ is unbiased, it comes

$$\mathbb{E}[\|\widehat{\beta} - \beta\|^2] = \sum_{j=1}^{2} \mathbb{V}\mathrm{ar}(\widehat{\beta}_j) = \sigma^2 \mathrm{Tr}((X^\top X)^{-1})) = \frac{\sigma^2}{D}\left(\|X_2\|^2 + \|X_1\|^2\right).$$

For the estimator $\widetilde{\beta} = (\widetilde{\beta}_1, 0)^\top$, we have

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\beta} - \beta\|^2] &= \mathbb{E}[(\widetilde{\beta}_1 - \beta_1)^2] + \beta_2^2 = \mathbb{E}[((X_1^\top X_1)^{-1} X_1^\top Y - \beta_1)^2] + \beta_2^2 \\
&= \mathbb{E}[((X_1^\top X_1)^{-1} X_1^\top (\beta_1 X_1 + \beta_2 X_2 + \varepsilon)\beta_1)^2] + \beta_2^2 \\
&= \left((X_1^\top X_1)^{-1} X_1^\top X_2\right)^2 \beta_2^2 + \sigma^2 (X_1^\top X_1)^{-1} + \beta_2^2 \\
&= \frac{\sigma^2}{\|X_1\|^2} + \beta_2^2\left(1 + \frac{(X_1^\top X_2)^2}{\|X_1\|^4}\right).
\end{aligned}$$

Proof

For $D > 0$, it comes that $D < \|X_1\|^2\|X_2\|^2$. Therefore, we get

$$\begin{aligned}\mathbb{E}[\|\widehat{\beta} - \beta\|^2] - \mathbb{E}[\|\widetilde{\beta} - \beta\|^2] &= \frac{\sigma^2}{D}\left(\|X_2\|^2 + \|X_1\|^2\right) - \frac{\sigma^2}{\|X_1\|^2} - \beta_2^2\left(1 + \frac{(X_1^\top X_2)^2}{\|X_1\|^4}\right) \\ &> \frac{\sigma^2\|X_1\|^2}{D} - \beta_2^2\left(1 + \frac{(X_1^\top X_2)^2}{\|X_1\|^4}\right).\end{aligned}$$

□

Model selection / Criterions

Choose a set of variables (say a **model**). While it may be easy to decide between two models, the question of model choice is more delicat.

- There is no natural order between the variables.
- There are many possible models. For example, if there are 8 possible variables in addition to the vector $\mathbf{1}_n$ (always take the intercept), then we have $\sum_{j=0}^{8} C_j^8 = 2^8 = 256$ possible models to compare.

We will focus on methods that rely on the following criterions:

1. Tests between nested models
2. R^2, R_a^2 adjusted
3. C_p of Mallows $\longrightarrow$ bon estimateur du risque de généralisation
4. AIC- criterion
5. BIC- criterion

$\left.\right\}$ vraisemblance maximisée : $f(SCR)$ + pénalité

Section 2

2. Notations and illustrative example

Notations and illustrative example

- Set $p = q + 1$ the number of explanatory variables (the intercept $\mathbf{1}_n$ included): $X = (\mathbf{1}_n, X_1, \ldots, X_q)$. Consider the framework of linear regression models.

$$Y = X\beta + \varepsilon, \quad \text{rang}(X) = p, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n).$$

- Denote $[m]$ any model of size m, $i.e.$ $m := \text{card}([m])$. Define for all model $[m]$:

$$\text{RSS}\ (m) = \|Y - P_m Y\|^2.$$

Comparaison of criteria

Consider two nested models $[m_0] \subset [m_1]$ such that $m_1 = m_0 + 1$. The model $[m_0]$ is composed by m_0 variables (the intercept $\mathbf{1}_n$ is considered to be in the model) and $[m_1]$ be a model with $m_1 = m_0 + 1$ variables such that

$$[m_1] = [m_0] \cup \{\text{one more variable} \notin [m_0]\}.$$

$$\mathcal{H}_0 : \text{ the model is } [m_0] \quad vs \quad \mathcal{H}_1 : \text{ the model is } [m_1]$$

Sudy: When $[m_0]$ is chosen at the expense of $[m_1]$, *i.e* we look for a test statistic

$$\{T \leq q\}$$

where $q = C_\alpha > 0$ is a constant which depends of the level $\alpha \in (0, 1)$ of the test.

Now, let's describe various criterions for choosing between these two nested models $[m_0]$ and $[m_1]$ in view of the data.

Section 3

3. criteria

3.1. Fisher test for nested models.

Theorem We assume $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ and X full rank. Let $\alpha \in]0,1[$. The statistics

$$T = \frac{\text{RSS}\ (m_0) - \text{RSS}\ (m_1)}{\text{RSS}\ (m_1)} \times (n - m_0 - 1)$$

allows us to test

$$\mathcal{H}_0 : \text{ "the model is } [m_0] \subset [m1] \text{"} \quad \text{vs} \quad \mathcal{H}_1 : \text{ "the model is } [m_1] \text{"}$$

Indeed, if $T \leq f_{1,n-m_0-1,1-\alpha}$ with $f_{1,n-m_0-1,1-\alpha}$ the quantile of order $(1-\alpha)$ of th Fisher law at $(1, n - m_0 - 1)$ ddl, then the model $[m_0]$ must be chosen at a level of risk α.

Proof: Trivial with the nested theorem.

3.2. The determination coefficient R^2

It is recalled that,for a model $[m]$ of size m

$$R^2(m) = 1 - \frac{\text{RSS }(m)}{\text{TSS}}.$$

Proposition $\mathcal{H}_0$: "$[m_0] \subset [m1]$" vs $\mathcal{H}_1$: "$[m_1]$"

$$T := R^2(m_1) - R^2(m_0) = \frac{\text{RSS }(m_0) - \text{RSS }(m_1)}{\text{TSS}} \geq 0.$$

Proof : Trivial.

Comment:

☛ In general, we do not use the R^2 as a selection criterion because it will always increase with the number of variables.

☛ Used to compare two models with the same number of variables.

3.3. The adjusted determination coefficient R_a^2

It is recalled that,for a model $[m]$ of size m

$$R_a^2(m) = 1 - \frac{(n-1)(1-R^2(m))}{n-m} = 1 - \frac{\text{RSS }(m)}{n-m} \times \frac{(n-1)}{\text{TSS}}.$$

Proposition $\mathcal{H}_0$: "$[m_0] \subset [m1]$" vs $\mathcal{H}_1$: "$[m_1]$"

$$R_a^2(m_0) \geq R_a^2(m_1) \iff T := \frac{\text{RSS }(m_0) - \text{RSS }(m_1)}{\text{RSS }(m_1)} \times (n - m_0 - 1) \leq 1.$$

Proof : trivial as

$$R_a^2(m_0) \geq R_a^2(m_1) \iff \frac{\text{RSS }(m_0)}{n - m_0} \leq \frac{\text{RSS }(m_1)}{n - m_0 - 1}$$

Comment:

☛ This helps to correct the disadvantages of the R^2 coefficient.

3.4. The C_p of Mallows

For all model $[m]$, we denote $\widehat{Y}_m = P_m Y$. It is recalled that $\text{RSS}\,(m) = \|P_m Y - Y\|^2$ and

$$\text{RSS}\,(m) = \|P_m Y - Y\|^2 \neq \|P_m Y - X\beta\|^2.$$

Definition Let $[m]$ be any model. The Mallows criterion associated with $[m]$ is defined by:

$$C_p(m) = \frac{\text{RSS}\,(m)}{\widehat{\sigma}^2} - n + 2m.$$

We can show that

(a) $\mathbb{E}[\text{RSS}\,(m)] = \mathbb{E}[\|\widehat{Y}_m - Y\|^2] = \|(\boldsymbol{I} - P_m)X\beta\|^2 + (n-m)\sigma^2.$

(b) $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2] = \|(\boldsymbol{I} - P_m)X\beta\|^2 + m\sigma^2.$

(c) $\mathbb{E}[C_p(m)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_m - X\beta\|^2].$

Proof : Will be proved in lecture class.

Comments

☛ **Unbiased estimator of the mean quadratic error:** We deduce from (c) that $C_p(m)\widehat{\sigma}^2$ is an unbiased estimator of the unknown mean quadratic prediction error $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2]$.

☛ **Minimisation of the criterion:** For any model $[m]$, the mean squared error of $\widehat{Y}_m$ is $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2]$. Ideally, it is a good criterion for estimating the estimator $\widehat{Y}_m$. Selecting a good $[m]$ model is like minimizing

$$m \longmapsto \mathbb{E}[\|\widehat{Y}_m - X\beta\|^2].$$

Unfortunately, this quantity depends on the unknown parameter β. We have at our disposal an unbiased estimator of this quantity. We could then minimize

$$m \longmapsto C_p(m)\widehat{\sigma}^2.$$

Since $\widehat{\sigma}^2$ does not depend on the model, it is natural, especially when trying to estimate $X\beta$ to minimize

$$m \longmapsto C_p(m).$$

Discussion {(A penalized criterion)}

We defined the C_p of Mallows criterion as follows

$$C_p(m)\widehat{\sigma}^2 = \text{RSS } (m) + 2m\widehat{\sigma}^2 - n\widehat{\sigma}^2 := \text{RSS } (m) + \text{pen}(m).$$

When studying the classic R^2, it appeared that the more variables were added, the more the RSS decreased:

$$m \text{ increases} \Rightarrow \text{RSS } (m) \text{ decreases.}$$

Adding a penalty $\text{pen}(m) := 2m\widehat{\sigma}^2$ to the RSS (m) in the criterion is an alternative way to the adjusted R^2 to counterbalance this effect

$$m \text{ increases} \Rightarrow \text{pen}(m) \text{ increases.}$$

We say that we **penalize the big models**.

Discussion {(Adding useless variables to the real model)}

Set that the "real" model denoted by $[m^*]$ is included in the model $[m_0]$, then

$$X\beta = P_{m_0}X\beta \Leftrightarrow X\beta - P_{m_0}X\beta = 0_n.$$

The equation (a) then becomes : $\mathbb{E}[\text{RSS }(m_0)] = \mathbb{E}[\|\widehat{Y}_{m_0} - Y\|^2] = (n - m_0)\sigma^2$ and we have

$$\text{RSS }(m_0) \approx (n - m_0)\widehat{\sigma}^2$$

Equations (b) and (c) give then: $\mathbb{E}[C_p(m_0)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2] = m_0\sigma^2$. Therefore,

$$C_p(m_0) \approx m_0.$$

Thus, if we add useless variables (increases m_0) to the true model (included in $[m_0]$), then RSS $(m_0) \approx (n - m_0)\sigma^2$ will not significantly decrease compared to the $C_p(m_0) \approx m_0$ which will increase more significantly.

Discussion {(Forgetting important variables to the real model)}

If the "real" model $[m^*]$ is not fully included in $[m_0]$ then

$$X\beta \neq P_{m_0}X\beta \Leftrightarrow X\beta - P_{m_0}X\beta = C.$$

So with the same reasoning as before we have:

(a) $\mathbb{E}[\text{RSS }(m_0)] = \mathbb{E}[\|\widehat{Y}_{m_0} - Y\|^2] = C + (n - m_0)\sigma^2$.

(b) $\mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2] = C + m_0\sigma^2$.

(c) $\mathbb{E}[C_p(m_0)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2]$.

We have then

$$\text{RSS }(m_0) \approx (n - m_0)\widehat{\sigma}^2 + C \quad \text{ et } C_p(m_0) \approx m_0 + C$$

where $C > 0$. In this case, $C_p(m_0) > m_0$.

To resume

- If we add useless variables to the "real" model, then $C_p(m_0) \approx m_0$.
- If we forget important variables to the "real" model, then $C_p(m_0) \approx m_0 + C$. where $C > 0$.

So if beyond the problem of estimating $X\beta$, we are interested, by the detection of the good variables, we will be interested in models $[m_0]$ such that $C_p(m_0) \leq m_0$.

☛ It should be noted that the previous interpretations are only true if the choice of the model (selection of the optimal $[m]$) is independent of the data (computation of $\widehat{Y}_m = P_m Y$), so we must cut the sample in 2:

- A sample for the learning to compute $\widehat{Y}_m$ for all $[m]$.
- Another sample for validation to select $[m_{optimal}]$.

C_p of Mallows

$$
\begin{aligned}
C_p(m_0) \leq C_p(m_1) \quad &\Longleftrightarrow \quad \frac{\text{RSS }(m_0)}{\widehat{\sigma}^2} \leq \frac{\text{RSS }(m_1)}{\widehat{\sigma}^2} + 2 \\
&\Longleftrightarrow \quad \frac{\text{RSS }(m_0) - \text{RSS }(m_1)}{\widehat{\sigma}^2} \leq 2.
\end{aligned}
$$

If $\widehat{\sigma}^2$ is replaced by RSS $(m_1)/(n - m_0 - 1)$, then the following condition appears

Proposition $\mathcal{H}_0$: "$[m_0] \subset [m1]$" vs $\mathcal{H}_1$: "$[m_1]$". We choose $[m_0]$ if

$$
T := \frac{\text{RSS }(m_0) - \text{RSS }(m_1)}{\text{RSS }(m_1)} \times (n - m_0 - 1) \leq 2.
$$

3.5. AIC/BIC criteria

Consider a linear regression model $Y = X\beta + \varepsilon$, where

$$\mathrm{rank}(X) = p, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{V}\mathrm{ar}(\varepsilon) = \sigma^2 I \quad \text{and } \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

The log of the likelihood of the model is :

$$\log L(Y, \beta) = -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

Let $\widehat{\beta} = (X^\top X)^{-1} X^\top Y = \widehat{\beta}^{MLE}$. Then, by définition the maximized likelihood (ML) is $L(Y, \widehat{\beta})$.

Proposition The model $[m]$ that maximizes the maximized likelihood on m is the model that minimizes

$$m \longmapsto \mathrm{RSS}\,(m).$$

Proof : Will be proved in Lecture class. $\square$

Comments

- We have seen that minimizing the RSS is not necessarily the best thing to do because it amounts to taking the largest model ($p = n = m$).
- As for the C_p of Mallows, we want to add a (positive) penalty to penalize the big models.

AIC/BIC criteria

Definition

- The AIC of a model $[m]$ is defined by

$$AIC(m) = \frac{n}{2}\log(\mathrm{RSS}\ (m)) + m.$$

- The BIC of a model $[m]$ is defined by

$$BIC(m) = n\log(\mathrm{RSS}\ (m)) + \log(n) \times m.$$

Comments:

☛ We choose the model $[m]$ that minimizes

$$m \longmapsto AIC(m) \quad \text{or} \quad m \longmapsto BIC(m)$$

☛ If $n > 7$ ($\Rightarrow \log(n) > 2$) then the BIC will tend to select models smaller than those selected by AIC.

AIC and BIC

Proposition $\mathcal{H}_0$: "$[m_0] \subset [m1]$" vs $\mathcal{H}_1$: "$[m_1]$". Asymptotically, when $n \to +\infty$,

$$\textbf{(AIC)} : T := \frac{\text{RSS}(m_0) - \text{RSS}(m_1)}{\text{RSS}(m_1)} \times (n - m_0 - 1) \leq \frac{2}{n} \times (n - m_0 - 1).$$

$$\textbf{(BIC)} : T := \frac{\text{RSS}(m_0) - \text{RSS}(m_1)}{\text{RSS}(m_1)} \times (n - m_0 - 1) \leq \frac{\log n}{n} \times (n - m_0 - 1).$$

Proof We minimize the function $C : m \longmapsto \log(\text{RSS}(m)) + f(n)m$ where $f_{AIC}(n) = 2/n$ and $f_{BIC}(n) = \log(n)/n$. Then, we have

$$\begin{aligned} C(m_0) \leq C(m_1) \quad &\Longleftrightarrow \quad \log(\text{RSS}(m_0)) - \log(\text{RSS}(m_1)) \leq f(n) \\ &\Longleftrightarrow \quad \frac{\text{RSS}(m_0) - \text{RSS}(m_1)}{\text{RSS}(m_1)} \times (n - m_0 - 1) \\ &\qquad \leq (e^{f(n)} - 1) \times (n - m_0 - 1). \end{aligned}$$

Section 4

4. Comparaison of criteria

Comparaison of criteria

For each of the 6 criteria, the study is roughly reduced to

$$T := \frac{\text{RSS}\,(m_0) - \text{RSS}\,(m_1)}{\text{RSS}\,(m_1)} \times (n - m_0 - 1) \leq q \quad \text{with}$$

- $q = 4$ for the Fisher test.
- $q = -\infty$ for the R^2 coefficient.
- $q = 1$ for the adjusted R_a^2 coefficient.
- $q = 2$ for the C_p of Mallows.
- $q = \frac{2}{n} \times (n - m_0 - 1)$ for the AIC.
- $q = \frac{\log n}{n} \times (n - m_0 - 1)$ for the BIC.

Comparaison of criteria : Comments

- This roughly orders each of the criteria: the most favorable to $[m_0]$ is the BIC criterion, the most favorable to $[m_1]$ is the R^2 coefficient. We must be wary of these comparisons because they still depend on the value of n, of $\widehat{\sigma}^2$,...
- It should be remembered that for the Fisher test, the criterion for two models can only be compared if one model contains the other (nested models).

Section 5

5. Step-by-step method

Step-by-step method

- Minimization of criterion can be a delicate task when p is high (2^{p-1} different models,all containing $\mathbf{1}_n$).
- Exhaustive search is not possible (either because we want to use Fisher's test, or because p is too big).
- We can use a step-by-step method combined with one of the 6 criteria previously studied.

Disadvantage: Do not test all possible combinations (Global minimum is not guaranteed}

Three famous step-by-step methods (intercept is alawys included) are :

- **Forward selection**
- **Backard selection**
- **Stepwise selection/both selction**

Forward selection

We start with the model resume to the intercept $\mathbf{1}_n$. At each step, a regressor/variable is added to the model, the one with the best contribution (*i.e.* the ones which improves the chosen criterion). We stop when the criterion can not be improved by adding a new regressor/variable.

Backard selection

We stat the "biggest" model whose intercept. At each step, a regressor/variable is removed to the model, the one which improves the chosen criterion. We stop when the criterion can not be improved by removing a new regressor/variable.

Stepwise selection/both selection

This is the same method as the Forward selection method, except that at each step, a regressor/variable present in the model can be challenged (removed or added).

Section 6

6. Illustrative example under R

Car consumption dataset

Explain and predict gas consumption (in liters per 100 km) of different automobile models based on the following variables:

- `Type` = Type of the vehicle.
- `Consommation` = Fuel consumption in liters per 100 km.
- `Prix` = Vehicle price in Swiss francs.
- `Cylindree` = Cylinder capacity in cm3.
- `Puissance` = Power in kW.
- `Poids` = Weight in kg.

Response variable Y is `Consommation`. The covariates X_j correspond to the other 4 variables.

Linear regression model with `lm`

$$Y = \beta_0 \mathbf{1}_n + \beta_1 X_1 + \cdots + \beta_1 X_4 + \mathbf{E} = X\beta + \varepsilon,$$

where $\beta := (\beta_0, \beta_1, \cdots, \beta_4)^\top$, $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ (**[P1]–[P4]** statisfied).

- By default **R** adds an intercept (a column of 1). Here, the design matrix X is a matrix of size $n \times p$ with $p = 5$.
- The order in which variables are entered gives the indice j of the regressor X_j.

```
reg = lm(Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit)
```

Question: : Are all the predictors relevant?

```
library(MASS)
```

Forward method

```
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids,
        trace=F,direction=c('forward'))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_v
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

Backward method

```
stepAIC(reg,~.,trace=F,direction=c("backward"))
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Puissance + Poids, data = conso_v
##
## Coefficients:
## (Intercept)         Prix    Puissance        Poids
##   2.499e+00    1.852e-05    2.013e-02    3.735e-03
```

Both method

```
stepAIC(reg0,Consommation~Prix+Cylindree+Puissance+Poids,
        trace=F,direction=c("both"))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_v
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

Step by step methodes BIC

The command `k=log(n)` has to be added if we want to use BIC criterion (AIC is by default).

```
dim(conso_voit)
```

```
## [1] 31  5
```

```
n=31
```

Forward method

```r
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids
        ,trace=F,direction=c('forward'),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_v
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

Backward method

```
stepAIC(reg,~.,trace=F,direction=c("backward"),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Puissance + Poids, data = conso_v
##
## Coefficients:
## (Intercept)         Prix    Puissance        Poids
##   2.499e+00    1.852e-05    2.013e-02    3.735e-03
```

Both method

```
stepAIC(reg0,Consommation~Prix+Cylindree+Puissance+Poids,
        trace=F,direction=c("both"),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_v
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

To conclude

Note that in our example the given result of the 3 methods is the same even for 2 differents criteria. It is not always the case.