MAP553 Regression

Chapter 5: Model selection

Contents

5.1. 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" 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).

Definition 1 *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:*

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

Comment:

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

Proposition 1 *For all $\theta \in \mathbb{R}^p$:*

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

Proof : Obvious. □

To illustrate the purpose of this chapter, let's do some calculations for the following example. We assume the model

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

where $X = [X_1 \ X_2]$ is a $n \times 2$ matrix of rank 2. Let $\beta = (\beta_1, \beta_2)^T \in \mathbb{R}^2$ and such that $\beta_2 \neq 0$. One may wonder if the X_2 varible is useful, and study the case where we would consider $\beta_2 = 0$ even if it is false, and look for when to omit an explanatory variable can be advantageous in terms of risk .

Let define the following model

$$Y = X_1\beta_1 + \varepsilon, \tag{2}$$

on which the OLSE is determined for the estimate of β_1, which is

$$\widetilde{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y,$$

where Y is defined by the model~(1), thus $\widetilde{\beta}_1$ is biased. Denote by $\hat{\beta}$, the OLSE of the estimation of β calculate from the model~(1). Thus, we have 2 estimators, one biased and the other one unbiased

$$\widetilde{\beta} = (\widetilde{\beta}_1, 0)^T \text{ and } \hat{\beta} = (X^T X)^{-1} X^T Y.$$

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

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

where D denote the determinant of the matrix $(X^TX)^{-1}$.

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

$$(X^TX)^{-1} = \frac{1}{D}\begin{pmatrix} \|X_2\|^2 & -X_1^TX_2 \\ -X_1^TX_2 & \|X_1\|^2 \end{pmatrix},$$

where $D := \|X_1\|^2\|X_2\|^2 - (X_1^TX_2)^2 > 0$.

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

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

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

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\beta}-\beta\|^2] &= \sum_{j=1}^{2}\mathbb{E}[(\widetilde{\beta}_j-\beta_j)^2] = \mathbb{E}[(\widetilde{\beta}_1-\beta_1)^2] + \beta_2^2 = \mathbb{E}[((X_1^TX_1)^{-1}X_1^TY-\beta_1)^2] + \beta_2^2 \\
&= \mathbb{E}[((X_1^TX_1)^{-1}X_1^T(\beta_1X_1+\beta_2X_2+\varepsilon)\beta_1)^2] + \beta_2^2 = \left((X_1^TX_1)^{-1}X_1^TX_2\right)^2\beta_2^2 + \sigma^2(X_1^TX_1)^{-1} + \beta_2^2 \\
&= \frac{\sigma^2}{\|X_1\|^2} + \beta_2^2\left(1 + \frac{(X_1^TX_2)^2}{\|X_1\|^4}\right).
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[(\hat{\beta}-\beta)^2] - \mathbb{E}[\|\widetilde{\beta}-\beta\|^2] &= \frac{\sigma^2}{D}\left(\|X_2\|^2 + \|X_1\|^2\right) - \frac{\sigma^2}{\|X_1\|^2} - \beta_2^2\left(1 + \frac{(X_1^TX_2)^2}{\|X_1\|^4}\right) \\
&> \frac{\sigma^2\|X_1\|^2}{D} - \beta_2^2\left(1 + \frac{(X_1^TX_2)^2}{\|X_1\|^4}\right).
\end{aligned}$$

□

In the following, we will be interested in methods to choose a set of variables (which we will call **model**). While it may be easy to decide between two models, the question of model choice is more delicat. Indeed,

- 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 $\mathbb{1}_n$ (always take the intercept), then we have $\sum_{j=0}^{8} C_j^8 = 2^8 = 256$ possible models to compare.

More specifically, we will focus on methods that rely on the following tools/criterions:

- Tests between nested models
- R^2
- R_a^2 adjusted
- C_p of Mallows
- AIC- criterion
- BIC- criterion

5.2. Notations and illustrative example

We note $p = q + 1$ the number of explanatory variables (the intercept $\mathbb{1}_n$ included). We define

$$X = (\mathbb{1}_n, X_1, \ldots, X_q).$$

We denote by $[m]$ any model of size m, *i.e.* $m := \text{card}([m])$. We consider the framework of linear regression models.

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

where $\text{rang}(X) = p, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{V}\text{ar}(\varepsilon) = \sigma^2 I.$

We define for all model $[m]$

$$RSS(m) = \|Y - P_m Y\|^2,$$

where P_m is the matrix of orthogonal projection into the space generated by thethe variables of $[m]$.

For $1 \leq m_0 \leq p-1$, we define $[m_0]$ a model composed by m_0 variables (the intercept $\mathbb{1}_n$ is considered to be in the model). Let $[m_1]$ be a model with $m_1 = m_0 + 1$ variables such that

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

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

5.3. criterions

5.3.1. Fisher test for nested models.

As part of this approach, two different test statistics are given:

$$\begin{aligned} F &= \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1), \\ \widetilde{F} &= \frac{RSS(m_0) - RSS(m_1)}{\hat{\sigma}^2} = \frac{RSS(m_0) - RSS(m_1)}{RSS} \times (n - p) \end{aligned}$$

Theorem 1 *We assume Y to be gaussian. Let* $\alpha \in]0,1[$. *The statistics F and* $\widetilde{F}$ *allow us to test*

$$H_0 : \text{"the model is } [m_0]\text{"} \quad cvs \quad H_1 : \text{"the model is} [m_1]\text{"}$$

Indeed, :

- *If* $F > f_{1,n-m_0-1,1-\alpha}$, *then the model* $[m_1]$ *must be chosen at a level of risk* α.
- *If* $\widetilde{F} > f_{1,n-p,1-\alpha}$, *then the model* $[m_1]$ *must be chosen at a level of risk* α.

Comment:

☛ Note that it is difficult to compare these two results. The previous theorem is only valid under the condition $[m_0] \subset [m_1]$.

Proof : We recall that

$$RSS = \|Y - P_X Y\|^2 = (n-p)\hat{\sigma}^2.$$

The Pythagore theorem gives by projecting $Y - P_{m_0}Y$ into $[m_1]$:

$$\|Y - P_{m_0}Y\|^2 = \|P_{m_1}Y - P_{m_0}Y\|^2 + \|Y - P_{m_1}Y\|^2,$$

which is equivalent to :

$$RSS(m_0) = \|P_{m_1}Y - P_{m_0}Y\|^2 + RSS(m_1).$$

Then, we deduce that

$$\begin{aligned} F &= \frac{\|P_{m_1}Y - P_{m_0}Y\|^2/1}{\|Y - P_{m_1}Y\|^2/(n-m_0-1)}, \\ \widetilde{F} &= \frac{\|P_{m_1}Y - P_{m_0}Y\|^2/1}{\|Y - P_X Y\|^2/(n-p)}. \end{aligned}$$

Using the quantiles of Fisher's law, Theorem~3 and the theorem~6, we obtain the result. □

5.3.2. The determination coefficient R^2

It is recalled that, in general, the definition of the coefficient of determination R^2 is

$$R^2 = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2},$$

with $\mathbb{1}_n = (1,1,\ldots,1)^T$, $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\widehat{Y} = X\widehat{\beta}$. Also, for any model $[m]$, we define

$$R^2(m) = \frac{\|P_m Y - \overline{Y}\mathbb{1}_n\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{\|Y - P_m Y\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$

Denoting

$$TSS = \|Y - \overline{Y}\mathbb{1}_n\|^2, \quad RSS(m) = \|Y - P_m Y\|^2,$$

we have :

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

We deduce the following property:

Proposition 3 *We have :*

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

Proof : Equality is obvious, inequality follows from Pythagore's theorem. □

Comment:

- ☛ In general, we do not use the R^2 as a selection criterion because it will always increase with the number of variables. But it is an indicative criterion when it remains constant and the number of variables is increased. It is also used to compare two models with the same number of variables

5.3.3. The adjusted determination coefficient R_a^2

It is recalled that, in general, the definition of the adjusted R^2 coefficient of determination is

$$R_a^2 = 1 - \frac{(n-1)\|Y - \widehat{Y}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{(n-1)(1-R^2)}{n-p}.$$

We set, for all model $[m]$,

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

The $R_a^2(m)$ is therefore a function of the sum of residual squares divided by the number of degrees of freedom. Note that if m increases then $RSS(m)$ decreases and $n-m$ decreases. This helps to correct the disadvantages of the R^2 coefficient.

5.3.4. The C_p of Mallows

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

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

Definition 2 *Let* $[m]$ *be any model. The Mallows criterion associated with* $[m]$ *is defined by:*

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

We can show that

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

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

(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 around the criterion:

- **A penalized criterion:** We defined the C_p of Mallows criterion as follows
$$C_p(m)\widehat{\sigma}^2 = RSS(m) + 2m\widehat{\sigma}^2 - n\widehat{\sigma}^2 := RSS(m) + \text{pen}(m).$$
When studying the classic R^2, it appeared that the more variables were added, the more the RSS decreased:
$$m \text{ increases} \Rightarrow RSS(m) \text{ decreases.}$$
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**.

- **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 : $\mathbb{E}[RSS(m_0)] = \mathbb{E}[||\widehat{Y}_{m_0} - Y||^2] = (n - m_0)\sigma^2$ and we have

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

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

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

(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

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

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 $C_p(m_0) \leq m_0$.

Important Comment:

- ☛ 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 $[m_{optimal}]$.

5.3.5. AIC/BIC criterion

Consider a linear regression model

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

where

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

The likelihood of the model is :

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

Let $\widehat{\beta} = (X^T X)^{-1} X^T Y$ be the OLSE which is in our gaussian setting the ordinary maximum likelihood estimator (OMLE). Then, by définition the OMLE $\widehat{\beta}$ maximize the likelihood. In others words, the maximized likelihood (ML) is

$$L(Y, \widehat{\beta}).$$

Proposition 4 *The model* $[m]$ *that maximizes the maximized likelihood (ML) on m is the model that minimizes*

$$m \longmapsto RSS(m).$$

Proof : Will be proved in Lecture class. □

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.

For the AIC criterion, the penalty is simply $\text{pen}(m) = m$ and for the BIC criterion, the penalty is $\text{pen}(m) = \frac{\log n}{2} \times m$.

Definition 3

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

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

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

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

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.

5.4. Comparaison of criterions

The purpose of this section is to compare the criterions in the previous section. For that we will consider two nested models $[m_0] \subset [m_1]$ such that $m_1 = m_0 + 1$. We set

$$H_0 : \text{ the model is } [m_0] \quad vs \quad H_1 : \text{ the model is } [m_1]$$

We study the cases where $[m_0]$ 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. In this section, we consider $\alpha = 0.05$ and $n - m_0 - 1 \geq 16$.

Fisher Test :

$$T := F = \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq 4$$

The R^2 coefficient :

In our setting, $[m_1]$ is always chosen at the expense of $[m_0]$.

The adjusted R^2 coefficient :

$$\begin{aligned} R^2_a(m_0) \geq R^2_a(m_1) \quad &\Longleftrightarrow \quad \frac{RSS(m_0)}{n - m_0} \leq \frac{RSS(m_1)}{n - m_0 - 1} \\ &\Longleftrightarrow \quad T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq 1. \end{aligned}$$

The C_p of Mallows :

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

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

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

AIC and BIC criterions:

In this case, we minimize the function

$$C : m \longmapsto \log(RSS(m)) + f(n)m$$

where for AIC criterion $f(n) = 2/n$ and for BIC criterion $f(n) = \log(n)/n$. Then, we have

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

Asymptotically, when $n \to +\infty$,

- **For AIC**

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

- **For BIC**

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

To resume

For each of the 6 criteria, we have roughly reduced ourselves to the study of

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

avec

- $q = 4$ for the Fisher test.
- $q = -\infty$ for the R^2 coefficient.
- $q = 1$ for the adjusted R^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.

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).

5.5. Step-by-step method

We have seen in the introduction that the minimization of criteria can be a delicate task when the number of explanatory variables is high. Indeed, if we have p variables (whose constant vector $\mathbb{1}_n$, the intercept), we have 2^{p-1} different models (all containing $\mathbb{1}_n$). When 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 criterions previously studied. The disadvantage is that it does not test all possible combinations. We are therefore not sure of obtaining a global minimum. The three famous step-by-step methods are the following:

- **Forward selection:** We start with the model resume to the intercept $\mathbb{1}_n$. 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 selction :** 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).

Comments:

☛ Remember that the $\mathbb{1}_n$ intercept/variable is always present for all models.

☛

5.6. Illustrative example under R

Comme back to the example of the Chapter 4.

- Y =`Consommation` = Fuel consumption in liters per 100 km.
- X_1 =`Prix` = Vehicle price in Swiss francs.
- X_2 =`Cylindree` = Cylinder capacity in cm3.
- X_3 =`Puissance` = Power in kW.
- X_4 =`Poids` = Weight in kg.

Consider the following linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \ \forall i = 1, \cdots, n. \tag{3}$$

We recall that we assume the model~(3), under the Rank assumption and under **[P1]**–**[P4]** where

- Errors are centered (linearity of the model in pratice) : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$.

- Errors have homosedasctic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- Errors are uncorralated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j)$.
- Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

First download the dataset.

```
conso_voit = read.table("conso.txt", header=TRUE, sep="\t", dec=",",row.names=1)
```

The linear regression of the `Consommation` variable on the other variables is done using the `lm` function.

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

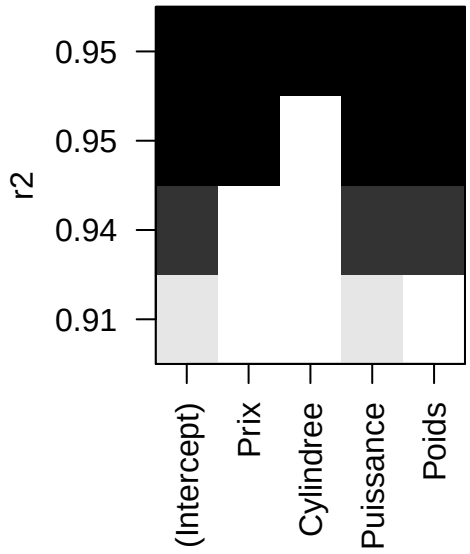
The question is : Are all the predictors relevant? To answer this question, if p is small we can procede as follows.

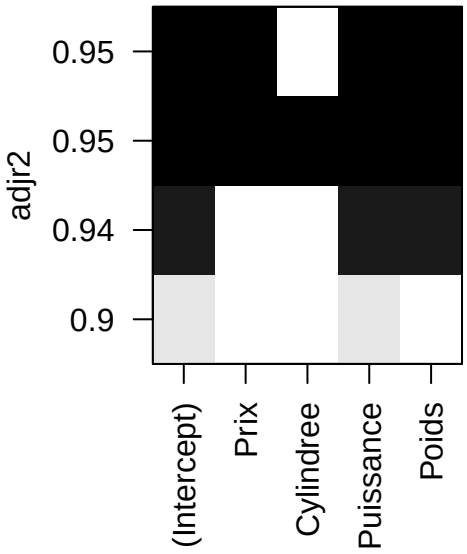
```
library(leaps)
# int=T (to include the intercept)
# nbest=1 (to select one model by dimension)
# nvmax=4 (the maximum number of regressors
   #(except the intercept) to include in a model)
choosen_model=regsubsets(Consommation~Prix+Cylindree+Puissance+Poids,int=T,
                         nbest=1,nvmax=4,method="exhaustive",data=conso_voit)

summary(choosen_model)
```

```
## Subset selection object
## Call: regsubsets.formula(Consommation ~ Prix + Cylindree + Puissance +
##     Poids, int = T, nbest = 1, nvmax = 4, method = "exhaustive",
##     data = conso_voit)
## 4 Variables  (and intercept)
##           Forced in Forced out
## Prix          FALSE      FALSE
## Cylindree     FALSE      FALSE
## Puissance     FALSE      FALSE
## Poids         FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          Prix Cylindree Puissance Poids
## 1  ( 1 ) " "  " "       "*"       " "
## 2  ( 1 ) " "  " "       "*"       "*"
## 3  ( 1 ) "*"  " "       "*"       "*"
## 4  ( 1 ) "*"  "*"       "*"       "*"
```

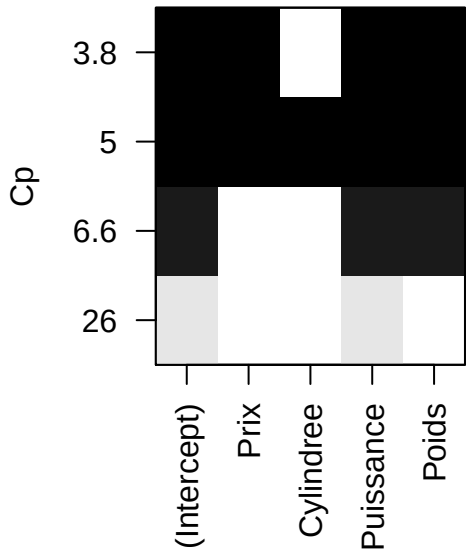
Plot make it easier to understand.

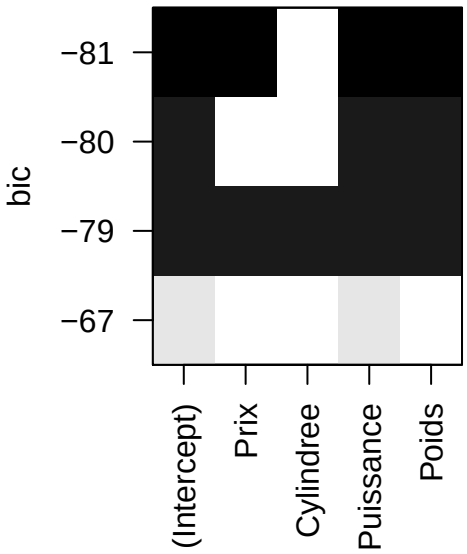
```
par(mfrow=c(1,2))
plot(choosen_model,scale="r2")
plot(choosen_model,scale="adjr2")
```






```
plot(choosen_model,scale="Cp")
plot(choosen_model,scale="bic")
```





5.6.1. Step by step methodes AIC

```
library(MASS)
```

```r
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=T,direction=c('forward'))
```

5.6.1.1. Forward method

```
## Start:  AIC=79.87
## Consommation ~ 1
##
##              Df Sum of Sq    RSS    AIC
## + Puissance  1    346.79  35.35  8.071
## + Cylindree  1    338.37  43.77 14.692
## + Prix       1    303.45  78.69 32.878
## + Poids      1    285.17  96.96 39.351
## <none>                   382.14 79.866
##
## Step:  AIC=8.07
## Consommation ~ Puissance
##
##              Df Sum of Sq    RSS     AIC
## + Poids      1   14.2733 21.077 -5.9605
## + Cylindree  1    3.0114 32.339  7.3104
## <none>                   35.350  8.0706
## + Prix       1    0.0002 35.350 10.0704
##
## Step:  AIC=-5.96
## Consommation ~ Puissance + Poids
##
##              Df Sum of Sq    RSS     AIC
## + Prix       1    3.2053 17.871 -9.0744
## <none>                   21.077 -5.9605
## + Cylindree  1    0.0580 21.019 -4.0460
##
## Step:  AIC=-9.07
## Consommation ~ Puissance + Poids + Prix
##
##              Df Sum of Sq    RSS     AIC
## <none>                   17.871 -9.0744
## + Cylindree  1   0.50652 17.365 -7.9657
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
```

```
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

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

5.6.1.2. Backward method

```
## Start:  AIC=-7.97
## Consommation ~ Prix + Cylindree + Puissance + Poids
##
##             Df Sum of Sq    RSS     AIC
## - Cylindree  1    0.5065 17.871 -9.0744
## <none>                   17.365 -7.9657
## - Prix       1    3.6537 21.019 -4.0460
## - Puissance  1    4.1792 21.544 -3.2805
## - Poids      1   14.9706 32.335  9.3075
##
## Step:  AIC=-9.07
## Consommation ~ Prix + Puissance + Poids
##
##             Df Sum of Sq    RSS     AIC
## <none>                   17.871 -9.0744
## - Prix       1    3.2053 21.077 -5.9605
## - Puissance  1    3.9434 21.815 -4.8934
## - Poids      1   17.4783 35.350 10.0704
##
## Call:
## lm(formula = Consommation ~ Prix + Puissance + Poids, data = conso_voit)
##
## Coefficients:
## (Intercept)         Prix    Puissance        Poids
##   2.499e+00    1.852e-05    2.013e-02    3.735e-03
```

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

5.6.1.3. Both method

```
## Start:  AIC=79.87
## Consommation ~ 1
##
##             Df Sum of Sq    RSS    AIC
```

```
## + Puissance  1    346.79  35.35  8.071
## + Cylindree  1    338.37  43.77 14.692
## + Prix       1    303.45  78.69 32.878
## + Poids      1    285.17  96.96 39.351
## <none>                   382.14 79.866
##
## Step:  AIC=8.07
## Consommation ~ Puissance
##
##             Df Sum of Sq    RSS    AIC
## + Poids      1     14.27  21.08 -5.961
## + Cylindree  1      3.01  32.34  7.310
## <none>                    35.35  8.071
## + Prix       1      0.00  35.35 10.070
## - Puissance  1    346.79 382.14 79.866
##
## Step:  AIC=-5.96
## Consommation ~ Puissance + Poids
##
##             Df Sum of Sq    RSS    AIC
## + Prix       1     3.205 17.871 -9.074
## <none>                   21.077 -5.961
## + Cylindree  1     0.058 21.019 -4.046
## - Poids      1    14.273 35.350  8.071
## - Puissance  1    75.888 96.964 39.351
##
## Step:  AIC=-9.07
## Consommation ~ Puissance + Poids + Prix
##
##             Df Sum of Sq    RSS     AIC
## <none>                   17.871 -9.0744
## + Cylindree  1    0.5065 17.365 -7.9657
## - Prix       1    3.2053 21.077 -5.9605
## - Puissance  1    3.9434 21.815 -4.8934
## - Poids      1   17.4783 35.350 10.0704
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

5.6.2. Step by step methodes BIC

Recall that the size of the sample is $n = 31$. Here note that `trace=F`, then the details won't appear. 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
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=F,direction=c('forward'),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## 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_voit)
##
## 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,data=conso_voit,
        trace=F,direction=c("both"),k=log(n))
```

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

5.6.3 To conclude

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