Méthodes usuelles de régression

R. Paegelow - MF Beclin

DO4

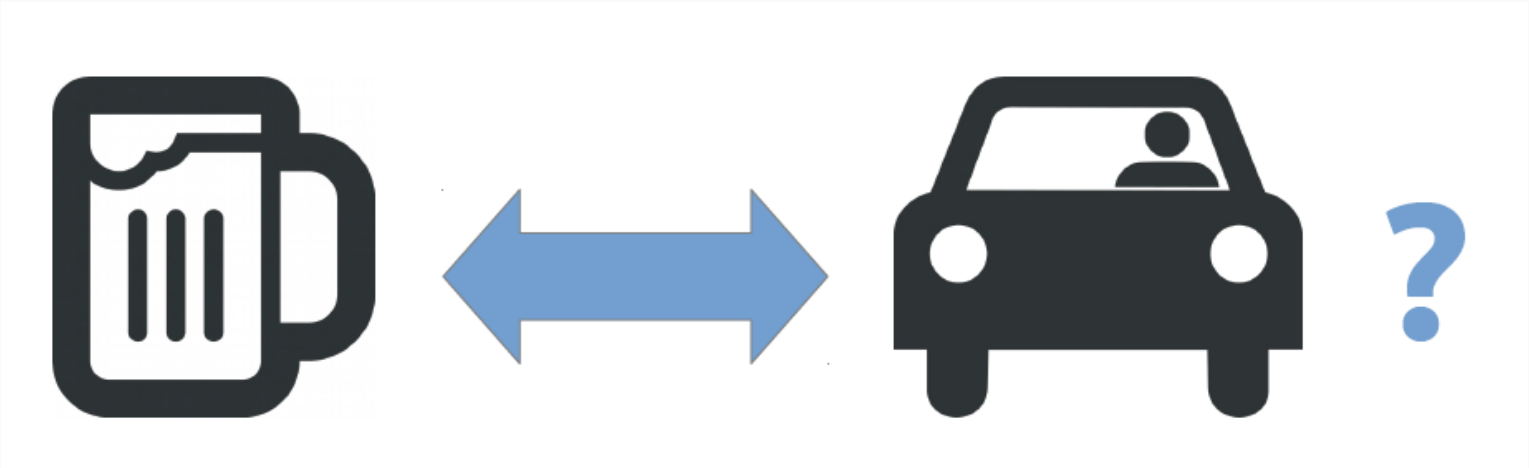
Décembre 2024

Ce chapitre expose quelques méthodes de régression permettant d'expliquer une variable Y en fonction de variables quantitatives et / ou qualitatives.

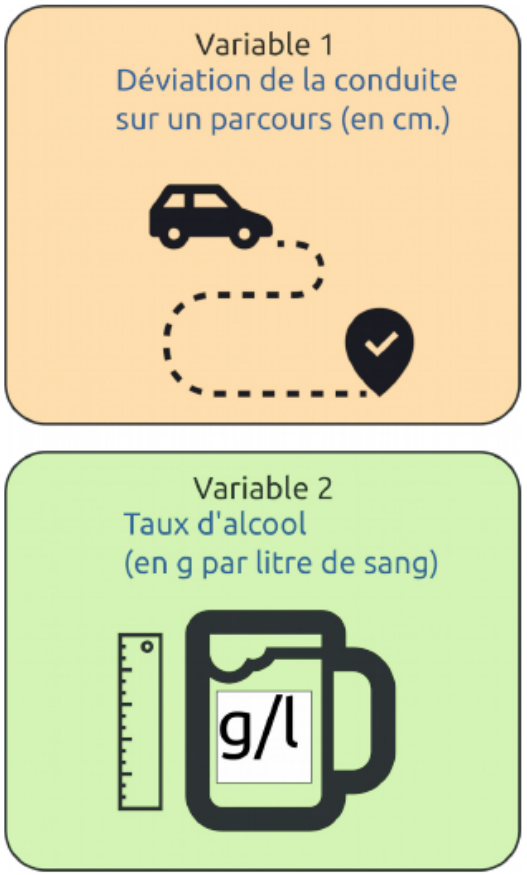
- Régression Simple : *une seule variable explicative quantitative.*
- Régression Multiple : *plusieurs variables explicatives quantitatives.*
- Régression Logistiques : *expliquer une variable qualitative Y, le plus souvent binaire, en fonction de variables quantitatives et qualitatives.*
- Arbres : *les arbres de régression ou de classification expliquent quant à eux une variable Y quantitative ou qualitative en fonction de variables quantitatives et / ou qualitatives.*

Problématique

- Existe-t-il une relation entre le taux d'alcool consomé et les sorties de routes en voiture ?

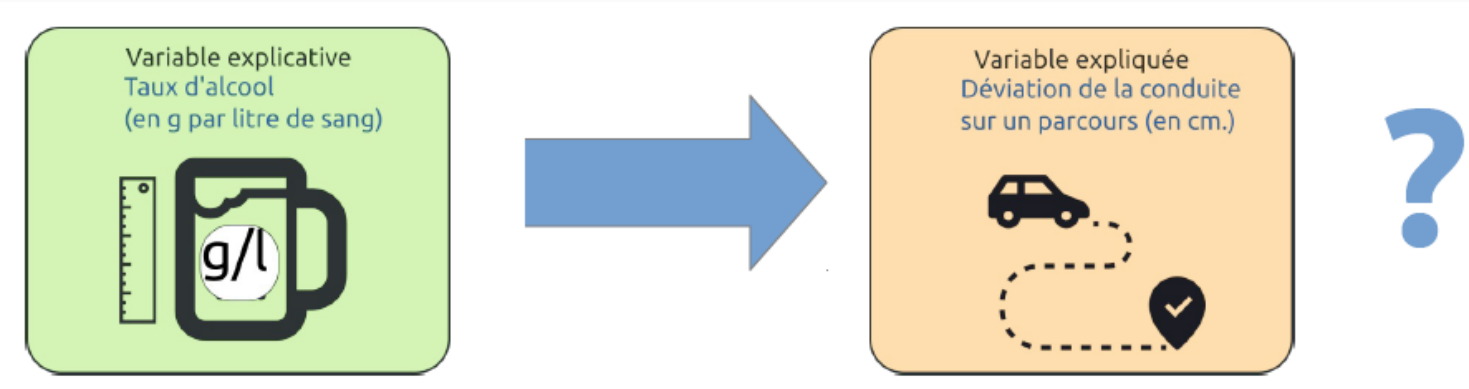


- Les variables :
  - On recueille deux variables continues au même moment
  - Une variable expliquée
  - Une variable explicative
  - Parfois arbitraire
- Question de recherche :
  - Les deux variables, sont-elles corrélées ?



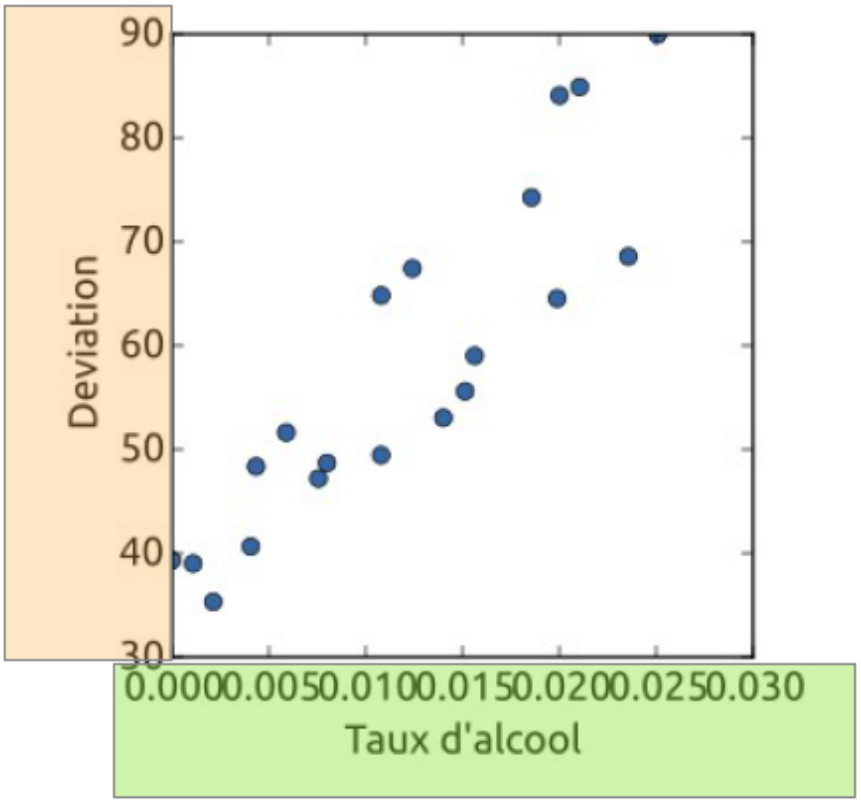
La problématique :

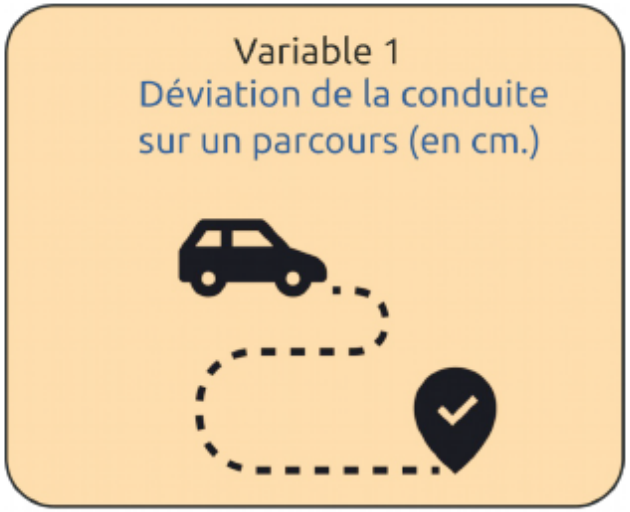
- Pourrait-on prédire la déviation sur un test de conduite à partir du taux d'alcool ?

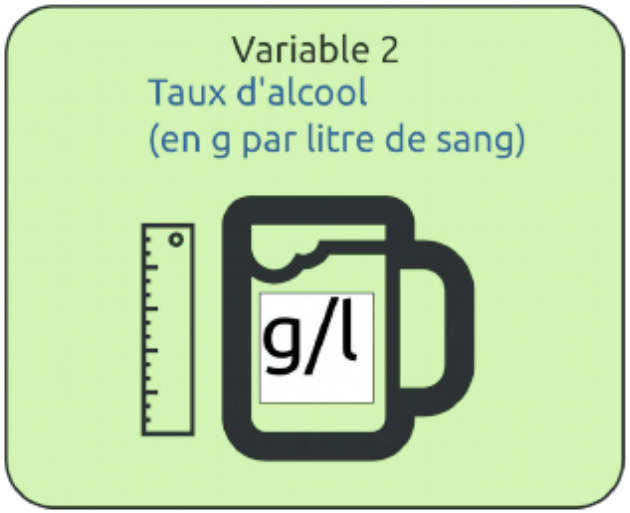


La représentation de données :

- Nuage de points $\rightarrow$ 1 point par sujet
- Abscisse : la variable explicative
- Ordonnée : la variable expliquée

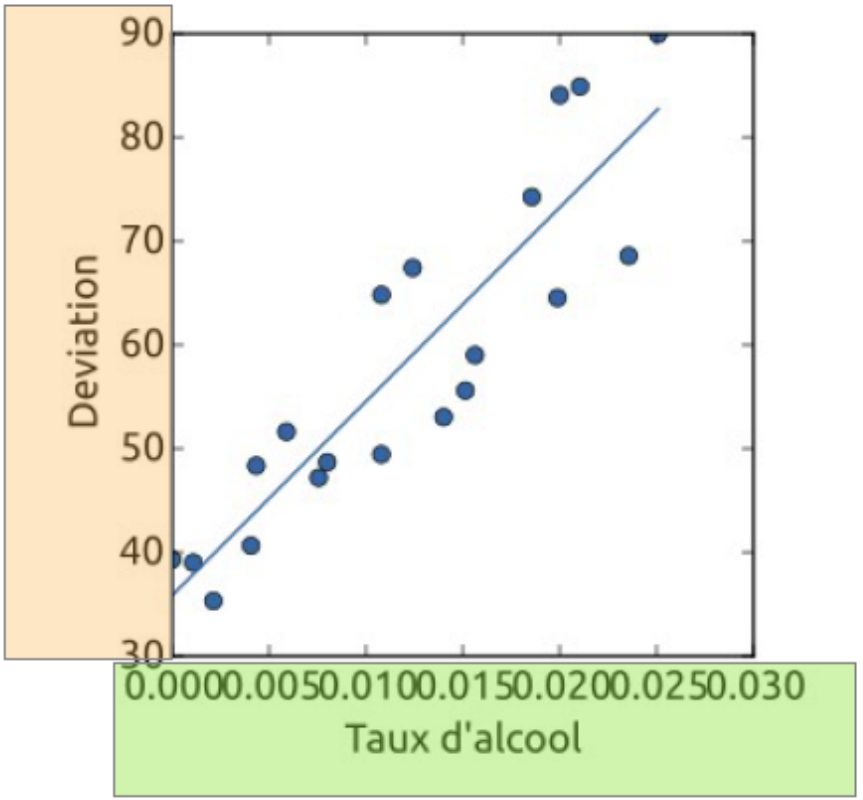


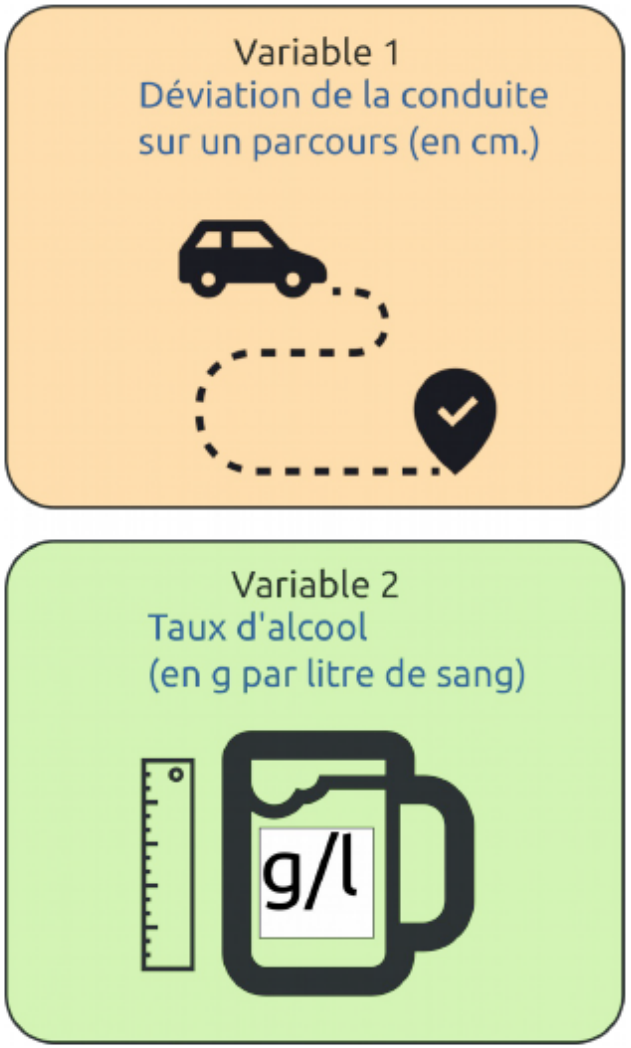




La régression linéaire :

- Peut-on prédire les valeurs sur l'ordonnée à partir des valeurs sur l'abscisse ?





Traitement de données

▶ Le problème de la pollution de l'air et son influence sur la santé publique.

▶ Des nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde de souffre (SO_2), le dioxyde d'azote (NO_2), l'azote (O_3) ou des particules sous forme de poussières contenues dans l'air.

Objectif

▶ Prévoir la concentration en ozone pour le lendemain afin d'avertir la population en cas de pic de pollution.

▶ Analyser la relation entre le maximum journalier de la concentration en ozone (en $\mu g/m^3$) et les données météorologiques.

Données ▶ 112 données relevées durant l'été 2001 à Rennes.

Question : Peut-on prévoir le taux d'ozone du lendemain ?

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
2001-06-01	87	15.6	18.5	18.4	4	4	8	0.69	-1.71	-0.69	84
2001-06-02	82	17.0	18.4	17.7	5	5	7	-4.33	-4.00	-3.00	87
2001-06-03	92	15.3	17.6	19.5	2	5	4	2.95	1.88	0.52	82
2001-06-04	114	16.2	19.7	22.5	1	1	0	0.98	0.35	-0.17	92
2001-06-05	94	17.4	20.5	20.4	8	8	7	-0.50	-2.95	-4.33	114
2001-06-06	80	17.7	19.8	18.3	6	6	7	-5.64	-5.00	-6.00	94
2001-06-07	79	16.8	15.6	14.9	7	8	8	-4.33	-1.88	-3.76	80
...	...	...									

Questions :

- Quelles variables peuvent jouer sur le maximum d'ozone ?
- Est-ce qu'il est possible de prévoir la concentration d'ozone ?

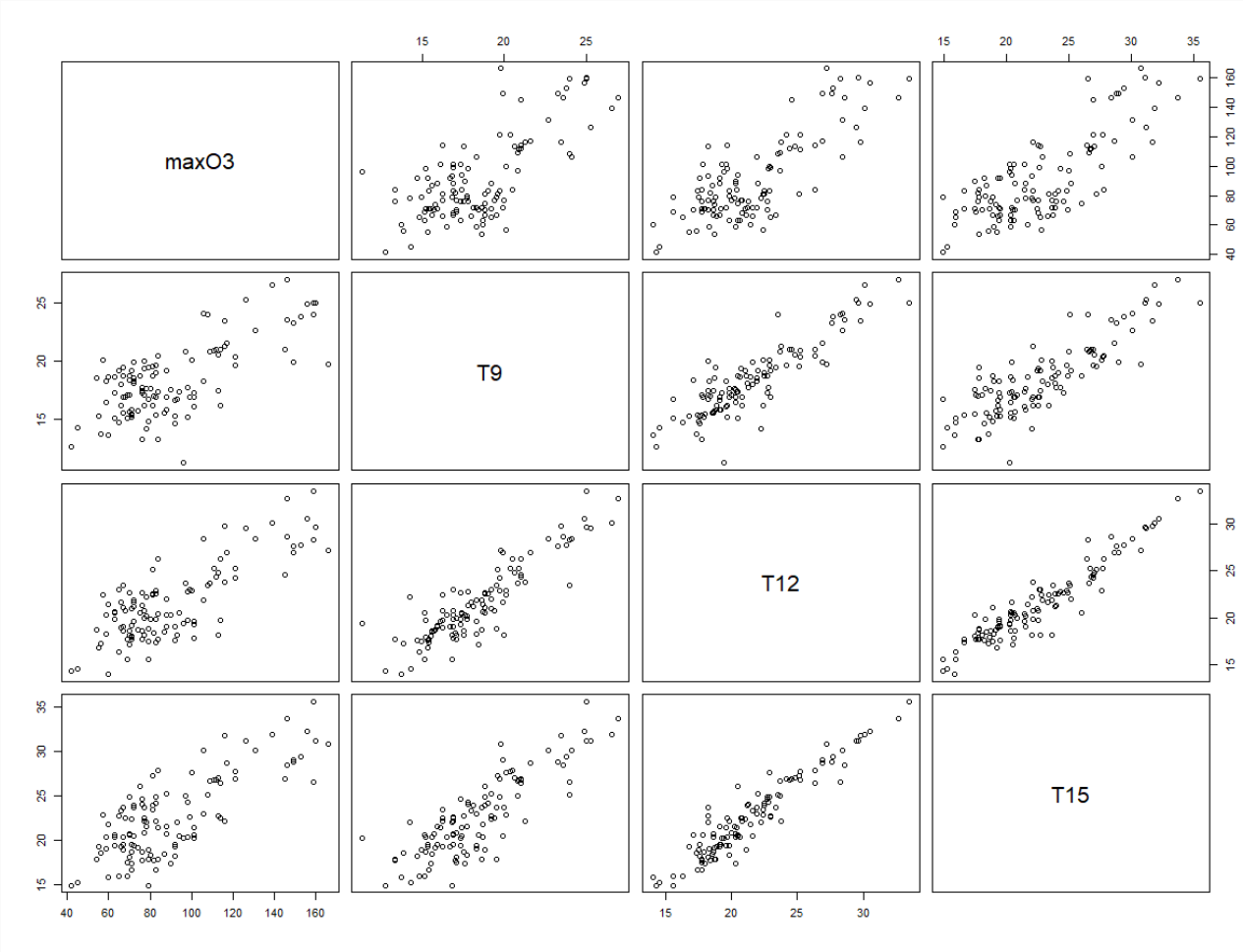
Objectifs :

- Objectif explicatif : Expliquer une variable quantitative Y en fonction de p variables quantitatives $X_1, ..., X_p$
- Objectif de prédiction : Prédire de nouvelles valeurs pour Y

Visualisation des données :

Questions :

- A quoi peuvent servir ces graphes ?
- Quelle est l'utilité de faire des graphes ?
- Qu'est ce qu'on peut voir avec ces graphes ?



En R, la fonction pairs permet de faire un nuage de points avec une variable en abscisses et une variable en ordonnée

Régressions linéaires (simples, multiples, généralisée) : Outil multitâche

- Faire de l'inférence : "tirer des conclusions concernant une population via une analyse statistique sur un échantillon de cette population" -¿ tests statistiques
- Prédiction : Prédire des nouvelles valeurs d'une variable en fonction des autres variables
- Régression linéaire généralisé : classification (régression logistique ou ordinale)
- C'est la base de beaucoup d'autre modèle : sur les séries temporelles par exemple, modèle de survie (estimée la survie des patients ....) ....

Une méthode statistique permet de modéliser la relation linéaire entre deux variables quantitatives dans un objectif explicatif et/ou prévisionnel.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

où ε est une variable aléatoire qui représente :

- le bruit ou l'erreur de mesure.
- l'erreur d'échantillonnage
- les facteurs mal contrôlés
- les oubli de facteur

★ β_0 et β_1 sont des inconnus.

▶ Ces inconnues sont estimées à partir d'un échantillon de n points $(x_1, y_1), ..., (x_n, y_n)$.

- le vecteur Y est un vecteur aléatoire

Le modèle s'écrit sous forme indicée :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, ..., n$$

- Les variables aléatoires ε_i sont i.i.d., $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}(\varepsilon_i) = \sigma^2$pour tout $i = 1, ..., n$.
- $\text{cov}(\varepsilon_i, \varepsilon_k) = 0$ pour tout $i = 1, ..., n$.

Estimation

On rappelle que le modèle s'écrit de la façon suivante :

$$y_i = \beta_0 + \beta_1 \, x_i + \varepsilon_i$$

où les ε_i sont appelées les erreurs, les résidus du modèle.

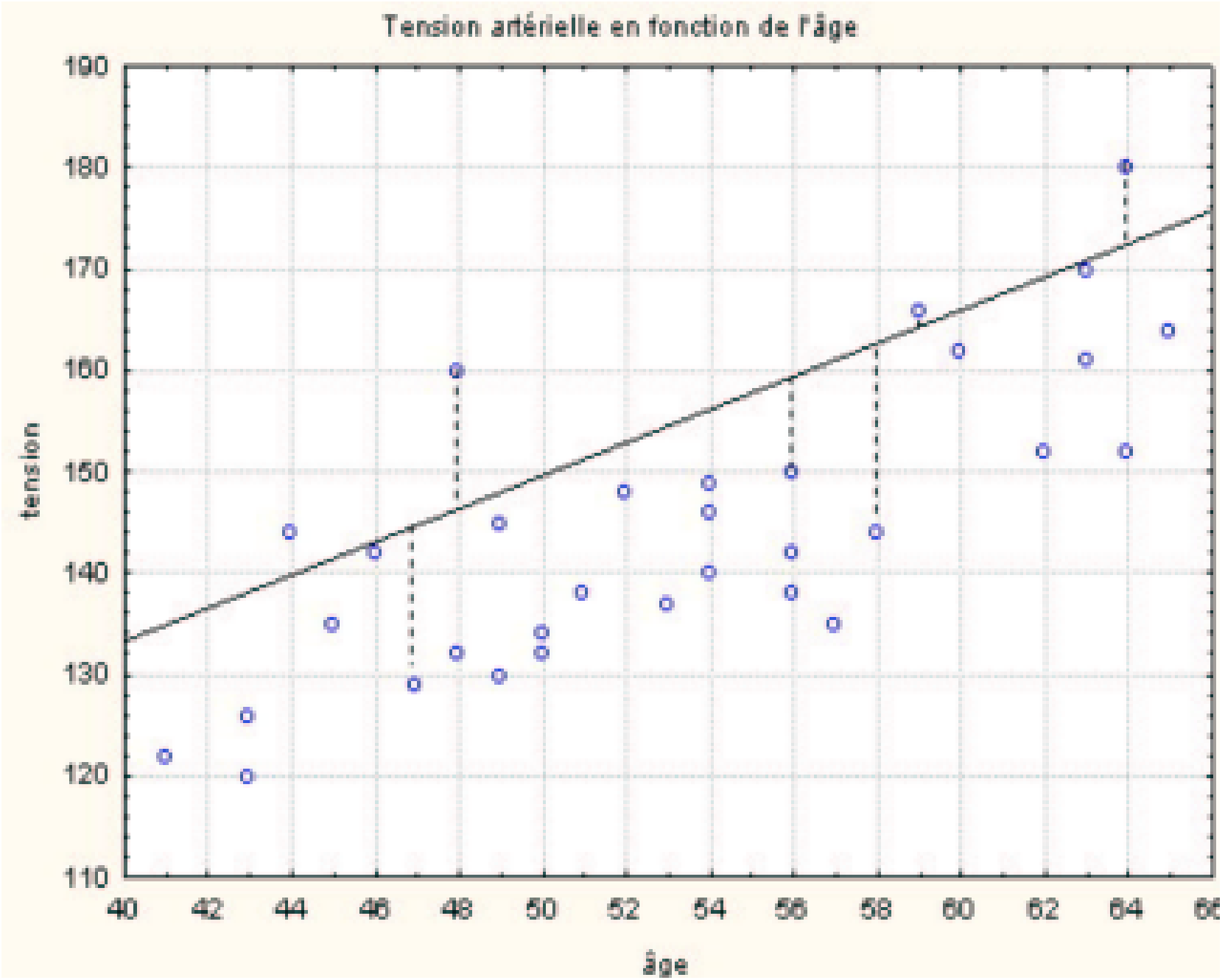
Comment choisir β_0 et β_1 ?

Le modèle sera bon si les erreurs $\varepsilon_i = y_i - (\beta_0 + \beta_1 \, x_i)$ sont petites.

Une idée est donc de choisir pour β_0 et β_1 ceux qui minimisent la quantité suivante :

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \, x_i))^2$$

C'est la méthode des moindres carrés.


Tension artérielle en fonction de l'âge
tension
âge
190
180
170
160
150
140
130
120
110
40
42
44
46
48
50
52
54
56
58
60
62
64

Estimateur par moindres carrés

On cherche donc et on note $\widehat{\beta}_0$ et $\widehat{\beta}_1$ ceux qui minimisent la quantité suivante :

$$\phi(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \, x_i))^2$$

Après calculs, on obtient :

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \, \overline{x} \qquad \text{et} \qquad \widehat{\beta}_1 = \frac{cov(x, y)}{var(x)}$$

avec $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i, \quad \overline{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i,$

$var(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2$ et $cov(x, y) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$

Définition 2.1 :

- $\mathrm{SCT} = \sum_{i}^{n} (y_i - \bar{y})^2$
- $\mathrm{SCE} = \sum_{i}^{n} (\widehat{y}_i - \bar{y})^2$
- $\mathrm{SCR} = \sum_{i}^{n} (\widehat{y}_i - y_i)^2$
- $R^2 := \frac{\mathrm{SCE}}{\mathrm{SCT}} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}}$

SCT Somme des carrées totale, SCE Somme des carrées expliqué (par la régression) SCR Somme des carrées des résidus

Propriété 2.1 :

$SCT = SCR + SCE$

R^2 est un indicateur de la qualité de la régression.

Écriture matricielle

- $X = (x_1, .., x_n) \in \mathbb{R}^n$: n observations de la variable explicative
- $Y = (y_1, .., y_n) \in \mathbb{R}^n$: n observations de la variable à expliquer
- $\mathbf{1} = (1, .., 1) \in \mathbb{R}^n$
- $\mathbf{1} = (\epsilon_1, .., \epsilon_n) \in \mathbb{R}^n$

Les données d'un point de vue géométrique

En utilisant l'interprétation graphique et en appliquant le théorème de Pythagore, on a :

$$\|Y - \bar{y}\mathbf{1}\|^2 = \|\tilde{Y} - \bar{y}\mathbf{1}\|^2 + \|Y - \tilde{Y}\|^2$$

Définition 2.2 :

- $\mathrm{SCT} = \|Y - \bar{y}\mathbf{1}\|^2$
- $\mathrm{SCE} = \|\tilde{Y} - \bar{y}\mathbf{1}\|^2$
- $\mathrm{SCR} = \|Y - \tilde{Y}\|^2$
- $R^2 := \frac{\mathrm{SCE}}{\mathrm{SCT}} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}}$

Interprétations de R^2

- R^2 correspond au cosinus de l'angle θ entre l'hypothénuse et la base du triangle $(Y, \tilde{Y}, \bar{y}\mathbf{1})$
- Si $R^2 = 1$ alors le modèle explique tout. On a alors $\theta = 0$. Les points de l'échantillon sont parfaitement alignés.
- Si $R^2 = 0$ on a alors pour tout $i \in 1, N, \tilde{Y}_i = \bar{y}$. On ne modélise rien de mieux que la moyenne. Le modèle est alors inadapté.
- En général le modèle interprète $100xR^2\%$ de la variance totale des données.

Regression linéaire Normal

En supposant ϵ_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, on obtient des résultats sur les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$. σ^2 devient aussi un paramètre à estimer.

Les estimateurs par moindre carrée (ou par maximum de vraisemblance, ce sont les mêmes) sont alors :

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}^2_{\widehat{\beta_0}}} \sim \mathcal{T}(n-1)$$

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}^2_{\widehat{\beta_1}}} \sim \mathcal{T}(n-1)$$

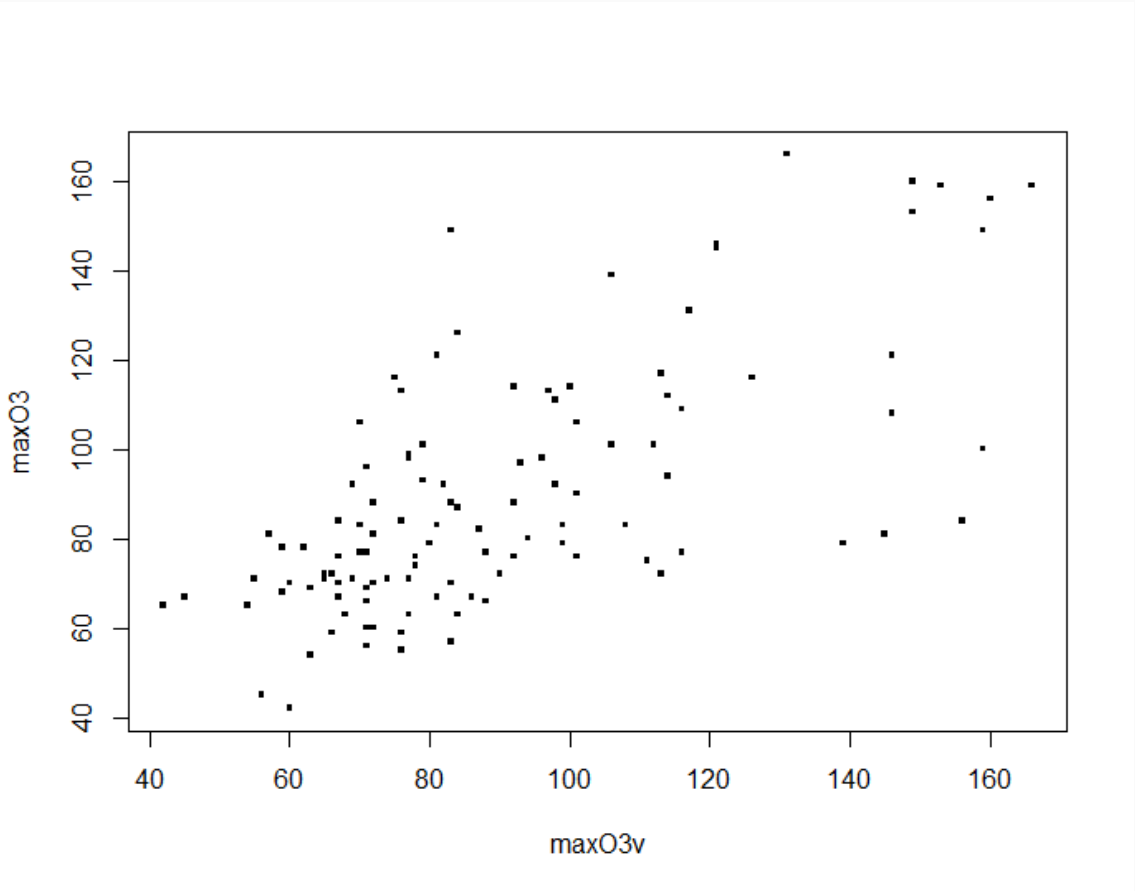
$$\frac{(n-2)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$$

On peut donc faire de l'inférence et tester des hypothèses.

Par exemple, on va tester l'hypothèse $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ par un test de Student.

Lors d'une régression sur R, les valeurs du test de Student et les p-valeurs associées sont sorties par la fonction de régression

Le nuage de points avec le maximum d'ozone en fonction du maximum d'ozone de la veille



plot(maxO3∼maxO3v, data = ozone, pch=15, cex=.5)

```
> #Estimer les parametres
> summary(lm(maxO3~maxO3v, data = ozone))

Call:
lm(formula = maxO3 ~ maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-50.949 -14.687  -0.767  11.545  63.863

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.50249    6.57153   4.337 3.21e-05 ***
maxO3v       0.68235    0.06929   9.848  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.64 on 110 degrees of freedom
Multiple R-squared:  0.4686,    Adjusted R-squared:  0.4637
F-statistic: 96.99 on 1 and 110 DF,  p-value: < 2.2e-16
```

Estimation des paramètres en pratique avec R

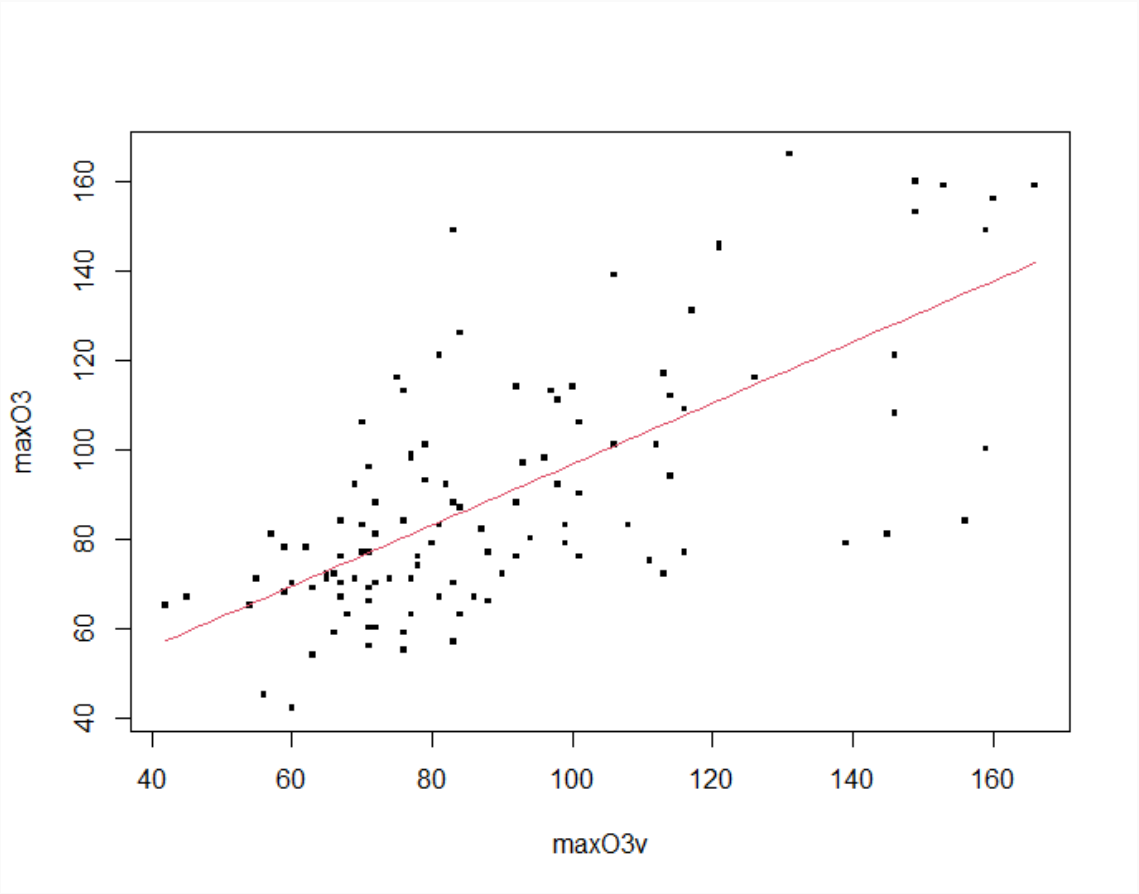
▶ (β_0,β_1) sont estimés ▶ Les tests de significativité des coefficients donnent des probabilités critiques d'environ ▶ la probabilité critique inférieure à 5% pour la constante indique que la constante doit apparaitre dans le modèle ($***$)

▶ La probabilité critique inférieure à 5% pour la pente indique une liaison significative entre maxO3 et maxO3v ($***$)

▶ l'estimation de l'écart type résiduel ▶ Le coefficient de détermination $R^2 = 46\%$

Droite de régression

- Une grille de points sur les abscisses
- Applique le modèle trouvé sur cette grille



Nuage de points et droite de régression

Le critère des moindres carrés, comme la vraisemblance appliquée à une distribution gaussienne douteuse, est très sensible à des observations atypiques, hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. On cherche à identifier les observations influentes c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer un valeur aberrante, corriger une erreur de mesure, ne rien faire..., cela dépend du contexte.

Effet levier : Écrivons les prédicteurs $\widehat{y_i}$ comme combinaisons linéaires des observations :

$$\widehat{y_i} = b_0 + b_1 x_i = \sum_{j=1}^{n} h_{ij} y_j \quad \text{avec} \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

en notant H la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement : $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de $\widehat{y_i}$.

Résidus : La différence entre la valeur observée et la valeur lissée (ou estimée) c'est le résidu estimé :$e_i = y_i - \widehat{y}_i$.

Résidus $_{(i)}$: $\mathrm{e}_{(i)i} = y_i - \widehat{y_{(i)i}} = \frac{e_j}{1-h_{ii}}$ où $\widehat{y_{(i)i}}$ est la prévision de y_i calculée sans la i ème observation (x_i, y_i).

Résidus standardisés : Les résidus n'ont pas la même variance : $E(e_i) = 0$ et $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$. On calcul alors des versions standardisées afin de les rendre comparables : $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$

Résidus studentisés : La standardisation ("interne") dépend de e_i dans le calcul de s estimation de $\mathrm{Var}(e_i)$. Une estimation non biaisée de cette variance est basée sur

$$s^2_{(i)} = \left[(n-2)s^2 - \frac{e_i^2}{1-h_{ii}}\right] / (n-3)$$

qui ne tient pas compte de la ième observation. On définit alors les résidus studentisés par :

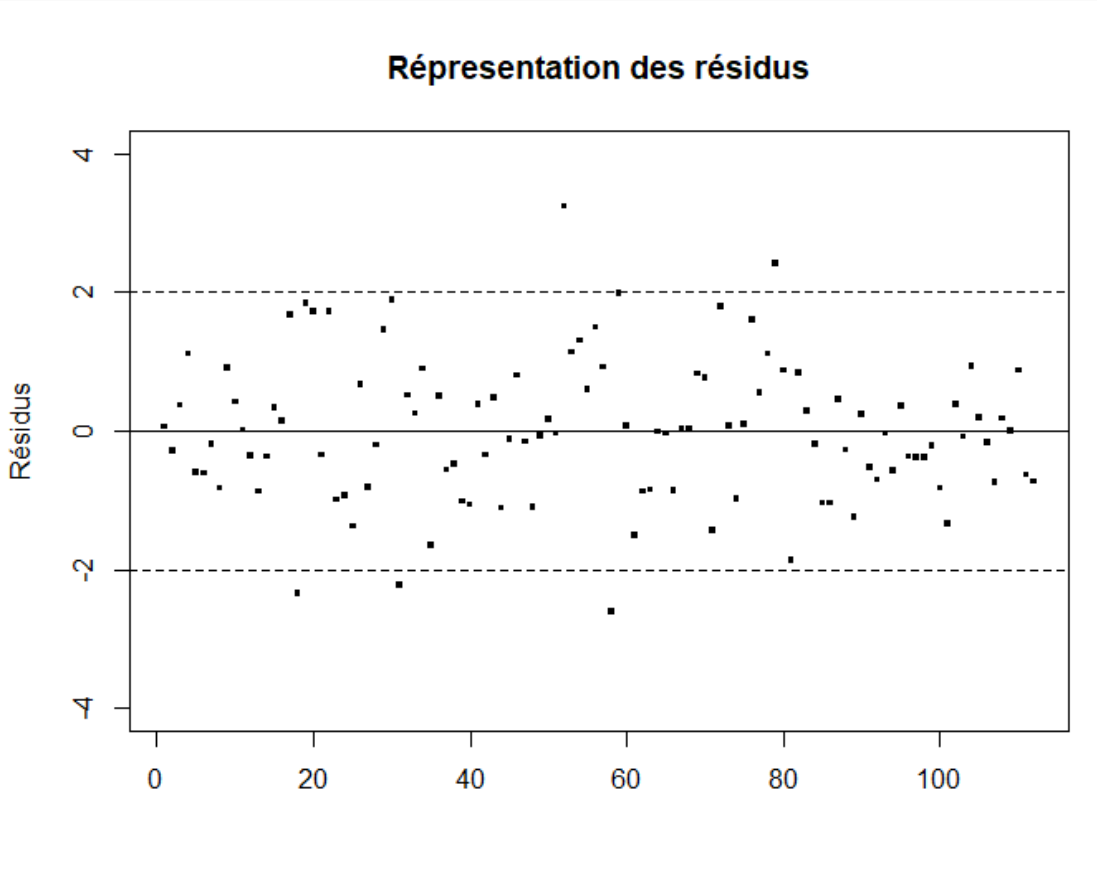
$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $n3$ degrés de liberté. Il est ainsi possible de construire un test afin tester la présence d'une observation atypique.

Plus concrètement, en pratique, les résidus studentisés sont comparés aux bornes ± 2.

Analyse les résidus

l'analyse des résidus est primordiale $\Rightarrow$ vérifier l'ajustement individuel du modèle (point aberrant).
Les résidus sont obtenus par la fonction residuals, cependant les résidus obtenus ne sont pas de même variance
$\Rightarrow$ On utilise des résidus "studentisés" (de même variance).



En théorie 95% (parmi 112 jours) des résidus studentisés se trouvent dans l'intervalle $[-2, 2]$. C'est le cas ici puisque 5 résidus seulement se trouvent à l'extérieur de cet intervalle.

Q-Q plot Un Q-Q plot (Quantile-Quantile plot) $\rightarrow$ comparer la distribution d'un échantillon avec une distribution théorique (par exemple, une distribution normale) ou de comparer deux échantillons entre eux.

Permet de vérifier si les données suivent une distribution spécifique, en particulier une distribution normale.

On peut utiliser un Q-Q plot pour vérifier la normalité des résidus ou si les résidus studentisées suivent bien une loi de Student.

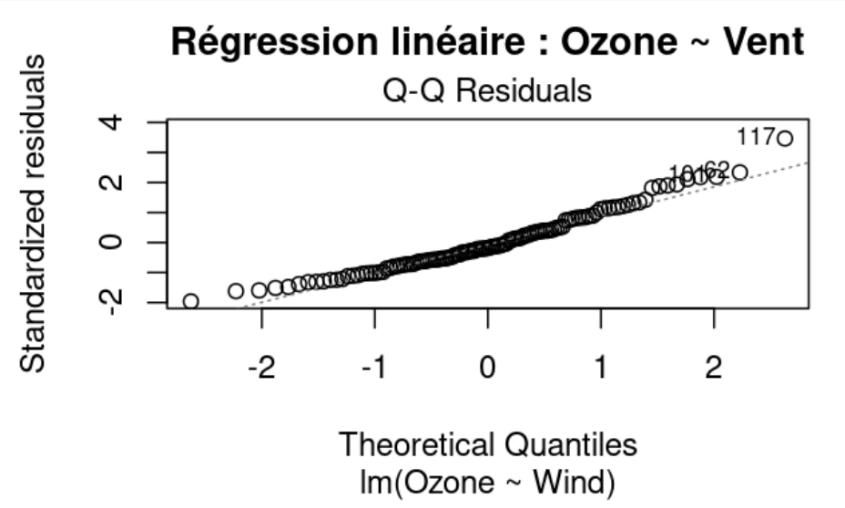


Figure 1: Q-Q plot résidus standardisés

Intervalle de confiance

$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2}$$

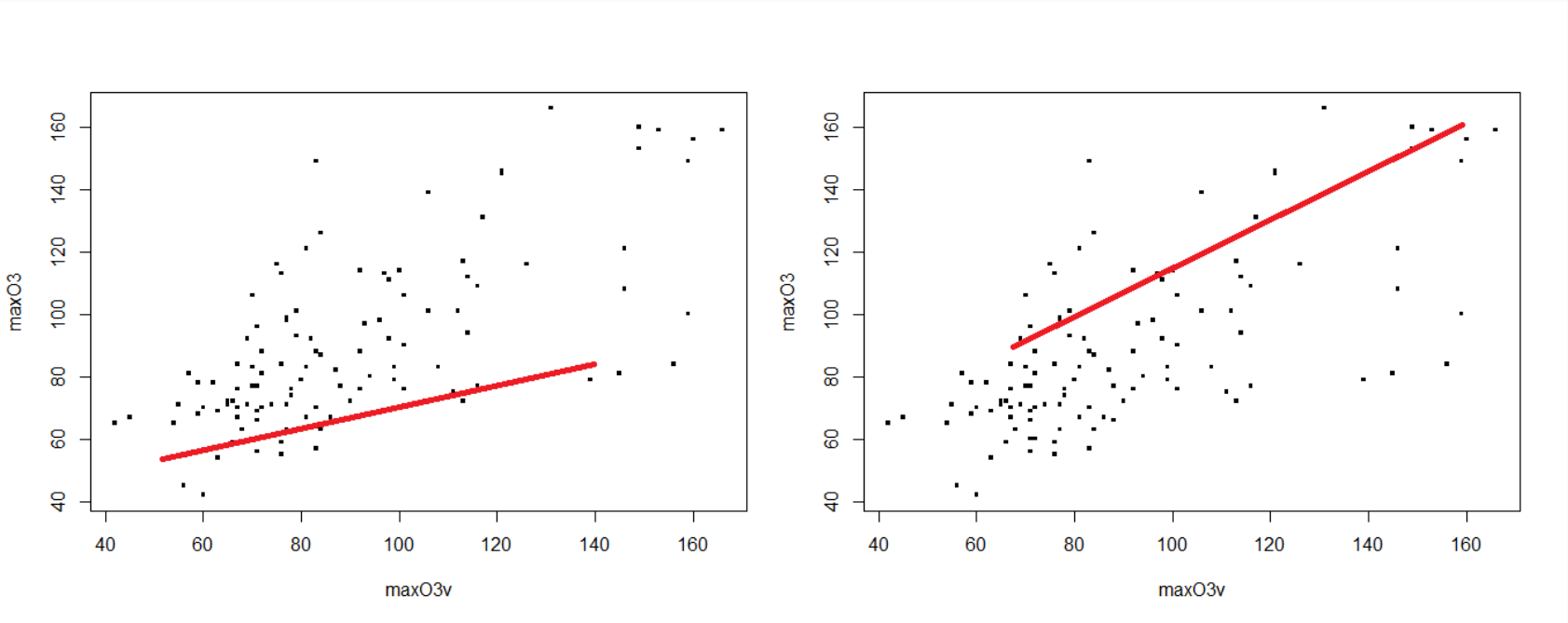
Prédiction

```
> xnew <- 73
> xnew <- as.data.frame(xnew)
> colnames(xnew) <- "max03v"
> predict(regS,xnew,interval="pred")
       fit      lwr      upr
1 78.31377 37.15394 119.4736
```

- Une nouvelle observation xnew, il suffit d'utiliser les estimations pour prévoir la valeur Y correspondante.

- L'argument xnew de la fonction predict doit être un data-frame avec les mêmes noms de variables explicatives.

  - Intervalle de prédiction $\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)^{1/2}$

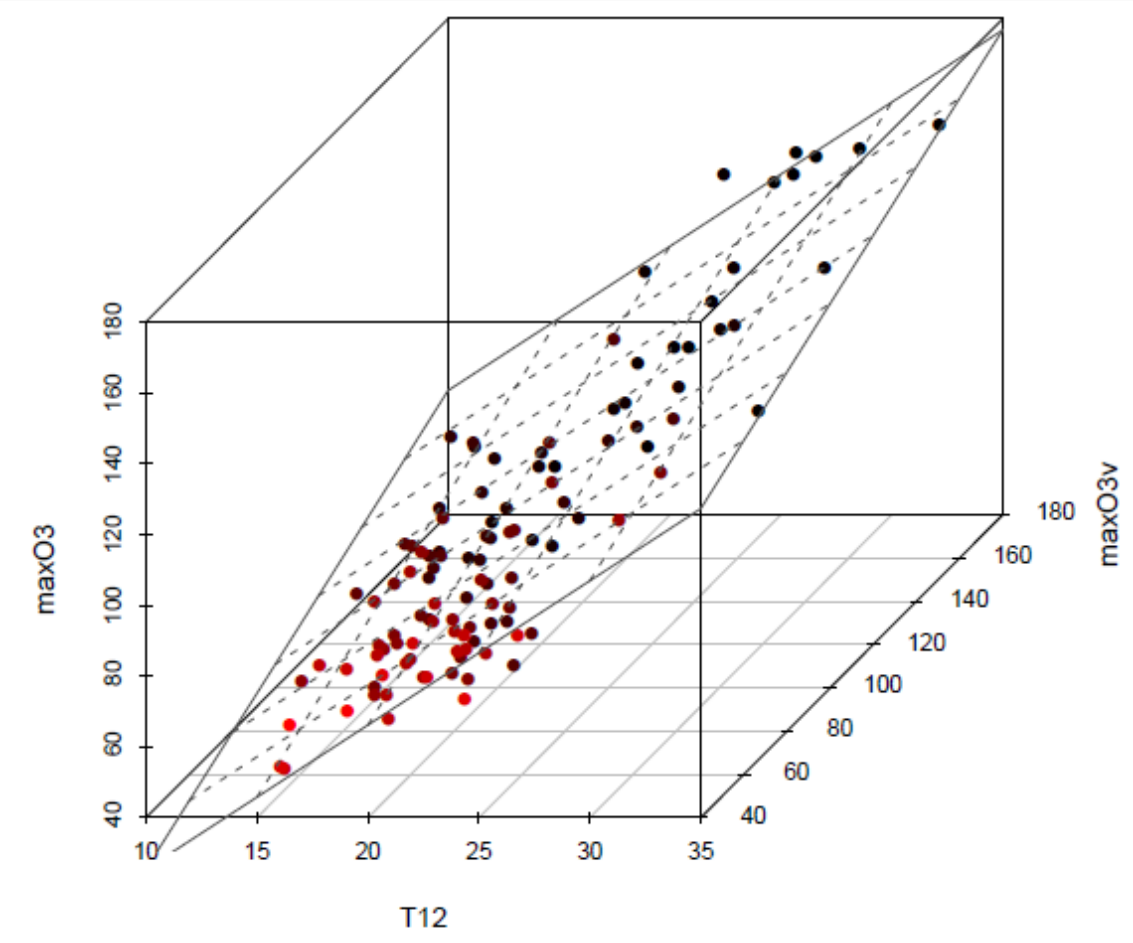


- La régression linéaire simple versus la régression linéaire multiple

★ « simple » indique → qu'on a une seule variable explicative.

★ « multiple » → indique qu'on a au moins deux variables explicatives.

$\Rightarrow$ La nécessité de prendre en compte d'autres variables pour expliquer le maximum d'ozone.



Prendre en compte l'effet des deux variables maxO3v et T12.

Modèle de Régression Multiple

Le modèle s'écrit sous forme indicée :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \forall i = 1, ..., n$$

- ε_i i.i.d., $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, ..., n.$
- $\text{cov}(\varepsilon_i, \varepsilon_k) = 0 \quad \forall i = 1, ..., n.$

Le modèle traduit l'influence de chaque variable sur Y

- Linéarité du modèle : linéarité par rapport aux paramètres
- Additivité : les effets des variables s'additionnent
- Modèle polynomial possible : $y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^2 + \varepsilon_i$

D'un point de vue matriciel :

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & & x_{ij} & & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbb{E}(\epsilon) = 0, \quad \mathbb{V}(\epsilon) = \sigma^2 \mathrm{Id}$$

$\Rightarrow$ Estimer le vecteur de paramètres β.

Estimation des paramètres du modèle

Critère des moindres carrés : Méthode de régression

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 = \arg\min_{\beta} \|Y - X\beta\|^2$$

$\Rightarrow$ Minimiser l'écart entre l'observation et la prévision par le modèle le tout au carré

$$\widehat{\beta} = ({}^tXX)^{-1}\ {}^tXY \quad \textbf{si } {}^tXX \textbf{ est inversible}$$

Prédiction et résidus

Valeurs prédites :

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}.$$

Résidus

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

Analyse de la variance et F-test Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients sont nuls (à l'exception du terme constant) contre le modèle complet avec toutes les variables.

Les tests de Fisher, également appelés tests F en régression linéaire, permettent de tester la nullité d'un ensemble de paramètres.

Considérons un modèle de régression linéaire multiple M_1 :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d + \epsilon,$$

où $X \in \mathbb{R}^d$ et $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ sont i.i.d.

On note SS_{M_1} l'erreur quadratique du modèle M_1 :
$SS_{M_1} = \frac{1}{n} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$.

Considérons le test d'hypothèse $\mathcal{H}_0^k$:
$\{j_1, \ldots, j_k\} \subset \{1, \ldots, d\}, \quad \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_k} = 0$. On note alors le modèle restreint M_2 : $y = \sum_{j \notin \{j_1, \ldots, j_k\}} \beta_j x_j + \epsilon$. La statistique F s'écrit alors :

$$F = \frac{\left(\sum_{i=1}^n (y_i - \widehat{y}_i^{M_2})^2 - \sum_{i=1}^n (y_i - \widehat{y}_i^{M_1})^2 \right) / k}{\sum_{i=1}^n (y_i - \widehat{y}_i^{M_1})^2 / (n - d - 1)}.$$

Sous l'hypothèse $\mathcal{H}_0^k$, on a $F \sim \mathcal{F}(k, n - d - 1)$.

En particulier, si on considère l'hypothèse $\mathcal{H}_0^{d-1} : \forall j \neq 0, \quad \beta_j = 0$, les estimateurs du modèle M_2 sont $\widehat{y}_i^{M_2} = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

D'après un résultat bien connu en régression linéaire, la variance totale est égale à la variance expliquée par le modèle additionné à la variance des résidus, c'est-à-dire :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\widehat{y}_i^{M_1} - \bar{y})^2 + \sum_{i=1}^n (y_i - \widehat{y}_i^{M_1})^2.$$

La statistique F s'écrit alors :

$$F \quad = \quad \frac{\left(\sum_{i=1}^n (\widehat{y}_i^{M_1} - \bar{y})^2\right)/(d-1)}{\left(\sum_{i=1}^n (y_i - \widehat{y}_i^{M_1})^2\right)/(n-d-1)},$$

et $F \sim \mathcal{F}(d-1, n-d-1)$. (loi de Fisher)

Le test de Fisher, permet alors de tester des modèles emboîtés.

Par exemple, on souhaite tester si ajouter une variable au modèle est intéressant en terme d'inférence.

Le modèle initial est : $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \epsilon_i$

le modèle complet est : $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon_i$ va permettre de tester si le modèle complet explique mieux la variance que le modèle initial

Ça peut être une façon de sélectionner les variables à mettre dans une régression multivariée.

Critère

- Le critère AIC (Akaike's Information Criterion) : Critère à minimiser
- Le critère BIC (Bayesian Information Criterion) : Critère à minimiser
- $R^2_{\text{ajusté}} = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right)$ avec p est le nombre de prédicteurs (variables explicatives) dans le modèle.
  - Pénalise l'ajout de nouvelles variables
  - le R^2 ajusté peut diminuer à l'ajout d'une nouvelle variable si elle
  - Comparaison avec R^2 :
  - Si $R^2_{\text{ajusté}} > R^2$, cela indique que les variables ajoutées sont utiles.
  - Si $R^2_{\text{ajusté}} < R^2$, cela indique que les variables ajoutées ne sont probablement pas significatives.

Méthodes algorithmiques de sélection - Forward

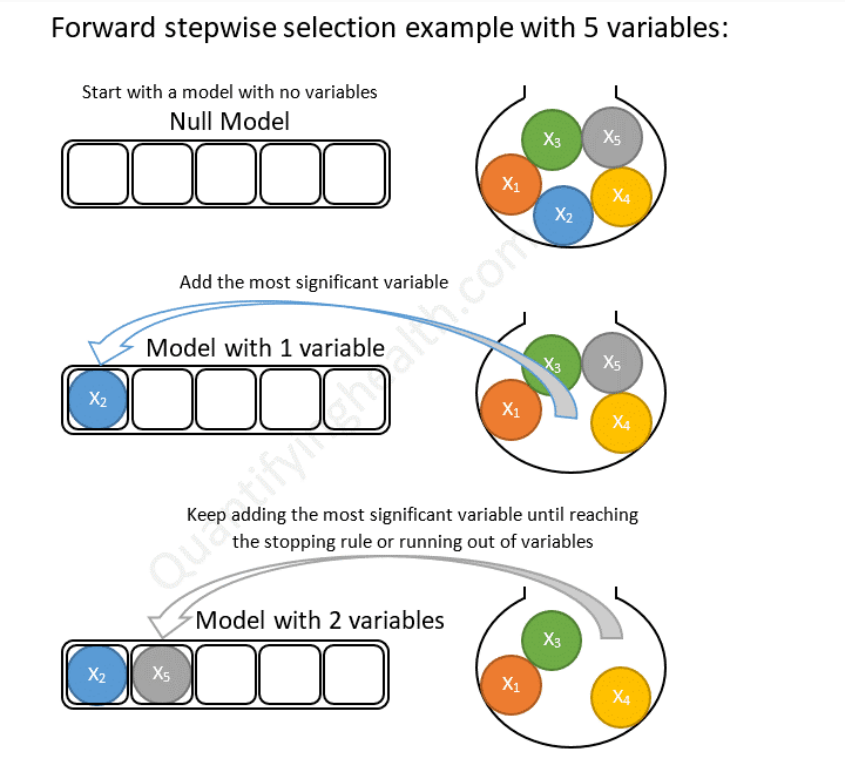


Figure 2: Procédure Forward

La variable la plus significative à ajouter peut-être sélectionnée parmi plusieurs méthodes :

- La plus petite p-valeur
- la variable qui augmente le plus la R2 ou le R2 ajusté
- la variable qui diminue le plus la somme des résidus au carré

La règle d'arrêt peut être calculé par un seuil sur les critères AIC ou BIC

Méthodes algorithmiques de sélection - Backward

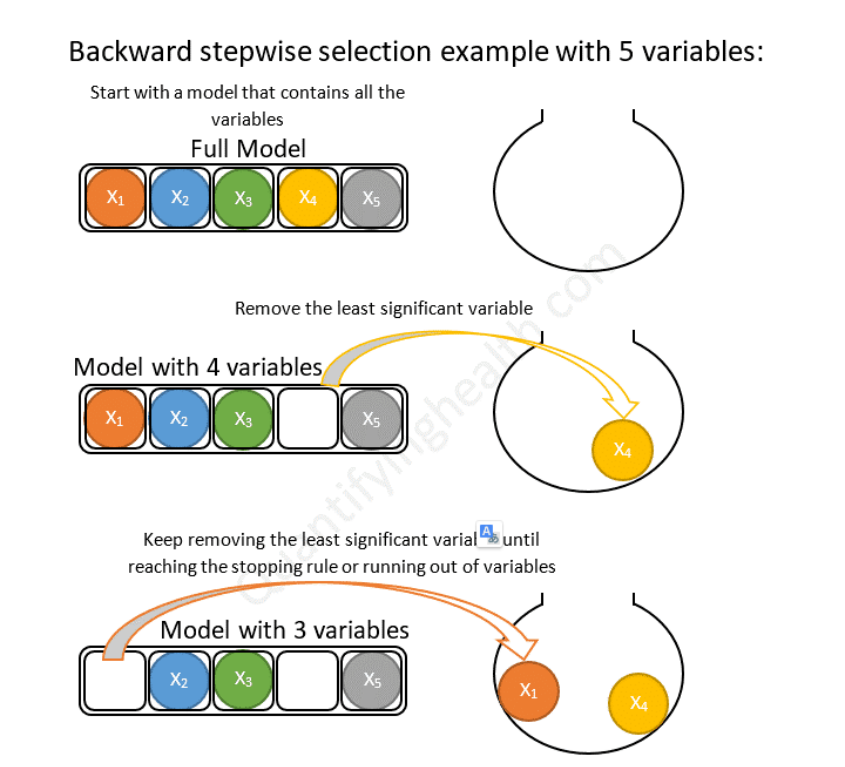


Figure 3: Procédure Backward

Conclusion : Il y a plusieurs façons de sélectionner des variables. Pour réaliser une sélection de variables, il faut alors choisir une méthode et s'y tenir en justifiant ses choix. Parfois plusieurs options sont possibles, les décisions prises peuvent dépendre du cas d'application et de l'utilisation du modèle.

D'autre type de régression permettent de séléctionner des variables dans le cas de grande dimension, par exemple pénalisation LASSO, Ridge , Elasticnet qui nous évoqueront dans le cours d'introduction au machine-learning.

Régression logistique : qu'est-ce que c'est ?

- Expliquer et prédire les réalisations d'une variable réponse Y de type catégorielle ou plus simplement binaire à partir de p variables explicatives $(X_1, \ldots, X_p)$ qualitatives et/ou quantitatives.

- Dans la suite on se concentrera sur le cas où Y est une variable binaire.

- Exemple 1 : Classer une tumeur bénigne (codée 0) ou maligne (codée 1) en fonction d'un certain nombre de caractéristiques de celle-ci.

- Exemple 2 : Classer un client mauvais (codé 0) ou bon (codé 1) en fonction d'un certain nombre de caractéristiques de celui-ci.

Vers les GLM

- Une approche modèle linéaire ? ? typiquement

$$\mathbb{E}(Y|X_1 = x_1, \ldots, X_n = x_n) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

mais Y est binaire *donc "loin" d'une loi normale* et

$$\mathbb{E}(Y|X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(Y = 1|X_1 = x_1, \ldots, X_n = x_n) \in [0, 1]$$

. . .*et rien n'assure que* $\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$ *sera bien dans* $[0, 1]$ *!*

$\hookrightarrow$ L'approche par modèle linéaire usuelle n'est pas adaptée au cas de réponses non-normales comme des binaires, des proportions, des comptages, des catégories . . .**nécessité de développer une généralisation pour de telles réponses : le modèle linéaire généralisé**.

- La variable à expliquer est donc de type qualitatif binaire. On va donc plutôt chercher à expliquer les probabilités

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})$$

ou tout du moins une transformation de celles-ci.

- Cela va se faire toujours à partir d'une combinaison linéaire de variables explicatives.

- On cherche donc une fonction g monotone de $[0, 1]$ dans $\mathbb{R}$ telle que

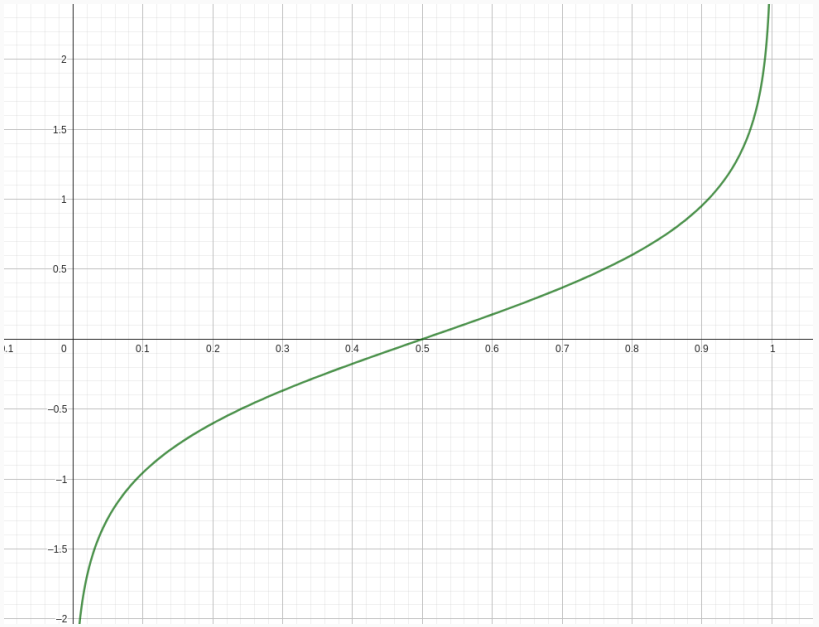
$$g(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$\hookrightarrow$ **modèle linéaire généralisé (GLM) avec une fonction de lien** g

La régression logistique a pour objectif d'expliquer et de prédire les valeurs d'une variable quantitative Y, le plus souvent binaire, à partir de variables explicatives $X = (X_1, \ldots, X_p)$ qualitatives et explicatives. Si on note 0 et 1 les modalités de Y, le modèle logistique s'écrit :

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

où $p(x)$ désigne la probabilité $\mathbb{P}(Y = 1 \mid X = x)$ et $x = (x_1, \ldots, x_p)$ est une réalisation de $X = (X_1, \ldots, X_p)$.



$\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)} = \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}$ est appelé la côte ou "odds" en anglais

Odds ratio ou rapport des côtes : C'est le rapport des cotes des probabilités que $Y|X=0$ et $Y|X=1$

$$RC = \frac{\frac{\mathbb{P}(Y=1|X=1)}{1-\mathbb{P}(Y=1|X=1)}}{\frac{\mathbb{P}(Y=1|X=0)}{1-\mathbb{P}(Y=1|X=0)}}$$

On a en régression logistique $\beta_1 = log(RC)$

On peut alors interpréter le coefficient de régression : $RC = e^{\beta_1}$

Si X est binaire, RC représente le rapport des côtes du risque de survenue de Y chez les individus exposé par rapport au non exposé

Si X est quantitative, RC représente le rapport des côtes du risque de survenue de Y pour une augmentation de X

- Par exemple, considèrons une maladie la probabilité de tomber malade dans la population est de $1/4$. La probabilité de ne pas être malade est $3/4$.
- La côte est de $1/3$.
- Maintenant, on considère qu'en présence d'un symptôme, la probabilité d'être malade est de $4/5$. La probabilité de ne pas être malade est de $1/5$.
- La côte en présence de symptôme est de 4
- le rapport des côtes est $4/(1/3) = 12$
- $RC > 1$, les patients qui ont le symptôme ont plus de risque d'être malade que les autres.
- $RC < 1$, les patients qui ont le symptôme ont moins de risque d'être malade que les autres.
  Par l'estimation des coefficient dans la régression logistique,

La fonction $\log\left(\frac{u}{1-u}\right)$ fait le lien entre la probabilité $p(x)$ et la combinaison linéaire des variables explicatives.
Du coups,

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}.$$

Les coefficients $\beta_1 + \cdots + \beta_p$ sont estimés par la méthode de vraisemblance à partir des observations.

Forest plot

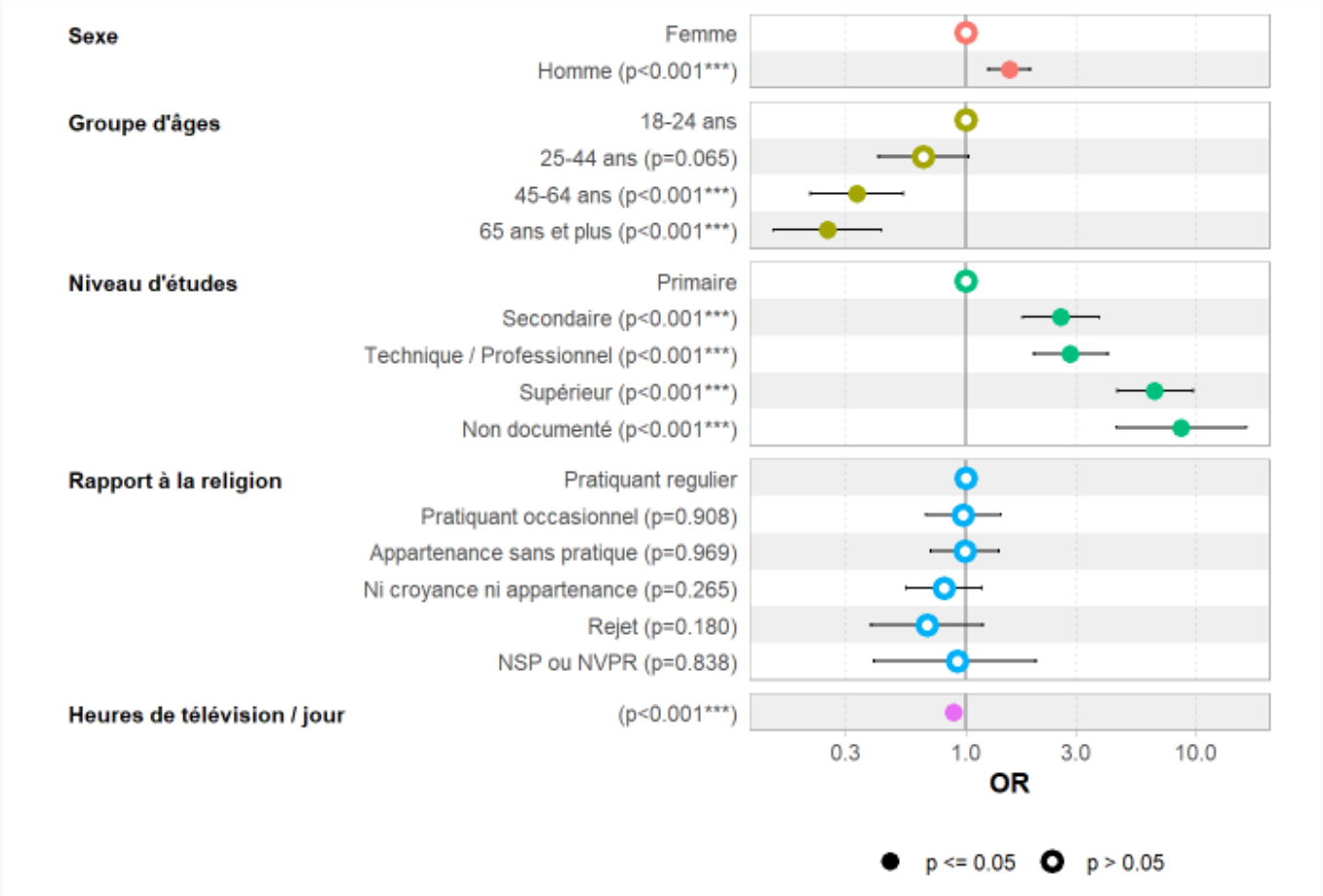


Figure 4: Forest plot

Seuil

Une fois les coefficients de la régression estimée par maximum de vraisemblance. On peut alors estimer la probabilité $\widehat{p(x)} = \widehat{P}(Y = 1|X = x)$

Pour classer les individus, il faut alors définir un seuil s, à partir duquel :

- si on a $\widehat{p(x_i)} \geq s$, alors $\widehat{y_i} = 1$
- si on a $\widehat{p(x_i)} < s$, alors $\widehat{y_i} = 0$

On peut alors calculer :

- les vrai positif : VP le nombre d'individu où l'évenement est prédit et ou l'évenement est en effet présent.
- les vrai positif : VN le nombre d'individu où l'évenement n'est pas prédit et ou l'évenement est en effet absent.
- les faux positif : FP le nombre d'individu où l'évenement est prédit et ou l'évenement est en réalité abscent.
- les faux : FN le nombre d'individu où l'évenement n'est pas prédit et ou l'évenement est en réalité présent.

Spécificité et Sensibilité

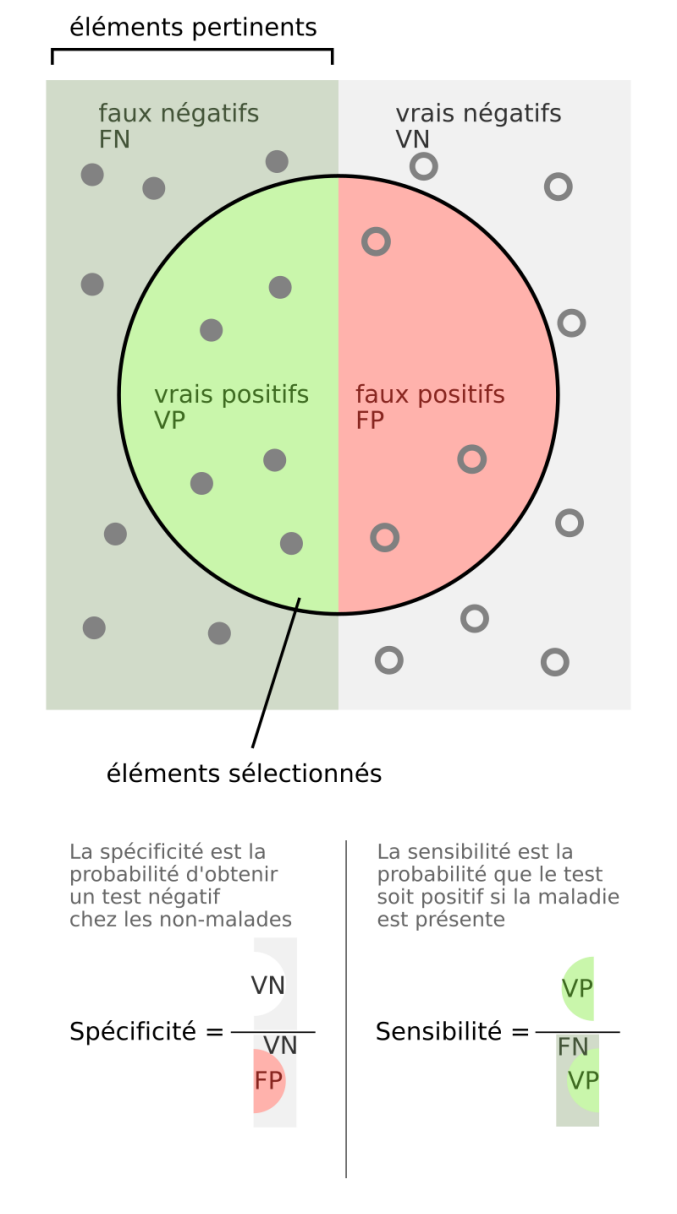


Figure 5: Spécificité et Sensibilité

Spécificité et Sensibilité

Sensibilité :

$$\frac{VP}{VP + FN}$$

C'est la probabilité de prédire l'évenement, si l'évenement est présent.

Spécificité :

$$\frac{VN}{VN + FP}$$

C'est la probabilité de ne pas prédire l'évenement, si l'évenement est absent.

En fonction de l'application, la mesure de sensibilité ou la mesure de spécificité peut-être plus ou moins importante.

Pour un test diagnostique d'une maladie, la sensibilité est très importante. On ne veut pas manquer un cas malade. Une bonne spécificité dans ce genre de test nous assurera de ne pas produire trop de faux positif, de ne pas desseller à tort des patients malades. Avoir une bonne sensibilité est plus important.

Précision et Rappel

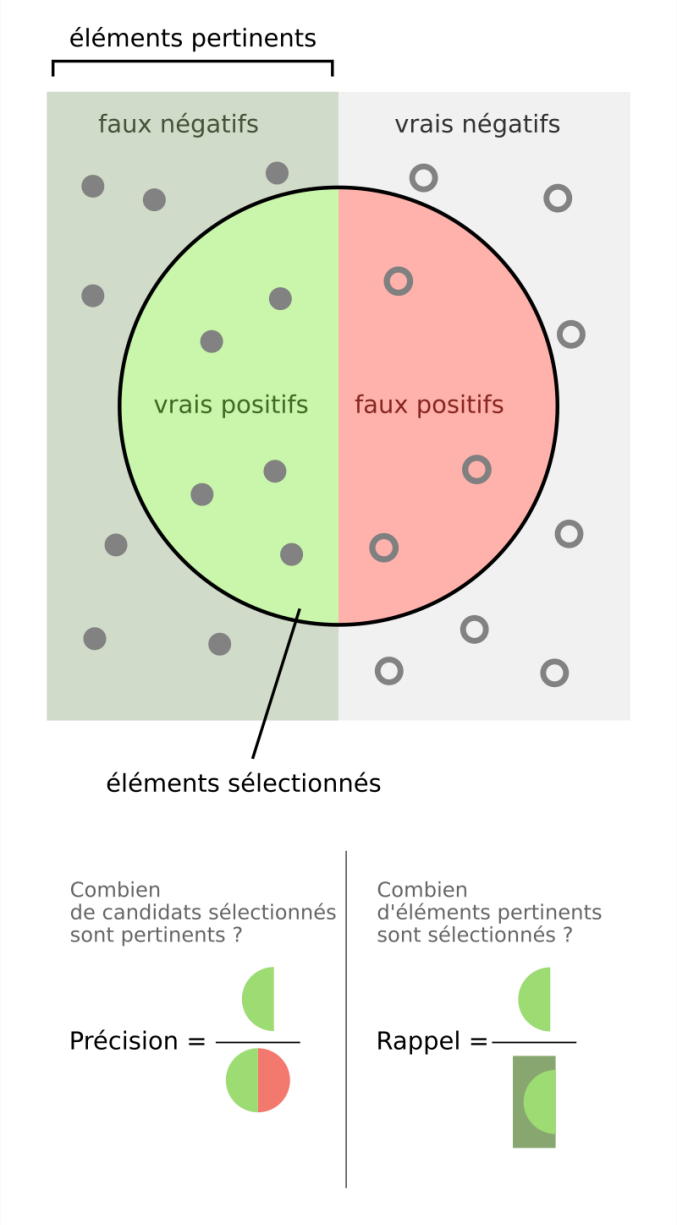


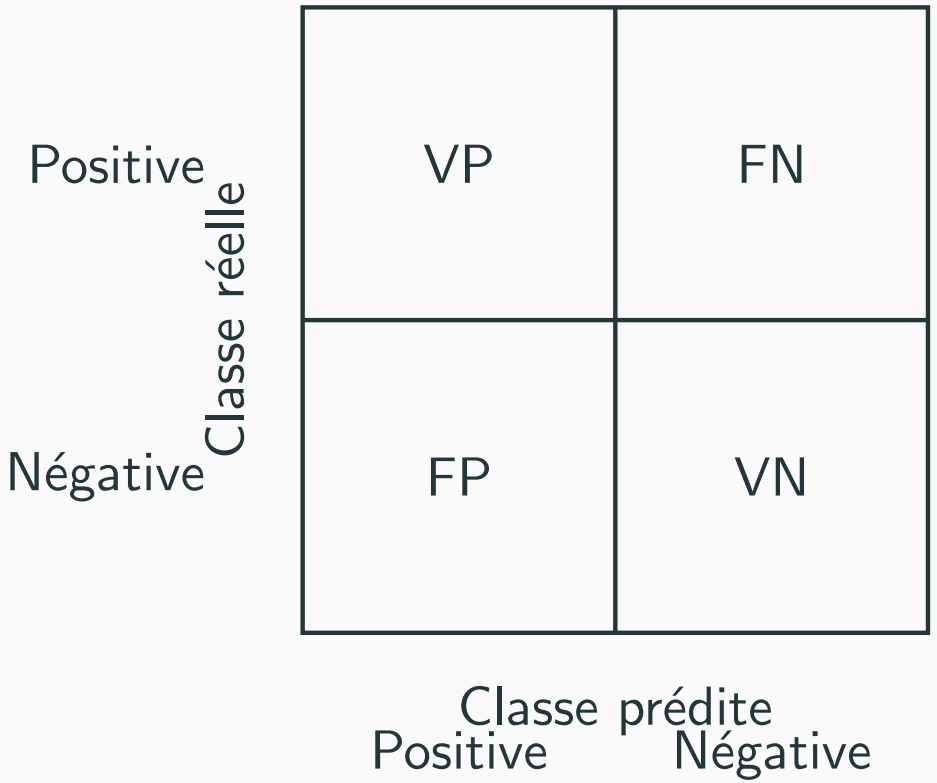
Figure 6: Précision et rappel

- Précision $= \frac{\text{VP}}{\text{VP}+\text{FP}}$
- Rappel $= \frac{\text{VP}}{\text{VP}+\text{FN}}$
- Accuracy $= \frac{\text{VP}+\text{VN}}{\text{Total}}$
- $F1 = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision}+\text{Rappel}}$

Matrice de confusion

La matrice de confusion est un outil permettant d'évaluer les performances d'un modèle de classification. Elle se présente généralement sous la forme suivante :

Voici une représentation graphique de la matrice de confusion :



Courbe ROC et AUC

Visualiser le compromis entre la **sensibilité** et la **spécificité** pour différents seuils de décision.

La courbe ROC est tracée en mettant FPR sur l'axe horizontal et TPR sur l'axe vertical.

L'**AUC** (*Area Under the Curve*) correspond à l'aire sous la courbe ROC. Elle quantifie les performances globales du modèle :

- AUC $= 1$: le modèle est parfait.
- AUC $= 0.5$: le modèle est équivalent à un tirage aléatoire.
- AUC < 0.5 : le modèle est pire qu'un tirage aléatoire.

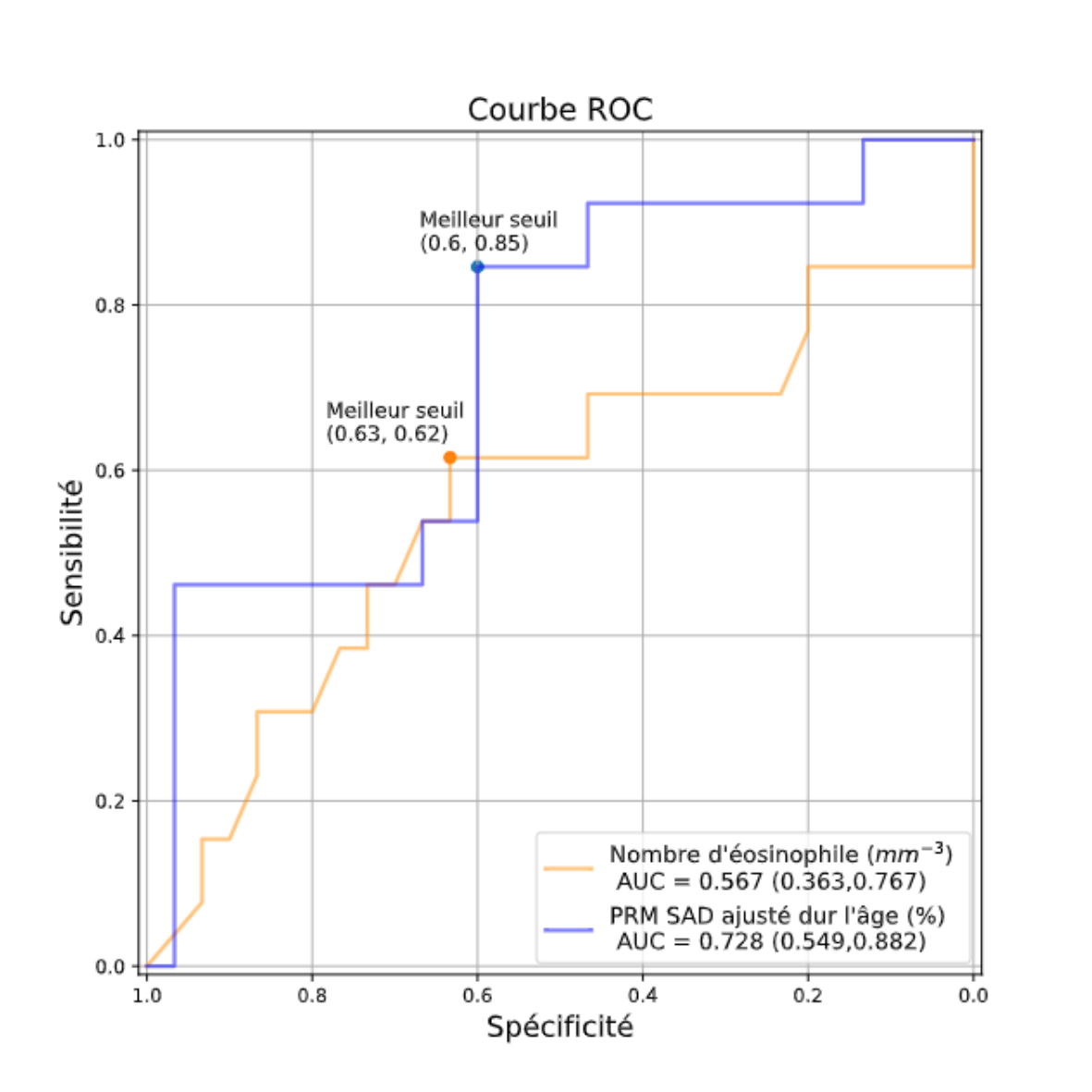


Figure 7: Courbe ROC : meilleur seuil décidé par l'Index de Youden

Le seuil est choisi à 0.5 par défaut mais peut-être modifié. On peut choisir un seuil qui optimise un critère en particulier selon les besoins de la classification : critère de Youden, Accuracy, F1 score, ect ...

Exemple :

Il s'agit d'expliquer et de prédire la réaction de clients suite à une campagne marketing effectué par téléphone. On cherche plus précisément à détecter les clients intéressés par le produit proposé en fonction d'informations sur le client (âge, emploi,...).

Le but est d'expliquer la variable binaire "YES" si le client a souscrit au produit, "NO" sinon.

Les étapes :

1) Importer les données
2) Construire le modèle
3) Sélectionner un modèle
4) Prévision
5) Évaluer la performance du modèle

Traitement de données :

Importer de données :

Les données sont issues de l'*UCI Machine learning Repository*. On dispose de 4119 individus (clients) sur lesquels on a observé la variable cible y ainsi que 20 autres variables potentiellement explicatives.

On importe et on résume les données du fichier *bank-additional.csv*.

```
> bank <- read.csv("bank-additional.csv", sep= ";")
> summary(bank)
      age            job              marital           education          default
 Min.   :18.00   Length:4119        Length:4119        Length:4119        Length:4119
 1st Qu.:32.00   Class :character   Class :character   Class :character   Class :character
 Median :38.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :40.11
 3rd Qu.:47.00
 Max.   :88.00
   housing              loan             contact             month           day_of_week
 Length:4119        Length:4119        Length:4119        Length:4119        Length:4119
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


    duration        campaign          pdays          previous       poutcome
 Min.   :   0.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   Length:4119
 1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   Class :character
 Median : 181.0   Median : 2.000   Median :999.0   Median :0.0000   Mode  :character
 Mean   : 256.8   Mean   : 2.537   Mean   :960.4   Mean   :0.1903
 3rd Qu.: 317.0   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
 Max.   :3643.0   Max.   :35.000   Max.   :999.0   Max.   :6.0000
  emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m      nr.employed        y
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.635   Min.   :4964   Length:4119
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.334   1st Qu.:5099   Class :character
 Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857   Median :5191   Mode  :character
 Mean   : 0.08497   Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5166
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228
```

2) Construire le modèle :

On divise tout d'abord aléatoirement la base de données en :

- Un échantillon d'apprentissage de taille 3000 qui sera utilisé pour estimer les modèles logistiques.
- Un échantillon test de taille 1119 qui sera utilisé pour mesurer la performance des modèles.

```
set.seed(5678)
perm <- sample(nrow(bank),3000)
training <-bank[perm,]
test <- bank[-perm,]
```

La régression logistique appartient à la famille des modèles linéaires généralisés.

La fonction glm : ajustement des modèles linéaires généralisés (similaire à la fonction lm). Il faut spécifier une famille de lois de probabilité. Dans le cadre de la régression logistique, il s'agit de la famille binomiale.

La fonction coef permet d'estimer les paramètres du modèle. Mais on doit vérifier si on a des données manquantes ou pas.

Cependant, lorsque certaines variables explicatives sont quantitatives, il est souvent préférable de tester la significativité de chaque variables dans le modèle plutôt que de regarder coefficient par coefficient

$\Rightarrow$ la fonction Anova du package **car**.

```
> Anova(logit, type =3, test.statistic = "LR", singular.ok = TRUE)
Analysis of Deviance Table (Type III tests)

Response: y
               LR Chisq Df Pr(>Chisq)
age                0.18  1   0.673914
job               13.95 11   0.235644
marital            4.08  3   0.252927
education          2.91  7   0.893276
default            1.73  2   0.421252
housing            0.10  1   0.754683
loan               0.04  1   0.838255
contact            8.70  1   0.003184 **
month             44.37  9  1.206e-06 ***
day_of_week        2.68  4   0.612370
duration         426.37  1  < 2.2e-16 ***
campaign           1.47  1   0.224731
pdays              1.29  1   0.256153
previous           0.34  1   0.562346
poutcome           7.29  2   0.026184 *
emp.var.rate       0.84  1   0.359647
cons.price.idx     0.04  1   0.842643
cons.conf.idx      2.10  1   0.147109
euribor3m          0.81  1   0.366655
nr.employed        1.59  1   0.207794
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3) Sélectionner un modèle :

Dans notre exemple, plusieurs variables sont considérées comme non significatives. La fonction step permet de sélectionner un modèle à l'aide d'un procédure pas à pas basée sur la minimisation du critère **AIC** (Akaike Information Criterion)

La procédure de sélection pas à pas effectuée par la fonction step est ici une procédure descendante $\text{DIRECTION} = \text{"BACKWARD"}$: à chaque étape, la variable dont le retrait du modèle conduit à la diminution la plus grande du critère AIC est enlevée. Le processus s'arrête lorsque toutes les variables sont retirées ou lorsque le retrait d'aucune variable ne permet de diminuer le critère.

```
logit_slect <- step(logit, direction = "backward")
Anova(logit_slect, type = 3, test.statistic = "LR")
```

```
Step:  AIC=1192.64
y ~ contact + month + duration + pdays + poutcome + cons.conf.idx +
    nr.employed

                Df Deviance    AIC
<none>               1158.6 1192.6
- pdays          1   1161.2 1193.2
- poutcome       2   1169.5 1199.5
- contact        1   1170.7 1202.7
- cons.conf.idx  1   1171.2 1203.2
- month          9   1209.0 1225.0
- nr.employed    1   1279.0 1311.0
- duration       1   1587.7 1619.7
```

4) Prévision :

▶ La fonction predict.glm (ou predict) : obtenir la probabilités $p(x) = \mathbb{P}(Y = yes \mid X = x)$ prédites par le modèle **logit-slect** des individus de l'échantillon test.

```
prev_slect <- predict(logit_slect, newdata = test, type = "respo
prev_slect
```

5) Évaluer la performance du modèle :

▶ On souhaite comparer les performances des modèles **logit** et **logit-slect** en estimant les taux de mal classés et les courbes **ROC** obtenues sur l'échantillon test.

▶ On obtient enfin les erreurs de classification estimées des 2 modèles en confrontant les valeurs prédites aux valeurs observées.