

LA RÉGRESSION LINÉAIRE

Exemple : On s'intéresse à la consommation en oxygène lors d'une course. Des données sont récoltées pour plusieurs athlètes indiquant leur temps de course, leur age, leurs poids, etc. et la consommation en oxygène lors de la course. Dans le cadre de la régression linéaire on s'intéresse à **modéliser** la consommation en oxygène pour pouvoir la **prévoir** pour un nouvel individu. La Figure ?? montre les 31 consommations en oxygène (*Oxygen*) récoltées en fonction du temps de course (*RunTime*). La consommation en oxygène est clairement dépendante du temps de course. L'objectif du chapitre est de proposer un modèle le plus pertinent possible entre les variables explicatives et la variable *Oxygen*, d'estimer les paramètres du modèle, de prédire le modèle en tenant compte des incertitudes d'estimation et de prédiction.

Age	Weight	Oxygen	RunTime	RestPulse	RunPulse	MaxPulse
44	89.47	44.609	11.37	62	178	182
40	75.07	45.313	10.07	62	185	185
44	85.84	54.297	8.65	45	156	168
42	68.15	59.571	8.17	40	166	172
38	89.02	49.874	9.22	55	178	180
47	77.45	44.811	11.63	58	176	176
40	75.98	45.681	11.95	70	176	180
43	81.19	49.091	10.85	64	162	170
44	81.42	39.442	13.08	63	174	176
38	81.87	60.055	8.63	48	170	186
44	73.03	50.541	10.13	45	168	168
45	87.66	37.388	14.03	56	186	192
45	66.45	44.754	11.12	51	176	176
47	79.15	47.273	10.6	47	162	164
54	83.12	51.855	10.33	50	166	170
49	81.42	49.156	8.95	44	180	185
51	69.63	40.836	10.95	57	168	172
51	77.91	46.672	10	48	162	168
48	91.63	46.774	10.25	48	162	164
49	73.37	50.388	10.08	67	168	168
57	73.37	39.407	12.63	58	174	176
54	79.38	46.08	11.17	62	156	165
52	76.32	45.441	9.63	48	164	166
50	70.87	54.625	8.92	48	146	155
51	67.25	45.118	11.08	48	172	172
54	91.63	39.203	12.88	44	168	172
51	73.71	45.79	10.47	59	186	188
57	59.08	50.545	9.93	49	148	155
49	76.32	48.673	9.4	56	186	188
48	61.24	47.92	11.5	52	170	176
52	82.78	47.467	10.5	53	170	172

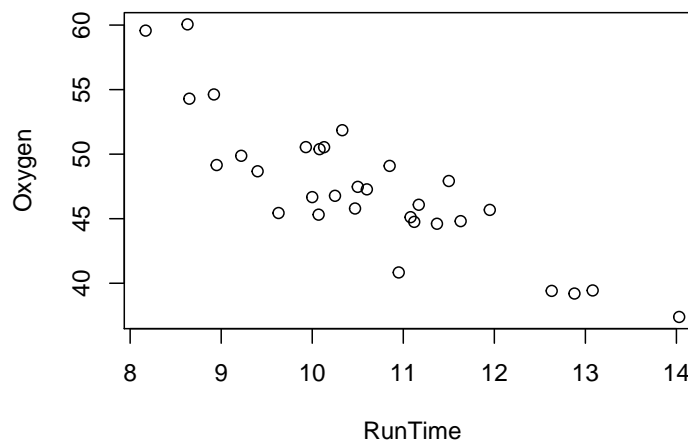


FIGURE 1 – Observations : consommation en O_2 en fonction du temps de course.

Contexte et notations

- X_1, \dots, X_p sont des variables explicatives (descripteurs)
- Y est la variable (quantitative) à expliquer (variable d'intérêt).

Objectif : on cherche à approcher la relation entre Y et (X_1, \dots, X_p) à partir d'un échantillon à n observations.

1 Introduction

Soit $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ le vecteur des observations. On note x_{ij} l'observation de la j ème variable explicative X_j pour la i ème observation.

1.1 Modèle probabiliste usuel

On suppose que y est la réalisation du modèle suivant

$$\mathbf{y}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i, \forall 1 \leq i \leq n$$

où β_0, \dots, β_p sont des nombres réels déterministes inconnus, $\epsilon_1, \dots, \epsilon_n$ sont n variables aléatoires non corrélées, de moyenne nulle $\mathbb{E}[\epsilon_i] = 0$ et de variance constante $\mathbb{V}(\epsilon_i) = \sigma^2$.

Remarque 1. Les remarques suivantes peuvent être faites :

- Ce modèle est l'extension naturelle du modèle de régression linéaire simple à la prise en compte de plusieurs variables explicatives.
- Il est linéaire par rapport aux paramètres $\beta_0, \beta_1, \dots, \beta_p$ inconnus.
- Parmi les variables X_j on peut trouver des transformations comme $\log X_1$, $\sqrt{X_1}$, X_1^2 et des termes d'interactions comme $X_1 * X_2$.
- $\beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ correspond à la partie déterministe du modèle, c'est l'espérance de la variable aléatoire \mathbf{y}_i , elle dépend de x_{i1}, \dots, x_{ip} . ϵ_i est un complément aléatoire et correspond à la part de \mathbf{y}_i qui ne dépend pas de x_{i1}, \dots, x_{ip} .
- Le modèle s'écrit aussi matriciellement (en gras les variables aléatoires, en rouge les paramètres inconnus) :

$$\begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} = X \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

où $X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$ $\mathbb{E} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = 0_n$ et $\mathbb{V} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \sigma^2 Id_n$

Le problème est donc d'estimer $\beta_0, \beta_1, \dots, \beta_p, \sigma^2$.

1.2 Approximation par moindres carrés

On cherche b_0, b_1, \dots, b_p tels que $C(b) = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip}))^2$ soit la plus petite possible.

Matriciellement, on obtient :

$$C(b) = {}^t(y - Xb)(y - Xb)$$

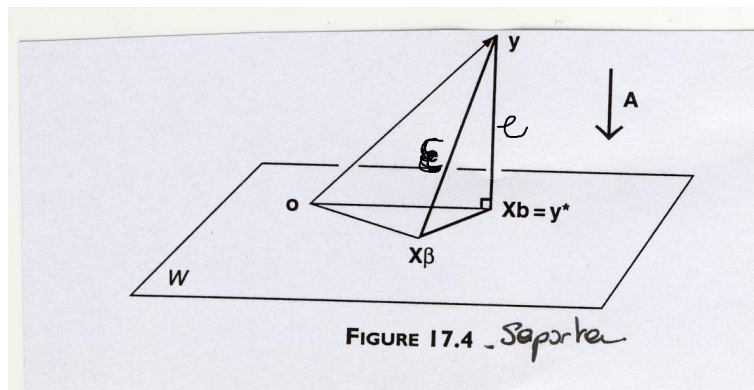
où $b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$. $C(b)$ est minimale au point $\hat{b} = ({}^tXX)^{-1}{}^tXy$.

Remarques :

- Chaque paramètre est une combinaison linéaire des observations.
- Le vecteur des prévisions, noté \hat{y} , s'exprime par :

$$\hat{y} = \hat{b}_0 1 + \hat{b}_1 x_1 + \dots + \hat{b}_p x_p = X\hat{b} = X({}^tXX)^{-1}{}^tXy.$$

Pour la suite on note $H = X({}^tXX)^{-1}{}^tX$. Il s'agit de la matrice de projection orthogonale sur $\text{vect}(1, x_1, \dots, x_p)$.



1.3 Propriétés de l'estimateur des moindres carrés

On note \mathbf{B} l'estimateur des moindres carrés. On a : $\mathbf{B} = ({}^tXX)^{-1}{}^tXy$. De plus, on note \hat{y} le vecteur aléatoire des prévisions, on a $\hat{y} = X\mathbf{B} = Hy$.

On a les propriétés suivantes :

Théorème 2. 1. \mathbf{B} est un estimateur sans biais de β .

2. \mathbf{B} est de tous les estimateurs sans biais de β de la forme Ay , celui de variance minimale. On a $\mathbb{V}(\mathbf{B}) = \sigma^2({}^tXX)^{-1}$

3. $\hat{\sigma}^2 = \frac{1}{n-p-1} \|y - \hat{y}\|^2$ est un estimateur sans biais de σ^2 .

2 Estimation et tests dans le cadre gaussien

2.1 Le modèle linéaire gaussien

Dans la suite pour réaliser des tests statistiques sur le modèle, on a besoin de l'hypothèse sur les lois des v.a..

On suppose alors que :

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = X\beta + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\text{où } \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \rightsquigarrow \mathcal{N}(0, \sigma^2 I_n)$$

Compléments de Statistique

Définition 3. Un vecteur aléatoire Z est un vecteur gaussien ssi toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne. Il est caractérisé par son espérance et sa matrice de variance-covariance.

Théorème 4. *Théorème de Cochran : Soit Z un vecteur gaussien centré réduit de \mathbb{R}^n et soit F un sous-espace de \mathbb{R}^n de dimension $p+1$. Soit P_F le projecteur orthogonal de \mathbb{R}^n sur F . Alors*

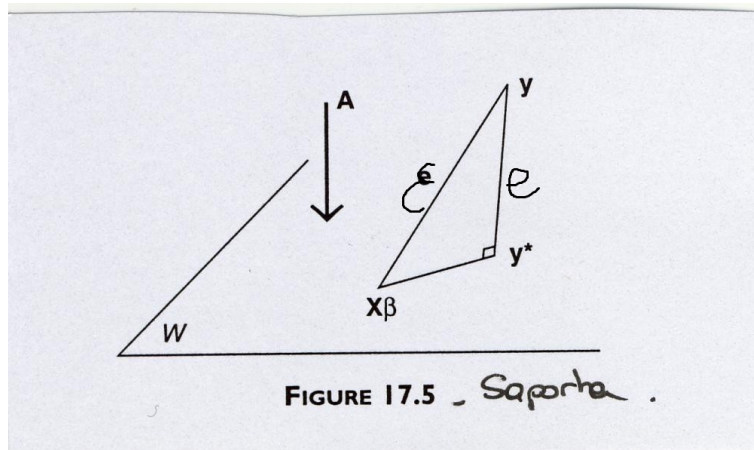
- $P_F Z \rightsquigarrow \mathcal{N}_n(0, P_F)$, $P_{F^\perp} Z \rightsquigarrow \mathcal{N}_n(0, P_{F^\perp})$ et ces projections sont des vecteurs aléatoires indépendants
- $\|P_F Z\|^2 \rightsquigarrow \chi_{p+1}^2$, $\|P_{F^\perp} Z\|^2 \rightsquigarrow \chi_{n-(p+1)}^2$ et ces variables aléatoires sont indépendantes

2.2 Estimation

Hypothèse : $\mathbf{y}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i$ où $\epsilon_1, \dots, \epsilon_n$ i.i.d. $\mathcal{N}_n(0, \sigma^2)$

Proposition 5. On a les résultats suivants :

1. \mathbf{B} est un vecteur gaussien multi-dimensionnel, $\mathbf{B} \rightsquigarrow \mathcal{N}_{p+1}(\beta, \sigma^2({}^t X X)^{-1})$
2. $\frac{1}{\sigma^2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \rightsquigarrow \chi_{n-(p+1)}^2$



2.3 Test de non régression

Proposition 6. Dans le cadre de la régression linéaire, on a la décomposition suivante :

$$\|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2$$

où $\bar{\mathbf{y}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i\right) \mathbf{1}_{\mathbb{R}^n}$

Définition 7. On appelle coefficient de détermination le réel R^2 défini par

$$R^2 = \frac{\text{Regression sum of square}}{\text{Total sum of squares}} = \frac{\sum (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^2}{\sum (\mathbf{y}_i - \bar{\mathbf{y}})^2}.$$

Remarque : R se nomme aussi le coefficient de corrélation multiple

En général, on regroupe les calculs dans une table d'analyse de la variance :

source de variation	Somme des carrés
Expliquée par la régression	$SSR = \sum (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^2$
Erreur Résiduelle	$SSE = \sum (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2$
Totale	$SST = \sum (\mathbf{y}_i - \bar{\mathbf{y}})^2$

Dans un deuxième temps, on procède au test de "significativité" du modèle dans son ensemble. Ce test teste l'hypothèse de non régression i.e $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$.

Si H_0 est vraie,

— $\frac{SSR}{\sigma^2} \rightsquigarrow \chi_p^2$

— $\frac{SSE}{\sigma^2} \rightsquigarrow \chi_{n-(p+1)}^2$

Soit F la v.a. définie par

$$F = \frac{SSR/p}{SSE/(n-(p+1))} = \frac{(n-p-1)R^2}{p(1-R^2)} = \frac{MSR}{MSE}$$

suit sous H_0 une loi de Fisher de degré de liberté $(p, n-p-1)$.

C'est pourquoi, la table d'analyse de la variance présentée ci-dessus est complétée par deux colonnes supplémentaires, pour préparer le calcul de la statistique de Fisher.

source de variation	Somme des carrés	DF	Carré moyen
Expliquée par la régression	$SSR = \sum (\hat{y}_i - \bar{y})^2$	p	$MSR = \frac{SSR}{DF}$
Résiduelle	$SSE = \sum (y_i - \hat{y}_i)^2$	n-p-1	$MSE = \frac{SSE}{DF}$
Totale	$SST = \sum (y_i - \bar{y})^2$	n-1	

Dans la pratique, on calcule $F_{obs} = \frac{MSR}{MSE}$ et on teste :

- si $F_{obs} < q_{1-\alpha}$ (où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher de degré de liberté $(p, n-p-1)$), on conserve H_0
- sinon on rejette H_0 .

Les logiciels renvoient $p = P_{H_0}(F > F_{obs})$

- si $p > 0,05$ on conserve H_0
- sinon on rejette H_0 .

Commandes sous R : `summary(lm)` ou `anova`.

2.4 Test d'influence d'un coefficient

Sous l'hypothèse de normalité des résidus, on rappelle que B est un vecteur gaussien multi-dimensionnel, $B \rightsquigarrow \mathcal{N}_{p+1}(\beta, \sigma^2(tXX)^{-1})$

Théorème 8. La variable $\frac{\mathbf{b}_i - \beta_i}{\sqrt{\hat{\sigma}^2[(tXX)^{-1}]_{ii}}}$ suit une loi de Student à $(n-p-1)$ degré de liberté.

On peut ensuite construire le test suivant $H_0 : \beta_i = 0$

Sous H_0 , la v.a. $\mathbf{T} = \frac{\mathbf{b}_i}{\sqrt{\hat{\sigma}^2[(tXX)^{-1}]_{ii}}}$ suit une loi de Student à $(n-p-1)$ degré de liberté, on calcule donc $t = T(\omega)$ et on la compare à la valeur de la table ou bien le logiciel calcule $P(|T| \geq |t|)$.

2.5 Cas - consommation en oxygène

Importation des données et Régression

```
>Fitness <- read.table('fitness.txt', header = TRUE)
>head(Fitness)
```

	Age	Weight	Oxygen	RunTime	RestPulse	RunPulse	MaxPulse
1	44	89.47	44.609	11.37	62	178	182
2	40	75.07	45.313	10.07	62	185	185
3	44	85.84	54.297	8.65	45	156	168
4	42	68.15	59.571	8.17	40	166	172
5	38	89.02	49.874	9.22	55	178	180
6	47	77.45	44.811	11.63	58	176	176

```
> summary(Fitness)
```

Age		Weight		Oxygen		RunTime	
Min.	:38.00	Min.	:59.08	Min.	:37.39	Min.	: 8.17
1st Qu.	:44.00	1st Qu.	:73.20	1st Qu.	:44.96	1st Qu.	: 9.78
Median	:48.00	Median	:77.45	Median	:46.77	Median	:10.47
Mean	:47.68	Mean	:77.44	Mean	:47.38	Mean	:10.59
3rd Qu.	:51.00	3rd Qu.	:82.33	3rd Qu.	:50.13	3rd Qu.	:11.27
Max.	:57.00	Max.	:91.63	Max.	:60.05	Max.	:14.03

RestPulse		RunPulse		MaxPulse	
Min.	:40.00	Min.	:146.0	Min.	:155.0
1st Qu.	:48.00	1st Qu.	:163.0	1st Qu.	:168.0
Median	:52.00	Median	:170.0	Median	:172.0
Mean	:53.45	Mean	:169.6	Mean	:173.8
3rd Qu.	:58.50	3rd Qu.	:176.0	3rd Qu.	:180.0
Max.	:70.00	Max.	:186.0	Max.	:192.0

Régression - commande summary

```
>lm1 <- lm(Oxygen~.,data = Fitness)
>summary(lm1)
```

```
Call:
lm(formula = Oxygen ~ ., data = Fitness)

Residuals:
    Min       1Q   Median       3Q      Max
-5.4026 -0.8991  0.0706  1.0496  5.3847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.93448   12.40326   8.299 1.64e-08 ***
Age          -0.22697    0.09984  -2.273  0.03224 *
Weight       -0.07418    0.05459  -1.359  0.18687
RunTime      -2.62865    0.38456  -6.835 4.54e-07 ***
RestPulse    -0.02153    0.06605  -0.326  0.74725
RunPulse     -0.36963    0.11985  -3.084  0.00508 **
MaxPulse      0.30322    0.13650   2.221  0.03601 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.317 on 24 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8108
F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09
```

3 Validation du modèle - Etude des résidus

3.1 Contrôle des hypothèses du modèle linéaire

Les propriétés des estimateurs de la régression linéaire viennent de :

- $\forall i \in \{1, \dots, n\}, E(\epsilon_i) = 0$ et $V(\epsilon_i) = \sigma^2 = cte$
- $\forall i \neq j, cov(\epsilon_i, \epsilon_j) = 0$

On vérifie ces deux hypothèses de manière empirique en traçant le graphe des résidus. En fonction du résultat, on peut proposer différentes modifications du modèle (quadratique, racine carrée, logarithme ou changement de variable pour rendre la variance constante).

L'autre hypothèse est la distribution de loi normale, on le vérifie avec le graphique des QQ-plot (cf Figure ??)

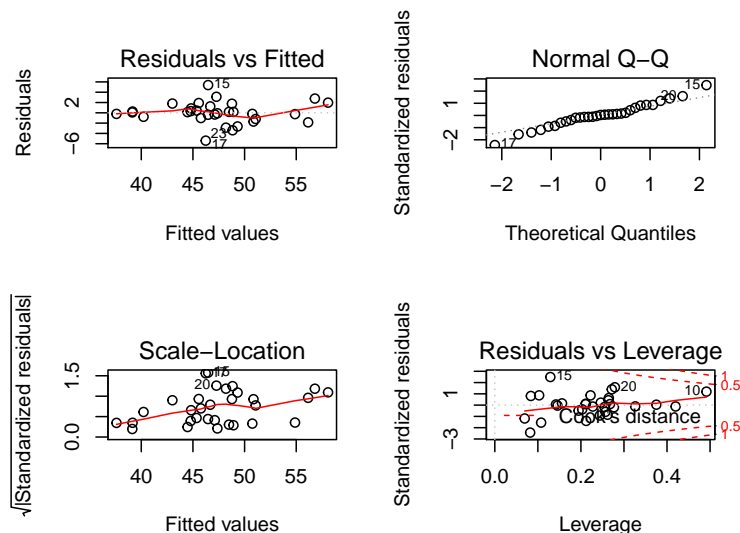


FIGURE 2 – Résidus du modèle $lm1$

3.2 Mesures d'influence

Attention : certaines observations peuvent être singulières, il s'agit :

- des observations influentes \rightarrow on calcule l'effet de levier,
- des observations aberrantes ou atypiques \rightarrow résidus extrêmes.

Observations influentes

Définition 9. Soit $H = X(X^t X)^{-1} X^t$, la matrice de projection orthogonale de \mathbb{R}^n dans $Vect(X)$ et h_i le i -ème terme de la diagonale de H . h_i s'appelle l'effet levier (leverage). Une grande valeur de h_i est due aux valeurs extrêmes de X , donc a une observation influente.

Observations aberrantes

Définition 10. On définit respectivement les **résidus** et les **résidus standardisés** :

$$\begin{aligned} \mathbf{e} &= (Id - H)\mathbf{y} \\ \mathbf{r} &= \text{diag}[\hat{\sigma}^2(1 - h_i)]^{-1/2} \mathbf{e} \end{aligned}$$

Si le modèle est correct, 95% des résidus standardisés, i.e. centrés réduits, ne doivent pas dépasser la valeur 3 en valeur absolue.

Les deux diagnostics précédents (effet levier et résidu élevé) sont combinés dans des mesures synthétiques. La plus utilisée est la distance de Cook's définie par

$$D_i = \frac{1}{\hat{\sigma}^2(p+1)} \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\|^2 = \left(\frac{h_i}{1 - h_i} \right) \frac{r_i^2}{p+1}.$$

Une distance $D_i > 1$ indique en général une **valeur influente anormale**.

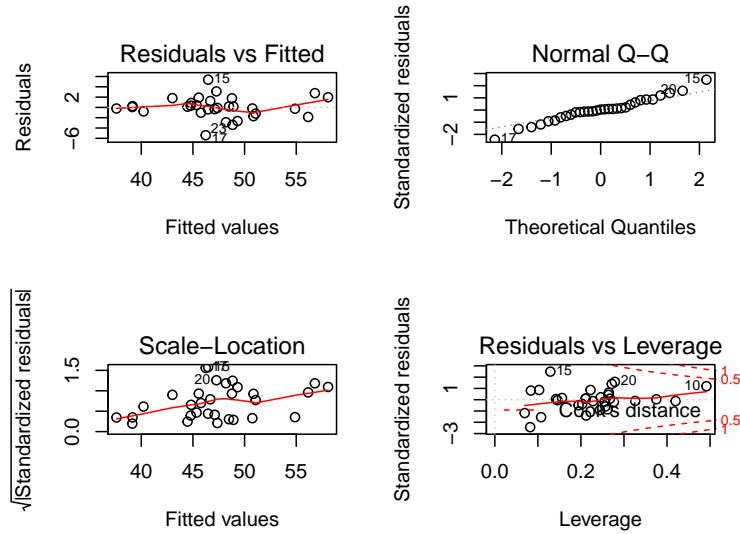


FIGURE 3 – Résidus du modèle $lm1$

3.3 Graphes

1. Graphe des résidus : Les points doivent être uniformément répartis, sans forme.
2. Droite des QQ-plot : L'hypothèse de normalité des résidus est rejetée si les points ne sont pas alignés.
3. Graphique de la racine des résidus en fonction des valeurs prédites (idem que le 1).
4. Graphique des résidus standardisés en fonction de l'effet levier avec les valeurs critiques de D_i .

3.4 Cas - consommation en oxygène

```
>plot(lm1)
```

4 Prédiction pour un nouvel individu

Soit $x_0 = (1, x_{01}, \dots, x_{0p})$ une nouvelle observation, on a les résultats suivants :

Théorème 11. — $\hat{y}_0 = {}^t x_0 \mathbf{B}$ est une v.a. $\mathcal{N}_n({}^t x_0 \beta; \sigma^2({}^t x_0 ({}^t X X)^{-1} x_0))$

— Intervalle de Confiance pour ${}^t x_0 \beta$:

$${}^t x_0 b \pm t_{\alpha/2; n-p-1} \cdot \sqrt{\hat{\sigma}^2 ({}^t x_0 ({}^t X X)^{-1} x_0)}$$

— Intervalle de Prédiction pour \mathbf{y}_0 :

$$\hat{y}_0 \pm t_{\alpha/2; n-p-1} \cdot \sqrt{\hat{\sigma}^2 (1 + {}^t x_0 ({}^t X X)^{-1} x_0)}$$

Cas - consommation en oxygène

```
>Fitness2 <-matrix(c(40,65,10, 60,170,185),1,6)
>colnames(Fitness2) <- c('Age','Weight','RunTime','RestPulse','RunPulse','MaxPulse')
>Fitness2
```

```
      Age Weight RunTime RestPulse RunPulse  MaxPulse
[1,]  40     65      10       60      170      185
```



```

>Fitness2 <- data.frame(Fitness2)

>predict(lm1,newdata =Fitness2)

      1
54.7139

>predict(lm1,newdata =Fitness2,interval = c( "confidence"),level = 0.95)

      fit      lwr      upr
1 54.7139 50.95357 58.47423

>predict(lm1,newdata =Fitness2,interval = c( "prediction"),level = 0.95)

      fit      lwr      upr
1 54.7139 48.63055 60.79724

```

5 Sélection de variables

5.1 Stabilité des coefficients

Rappel : $\mathbf{B} \rightsquigarrow N(\beta, \sigma^2({}^tXX)^{-1})$ et aussi $\mathbf{b}_i \rightsquigarrow N(\beta_i, \sigma^2({}^tXX)_{ii}^{-1})$ Si les variables x_1, \dots, x_p sont redondantes (colinéarité entre les variables explicatives) alors :

- $\mathbf{b}_1, \dots, \mathbf{b}_p$ sont fortement corrélés
- tXX est mal conditionnée donc $({}^tXX)^{-1}$ a des termes élevés (forte variance pour certains coefficients).

Attention : la qualité de l'estimation dépend du plan d'expériences!!!

cas - consommation en oxygène.

```

> vp = eigen(t(as.matrix(Fitness[, -3]))%*%as.matrix(Fitness[, -3]))$values
> max(vp)/min(vp)

[1] 59800.16

> corbeta

      Age      Weight    RunTime    RestPulse    RunPulse    MaxPulse
Age      1.000000000 -0.006437788 -0.3732403  0.14002167  0.05497320 -0.15999696
Weight  -0.006437788  1.000000000 -0.1764922  0.06093716  0.20807273 -0.37338795
RunTime  -0.373240335 -0.176492158  1.0000000 -0.41270841 -0.23391674  0.19468469
RestPulse 0.140021667  0.060937164 -0.4127084  1.00000000 -0.06466442 -0.05936214
RunPulse  0.054973197  0.208072734 -0.2339167 -0.06466442  1.00000000 -0.96035444
MaxPulse -0.159996964 -0.373387952  0.1946847 -0.05936214 -0.96035444  1.00000000

```

Il est donc important de visualiser ou mesurer l'effet de la colinéarité entre les prédicteurs.

5.2 L'objectif de la sélection de variable

L'objectif de la régression linéaire est d'obtenir un modèle **prédictif**. Or si le conditionnement est mauvais (redondance d'information parmi les prédicteurs), la variance de prédiction est grande, le modèle est donc peu prédictif. C'est le contexte du surapprentissage (trop de paramètres à estimer). On cherche alors à simplifier le modèle pour gagner en prédiction (compromis biais variance).

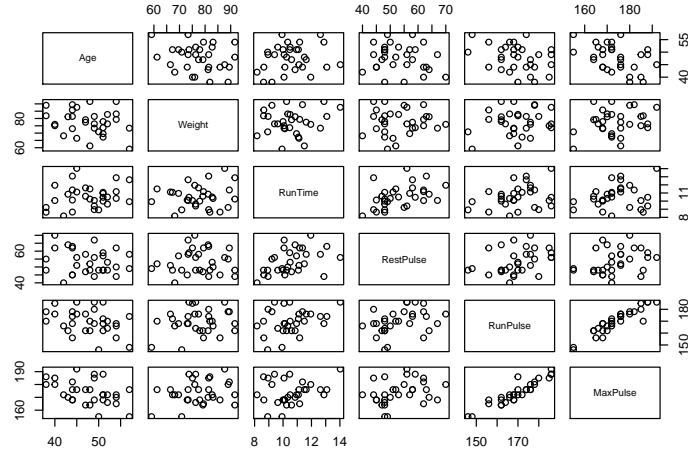


FIGURE 4 – Tracer du nuage de points.

5.3 Critères

Les critères suivants peuvent être utilisés pour sélectionner un modèle plus parcimonieux :

- R^2 : uniquement pour comparer des modèles comprenant le même nombre de variables
- Rsquare ajusté : $R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$
- Statistique de Student : attention à la corrélation entre les prédicteurs
- Critère d'Akaike : $AIC = -2 * \log L + 2(p + 1)$ dans R
- Critère Bayésien : $BIC = -2 * \log L + \ln(n)(p + 1)$ dans R

5.4 Algorithme de Sélection

Pas à pas sur le test de Student :

- forward (Sélection) A chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur de $P(> |t|)$ (test de student) est minimum (variable la plus influente). La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque $P(> |t|)$ reste plus grande qu'une valeur seuil fixée par défaut à 0.05.
- backward (Élimination) L'algorithme démarre cette fois du modèle complet. A chaque étape, la variable associée la plus grande valeur de $P(> |t|)$ est éliminée du modèle. La procédure s'arrête lorsque les variables restant dans le modèle ont des valeurs de $P(> |t|)$ plus petite qu'une valeur seuil fixée par défaut à 0.05.

Pas à pas sur le critère AIC (resp. BIC) :

- forward (Sélection) A chaque pas, une variable est ajoutée au modèle. C'est celle qui fait le plus décroître le critère AIC (resp. BIC). La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque quelque soit la variable à ajouter le critère AIC (resp. BIC) remonte.
- backward (Élimination) L'algorithme démarre cette fois du modèle complet. A chaque étape, la variable associée la plus grande diminution de l'AIC est éliminée du modèle. La procédure s'arrête lorsque quelque soit la variable à retirer le critère AIC (resp. BIC) remonte.

5.5 cas- consommation en oxygène

Commande Summary

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.93448    12.40326   8.299 1.64e-08 ***
```

```

Age          -0.22697    0.09984   -2.273   0.03224 *
Weight       -0.07418    0.05459   -1.359   0.18687
RunTime      -2.62865    0.38456   -6.835  4.54e-07 ***
RestPulse    -0.02153    0.06605   -0.326   0.74725
RunPulse     -0.36963    0.11985   -3.084   0.00508 **
MaxPulse      0.30322    0.13650    2.221   0.03601 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.317 on 24 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8108
F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09

```

Sans *RestPulse*

```

Call:
lm(formula = Oxygen ~ . - RestPulse, data = Fitness)

Residuals:
    Min       1Q   Median       3Q      Max
-5.4724 -0.8476  0.0094  0.9976  5.3807

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.20428   11.97929   8.532 7.13e-09 ***
Age          -0.21962    0.09550   -2.300 0.03010 *
Weight       -0.07230    0.05331   -1.356 0.18714
RunTime      -2.68252    0.34099   -7.867 3.19e-08 ***
RunPulse     -0.37340    0.11714   -3.188 0.00383 **
MaxPulse      0.30491    0.13394    2.277 0.03164 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.275 on 25 degrees of freedom
Multiple R-squared:  0.848,    Adjusted R-squared:  0.8176
F-statistic: 27.9 on 5 and 25 DF,  p-value: 1.811e-09

```

Sans *Weight*

```

Call:
lm(formula = Oxygen ~ . - Weight - RestPulse, data = Fitness)

Residuals:
    Min       1Q   Median       3Q      Max
-4.9685 -1.1654 -0.0636  1.2004  4.7726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.14789   11.78569   8.328 8.26e-09 ***
Age          -0.19773    0.09564   -2.068 0.04877 *
RunTime      -2.76758    0.34054   -8.127 1.31e-08 ***
RunPulse     -0.34811    0.11750   -2.963 0.00644 **
MaxPulse      0.27051    0.13362    2.024 0.05330 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.312 on 26 degrees of freedom
Multiple R-squared:  0.8368,    Adjusted R-squared:  0.8117
F-statistic: 33.33 on 4 and 26 DF,  p-value: 6.91e-10

```

Sans *MaxPulse*

```

Call:
lm(formula = Oxygen ~ . - Weight - RestPuls\e - MaxPulse, data = Fitness)

Residuals:
    Min       1Q   Median       3Q      Max
-4.8752 -1.2493  0.2606  1.0324  4.8994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.71806   10.23509  10.915 2.10e-11 ***
Age          -0.25640    0.09623   -2.664 0.0129 *
RunTime      -2.82538    0.35828   -7.886 1.77e-08 ***
RunPulse     -0.13091    0.05059   -2.588 0.0154 *
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.441 on 27 degrees of freedom
Multiple R-squared: 0.8111, Adjusted R-squared: 0.7901
F-statistic: 38.64 on 3 and 27 DF, p-value: 6.557e-10

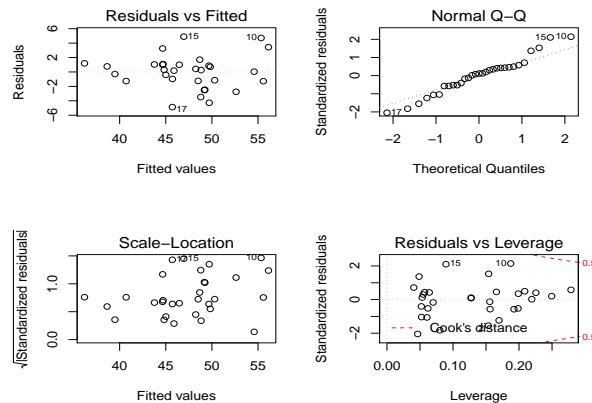


FIGURE 5 – Modèle réduit 2

Sélection avec le Critère AIC ou BIC

```
>slm_AIC <- step(lm1,direction="backward",k = 2)
>slm_BIC <- step(lm1,direction="backward",k = log(nrow(Fitness)))
>coef(slm_AIC)
```

(Intercept)	Age	Weight	RunTime	RunPulse	MaxPulse
102.20427520	-0.21962138	-0.07230234	-2.68252297	-0.37340085	0.30490783

```
>coef(slm_BIC)
```

(Intercept)	Age	RunTime	RunPulse	MaxPulse
98.1478880	-0.1977347	-2.7675788	-0.3481079	0.2705130

6 Extension : estimation par maximum de vraisemblance

6.1 Introduction

Une des méthodes les plus utilisées pour estimer les paramètres d'un modèle probabiliste est la méthode du maximum de vraisemblance. L'idée est de choisir les paramètres qui rendent les observations les plus vraisemblables.

Définition 12. La vraisemblance d'un échantillon (X_1, \dots, X_n) de la v.a. X est la famille indicée par $\theta \in \Theta$ des lois du vecteur (X_1, \dots, X_n) sous l'hypothèse que X suit la loi $\mathcal{L}(\theta)$.

6.2 Cas discret

Définition 13. On suppose que, pour tout $\theta \in \Theta$, X suit une loi discrète prenant ses valeurs dans un ensemble dénombrable D_θ . La vraisemblance de l'échantillon (X_1, \dots, X_n) est la famille indicée par $\theta \in \Theta$, des applications, $\mathcal{L}(\cdot; \theta) : D_\theta^n \rightarrow [0, 1]$ définies pour tout $(x_1, \dots, x_n) \in D_\theta^n$ sous l'hypothèse que X suit la loi $\mathcal{L}(\theta)$, par

$$\mathcal{L}(x_1, \dots, x_n; \theta) = P[X = x_1] \dots P[X = x_n].$$

6.3 Cas continu

Définition 14. On suppose que, pour tout $\theta \in \Theta$, X suit une loi de densité f_θ . La vraisemblance de l'échantillon (X_1, \dots, X_n) est la famille indicée par $\theta \in \Theta$, des applications, $\mathcal{L}(\cdot; \theta) : \mathbb{R}^n \rightarrow \mathbb{R}^+$ définies pour tout $(x_1, \dots, x_n) \in \mathbb{R}^n$ sous l'hypothèse que X suit la loi $\mathcal{L}(\theta)$, par

$$\mathcal{L}(x_1, \dots, x_n; \theta) = f_\theta(x_1) \dots f_\theta(x_n).$$

La méthode du maximum de vraisemblance consiste à déterminer, pour tout (x_1, \dots, x_n) , un paramètre réalisant le suprémum de $\mathcal{L}(x_1, \dots, x_n; \theta)$ sur tous les paramètres possibles $\theta \in \Theta$. Puisqu'un tel paramètre dépend a priori de (x_1, \dots, x_n) , on peut le noter $g(x_1, \dots, x_n)$. Un estimateur du maximum de vraisemblance est alors un estimateur T s'écrivant $T = g(X_1, \dots, X_n)$, où l'on admet que g est mesurable.

6.4 Exemple d'un échantillon gaussien

Soit un échantillon (X_1, \dots, X_n) de la v.a. X de loi $\mathcal{N}(m, \sigma^2)$, trouver l'estimateur du maximum de vraisemblance. Dans le cas gaussien, la vraisemblance a l'expression suivante :

$$\mathcal{L}(x_1, \dots, x_n; m, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right)$$

Il est souvent plus simple d'optimiser le logarithme de la vraisemblance $-\ln(\mathcal{L}(x_1, \dots, x_n; m, \sigma^2))$, où :

$$\begin{aligned} l(x_1, \dots, x_n; m, \sigma^2) &= -\ln(\mathcal{L}(x_1, \dots, x_n; m, \sigma^2)) \\ &= \frac{n}{2} \ln(2\pi\sigma^2) + \sum_{i=1}^n \frac{(x_i - m)^2}{2\sigma^2} \end{aligned}$$

La fonction étant convexe, l'optimum existe et est obtenu au point où le gradient est nul. On obtient :

$$\begin{aligned} \hat{m} &= \frac{\sum_{i=1}^n x_i}{n} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{m})^2 \end{aligned}$$

6.5 Exercice

1. Montrer que \mathbf{B} , l'estimateur de β par moindres carrés, correspond à l'estimateur du maximum de vraisemblance. Pour ce faire, on écrira la vraisemblance du vecteur aléatoire ${}^t(\mathbf{y}_1, \dots, \mathbf{y}_n)$ où \mathbf{y}_i est gaussien de moyenne $\beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ et de variance σ^2 .
2. Trouver l'estimateur de σ^2 par maximum de vraisemblance.