Modélisation statistique
Régression multivariée

aurore.lavigne@univ.lille.fr

Partie 2 : Régression linéaire multiple

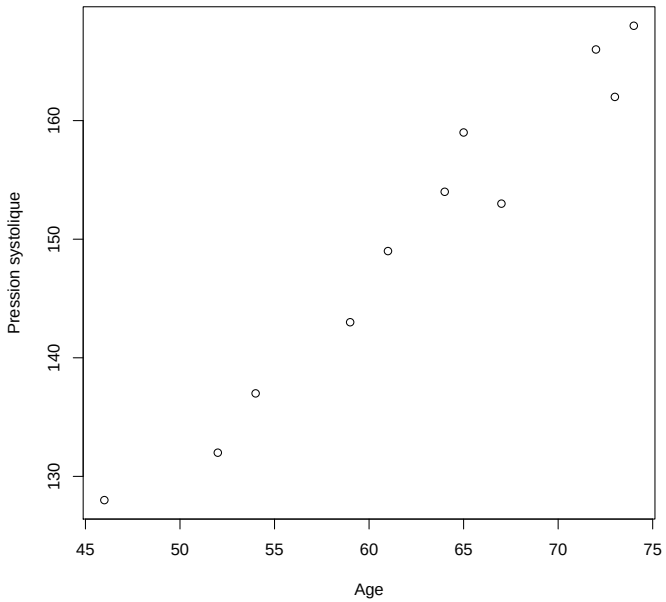
Objectifs

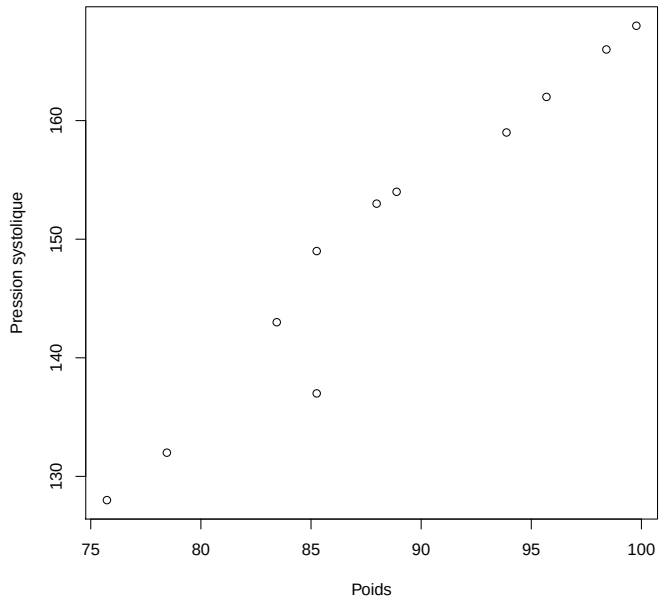
- L'objectif de la régression linéaire simple est d'expliquer une variable Y par une fonction affine d'une variable X.
- On cherchera à répondre aux questions suivantes :
  - Quel est le pouvoir explicatif du modèle ?
    - La variable X a t-elle un apport significatif dans l'explication des valeurs de Y ?
    - Cet apport est-il suffisamment grand pour être transposable à la population ?
  - Quelles sont les propriétés (notamment la précision) des paramètres estimés du modèle (biais, variance)
  - Quelle sera la qualité de la prédiction des valeurs de Y à partir de l'observation des valeurs de X ?

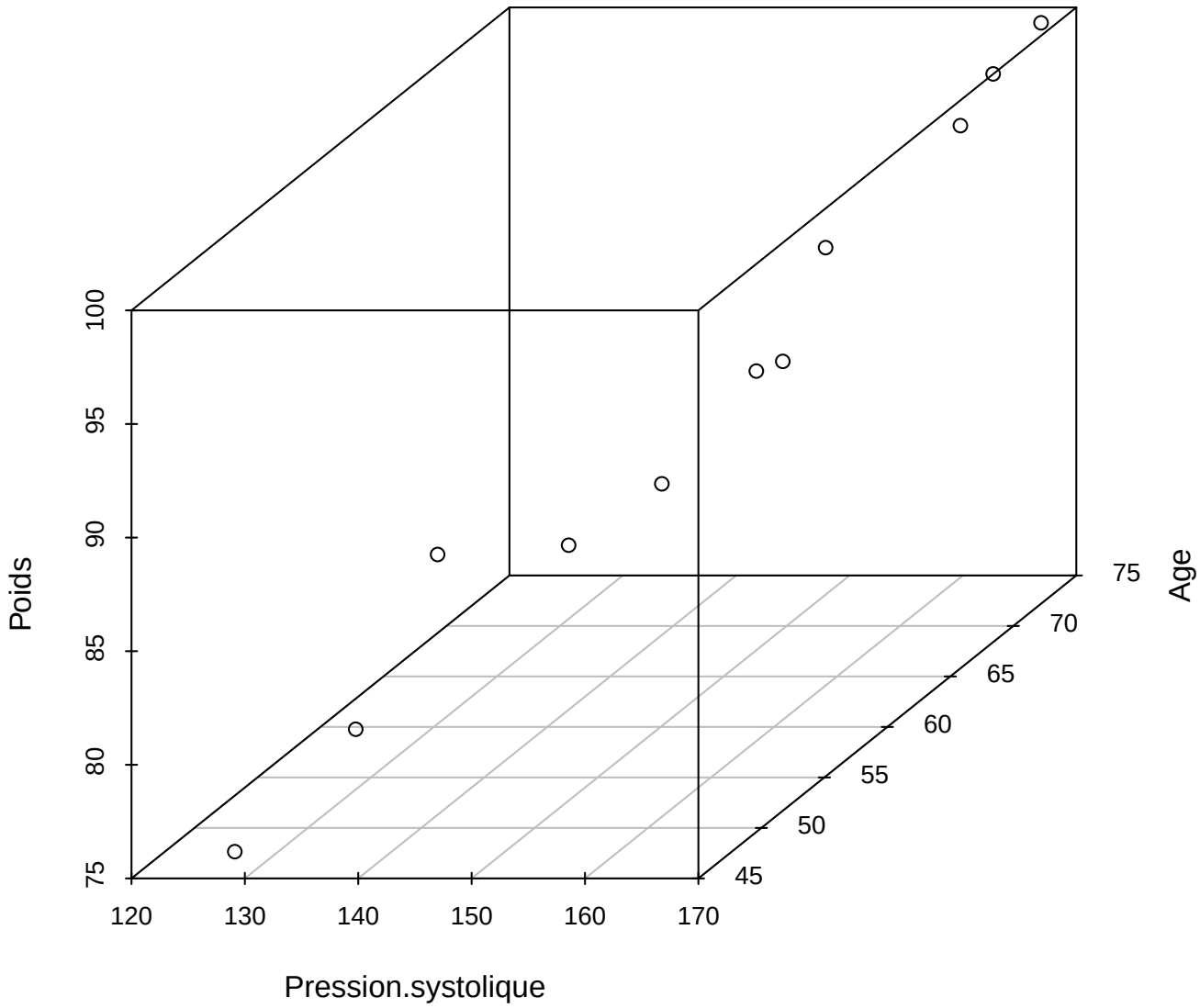
Plan

Exemple : pression systolique

	Pression.systolique	Age	Poids
1	132	52	78.46
2	143	59	83.44
3	153	67	87.98
4	162	73	95.69
5	154	64	88.89
6	168	74	99.77
7	137	54	85.26
8	149	61	85.26
9	159	65	93.87
10	128	46	75.73
11	166	72	98.41


Pression systolique
130
140
150
160
45
50
55
60
65
70
75
Age


Pression systolique
130
140
150
160
75
80
85
90
95
100
Poids


Poids
100
95
90
85
80
75
120
130
140
150
160
170
Pression.systolique
45
50
55
60
65
70
75
Age

Le modèle étudié

- Le modèle de régression multiple est une généralisation à plusieurs facteurs (p) du modèle simple. Il s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_1 x_{ip} + \varepsilon_i, i = 1 \ldots, n. \quad (1)$$

- La terminologie reste la même et on suppose que $n > p + 1$.
  - $\mathbf{Y}$: Variable d'intérêt
  - $\mathbf{x}_1, \ldots, \mathbf{x}_p$: variables explicatives
  - $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$: Vecteur des coefficients ou vecteur des paramètres.

- Chaque individu est décrit par p variables, formant un vecteur de dimension p (matrice $p \times 1$), appelé vecteur individu ou vecteur des observations.

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p.$$

1. Les ε_i sont i.i.d ; pour tout $i = 1, \ldots, n$ $\mathbb{E}(\varepsilon_i) = 0$
2. pour tout $i = 1, \ldots, n$ $\mathbb{V}(\varepsilon_i) = \sigma^2$; $\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$.
3. pour tout $i = 1, \ldots, n$ $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Remarque

La linéarité et le caractère Gaussien du modèle sont des hypothèses qui doivent être validées. Pour les vérifier, on peut soit utiliser la connaissance a priori que l'on a du modèle, soit réaliser un test statistique.

- Le modèle (1) peut s'écrire sous la forme :

$$Y_i = \left[1, \mathbf{x}_i^T\right] \boldsymbol{\beta} + \varepsilon_i, i = 1 \ldots, n \tag{2}$$

Exemples

1. On veut estimer le prix d'un appartement en fonction de sa situation, de sa superficie, de son standing, sa localisation, son ancienneté.
2. On dispose de données concernant l'âge, le kilométrage en milliers de kilomètres et le prix en milliers d'euros pour un échantillon de voitures d'occasion du même type.

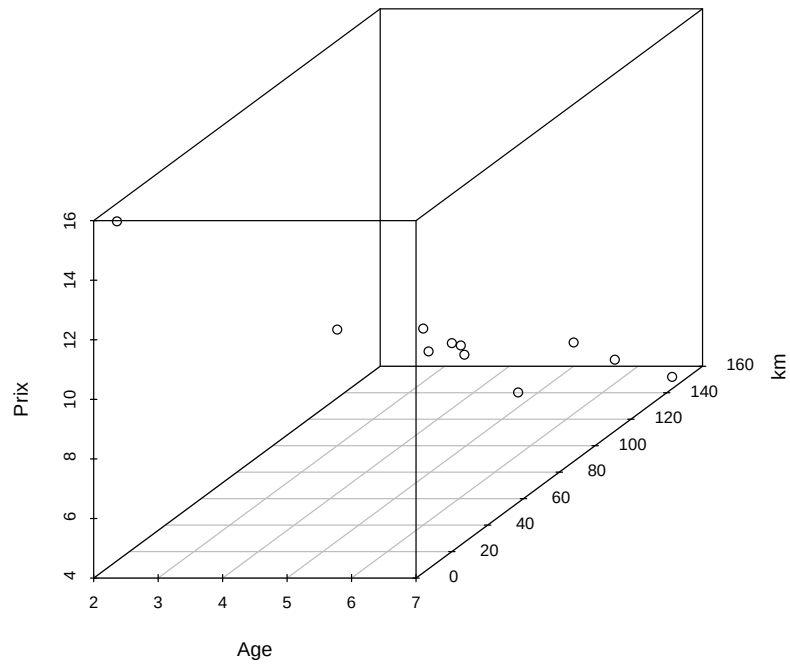
Age	5	4	6	5	5	5	6	6	2	7	7
Km	92	64	124	97	79	76	93	63	13	111	143
Prix	7.8	9.5	6.4	7.5	8.1	9	6.1	8.7	15.4	6.4	4.4

- On est en présence d'un exemple à 2 facteurs auquel on associe le modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

- Ou bien :

$$Y_i = [1 \; x_{i1} \; x_{i2}] \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i$$



- La vérification visuelle semble effectivement proche d'un plan.
- Cependant l'interprétation est délicate en présence de 3 facteurs ou plus.

Notation matricielle

- On peut écrire le modèle en notations matricielles. Ainsi en déroulant l'équation (2), on obtient :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

avec :

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} ; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} ; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

et

$$\mathbf{X} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}}_{\text{matrice } n \times (p+1)} = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \\ 1 & x_i^T \\ \vdots & \\ 1 & x_n^T \end{bmatrix}.$$

Remarque

- Comme la matrice $\mathbf{X}$ est de taille $n \times (p+1)$ et que $n > p+1$ ceci implique que $\mathbf{X}$ est de rang maximum $(p+1)$ si les $p+1$ colonnes de $\mathbf{X}$ sont linéairement indépendantes.
- Autrement dit, on suppose qu'aucune variable explicative ne peut être exprimée comme fonction linéaire des $p-1$ autres variables explicatives (on parle d'absence de co-linéarité entre les co-variables).
- Ceci correspond à un problème d'identifiabilité sur le problème posé.
- Dans le cas contraire, le problème serait mal posé.

Exemple : prenons le cas $p = 2$ et $x_2 = \alpha_1 x_1 + \alpha_0$. Alors

$$Y_i = \beta_0 + \beta_2 \alpha_0 + (\beta_1 + \alpha_1 \beta_2) x_{i1}.$$

On est ainsi ramené à un problème de régression à un facteur.

Plan

Estimation de $\boldsymbol{\beta}$ par la méthode des MCO.

- La distance verticale entre le nuage de points dans $\mathbb{R}^{p+1}$ et l'hyperplan de régression est :

$$F(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

But :

Trouver $\widehat{\boldsymbol{\beta}}$ tel que $F(\widehat{\boldsymbol{\beta}})$ soit le minimum de $F(\boldsymbol{\beta})$, i.e.

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|Y - X\boldsymbol{\beta}\|^2.$$

Dérivation matricielle

Définition

Soit $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ une application linéaire et $\mathbf{x} = (x_1, \ldots, x_k)^T$ un vecteur de $\mathbb{R}^k$. On appelle dérivée de f en $\mathbf{x}$ le vecteur colonne défini par

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \cdots \frac{\partial f(\mathbf{x})}{\partial x_i} \cdots \frac{\partial f(\mathbf{x})}{\partial x_k} \right]^T .$$

- On peut montrer que pour tout $\mathbf{d} \in \mathbb{R}^k$, on a :

$$\left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right]^T \mathbf{d} = \underbrace{\lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{d}) - f(\mathbf{x})}{h}}_{\text{dérivée de } f \text{ en } \mathbf{x} \text{ dans la direction } \mathbf{d}} .$$

Propriétés

- Dérivée d'une forme linéaire : si $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ où $\mathbf{a}$ est un vecteur de $\mathbb{R}^k$, alors

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} = \frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}}.$$

- Dérivée d'une forme quadratique : si $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ où A est une matrice carrée symétrique d'ordre k, alors

$$\frac{\partial \mathbf{x}^T A\mathbf{x}}{\partial \mathbf{x}} = 2A\mathbf{x} \ (= 2\mathbf{x} \text{ si } A = I_k).$$

Exercices

1. En prenant $k = 1$ retrouver des résultats connus sur les dérivées.
2. Prouver les propriétés ci-dessus.

Indication : On peut utiliser les dérivées directionnelles.

Application à l'estimation par MCO

- On a :

$$\begin{aligned} F(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = & \quad (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ = & \quad \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \ . \\ = & \quad \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

- Ainsi en utilisant la dérivation matricielle :

$$\frac{\partial \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}.$$

- De même, on obtient aussi :

$$\frac{\partial^2 \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{\partial \boldsymbol{\beta}^2} = 2\mathbf{X}^T\mathbf{X}.$$

Exercice.

1. Quel est le rang de la matrice $\mathbf{X}^T\mathbf{X}$?
2. Montrer que $\mathbf{X}^T\mathbf{X}$ est symétrique définie positive.
3. Calculer les matrices $\mathbf{X}^T\mathbf{X}$ et $\mathbf{X}^T\mathbf{Y}$ lorsque $p = 1$ et lorsque $p = 3$.

Indication : Pour toute matrice $\mathbf{X}$, on a :

$$\text{rang}(\mathbf{X}) = \text{rang}\left(\mathbf{X}^T\right) = \text{rang}\left(\mathbf{X}^T\mathbf{X}\right) = \text{rang}\left(\mathbf{X}\mathbf{X}^T\right).$$

Théorème

L'EMCO de $\boldsymbol{\beta}$ est obtenu en résolvant les équations :

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T Y = 0.$$

Il est donné par la formule

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

- Comme par hypothèse $\text{rang}(\mathbf{X}) = p+1$ c'est à dire $\mathbf{X}$ est de rang maximum, alors $\mathbf{X}^T\mathbf{X}$ est une matrice carré de rang maximum, alors elle est inversible.

Exercices

1. Montrer que l'expression de $\widehat{\boldsymbol{\beta}}$ correspond bien à ce qu'on avait trouvé pour $p = 1$.
2. Retrouver les résultats obtenus dans l'exemple salaire-années.
3. Pour $p = 2$, donner les formes explicites des EMCO en fonction de $(x_{i1}, x_{i2}, Y_i), i = 1, \ldots, n$.
4. Vérifier que pour l'exemple des prix des maisons : $\widehat{\boldsymbol{\beta}} = (16.73, -0.05, -0.87)^T$.

Propriétés de l'EMC

Théorème

(i) $\widehat{\boldsymbol{\beta}}$ est un e.s.b. de $\boldsymbol{\beta}$. Sa matrice de variance-covariance est :

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

(ii) $\widehat{\boldsymbol{\beta}}$ est le meilleur estimateur linéaire sans biais de $\boldsymbol{\beta}$ au sens où sa variance est minimale parmi tous les estimateurs linéaires sans biais de $\boldsymbol{\beta}$.

(iii) Si de plus les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, alors $\widehat{\boldsymbol{\beta}}$ correspond à l'estimateur du "Maximum de vraisemblance" de $\boldsymbol{\beta}$ et

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

Preuve.

Plan

Étude des résidus (erreurs d'ajustement)

- On définit les résidus du modèle par :

$$\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}, i = 1, \ldots, n,$$

que l'on note :

$$\widehat{\mathbf{e}} = \begin{pmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_i \\ \vdots \\ \widehat{e}_n \end{pmatrix} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y} - \widehat{\mathbf{Y}}$$

- En utilisant le fait que

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|\mathbf{Y} - \bar{Y}\mathbf{1}_n\|^2, \text{ avec } \mathbf{1}_n = [\underbrace{11\ldots1}_{n \text{ fois}}]^T,$$

on a :

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = & \left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right)^T\left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right) \\
= & \left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^T\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) + 2\underbrace{\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^T\left(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}_n\right)}_{=0(\text{exercice})} \\
& + \left(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}_n\right)^T\left(\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}_n\right) \\
= & \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2.
\end{aligned}$$

- On retrouve ainsi la décomposition vue dans le cadre de la régression simple :

$$SCT = SCM + SCR.$$

Exercice. On considère les matrices :

$$H_X = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \text{ et } J_n = \frac{1}{n}\underbrace{[\mathbf{1}_n\mathbf{1}_n\ldots\mathbf{1}_n]}_{n \text{ fois}}.$$

1. Montrer que H_X est un projecteur orthogonal. Quel est le rang de H_X ?
2. Montrer que $H_X\mathbf{X} = \mathbf{X}$ et que $H_X\mathbf{Y} = \widehat{\mathbf{Y}}$. Interpréter ces résultats. En déduire que $H_X J_n = J_n$.
3. Vérifier que $J_n\mathbf{Y} = \bar{Y}\mathbf{1}_n$. Montrer que

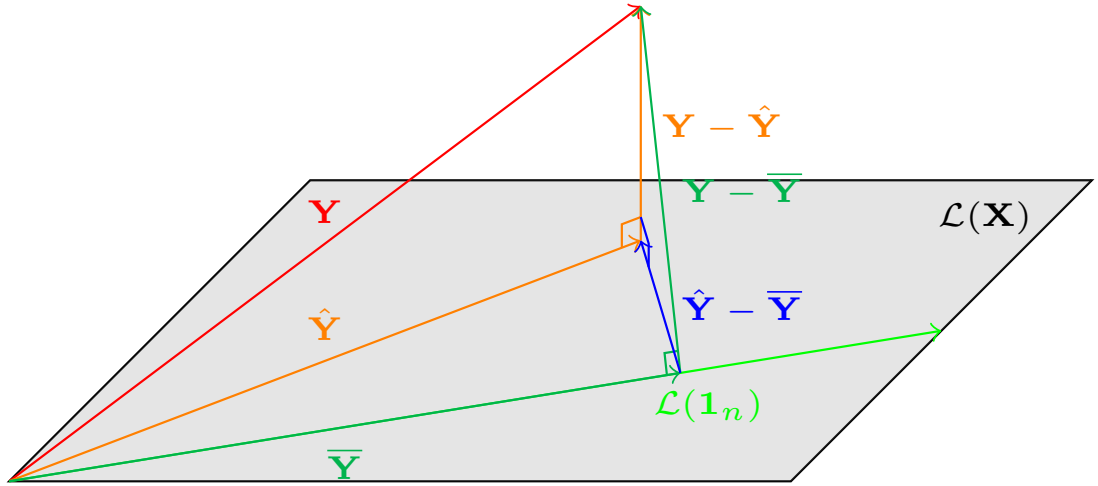
$$\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^T\left(\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}_n\right) = \mathbf{Y}^T\left(I_n - H_X\right)\left(H_X - J_n\right)\mathbf{Y}.$$

4. En déduire que :

$$\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^T\left(\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}_n\right) = 0.$$

Conclusions de l'exercice

- $\hat{\mathbf{Y}}$ est le projeté orthogonal de Y dans $\mathcal{L}(\mathbf{X})$, le s-e-v de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$.
- $\overline{Y}\mathbf{1}_n$ est le projeté orthogonal de $\mathbf{Y}$ dans $\mathcal{L}(\mathbf{1}_n)$, la droite vectorielle de $\mathbb{R}^n$ engendré par $\mathbf{1}_n$.
- $\mathbf{Y} - \hat{\mathbf{Y}}$ et $\hat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n$ sont orthogonaux.



Estimation de σ^2

Théorème

On considère l'estimateur

$$\widehat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{SCR}{n-p-1}$$

(i) $\widehat{\sigma^2}$ est un e.s.b de σ^2 indépendant de $\widehat{Y}$.

(ii) De plus si les $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ alors :

- $\frac{(n-p-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$.
- Pour tout $j = 0,\ldots,p$, on a $\dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma^2}\,(X^TX)^{-1}_{j+1,j+1}}} \sim \mathcal{T}_{n-p-1}$.

Preuve

Considérons la décomposition de $\mathbb{R}^n$ suivante : $\mathbb{R}^n = \mathcal{L}(\mathbf{X}) \oplus \mathcal{L}(\mathbf{X})^\perp$, où $\mathcal{L}(\mathbf{X})^\perp$ est le supplémentaire orthogonal de $\mathcal{L}(\mathbf{X})$ dans $\mathbb{R}^n$. Appliquons le théorème de Cochran à $\frac{1}{\sigma}\boldsymbol{\varepsilon} = \frac{\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}}{\sigma} \sim \mathcal{N}(0, \mathbf{I}_n)$.

- $\frac{1}{\sigma^2}||H_X\boldsymbol{\varepsilon}||^2 \sim \chi^2_{p+1}$, ce résultat n'est pas utile en pratique car $\boldsymbol{\varepsilon}$ est inconnu.
- $\frac{1}{\sigma^2}||(\mathbf{I}_n - H_X)\boldsymbol{\varepsilon}||^2 \sim \chi^2_{n-p-1}$. Or, $(\mathbf{I}_n - H_X)\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ donc

$$\frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{\sigma^2} = \frac{SCR}{\sigma^2} = \frac{(n-p-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$$

- $\frac{1}{\sigma^2}H_X\boldsymbol{\varepsilon}$ et $\frac{1}{\sigma^2}(\mathbf{I}_n - H_X)\boldsymbol{\varepsilon}$ sont indépendants. Or $H_X(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}$ et $\mathbf{X}\boldsymbol{\beta}$ n'est pas aléatoire donc $\hat{\mathbf{Y}}$ et $\widehat{\sigma^2}$ sont indépendants.

Mesure de la qualité d'ajustement d'un modèle

- Le coefficient de détermination :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}.$$

- Le R^2 ajusté qui tient compte du nombre d'observations et du nombre de variables explicatives :

$$R^2_{aj} = 1 - \frac{SCR/(n-p-1)}{SCT/n-1} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

- Le carré moyen des erreurs (Mean Squared Error) :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{SCR}{n}.$$

- L'erreur moyenne absolue (Mean Absolute Error) :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{e}_i|.$$

- L'erreur moyenne absolue en pourcentage (Mean Absolute Percentage Error) :

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{e}_i}{Y_i}\right|.$$

Plan du cours

- Tests sur les paramètres et intervalles de confiance
  1. **Sur les coefficients β_j.**
  2. Sur une combinaison linéaire de β_j.
  3. Sur plusieurs coefficients β_j simultanément.

Intervalles de confiance

- Si les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ alors les intervalles de confiance au niveau $1-\alpha$ des β_j sont donnés par :

$$IC(\beta_j) = \left[\widehat{\beta}_j - t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\widehat{\beta}_j}},\ \widehat{\beta}_j - t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\widehat{\beta}_j}}\right],$$

avec

$$\widehat{\sigma_{\widehat{\beta}_j}} = \sqrt{\widehat{\sigma^2}\,(X^TX)^{-1}_{j+1,j+1}} : \text{ erreur standard de } \widehat{\beta}_j.$$

Remarque

Remarque : Si les ε_i ne sont pas supposés gaussiens, alors l'intervalle de confiance précédent est un intervalle de confiance asymptotique, en remplaçant $t_{n-p-1,1-\alpha/2}$ par le fractile $z_{1-\alpha/2}$ de la loi normale.

Tests d'hypothèses : signification d'un coefficient β_j

- Pour $j = 0, \ldots, p$, on souhaite réaliser un test concernant la vraie valeur de β_j.
- Les hypothèses :

$$H_0 : \beta_j = b_j \text{ contre } H_1 : \beta_j \neq b_j, \ b_j \text{ une valeur fixée.}$$

- En particulier $b_j = 0$, on teste l'effet de la variable explicative x_j sur la variable Y
- Réaliser ce test consiste à se demander s'il faut exclure ou non la variable x_j du modèle (on parle de test d'exclusion).
- Si H_0 est rejetée au profit de H_1, on dit que le coefficient β_j est significatif.

Tests sur les paramètres

- Les hypothèses du test :

$$H_0 : \beta_j = b_j \text{ contre } H_1 : \beta_j \neq b_j, \ b_j \text{ une valeur fixée.}$$

- La statistique de test :

$$\frac{\widehat{\beta} - b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} \sim \mathcal{T}_{n-p-1} \text{ sous } H_0.$$

- La règle de décision :
  - si $\left| \frac{\widehat{\beta}-b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} \right| > t_{n-p-1,1-\alpha/2}$, alors je rejette H_0 au profit de H_1.
  - si $\left| \frac{\widehat{\beta}-b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} \right| \leq t_{n-p-1,1-\alpha/2}$, alors je ne rejette pas H_0 au profit de H_1.

Remarque

- Le cas particulier $\beta_j = 0$ correspond à la statistique de test

$$t = \frac{\widehat{\beta}}{\widehat{\sigma_{\widehat{\beta}_j}}}.$$

- Cette quantité s'appelle statistique t (t-statistic) et est calculée automatiquement par la plupart des logiciels statistiques.
- Notons que la règle de décision ci-dessus peut être étendue aux cas des tests unilatéraux de la manière suivante :
  - $H_0 : \beta_j = b_j$ contre $H_1 : \beta_j > b_j$: on rejette H_0 lorsque $\frac{\widehat{\beta} - b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} > t_{n-p-1,1-\alpha}$ et on ne rejette pas H_0 sinon.
  - $H_0 : H_0 : \beta_j = b_j$ contre $H_1 : \beta_j < b_j$ one rejette H_0 lorsque $\frac{\widehat{\beta} - b_j}{\widehat{\sigma_{\widehat{\beta}_j}}} < t_{n-p-1,1-\alpha}$ et on ne rejette pas H_0 sinon.

Notion de p-valeur

- Les logiciels fournissent la probabilité critique du test aussi appelé $p-$valeur ou degré de signication.
- La p-valeur correspond au niveau de risque de première espèce α^* pour lequel on hésiterait entre les deux décisions : rejeter ou ne pas rejeter H_0. C'est la plus petite valeur du risque de première espèce telle qu'on rejette H_0.
- Ainsi, on peut conclure un test en utilisant la p-valeur avec la règle de décision suivante :
  - si $\alpha^* < \alpha$, alors je rejette H_0 au profit de H_1.
  - si $\alpha^* \geq \alpha$, alors je ne rejette pas H_0 au profit de H_1.

Remarque

- Il faut être prudent dans l'interprétation de la statistique t. Il ne faut pas supprimer aveuglement d'un modèle une variable dont le coefficient n'est pas significatif.
- On peut avoir de bonnes raisons pratiques pour conserver la variable même si son impact semble faible d'un point de vue statistique.
- Ce test ne permet de conclure que sur un seul coefficient β_j et non sur l'élimination simultanée de plusieurs coefficients.

Test sur la régression générale et estimation

```
> reg=lm(Pression.systolique~Age + Poids, data=data)
> summary(reg)

Call:
lm(formula = Pression.systolique ~ Age + Poids, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4640 -1.1949 -0.4078  1.8511  2.6981
```

```
   Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.9941    11.9438   2.595  0.03186 *
Age           0.8614     0.2482   3.470  0.00844 **
Poids         0.7384     0.2881   2.563  0.03351 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.318 on 8 degrees of freedom
Multiple R-squared:  0.9768,Adjusted R-squared:  0.9711
F-statistic: 168.8 on 2 and 8 DF,  p-value: 2.874e-07
```

Intervalles de confiance sur les paramètres

```
> confint(reg)
                2.5 %     97.5 %
(Intercept) 3.45169598 58.536510
Age         0.28899203  1.433837
Poids       0.07395285  1.402824
```

Plan du cours

- Tests sur les paramètres et intervalles de confiance
  1. Sur les coefficients β_j.
  2. **Sur une combinaison linéaire de β_j.**
  3. Sur plusieurs coefficients β_j simultanément.

Intervalle de confiance d'une combinaison linéaire des coefficients

- On se donne un vecteur $a = (a_0, \ldots, a_p)^T$ et on cherche à estimer $a^T\beta$.
- Exemple : si $a^T = \left(1, x_1^*, \ldots, x_p^*\right)$ où les $x_j^*, j = 1, \ldots, p$ représentent une nouvelle observation des variables explicatives, alors $a^T\beta = \mathbb{E}(Y^*)$.
- On a $a^T\widehat{\beta}$ est un e.s.b de $a^T\beta$, de plus :

$$a^T\widehat{\beta} \sim \mathcal{N}\left(a^T\beta,\ \sigma^2 a^T\left(X^TX\right)^{-1}a\right),$$

ce qui implique :

$$\frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}} \sim \mathcal{T}_{n-p-1}.$$

- Ainsi, un intervalle de confiance au niveau $1-\alpha$ pour $a^T\beta$ est donné par :

$$IC\left(a^T\beta\right) = \left[a^T\widehat{\beta} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}\right].$$

- Dans le cas où $a^T = \left(1, x_1^*, \ldots, x_p^*\right)$ où les x_j^* est une nouvelle observation des variables explicatives, on a :

$$IC\left(\mathbb{E}\left(Y^*\right)\right) = \left[\widehat{Y^*} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^T\left(X^TX\right)^{-1}a}\right].$$

Test d'une contrainte linéaire sur les coefficients

- On souhaite tester une restriction linéaire de la forme :

$$\begin{cases} H_0 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p = b \\ \text{contre} \\ H_1 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p \neq b \end{cases}$$

- Exemples :

$$(1)\begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ \text{contre} \\ H_1 : \beta_1 + \beta_2 \neq 1 \end{cases} ; (2)\begin{cases} H_0 : \beta_2 - \beta_3 = 0 \\ \text{contre} \\ H_1 : \beta_2 - \beta_3 \neq 0 \end{cases}.$$

- Les hypothèses de test peuvent s'écrire sous forme matricielle :

$$(1) \begin{cases} H_0 : a^T\beta = b \\ \text{contre} \\ H_1 : a^T\beta \neq b \end{cases} \quad \text{avec } a^T = (a_0, \ldots, a_p)\,.$$

- Rappel :

$$\frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\widehat{\sigma^2} a^T \left(X^T X\right)^{-1} a}} \sim \mathcal{T}_{n-p-1}.$$

- Ainsi on peut établir la règle de décision suivante :

- si

$$\left| \frac{a^T \widehat{\beta} - b}{\sqrt{\widehat{\sigma^2} a^T \left(X^T X\right)^{-1} a}} \right| > t_{n-p-1,1-\alpha/2},$$

  alors je rejette H_0 au profit de H_1.

- si

$$\left| \frac{a^T \widehat{\beta} - b}{\sqrt{\widehat{\sigma^2} a^T \left(X^T X\right)^{-1} a}} \right| \leq t_{n-p-1,1-\alpha/2},$$

  alors je ne rejette pas H_0 au profit de H_1.

Exemple.

On considère le modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

avec

$$n = 40, \widehat{\beta}_0 = 1.37, \widehat{\beta}_1 = 0.632, \widehat{\beta}_2 = 0.452$$

et

$$\widehat{\mathbb{V}(\widehat{\beta}_1)} = 0.066, \widehat{\mathbb{V}(\widehat{\beta}_2)} = 0.048, \widehat{\mathbb{C}ov\left(\widehat{\beta}_1, \widehat{\beta}_2\right)} = -0.044.$$

Réaliser les tests :

$$(1)\begin{cases} H_0 : \beta_1 = \beta_2 \\ \text{contre} \\ H_1 : \beta_1 \neq \beta_2 \end{cases}, (2)\begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ \text{contre} \\ H_1 : \beta_1 + \beta_2 \neq 1 \end{cases}, (3)\begin{cases} H_0 : \beta_2 = 0 \\ \text{contre} \\ H_1 : \beta_2 \neq 0 \end{cases}.$$

Plan

Test de signification de plusieurs coefficients

On rencontre deux cas lorsqu'on souhaite comparer des modèles :

- les modèles sont emboités (l'un est un cas particlieu de l'autre)
- les modèles ne sont pas emboités.

Lorsque les modèles sont emboités : la prise de décision repose généralement sur un test qui consiste à comparer simultanément plusieurs paramètres.

- On peut tester la nullité de tous les coefficients on parle alors de test de validité globale du modèle.
- On peut aussi tester la nullité d'une partie de ces coefficients.

Lorsque les modèles ne sont pas emboités on utilisera des critères dits "d'ajustement pénalisé".

Comparaison de modèles

Définition

Deux modèles sont dits emboités si l'un d'entre-eux s'obtient comme un cas particulier de l'autre. Ils expliquent la même variable Y et les variables explicatives du "petit" apparaissent toutes dans le "grand".

Soit $r < p$, on peut toujours réordonner les colonnes de $\mathbf{X}$. On pose alors :

Petit modèle

$$\mathbf{Y} = \mathbf{X}_r \boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_r$$

avec

$$\boldsymbol{\beta}_r = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_r)^T$$

Grand modèle

$$\mathbf{Y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_p$$

avec

$$\boldsymbol{\beta}_p = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_r \quad \beta_{r+1} \quad \cdots \quad \beta_p)^T$$

Test de comparaison de deux modèles emboités

- Les hypothèses du test :

$$\begin{cases} H_0 : \beta_{r+1} = \beta_{r+2} = \cdots = \beta_p = 0 \\ \text{contre} \\ H_1 : \exists j \in \{r+1, \cdots, p\} : \beta_j \neq 0 \end{cases}$$

- La statistique de test :

$$\begin{aligned} F = & \frac{[SCR_q - SCR_p]/(p-q)}{SCR_p/n-p-1} = \frac{[SCM_p - SCM_q]/(p-q)}{SCM_p/n-p-1} \\ = & \frac{\left(R_p^2 - R_q^2\right)(n-p-1)}{q(1-R_p^2)} \sim \mathcal{F}_{p-q,n-p-1}, \text{ sous } H_0. \end{aligned}$$

- La règle de décision :
  - si $F > f_{p-q,n-p-1,1-\alpha}$, alors je rejette H_0 au profit de H_1.
  - si $F \leq f_{p-q,n-p-1,1-\alpha}$, alors je ne rejette pas H_0.

Remarques

- Dans le cas $q = 0$ le petit modèle est $Y_i = \beta_0 + \varepsilon_i$ et ce test s'appelle le test de signification globale du modèle.
- De manière plus générale, deux modèles sont emboités si le s-e-v engendré par les colonnes du petit modèle est inclus dans le s-e-v engendré par les colonnes du grand. Les résultats restent valides, $p + 1$ étant la dimension du grand s-e-v et $r + 1$ la dimension du petit.

Preuve

- Soit $\mathcal{L}(\mathbf{X}_p) \cap \mathcal{L}(\mathbf{X}_q)^\perp$ le s-e-v de $\mathcal{L}(\mathbf{X}_p)$ qui est orthogonal à $\mathcal{L}(\mathbf{X}_q)$. C'est le s-e-v engendré par les colonnes $r+1$ à p de $\mathbf{X}_p$.
- Décomposons maintenant $\mathbb{R}^n$ de la façon suivante :
  $\mathbb{R}^n = \mathcal{L}(\mathbf{X}_q) \oplus \mathcal{L}(\mathbf{X}_p) \cap \mathcal{L}(\mathbf{X}_q)^\perp \oplus \mathcal{L}(\mathbf{X}_p)^\perp$
- Selon le petit modèle, $\frac{\boldsymbol{\varepsilon}_q}{\sigma^2} \sim \mathcal{N}(0, \mathbf{I}_n)$.
- Soit H_q et H_p les projecteurs orthogonaux, respectivement dans $\mathcal{L}(\mathbf{X}_q)$ et $\mathcal{L}(\mathbf{X}_p)$. Les projecteurs orthogonaux dans les s-e-v $\mathcal{L}(\mathbf{X}_q)$, $\mathcal{L}(\mathbf{X}_p) \cap \mathcal{L}(\mathbf{X}_q)^\perp$ et $\mathcal{L}(\mathbf{X}_p)^\perp$ sont respectivements H_q, $H_p - H_q$ et $\mathbf{I}_n - H_p$.
- On applique le théoroème de Cochran en projetant $\frac{\boldsymbol{\varepsilon}_q}{\sigma^2}$ dans ces trois s-e-v orthogonaux. Sous l'hypotèse où le petit modèle est valide
  - $(H_p - H_q)\boldsymbol{\varepsilon}_p = (Y - \widehat{Y}_q) - (Y - \widehat{Y}_p)$

Test de signification globale du modèle

- Les hypothèses du test :

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_p = 0 \\ \text{contre} \\ H_1 : \exists j \in \{1, \ldots, p\} \text{ tq } \beta_j \neq 0 \end{cases}$$

- La statistique de test :

$$F = \frac{SCM/p}{SCR/n-p-1} = \frac{R^2(n-p-1)}{p(1-R^2)} \sim \mathcal{F}_{p,n-p-1} \text{ sous } H_0.$$

- La règle de décision :
  - si $F > f_{p,n-p-1,1-\alpha}$, alors je rejette H_0 au profit de H_1.
  - si $F \leq f_{p,n-p-1,1-\alpha}$, alors je ne rejette pas H_0 au profit de H_1.

- Réaliser ce test revient à comparer le modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

avec le modèle

$$Y_i = \beta_0 + \varepsilon_i.$$

- Si l'hypothèse H_0 n'est pas rejetée, cela veut dire que le modèle sans variable explicative est préférable au modèle de départ.
- Les variables x_j expliquent très peu les variations de Y, qu'il vaut mieux les éliminer.
- Lorsqu'on rejette H_0 au profit de H_1, on dit que le modèle est globalement significatif.

- Comme pour la statistique t les logiciels statistiques fournissent la valeur de la statistique F, ainsi que la p-valeur qui lui est associée.
- On présente parfois le calcul de F dans un tableau appelé ANOVA qui fait apparaitre plusieurs termes :

Source	Df	Sum Sq	Mean Sq	F value
Regression	p	SCM	$\frac{SCM}{p}$	$F = \frac{SCM/p}{SCR/n-p-1}$
Residuals	$n-p-1$	SCR	$\frac{SCR}{n-p-1}$	

- Avec le logiciel R :

```
> reg <-lm (Y ~ x_1+x_2+...+x_p,data=data)
> anova(reg)
```

Tests sur les modèles emboités

```
  Analysis of Variance Table

Response: Pression.systolique
          Df  Sum Sq Mean Sq  F value    Pr(>F)
Age        1 1778.62 1778.62 330.9619 8.566e-08 ***
Poids      1   35.29   35.29   6.5673   0.03351 *
Residuals  8   42.99    5.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Critères de comparaison de modèles (emboités ou pas)

- La comparaison de deux modèles se fera avec le R^2_{aj}. Un modèle est donc préférable à un autre si son R^2_{aj} est supérieur à celui de l'autre modèle.
- Il existe d'autres critères du même type permettant de comparer la qualité d'ajustement de 2 modèles relatifs à la même variable dépendante :
  - Le critère d'information d'Akaike (Akaike Information Criterion) :

$$AIC = n \ln\left(\frac{SCR}{n}\right) + 2(p+1).$$

  - Le critère de Bayes de Schwarz (Schwarz Bayesian Information Criterion) :

$$BIC = n \ln\left(\frac{SCR}{n}\right) + (p+1)\ln n.$$

- Contrairement au R^2_{aj}, les critères AIC et BIC sont des fonctions croissantes de SCR.
- Un modèle est donc préférable à un autre au sens du AIC (resp. BIC) si son critère AIC (reps. BIC) est inférieur à celui de l'autre modèle.
- Certains logiciels utilisent une formulation légèrement différente des critères AIC et BIC mais les interprétations restent les mêmes.
- Avec le logiciel **R**, les critères précédents s'obtient avec la fonction `AIC`, qui dépend d'un paramètre k.
- Ce paramètre vaut 2 et $2\ln n$ pour les critères AIC et BIC respectivement.

Plan

Intervalle de prévision

- On cherche maintenant un intervalle pour Y^* qui est une variable aléatoire. On a :

$$Y^* = a^T\beta + \varepsilon^* \text{ avec } a^T = \left(1, x_1^*, \ldots, x_p^*\right) \text{ et } \varepsilon^* \sim \mathcal{N}(0, \sigma^2),$$

ε^* indépendant des $\varepsilon_i, i = 1, \ldots, n$.

- Ainsi

$$Y^* \sim \mathcal{N}\left(a^T\beta, \sigma^2\right) \text{ et } \widehat{Y^*} = a^T\widehat{\beta} \sim \mathcal{N}\left(a^T\beta, \sigma^2 a^T\left(X^TX\right)^{-1}a\right),$$

ce qui implique

$$\widehat{e^*} = Y^* - \widehat{Y^*} \sim \mathcal{N}\left(0, \sigma^2\left[1 + a^T\left(X^TX\right)^{-1}a\right]\right).$$

- D'où

$$\frac{Y^* - \widehat{Y^*}}{\sqrt{\widehat{\sigma^2}\left[1 + a^T \left(X^T X\right)^{-1} a\right]}} \sim \mathcal{T}_{n-p-1}.$$

- L'intervalle de prévision est ainsi donnée par :

$$IP\left(Y^*\right)\left[\widehat{Y^*} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}\left[1 + a^T \left(X^T X\right)^{-1} a\right]}\right].$$

Remarque

La quantité

$$a^T \left(X^T X\right)^{-1} a$$

représente le levier de l'observation $x^* = \left(x_1^* \ldots, x_p^*\right)$.

Prédictions

```
> predict(reg)
       1        2        3        4        5        6        7
133.7183 143.4317 153.6716 164.5327 151.7570 168.4078 140.4640 146.4
```

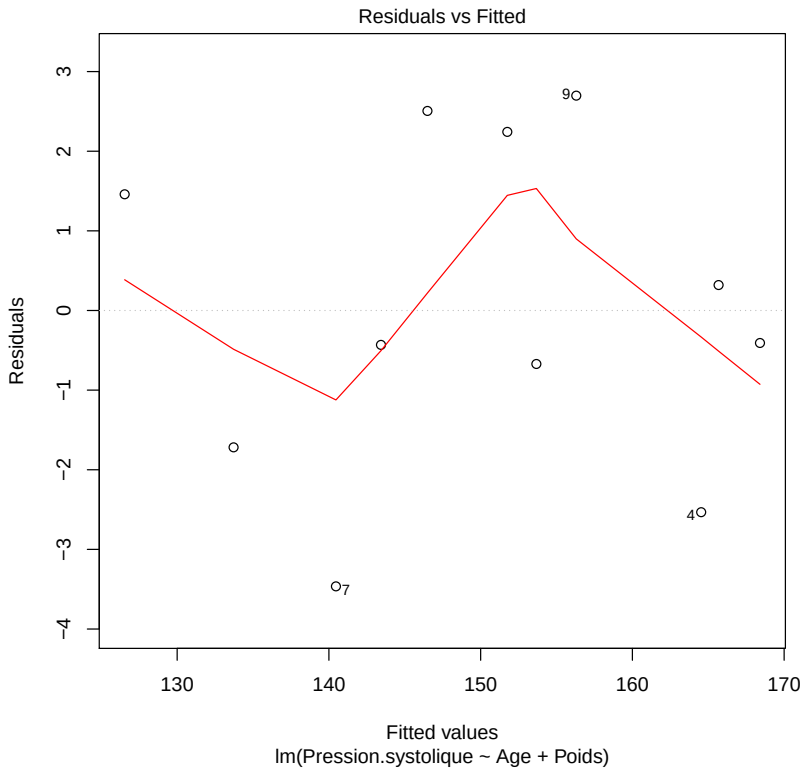
Plan

Analyse des résidus

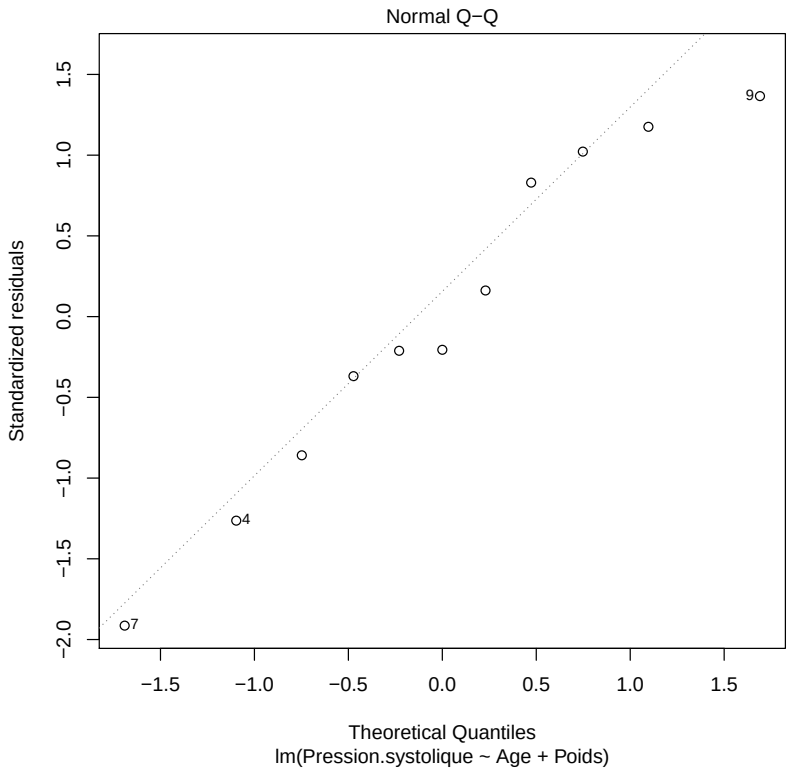
- L'étude des résidus est fondamentale. Elle permet de repérer des données atypiques, qui peuvent éventuellement être aberrantes :
  - Des observations mal reconstituées par le modèle
  - ou bien des observations qui jouent un rôle important dans l'estimation de la régression.
- De plus l'étude des résidus est souvent la seule façon de vérifier empiriquement le bien fondé des hypothèses du modèle.
- La représentation des résidus se fait en fonction de i (ou bien en fonction de $\widehat{Y}_i$).
- Cette représentation ne doit pas présenter une tendance particulière.

- Une courbure dans la forme des résidus en fonction des $\widehat{Y}_i$, suggère que l'hypothèse de linéarité n'est pas adaptée.
- Un comportement monotone de la variabilité des résidus en fonction des $\widehat{Y}_i$ peut indiquer une variance non constante.

Exemple : Graphe des résidus



Exemple : Quantile-quantile plot



On note $\hat{E}$ le vecteur des résidus $\hat{E} = (Y - \hat{Y})$.

Propriété

$$\hat{E} \sim \mathcal{N}(0, \sigma^2(I_n - H_x))$$

Remarques

- H_x n'est a priori pas diagonale donc les résidus sont en général corrélés entre-eux.
- σ^2 n'est pas connu.

Notion de levier

Définition

On appelle matrice chapeau (hat matrix) associée à X, la matrice H_X définie par

$$H_X = X \left(X^T X\right)^{-1} X^T.$$

Remarque

Le i-ème élément de la diagonale de H_X est le levier de l'observation numéro i, i.e :

$$h_{ii} = a_i^T \left(X^T X\right)^{-1} a_i,$$

où a_i^T est la ligne numéro i de la matrice X, donnée par :

$$a_i^T = [1 \; x_{i1} \; x_{i2} \ldots x_{ip}] .$$

Définition

Le levier se définit comme la distance de Mahalanobis entre un point (défini par ses p variables explicatives uniquement) et le centre de gravité du nuage de point.
Remarque : La distance de Mahalanobis tient compte de la forme du nuage de points.

Propiétés

(i) Pour tout $i = 1 \ldots, n$, on a :

$$\frac{1}{n} \leq h_{ii} \leq 1 \text{ et } \sum_{i=1}^{n} h_{ii} = p + 1.$$

(ii) Pour tout $i = 1 \ldots, n$, on a :

$$\mathbb{V}(\widehat{e}_i) = \sigma^2(1 - h_{ii}).$$

(iii) Dans le cas de la régression simple ($p = 1$), on a :

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n s_{xx}}$$

En pratique

- En pratique, les points présentants un fort levier sont problématiques car nous allons voir que plus le levier est grand, plus le point peut avoir une influence prédominante sur l'estimation de $\boldsymbol{\beta}$ ou sur sa propre prédiction.
- Puisque le levier moyen est $\frac{p+1}{n}$, on considèrera que le levier est fort dès lors qu'il est supérieur à $\frac{2(p+1)}{n}$.
- On peut aussi étudier les leviers en les triant par ordre décroissant et en remarquant les points dont le levier se détache des autres.

Observations mal reconstituées par le modèle

Pour savoir si, une observation est bien reconstituée par le modèle, on étudie son **résidu standardisé** (si n grand) ou **studentisé** (si n petit).

- Un fort résidus standardisé/studentisé indique que l'observation est mal reconstituée par le modèle. Il faut alors étudier la donnée attentivement, pour déterminer s'il s'agit d'une valeur aberrante.
- Mais il faut faire attention au fait qu'une valeur peut être aberrante sans être pour autant influente sur la régression (et vice versa).

Résidus standardisés

Définition

On définit les résidus standardisés par

$$\widehat{e_i^{std}} = \frac{\widehat{e_i}}{\sqrt{\widehat{\sigma^2}\,(1 - h_{ii})}}.$$

- Lorsque n est grand les résidus standardisés doivent rester compris entre -2 et 2.

Résidus studentisés

Définition

On définit les résidus studentisés par

$$\widehat{t}_i^* = \frac{\widehat{e}_i}{\sqrt{\widehat{\sigma^2_{(-i)}}\,(1-h_{ii})}},$$

où $\widehat{\sigma^2_{(-i)}}$ est l'estimateur de σ^2 obtenue sans utiliser l'observation numéro i

$$\widehat{\sigma^2_{(-i)}} = \frac{SCM_{(-i)}}{n-p-2} = \frac{1}{n-p-2}\sum_{\substack{k=1\\k\neq i}}^{n-1}\widehat{e}_k^2 = \frac{\widehat{\sigma^2}\,(n-p-1)-\widehat{e}_i^2}{n-p-2}.$$

- Un résidu est jugée fort au niveau de confiance $1-\alpha$, lorsque

$$\left|\widehat{t}_i^*\right| > t_{n-p-2,1-\alpha/2}.$$

Etude de l'influence de chaque observation sur sa propre prédiction

On appelle résidu prédit l'écart $y_i - \hat{y}_{(-i)}$, avec $\hat{y}_{(-i)}$ la pévision obtenue avec un l'échantillon de taille $n-1$ excluant la i^e observation.

Propriété

On peut montrer que

$$y_i - \hat{y}_{(-i)} = \frac{y_i - \hat{y}_i}{1 - h_i}$$

Remarque

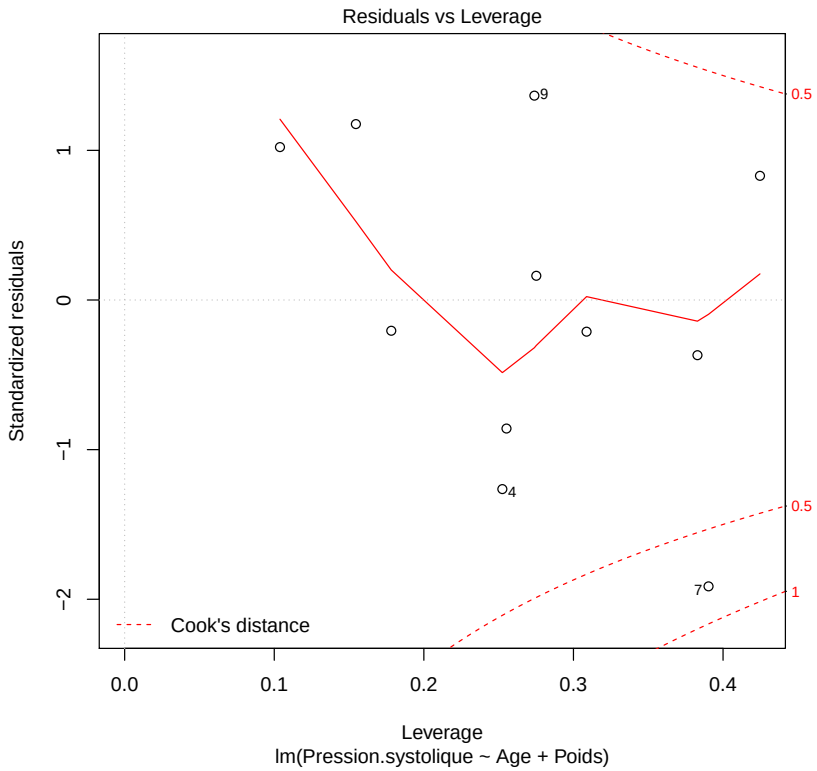
Le résidu prédit peut être grand car

- le résidu standardisé est grand,
- le levier est grand

Remarque

- La variance du résidu estimé est d'autant plus faible que son levier h_{ii} est grand.
- Le résidu prédit est d'autant plus grand que le levier est grand.
- En pratique un levier supérieur à $\frac{2(p+1)}{n}$ est considéré comme important.
- Dans ce cas on considère que l'observation qui lui est associé joue un rôle important dans la détermination de l'hyperplan de régression car elle est éloignée du centre de gravité du nuage de points.

Exemple : Levier



Distance de Cook

- La distance de Cook permet de mesurer l'influence d'une observation sur les l'estimation de β.

Définition

La distance de Cook associée à l'observation numéro i se définit par :

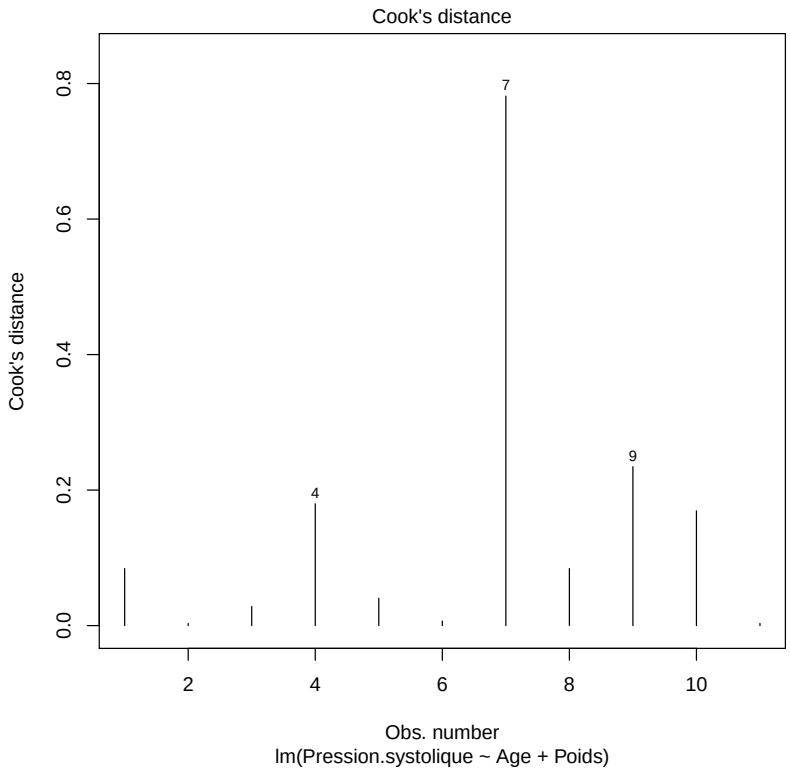
$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)\hat{\sigma}^2}$$

On peut montrer que

$$D_i = \frac{\widehat{e_i^{std}}^2 h_{ii}}{(p+1)(1-h_{ii})}.$$

- Une distance de Cook supérieure à 1 indique en général une influence anormale de l'observation Y_i sur la régression effectuée.

Exemple : Distance de Cook



Sélection de modèle

Objectif

Déterminer le modèle qui explique le mieux la variabilité des données tout en limitant au maximum le nombre de variables

Une variable peut ne pas être utile dans la régression car :

- elle n'explique pas la réponse / pas de lien avec la réponse
- elle apporte une information qui est déjà contenue dans une ou plusieurs variables du modèle. On parle alors de colinéarité.

Colinéarité

On parle de colinéarité si la corrélation entre une variable et une autre variable ou une combinaison linéaire d'autres variables du modéle est supérieure à 0.8.

Ce qui peut-être problématique avec la colinéarité

- le signe du coefficient est contradictoire avec celui de la corrélation
- la variance des estimateurs est très grande
- les variables peuvent être rejetées par le test de Student
- les résultats sont instables et sensibles à l'ajout d'une ou de quelque données

Détection de la colinéarité

- Calcul des coefficients de corrélation entre variables explicatives. $=>$ PBLM : il ne détecte pas la colinéarité avec une comb. lin. de variables
- Calcul du VIF : facteur d'inflation de la variance

$$v_j = \frac{1}{1 - R_j^2}$$

avec R_j^2 le coefficient de détermination de la j^e variable explicative en fonction des p-1 autres. On a

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n} v_j$$

On dit que v_j est trop grand si $v_j > 5$.

- Comparaison des signes du coefficient $\hat{\beta}_j$ et de la corrélation entre la j^e variable explicative et la réponse.

Corrélation partielle

- Deux variables peuvent être fortement corrélées sont pour autant avoir quelque chose à voir. C'est par exemple le cas de la vente de glaces et de la vente de lunettes de soleil ou de la taille des personnes et de leur longueur de cheveux.
- Dans ces deux cas, il existe une troisième variable qui influe directement sur les deux autres. C'est par exemple le cas de la température pour les ventes de lunettes de soleil et de glaces ou le sexe pour la taille et la longueur des cheveux.

La corrélation partielle mesure le lien entre deux variables en controllant l'effet d'une troisième.

$$\rho_{x,y|z} = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{1-\rho_{x,z}^2}\sqrt{1-\rho_{y,z}^2}}$$

Corrélation partielle et colinéarité

L'apport de l'addition d'une variable dans un modèle doit se mesurer en tenant compte des variables explicatives déjà présentes. C'est donc la corrélation partielle entre la nouvelle variable explicative et la réponse connaissant les autres variables explicatives qui doit être considéré pour mesurer la pertinence de la variable plutôt que sa corrélation avec la réponse.

Le test de Ficher qui compare le modèle $p+1$ variables explicatives contre un modèle emboité à p variables explicatives, revient à tester la corrélation partielle d'ordre p. On mesure l'information **additionnelle** apportée par la nouvelle variable.

Procédures automatisées de sélection de modèle

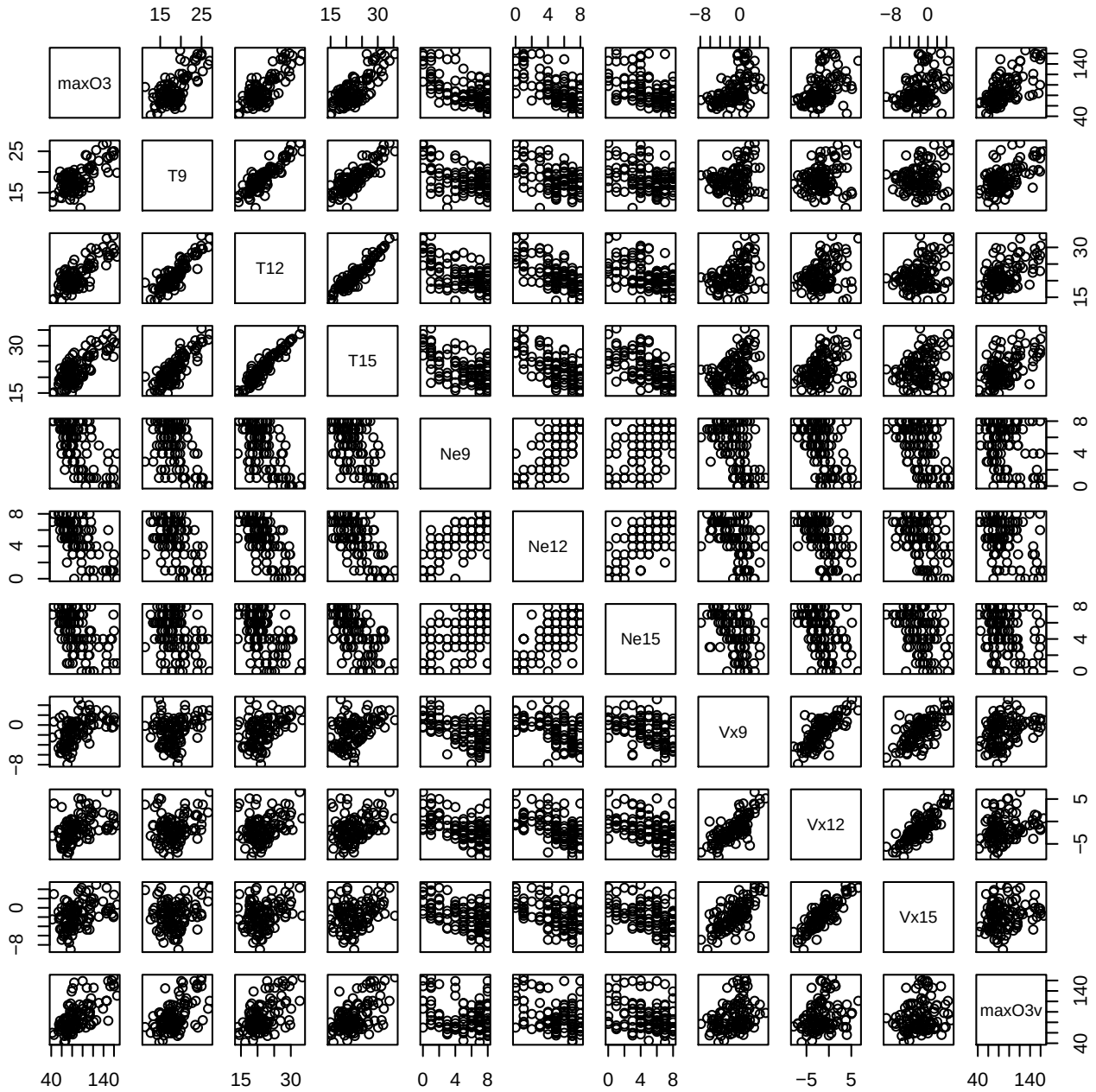
- La sélection se fait sur un critère : AIC, BIC, R^2_{aj}, test de Fisher, ...
- On dispose de k variables explicatives candidates
- Dans l'idéal, il faudait estimere et comparer tous les modèles réalisables avec ces variables, il y en a 2^k. Ca en fait beaucoup !
- En pratique on utilisera les procédures suivantes :
  - Forward On part du modèle ne contenant que la constante, et on ajoute la variable qui optimise le critère. On continue ainsi, jusqu'à ce que le critère soit optimisé lorsqu'on n'ajoute plus de variable.
  - Backward On part du modèle complet, et à chaque étape on supprime la variable qui optimise le critère. On s'arrête quand le critère est optimisé en ne faisant rien.
  - Stepwise On part du modèle à une constante, on ajoute une, puis deux variables. Avant d'ajouter une variable suplémentaire on vérifie qu'un meilleur modèle ne s'obtient pas en supprimant une des variables déjà présentes.

Exemple : Ozone

"L'ozone est un polluant atmosphérique bien connu pour ses effets délétères sur la santé humaine et sur la végétation. Il se forme dans les basses couches de l'atmosphère en période estivale, sous l'effet du rayonnement solaire et lorsque les températures sont élevées. " INERIS

On dispose des variables suivantes

- `max03` Maximum d'ozone relevé dans journée
- `T9` Température à 9 h
- `T12` Température à 12 h
- `T15` Température à 15 h
- `Ne9` Nébulosité à 9 h
- `Ne12` Nébulosité à 12 h
- `Ne5` Nébulosité à 15 h
- `Vx9` Vitesse du vent à 9 h
- `Vx12` Vitesse du vent à 12 h
- `Vx15` Vitesse du vent à 15 h
- `max03v` Maximum d'ozone relevé la veille


maxO3
T9
T12
T15
Ne9
Ne12
Ne15
Vx9
Vx12
Vx15
maxO3v