Statistique de la donnée

Chapitre 4 :
Modèle de régression linéaire multiple

Ali JAGHDAM
ESILV - 2020

Exemple introductif

On cherche à modéliser la relation entre poids des bébés à naissance et l'âge, le poids et le statut tabagique de la mère durant la grossesse. On pose :

- y = poids de naissance en grammes (bwt),
- x_1 = âge de la mère (age),
- x_2 = poids de la mère en kilos (weight),
- x_3 = statut tabagique de la mère pendant la grossesse (smoke) codée 1=oui et 0=non.

On suppose que cette relation est linéaire de la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

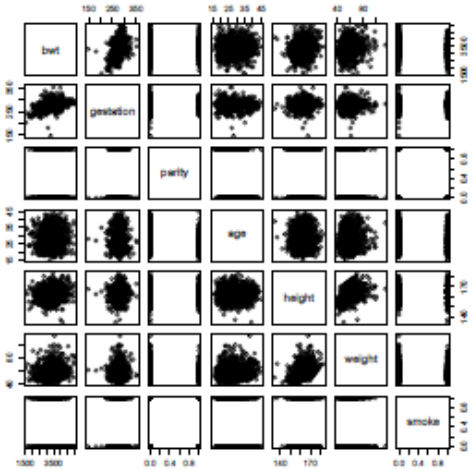
- On veut estimer cette relation avec un modèle de régression multiple.
- On utilise un échantillon de $n = 1174$ naissances pour lesquelles le poids du bébé, l'âge, le poids et le statut tabagique de la mère, ont été mesurés.

Exemple introductif

```
load("poids.RData")
print(data[1:5,c("bwt","age","weight","smoke")],digits=4)
```

```
##    bwt age weight smoke
## 1 3402  27  45.36     0
## 2 3203  33  61.23     0
## 3 3629  28  52.16     1
## 4 3062  23  56.70     1
## 5 3856  25  42.18     0
```

```
pairs(data) #diagrammes de dispersion
```




```
modele <- lm(bwt~ age+weight+smoke,data=data)
modele$coefficients
```

```
## (Intercept)         age      weight       smoke
##  3050.56238    -0.91802     7.90266  -254.25425
```

Le modèle

On cherche à modéliser la relation entre plus de 2 variables quantitatives.

Un modèle de régression linéaire multiple est de la forme suivante :

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon \tag{1}$$

où :

- y est la variable à expliquer (à valeurs dans $\mathbb{R}$) ;
- $x_1, \ldots, x_p$ sont les variables explicatives (à valeurs dans $\mathbb{R}$) ;
- ε est le terme d'erreur aléatoire du modèle ;
- $\beta_0, \beta_1, \ldots, \beta_p$ sont les paramètres à estimer.

Commentaires :

- La désignation "multiple" fait référence au fait qu'il y a plusieurs variables explicatives x_j pour expliquer y.
- La désignation "linéaire" correspond au fait que le modèle (1) est linéaire.

Pour n observations, on peut écrire le modèle de régression linéaire multiple sous la forme :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad \text{pour } i = 1, \ldots, n. \tag{2}$$

Dans ce chapitre, on suppose que :

- ε_i est une variable *aléatoire*, non observée,
- x_{ij} est observé et *non aléatoire*,
- y_i est observé et *aléatoire*.

On fait les trois **hypothèses additionnelles** suivantes :

(A1) $\mathbb{E}[\varepsilon_i] = 0, \ \forall i = 1, \ldots, n,$

ou de manière équivalente :

$\mathbb{E}[y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \ \forall i = 1, \ldots, n.$

Commentaire sur l'hypothèse (A1) : elle indique que les erreurs sont centrées

(A2) $\mathbb{V}(\varepsilon_i) = \sigma^2, \ \forall i = 1, \ldots, n,$

ou de manière équivalente :

$\mathbb{V}(y_i) = \sigma^2, \ \forall i = 1, \ldots, n.$

Commentaires sur l'hypothèse (A2) :

- On parle d'hypothèse d'homoscédasticité ($\simeq$ homogénéité des variances).
- Cette variance σ^2 est un paramètre du modèle qu'il faudra estimer.

(A3) $\mathrm{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0, \ \forall i \neq i'$

ou de manière équivalente :

$\mathrm{Cov}(y_i, y_{i'}) = 0, \ \forall i \neq i'.$

Commentaire sur l'hypothèse (A3) :

- Sous cette hypothèse, les termes d'erreur ε_i sont non corrélés.

On peut écrire matriciellement le modèle (2) de la manière suivante :

$$Y = X\beta + \epsilon \tag{3}$$

où

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{et} \quad \epsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

- Y désigne le vecteur à expliquer de taille n,
- X la matrice explicative de taille $n \times (p+1)$,
- ϵ le vecteur d'erreurs de taille n.

Exercice : Touver X et Y pour les données sur les appartements.

Les hypothèses peuvent alors s'écrire sous forme matricielle :

(A1') $\mathbb{E}(\epsilon) = 0_n$

ou de manière équivalente :

$\mathbb{E}(Y) = X\beta \in \mathbb{R}^n$.

(A2') $\mathbb{V}(\epsilon) = \sigma^2 I_n$

ou de manière équivalente :

$\mathbb{V}(Y) = \sigma^2 I_n$.

Dans la suite de ce chapitre, on suppose que

$$n > (p+1) \text{ et } \text{rang}(X) = p+1$$

On a donc plus d'observations que de variables et il n'existe pas de liaison linéaire entre les variables explicatives x_j c'est à dire pas de multicolinéarité.

Remarque.

Il est important de bien faire la différence entre

- l'expression $\mathbb{E}(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ (qui désigne l'espérance d'une variable aléatoire scalaire), et l'expression $\mathbb{E}(Y) = X\beta$ (qui désigne l'espérance d'une variable aléatoire vectorielle) : on obtient dans un cas un scalaire, dans l'autre cas un vecteur de $\mathbb{R}^n$.

- l'expression $\mathbb{V}(y_i) = \sigma^2$ (qui désigne la variance d'une variable aléatoire scalaire), et l'expression $\mathbb{V}(Y) = \sigma^2 I_n$ (qui désigne la covariance d'une variable aléatoire vectorielle) : on obtient dans un cas un scalaire (σ^2), dans l'autre cas une matrice carrée ($\sigma^2 I_n$) de dimension $n \times n$.

Estimation des paramètres $\beta_1, \beta_2, \ldots, \beta_p$ et σ^2

A partir de l'echantillon (aléatoire) de n observations

$$\{(x_{i1}, \ldots, x_{ip}, y_i),\ i = 1, \ldots, n\},$$

on veut estimer les paramètres

$$\beta_0, \beta_1, \ldots, \beta_p \text{ et } \sigma^2.$$

- Pour estimer $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, on peut utiliser la méthode des moindres carrés qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de ε_i, contrairement à la méthode du maximum de vraisemblance qui est fondée sur la normalité de ε_i.
- La méthode des moindres carrés ne fournit pas un estimateur de σ^2.

Estimation de β par les moindres carrés

On cherche $\widehat{\beta} \in \mathbb{R}^{p+1}$ qui minimise la somme des erreurs quadratiques

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

On doit donc résoudre le problème d'optimisation suivant :

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} [y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]^2. \tag{4}$$

Vocabulaire :

- $\hat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}$ est appelé la valeur prédite.
- $\hat{\varepsilon}_i = y_i - \hat{y}_i$ est appelé le résidu.

En notant $x_i^T = (1, x_{i1}, \ldots, x_{ip})$, la valeur prédite $\hat{y}_i$ s'écrit

$$\hat{y}_i = x_i^T \widehat{\beta}.$$

Résolution du problème d'optimisation

Le problème d'optimisation est :

$$\min_{\beta \in \mathbb{R}^{p+1}} F(\beta),$$

avec

$$\begin{aligned} F(\beta) &= \sum_{i=1}^{n} [y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]^2 \\ &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta \end{aligned}$$

Le minimum est atteint pour

$$\frac{\partial F(\beta)}{\partial \beta} = 0.$$

Rappels. Soient a et x deux vecteurs de dimension K, et soit A une matrice de dimension $K \times K$. On a :

$$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a \quad \text{et} \quad \frac{\partial x^T A x}{\partial x} = 2Ax \text{ si } A \text{ est symétrique,.}$$

Solution du problème d'optimisation

On en déduit après quelques manipulations :

$$\widehat{\beta} = (X^T X)^{-1} X^T Y, \tag{5}$$

sous réserve que $X^T X$ soit inversible.

Commentaires

- Le minimum de F est égal à $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$. Ce minimum est appelé la somme des carrés des résidus (SCR).
- La valeur prédite $\widehat{y_i}$ estime $\mathbb{E}[y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ et non pas y_i. Une meilleure notation serait $\widehat{\mathbb{E}[y_i]}$.
- Aucune des hypothèses n'a été utilisée ici pour obtenir $\widehat{\beta}$.

Propriétés de $\widehat{\beta}$

Sous les hypothèses (A1') et (A2'), on peut montrer que

- $\mathbb{E}[\widehat{\beta}] = \beta$,
- $\mathbb{V}(\widehat{\beta}) = \sigma^2(X^T X)^{-1}$

Commentaires

- L'estimateur $\widehat{\beta}$ est sans biais.
- Il est aussi de variance minimale parmi tous les estimateurs linéaires par rapport à Y) sans biais (propriété dite de Gauss-Markov).

Estimation de σ^2

Le paramètre σ^2 est défini par

$$\sigma^2 = \mathbb{V}(\varepsilon_i) = \mathbb{V}(y_i) = \mathbb{E}\left[(y_i - \mathbb{E}[y_i])^2\right].$$

En prenant $\hat{y}_i = x_i^T \widehat{\beta}$ comme estimateur de $\mathbb{E}[y_i]$, il apparaît naturel d'estimer σ^2 par

$$s^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\sum_{i=1}^{n} (\widehat{\varepsilon}_i)^2}{n-p-1} = \frac{SCR}{n-p-1}.$$

Commentaires

- s^2 est un estimateur sans biais de σ^2
- La perte de $p+1$ degrés de liberté dans l'expression de s^2 est le "coût" de l'estimation de $\beta_0, \beta_1, \ldots, \beta_p$ nécessaire pour obtenir les $\hat{y}_i$.

Sorties R des données poids de naissance

```
summary(modele)


##
## Call:
## lm(formula = bwt ~ age + weight + smoke, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1961   -308     11    309   1487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3050.562    108.861   28.02  < 2e-16 ***
## age           -0.918      2.535   -0.36     0.72
## weight         7.903      1.568    5.04  5.4e-07 ***
## smoke       -254.254     29.939   -8.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 499 on 1170 degrees of freedom
## Multiple R-squared:  0.081,Adjusted R-squared:  0.0786
## F-statistic: 34.4 on 3 and 1170 DF,  p-value: <2e-16
```

Tests d'hypothèses et intervalles pour les paramètres β_j

On veux maintenant tester la nullité des coefficients β_j du modèle de régression.

Pour faire ces tests, il est nécessaire de faire une **hypothèse supplémentaire** :

(A3)' $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$

ou de manière équivalente

$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.

Commentaire. L'unique "nouveauté" ici est la normalité.

Test de signification du modèle

Typiquement, on commence par tester :

$$\mathcal{H}_0 : \text{“}\beta_1 = \ldots = \beta_p = 0\text{”} \text{ contre } \mathcal{H}_1 : \text{“}\exists j \in \{1, \ldots, p\},\ \beta_j \neq 0\text{”}.$$

On utilise la statistique suivante :

$$F_n = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 / p}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-p-1)} = \frac{SCE/p}{SCR/(n-p-1)}$$

qui est distribuée sous $\mathcal{H}_0$ selon une loi de Fisher à p et $n-p-1$ degrés de libertés. On rejette $\mathcal{H}_0$ avec un risque $0 \leq \alpha \leq 1$ si

$$F_n \geq f_{1-\alpha}(p, n-p-1)$$

où $f_{1-\alpha}(p, n-p-1)$ est le fractile d'ordre $1-\alpha$ de la loi $F(p, n-p-1)$.

Table d'analyse de la variance (ANOVA) :

Source de variation	Somme des carrés	ddl	carré moyen	F
régression (expliquée)	$SCE = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_n)^2$	p	$\frac{1}{p}\sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_n)^2$	$\frac{SCE/p}{SCR/(n-p-1)}$
Résiduelle	$SCR = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	n-(p+1)	$\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	
Totale	$SCT = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$	n-1	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$	

Remarques :

- On retrouve la statistique dite de Fisher F_n qui permet de tester l'ajustement du modèle.
- On retrouve la propriété fondamentale $SCT = SCE + SCT$ qui permet de mesurer l'ajustement du modèle par le coefficient de détermination

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

- Le coefficient R^2 donne la proportion de variabilité de y qui est expliquée par le modèle. Plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données.

Test de significativité d'un paramètre β_j

On désire maintenant tester :

$$\mathcal{H}_0 : \text{“}\beta_j = 0\text{”} \quad \text{contre} \quad \mathcal{H}_1 : \text{“}\beta_j \neq 0\text{”}$$

Nouvelles propriétés pour les estimateurs $\widehat{\beta}_j$ et s^2

Sous les hypothèses (A1')-(A3'), on a :

(a) $\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 c_{jj}\right)$
où c_{jj} est le terme $(j+1, j+1)$ de la matrice $(X^T X)^{-1}$

(b) $\dfrac{(n-p-1)s^2}{\sigma^2} \sim \chi^2(n-p-1)$

(c) $\widehat{\beta}_j$ et s^2 sont indépendants

Un rappel de probabilité

Si $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2(\nu)$ et U est indépendant de V, alors $\dfrac{U}{\sqrt{\frac{V}{\nu}}} \sim T(\nu)$.

On déduit alors des propriétés (a)-(c) que

$$\frac{\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}}{\sqrt{\frac{\frac{(n-p-1)s^2}{\sigma^2}}{n-p-1}}} = \frac{\widehat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim T(n-p-1).$$

On utilisera donc la statistique suivante :

$$T_n = \frac{\widehat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}},$$

qui est distribuée selon une loi de Student à $n-p-1$ degrés de libertés.

Test de $\mathcal{H}_0$ contre $\mathcal{H}_1$

Sous l'hypothèse $\mathcal{H}_0$: "$\beta_j = 0$", on a

$$T_n = \frac{\widehat{\beta}_j}{s\sqrt{c_{jj}}} \sim T(n-p-1). \tag{6}$$

Pour une hypothèse alternative $\mathcal{H}_1$: "$\beta_j \neq 0$" bilatérale, on rejette $\mathcal{H}_0$ avec un risque $0 \leq \alpha \leq 1$ si

$$|t| \geq t_{1-\alpha/2}(n-p-1)$$

où t est la réalisation de T_n et $t_{1-\alpha/2}(n-p-1)$ est le fractile d'ordre $1-\alpha/2$ de la loi $T(n-p-1)$.

Remarques.

Pour réaliser ce test, on peut également :

- regarder la *p*-valeur aussi appelée niveau de signification du test : si *p*-valeur $\leq \alpha$, on rejette $\mathcal{H}_0$. Dans le cas d'un test bilatéral ($\mathcal{H}_1$: "$\beta_1 \neq 0$"), on a :

$$p\text{-valeur} = \mathbb{P}(|T_n| > |t| \,/\, \mathcal{H}_0). \tag{7}$$

  On rejette $\mathcal{H}_0$ si *p*-valeur $\leq \alpha$.

- construire l'intervalle de confiance de β_j :

$$[\widehat{\beta}_j \pm t_{1-\alpha/2}(n-p-1)s\sqrt{c_{jj}}].$$

  On rejette $\mathcal{H}_0$ si 0 n'appartient pas à cet intervalle.

Rejeter $\mathcal{H}_0$ signifie :

- que le coefficient β_j est significativement non nul,
- que β_j s'interprète comme le taux d'accroissement moyen de y en fonction d'une variation de x_j lorsque tous les autres régresseurs $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ restent fixés.

Exemple des données poids de naissance.

```
summary(modele)$coefficients

##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 3050.56238   108.8612 28.0225 1.2569e-132
## age           -0.91802     2.5353 -0.3621  7.1734e-01
## weight         7.90266     1.5675  5.0414  5.3511e-07
## smoke       -254.25425    29.9388 -8.4925  6.0621e-17

confint(modele)

##                 2.5 %    97.5 %
## (Intercept) 2836.9775 3264.1473
## age           -5.8922    4.0562
## weight         4.8271   10.9782
## smoke       -312.9940 -195.5145
```

Contribution jointe d'un ensemble de régresseurs

On peut maintenant tester la nullité de $q \leq p$ paramètres :

$$\mathcal{H}_0 : \text{“}\beta_1 = \ldots = \beta_q = 0\text{”} \quad \text{contre} \quad \mathcal{H}_1 : \text{“}\exists j \in \{1, \ldots, q\},\ \beta_j \neq 0\text{”}.$$

Cela revient à comparer deux modèles :

- le modèle complet à p regresseurs (modèle 1) pour lequel on évalue la somme des carrés des résidus SCR_1,
- le modèle réduit à $p - q$ regresseurs (modèle 0) pour lequel on évalue la somme des carrés des résidus SCR_0.

On peut montrer que sous $\mathcal{H}_0$:

$$\frac{(\mathrm{SCR}_0 - \mathrm{SCR}_1)/q}{\mathrm{SCR}_1/(n-p-1)} \sim F(q, n-p-1).$$

La zone de rejet associée à cette statistique de test est donc :

$$\mathcal{R} =]f_{1-\alpha}(q,\, n-p-1),\, +\infty[.$$

Rejeter $\mathcal{H}_0$ signifie qu'au moins un des q coefficients est non nul.

Exemple des données poids de naissance.

```
modele0 <- lm(bwt~ smoke,data=data)
modele1 <- lm(bwt~ age+weight+smoke,data=data)
anova(modele0,modele1)


## Analysis of Variance Table
##
## Model 1: bwt ~ smoke
## Model 2: bwt ~ age + weight + smoke
##   Res.Df       RSS Df Sum of Sq    F  Pr(>F)
## 1   1172 297411671
## 2   1170 291055628  2   6356043 12.8 3.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On désire prévoir à l'aide du modèle la valeur de la variable y pour des observations futures $(x_{1,0}, \ldots, x_{p,0})$ des p variables explicatives.

Posons

$$x_0 = (1, x_{1,0}, \ldots, x_{p,0})^T \in \mathbb{R}^{p+1}$$

D'après le modèle on a :

$$y_0 = x_0{}^T \beta + \varepsilon_0,$$

et la prédiction est :

$$\widehat{y_0} = \widehat{\mathbb{E}[y_0]} = x_0{}^T \widehat{\beta}.$$

L'erreur de prédiction est définie par $\widehat{y_0} - y_0$ et on peut montrer que sous les hypothèses du modèle (incluant l'hypothèse de normalité), on a :

$$\widehat{y_0} - y_0 \sim \mathcal{N}\left(0, \sigma^2 \left(1 + x_0^T (X^T X)^{-1} x_0\right)\right). \tag{8}$$

On en déduit que :

$$\frac{y_0 - \hat{y}_0}{\sigma\sqrt{1 + x_0^T (X^T X)^{-1} x_0}} \sim \mathcal{N}(0, 1).$$

On peut montrer que :

$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + x_0^T (X^T X)^{-1} x_0}} \sim T(n - p - 1).$$

On utilise ce résultat pour construire un intervalle de prédiction pour y_0, c'est à dire l'intervalle $[A, B]$ tel que

$$\mathbb{P}(A \leq y_0 \leq B) = 1 - \alpha.$$

Ici, y_0 est une variable aléatoire et non pas un paramètre. L'intervalle de prédiction est donc un intervalle dans lequel une future observation y_0 va tomber avec une certaine probabilité (différent d'un intervalle de confiance).

On en déduit l'intervalle de prédiction pour y_0 au niveau de confiance $1-\alpha$ suivant :

$$\left[\hat{y}_0 \pm t_{1-\alpha/2}(n-p-1)s\sqrt{1+x_0^T(X^TX)^{-1}x_0}\right]$$

On peut aussi construire un intervalle de confiance de la valeur moyenne

$$\mathbb{E}[y_0] = x_0^T\beta,$$

qui est cette fois un paramètre. On va donc chercher l'intervalle aléatoire $[A, B]$ tel que

$$\mathbb{P}(A \leq \mathbb{E}[y_0] \leq B) = 1-\alpha.$$

Pour construire cet intervalle, on montre que :

$$\hat{y}_0 \sim \mathcal{N}\left(x_0'\beta, \sigma^2 x_0^T (X^T X)^{-1} x_0\right),$$

$$\frac{\hat{y}_0 - x_0^T \beta}{s\sqrt{x_0^T (X^T X)^{-1} x_0}} \sim T(n-p-1).$$

On en déduit l'intervalle de confiance de $\mathbb{E}[y_0] = x_0^T \beta$ suivant :

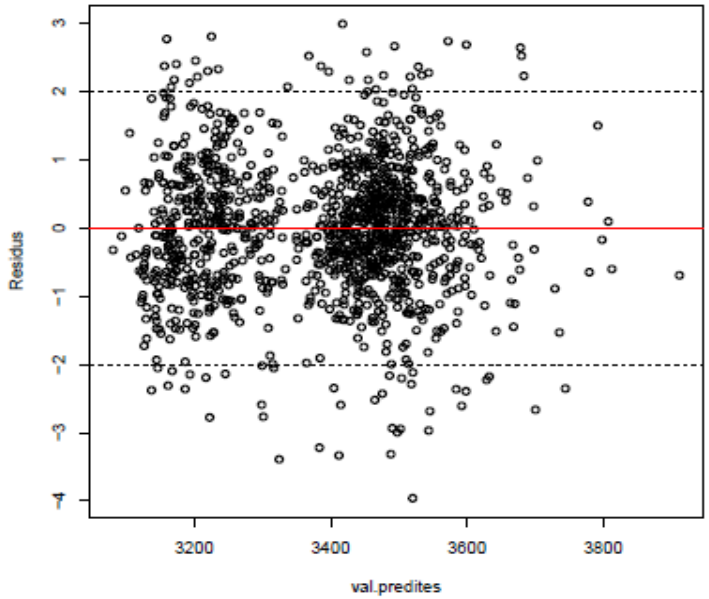
$$\left[\widehat{y_0} \mp t_{1-\alpha/2}(n-p-1)\; s\; \sqrt{x_0^T (X^T X)^{-1} x_0}\right].$$

Exemple des données poids de naissance.

```
#prevision de l'age du bebe d'une femme de 30 ans, 50 kg et fumeuse
predict(modele,data.frame(age=30,weight=50,smoke=1),interval="prediction")


##      fit    lwr  upr
## 1 3163.9 2183.8 4144


predict(modele,data.frame(age=30,weight=50,smoke=1),interval="confidence")


##      fit    lwr    upr
## 1 3163.9 3109.1 3218.7
```

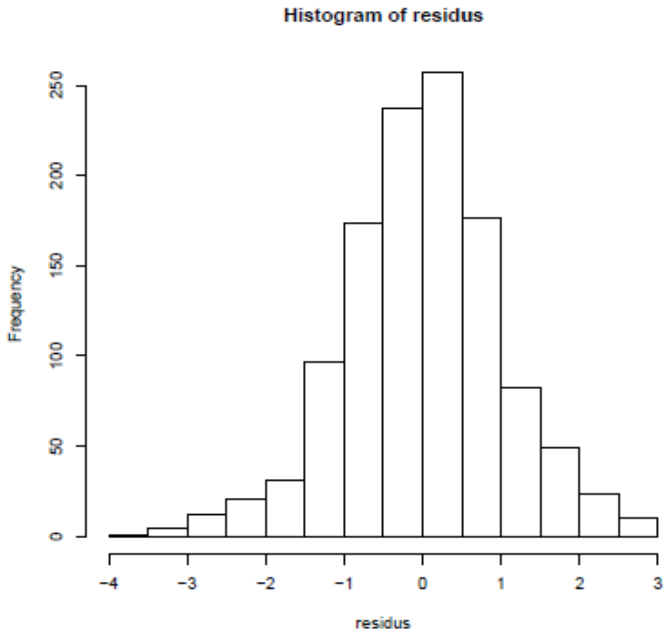
Exemple des données poids de naissance.

```
#On calcule les residus studentises
residus=rstudent(modele)

#On calcule les valeurs predites
val.predites <- predict(modele)

#Graphique predictions-residus
plot(val.predites ,residus, xlab="val.predites", ylab="Residus")
abline(h=c(-2,0,2), lty=c(2,1,2),col=c(1,2,1))
```



Analyse des résidus (Complément 1)

```
#normalite des residus
shapiro.test(residus)


##
##  Shapiro-Wilk normality test
##
## data:  residus
## W = 0.993, p-value = 0.000022

hist(residus)
```



Sélection de variables (Complément 2)

Il s'agit maintenant de sélectionner parmi les p variables explicatives, les $q \leq p$ variables qui donnent le "meilleur" modèle pour prédire y.

Il faut donc :

- un critère de qualité d'un modèle afin de comparer deux modèles n'ayant pas nécessairement le même nombre de variables explicatives.
- une procédure qui permet de choisir parmi tous les modèles, le meilleur au sens de ce critère. On parle de procédure de choix de modèle.

Un problème de complexité :

- Le nombre de modèles à considérer est $\sum_{q=1}^{p} C_p^q = 2^p - 1$. Ce nombre croît exponentiellement avec p. Par exemple, si $p = 30$, on devrait considérer $2^{30} = 10^9$ modèles...
- En pratique, on utilise donc des heuristiques dont les plus simples sont les procédures pas à pas ascendante ou descendante.

Les critères R^2 et R^2 ajusté

- Le coefficient $R^2 = 1 - \frac{SCR}{SCT}$
  - mesure l'ajustement du modèle aux données,
  - augmente lorsque le nombre de variables incluses dans le modèle augmente,
  - permet de comparer des modèles ayant le même nombre de variables.
- Le coefficient $R^2_{\text{ajuste}} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)}$
  - estime le $R^2_{\text{population}} = 1 - \frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(Y)} = 1 - \frac{\sigma^2}{\sigma^2_Y}$,
  - n'augmente pas forcément lorsque le nombre de variables introduites dans le modèle augmente,
  - permet de comparer des modèles ayant un nombre de variables différent.

Les critères AIC et BIC.

Ce sont deux critères de vraisemblance pénalisées définis par :

- $AIC = -2ln(L) + 2k$: Akaike Information Criterion
- $BIC = -2ln(L) + kln(n)$: Bayesian Information Criterion

où L est la vraisemblance maximisée et k est le nombre de paramètres libres du modèle.

En régression multiple :

- il y a $q+2$ paramètres $\beta_0, \beta_1, \ldots, \beta_q, \sigma$ et une equation donc $k = q+1$ paramètres libres.
- la vraisemblance est définie comme la densité conjointe des y_i et son expression est

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right)$$

Les estimateurs du maximum de vraisemblance sont $\tilde{\beta} = (X^TX)^{-1}X^TY$ et $\tilde{\sigma}^2 = \frac{SCR}{n}$. La vraisemblance maximisée est $L = L(\tilde{\beta}, \tilde{\sigma}^2)$ et on obtient :

$$-2ln(L) = n(ln(2\pi\tilde{\sigma}^2) + 1)$$

Ecriture simplifiée en régression multiple.

$$AIC = n\ ln(SCR) + 2k + cste$$
$$BIC = n\ ln(SCR) + k\ ln(n) + cste$$

Ces critères doivent être minimisés dans une procédure de choix de modèle.

Procédure pas à pas ascendante (forward stepwise).

- On part du modèle nul sans variable.
- On effectue p régressions linéaires simples et on sélectionne le modèle qui minimise le critère AIC.
- On effectue $p-1$ régressions linéaires avec 2 variables explicatives et on sélectionne le modèle qui minimise le critère AIC.
- On recommence jusqu'à ce que le critère AIC ne diminue plus.

Procédure pas à pas descendante (backward stepwise).

On part cette fois du modèle complet à p variables explicatives et on supprime pas à pas les variables. Le test d'arrêt et le critère sont les mêmes que pour la procédure ascendante.

Exemple des données poids de naissance.

```
full <- lm(bwt ~ gestation + age + weight + smoke, data=data)
null <- lm(bwt ~ 1, data=data)
back <- step(full, direction="backward")
```

```
## Start:  AIC=14379
## bwt ~ gestation + age + weight + smoke
##
##             Df Sum of Sq       RSS   AIC
## - age        1     60465 242763748 14377
## <none>                   242703282 14379
## - weight     1   5352463 248055745 14402
## - smoke      1  14379595 257082877 14444
## - gestation  1  48352346 291055628 14590
##
## Step:  AIC=14377
## bwt ~ gestation + weight + smoke
##
##             Df Sum of Sq       RSS   AIC
## <none>                   242763748 14377
## - weight     1   5637978 248401726 14402
## - smoke      1  14556053 257319800 14443
## - gestation  1  48324498 291088245 14588
```

```
formula(back)
```

```
## bwt ~ gestation + weight + smoke
```

Sélection de variables (Complément 2)

```
forw <- step(null, scope=list(lower=null,upper=full), direction="forward", trace = 1)


## Start:  AIC=14683
## bwt ~ 1
##
##             Df Sum of Sq       RSS   AIC
## + gestation  1  52601391 264100599 14472
## + smoke      1  19290318 297411671 14611
## + weight     1   7699680 309002310 14656
## <none>                   316701990 14683
## + age        1    230584 316471406 14684
##
## Step:  AIC=14472
## bwt ~ gestation
##
##          Df Sum of Sq       RSS   AIC
## + smoke   1  15698873 248401726 14402
## + weight  1   6780798 257319800 14443
## + age     1    754996 263345603 14471
## <none>                264100599 14472
##
## Step:  AIC=14402
## bwt ~ gestation + smoke
##
##          Df Sum of Sq       RSS   AIC
## + weight  1   5637978 242763748 14377
## <none>                248401726 14402
## + age     1    345981 248055745 14402
##
## Step:  AIC=14377
## bwt ~ gestation + smoke + weight
##
##        Df Sum of Sq       RSS   AIC
## <none>              242763748 14377
## + age   1     60465 242703282 14379
```

Sélection de variables (Complément 2)

```
formula(forw)


## bwt ~ gestation + smoke + weight


summary(lm(forw,data=data)) # idem forw


##
## Call:
## lm(formula = forw, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1471.9  -305.0    -7.9   276.2  1455.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -499.702    246.393   -2.03    0.043 *
## gestation     12.703      0.832   15.26  < 2e-16 ***
## smoke       -229.004     27.341   -8.38  < 2e-16 ***
## weight         7.386      1.417    5.21  2.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 456 on 1170 degrees of freedom
## Multiple R-squared:  0.233,Adjusted R-squared:  0.231
## F-statistic:  119 on 3 and 1170 DF,  p-value: <2e-16
```