

UNIVERSITÉ LAVAL
ÉCOLE D'ACTUARIAT

ACT 2003
Notes de cours
Modèles linéaires en actuariat

David Beauchemin

Automne 2017

© 2017 David Beauchemin



Cette création est mise à disposition selon le contrat [Attribution-Partage dans les mêmes conditions 4.0 International](#) de Creative Commons. En vertu de ce contrat, vous êtes libre de :

- **partager** — reproduire, distribuer et communiquer l’œuvre ;
- **remixer** — adapter l’œuvre ;
- utiliser cette œuvre à des fins commerciales.

Selon les conditions suivantes :



Attribution — Vous devez créditer l’œuvre, intégrer un lien vers le contrat et indiquer si des modifications ont été effectuées à l’œuvre. Vous devez indiquer ces informations par tous les moyens possibles, mais vous ne pouvez suggérer que l’offrant vous soutient ou soutient la façon dont vous avez utilisé son œuvre.



Partage dans les mêmes conditions — Dans le cas où vous modifiez, transformez ou créez à partir du matériel composant l’œuvre originale, vous devez diffuser l’œuvre modifiée dans les mêmes conditions, c’est-à-dire avec le même contrat avec lequel l’œuvre originale a été diffusée.

Résumé

abstrat

Remerciements

Le document a été bâti à partir des notes de cours ACT-2003 manuscrites rédigées par Frédéric Guillot. Ce document est une reproduction améliorée de celle-ci. Je suis grandement reconnaissant de la confiance de Frédéric Guillot pour l'exécution de cette initiative de ma part.

Je remercie Thomas Landry de m'avoir laissé utiliser ses notes pour l'utilisation et l'explication de certains concepts du cours ACT-2000. De plus, je remercie Samuel Cabral Cruz pour le code \LaTeX des bulles d'informations qui à mon avis améliore la beauté de ce document. Finalement, je remercie Kaesey-Andrew Lépine qui a pris le temps de relire le document et de trouver de nombreuses erreurs typographiques.

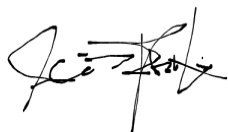
A handwritten signature in black ink, appearing to be 'F. Guillot', with a stylized flourish at the end.

Table des matières

1	Introduction	4
2	Régression linéaire simple	5
2.1	Introduction	5
2.1.1	Régression linéaire simple	6
2.1.2	Régression linéaire multiple	6
2.1.3	Régression exponentielle	8
2.1.4	Régression quadratique	8
2.2	Le modèle de régression linéaire simple	11
2.2.1	Coefficients de régression	13
2.2.2	Caractéristiques du terme d'erreur	20
2.3	Propriétés de l'estimateur des moindres carrés (EMC)	21
2.3.1	Estimateur sans biais	21
2.3.2	Variances et covariances des estimateurs	22
2.3.3	Optimalité	25
2.4	Régression passant par l'origine	25
2.5	Analyse de la variance	27
2.5.1	Notions préliminaires : Somme des carrés	28
2.5.2	Notions préliminaires : Degrés de liberté	31
2.5.3	Tableau d'analyse de la variance	32
2.6	Intervalles de confiance (I.C.) et test d'hypothèses	33
2.6.1	Distribution des variables aléatoires	33
2.6.2	Intervalle de confiance pour β_1	34
2.6.3	Intervalle de confiance pour β_0	35
2.6.4	Test d'hypothèses sur les paramètres	35
2.6.5	Test de la validité <i>globale</i> de la régression	41
2.7	Prévisions et intervalles de confiance	42
2.7.1	I.C. pour la prévision de type I (Valeur moyenne)	44
2.7.2	I.C. pour la prévision de type II (Vraie valeur)	46
3	Régression multiple	49
3.1	Le modèle sous forme matricielle	51
3.1.1	Estimateur des moindres carrés (EMC)	53

3.1.2	Résidus et tableau ANOVA	55
3.1.3	Estimateur de σ^2	56
3.1.4	Intervalle de confiance et tests d'hypothèses	56
3.1.5	Test de Student sur un seul paramètre	57
3.1.6	Test de Fisher pour la validité globale de la régression	58
3.1.7	Test de Fisher partiel	58
3.2	Sélection d'un modèle optimal	59
3.2.1	Technique 1 : Essai de tous les modèles	60
3.2.2	Technique 2 : Élimination régressive (<i>Backward elimination</i>)	61
3.2.3	Technique 3 : Sélection progressive (<i>forward selection</i>)	62
3.2.4	Technique 4 : Régression pas à pas (<i>stepwise regression</i>)	62
3.3	Régression avec variables indicatrices	66
3.4	Analyse qualitative des résidus	68
3.4.1	Problèmes possibles dans la distribution des résidus	68
3.4.2	Quantiles normaux	71
3.4.3	Exemple complet	72
4	Les modèles linéaires généralisés	80
4.1	Introduction	80
4.2	Notions préliminaires : La famille exponentielle	81
4.2.1	Loi Normale	81
4.2.2	Loi Poisson	82
4.2.3	Loi Bernoulli	82
4.2.4	Autres lois	83
4.3	Généralités sur les modèles de régression avec la famille exponentielle	83
4.3.1	Contexte	83
4.3.2	Autres lois	83
4.3.3	Structure du modèle	83
4.3.4	Propriété de la fonction de lien	84
4.4	Approche générale	84
4.4.1	Procédure avec les GLM	84
4.4.2	Estimation des paramètres	85
4.4.3	Validation <i>global</i> du modèle avec la <i>déviance</i>	86
A	Code source de l'exemple chapitre 3	87

Chapitre 1

Introduction

L'établissement de prévisions joue un rôle central dans notre vie de tous les jours (prévisions météorologiques, horoscope, etc.), et plus particulièrement dans celle des actuaires.

Objectifs de la régression

Régulièrement en actuariat, on se questionne sur les effets de différentes variables sur d'autres. Par exemple,

- Quel est l'effet de l'âge sur la fréquence des sinistres automobiles ?
- Quel est l'effet du sexe sur la mortalité ?

On cherche à étudier et déterminer les relations entre des variables mesurables à partir de données.

Deux grandes classes de variables mesurables :

- Qualitatives : basées sur des opinions et/ou des intuitions.
- Quantitatives : basées sur des observations, un modèle et des arguments mathématiques.

Deux *grandes étapes* pour établir des prévisions quantitatives

1. Bâtir le modèle et estimer les paramètres :
ex : $F = M \times a$ Qui représente un modèle déterministe
ex : $Y = 3 \times X + 6 + \epsilon_t$; où $\epsilon_t \sim N(0, 10)$ Qui représente un modèle probabiliste
2. Calculer les prévisions à partir du modèle.

Dans le cadre du cours, seulement les modèles probabilistes linéaires seront étudiés.

Chapitre 2

Régression linéaire simple

2.1 Introduction

De façon générale, en régression, nous avons :

Y	Variable dépendante, ou de réponse	Output
X_1, X_2, \dots, X_n	Soit n variables indépendantes ou explicatives, ou exogènes ¹	Input
$\beta_0, \beta_1, \dots, \beta_n$	Les paramètres à estimer	

Voici une illustration du concept de régression linéaire

Étape 1

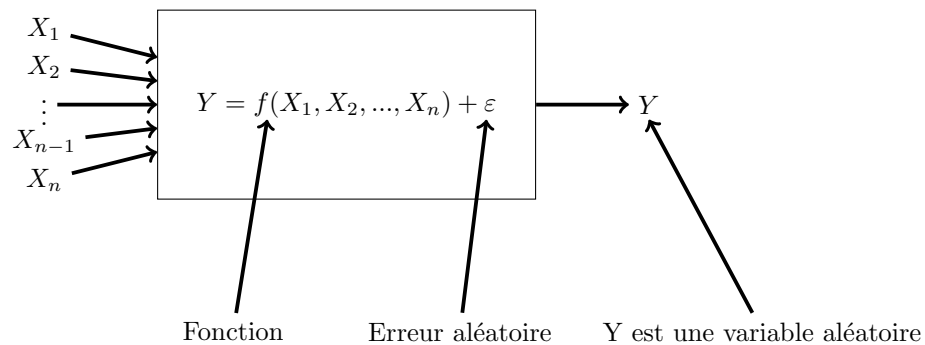
Observation
des X_i

Étape 2

Modèle de ré-
gression

Étape 3

Prévision de Y



1. Les variables X_i sont indépendantes par rapport à y, mais pas nécessairement entre elles.

2.1.1 Regression linéaire simple

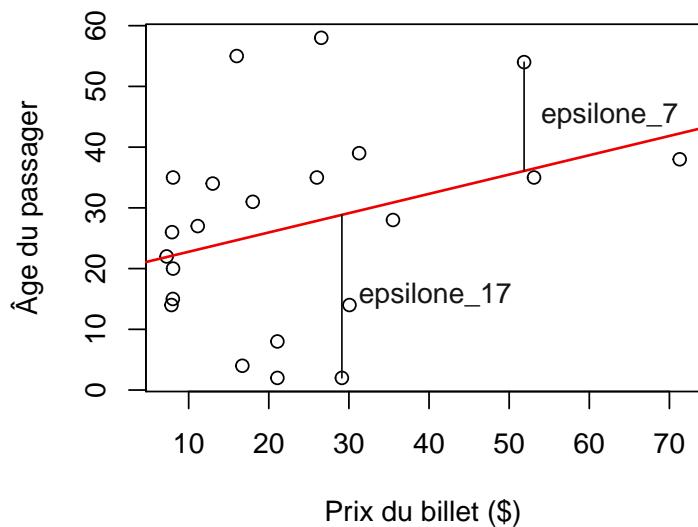
On cherche à prédire l'âge des passagers du Titanic selon le prix du billet à l'aide du modèle linéaire suivant,

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

↑
↗
↖

Âge du passager
Prix du billet
Erreur aléatoire

Âge prédit des passagers du Titanic



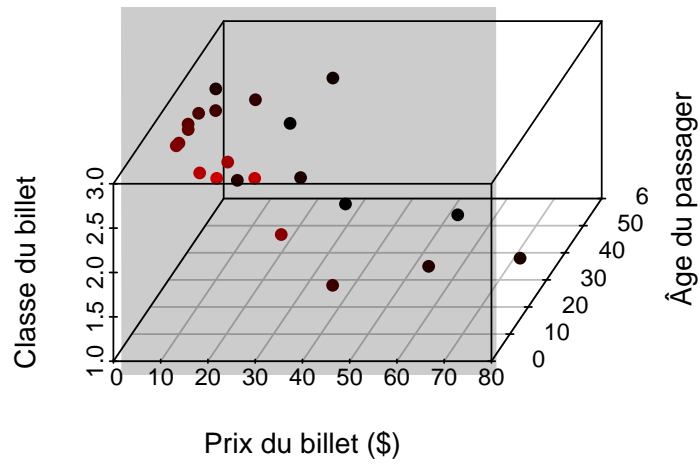
2.1.2 Régression linéaire multiple

On cherche à prédire l'âge des passagers du Titanic selon le prix du billet et son sexe à l'aide du modèle linéaire suivant,

The diagram illustrates a linear regression model for predicting flight price. At the top, the regression equation is displayed: $Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \varepsilon$. Below this equation, four labels are positioned with arrows pointing to specific components of the model:

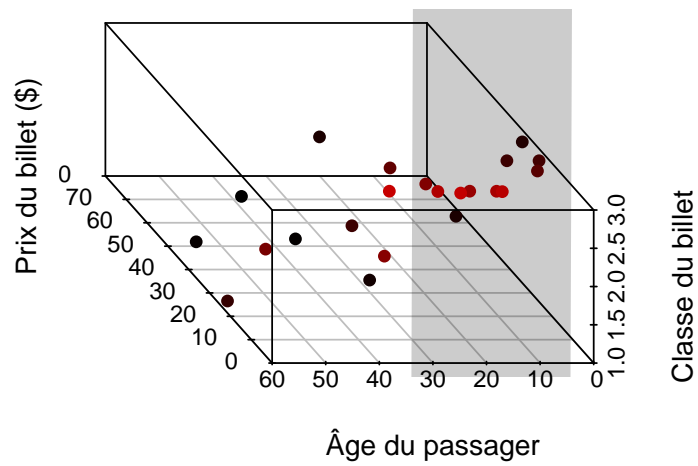
- Âge du passager**: An arrow points from this label to the coefficient β_1 .
- Prix du billet**: An arrow points from this label to the coefficient β_2 .
- Type de billet**: An arrow points from this label to the variable X_2 .
- Erreur aléatoire**: An arrow points from this label to the error term ε .

Âge prédit des passagers du Titanic



Voici la régression sous un autre angle, on voit la surface plane de régression.

Âge prédit des passagers du Titanic



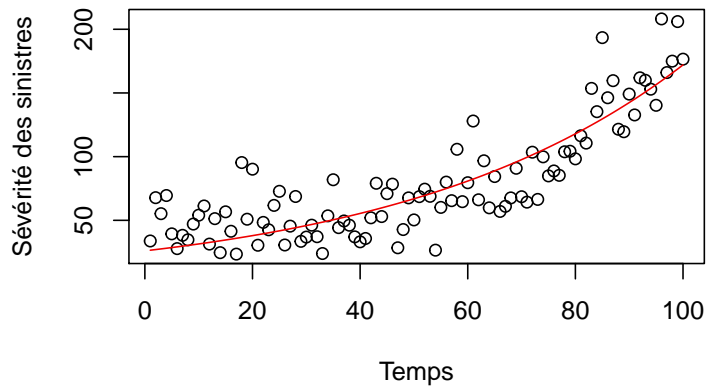
2.1.3 Régression exponentielle

On cherche à prédire la sévérité d'un sinistre automobile en fonction du temps à l'aide du modèle exponentiel suivant,

$$Y = \beta_0 \times e^{\beta_1 \times X} \times \varepsilon$$

Sévérité du sinistre Temps Erreur aléatoire

Modèle de prédiction de la sévérité des sinistres



Note

On remarque que la régression exponentielle est similaire à une régression linéaire simple.

$$\ln(Y) = \ln(\beta_0) + \beta_1 \times X + \ln(\varepsilon)$$
$$Y^* = \beta_0^* + \beta_1 \times X + \varepsilon^*$$

Qu'on appelle aussi une régression multiplicative ou log linéaire.

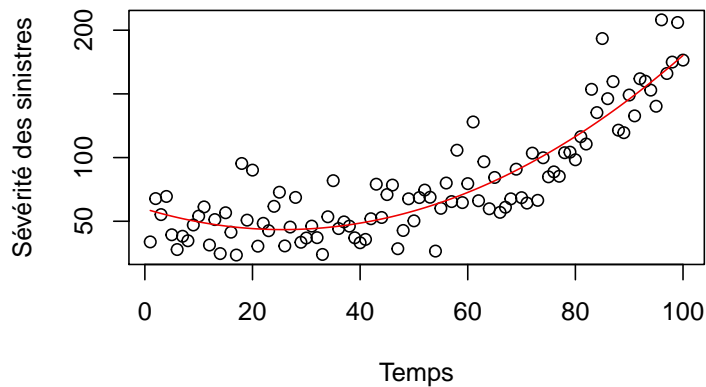
2.1.4 Régression quadratique

On cherche à prédire la sévérité d'un sinistre automobile en fonction du temps et du temps au carré à l'aide du modèle quadratique suivant,

$$Y = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \varepsilon$$

\uparrow Sévérité du sinistre \nwarrow Temps \nearrow Erreur aléatoire

Modèle de prédiction de la sévérité des sinistres



Note

On remarque que la régression quadratique est similaire à une régression linéaire multiple. En posant $X_1 = X$ et $X_2 = X^2$

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \varepsilon$$

Soit une régression linéaire multiple.

Dans le cadre du cours, seulement les modèles linéaires seront à l'étude pour les différentes raisons suivantes

- Plus simples
- Plusieurs modèles peuvent se ramener à un modèle linéaire simple ou multiple. (voir [2.1.3](#) et [2.1.4](#))
- Constituent souvent une très bonne approximation de la réalité qui peut être très complexe, telle que l'assurance.
- Se généralisent facilement, tels que les *Generalized Linear Models*.

Le principal problème de la modélisation linéaire est de trouver les différents paramètres $\beta_0, \beta_1, \dots, \beta_n$ de telle sorte que

$$\varepsilon = Y - f(X_1, \dots, X_n; \beta_0, \beta_1, \dots, \beta_n) \quad (2.1)$$

soit minimisé.

Il existe plusieurs méthodes pour calcul l'erreur. Soit les erreurs suivantes :

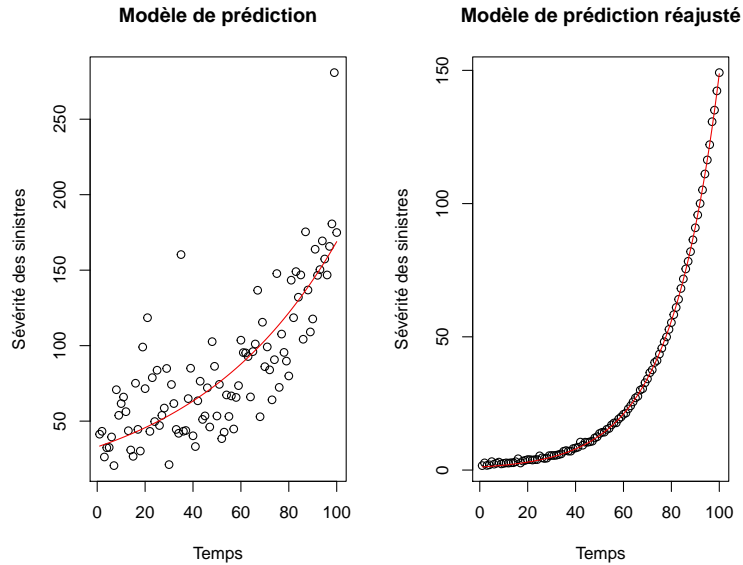
- Erreur totale
- Erreur absolue
- Erreur quadratique

Quel type d'erreur est suffisante pour déterminer ε ?

2.1.4.1 Erreur totale

$$\sum_{t=1}^n \varepsilon_t = \sum_{t=1}^n \left(Y_t - (\beta_0 + \beta_1 \times X_t) \right) \quad (2.2)$$

- Facile à mettre à 0
- Manque de fiabilité à cause de la mise à zéro



2.1.4.2 Erreur absolue

$$\sum_{t=1}^n |\varepsilon_t| = \sum_{t=1}^n \left| Y_t - (\beta_0 + \beta_1 \times X_t) \right| \quad (2.3)$$

- Très robuste
- Très compliquée mathématiquement, pour minimiser $\sum_{t=1}^n |\varepsilon_t|$ cela implique de dériver la fonction.

2.1.4.3 Erreur quadratique

$$\sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n \left[Y_t - (\beta_0 + \beta_1 \times X_t) \right]^2 \quad (2.4)$$

- Mathématiquement plus simple que l'erreur absolue
- Donne beaucoup de poids aux grandes erreurs

L'erreur quadratique semble donc l'option la plus simple due à la facilité mathématique et sa fiabilité.

2.2 Le modèle de régression linéaire simple

Le modèle de régression linéaire simple tente d'expliquer le mieux possible la variable **dépendante**² Y à l'aide d'une variable **indépendante**³ X .

Si on dispose de n paires d'observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ alors, le modèle s'exprime comme suit :

$$Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i, i = 1, \dots, n. \quad (2.5)$$

Où β_0 est le paramètre associé à l'ordonnée à l'origine du modèle ; β_1 est le paramètre associé à la pente de la droite ; et ε est le terme d'erreur.

Quelques remarques sur le modèle

Dans l'équation 2.5 du modèle, on remarque que :

- Les observations de Y_i sont tirées d'une variable aléatoire ;
- Les observations de X_i sont considérées comme des valeurs connues et non aléatoires ;
- Les paramètres β_0 et β_1 sont inconnus au départ. Ils doivent être estimés ;
- ε_i sont des réalisations inconnues d'une variable aléatoire.

Exemple d'un modèle de régression

X_t : Nombre d'années de scolarité de l'actuaire t

Y_t : Salaire de l'actuaire t

2. On appelle parfois la variable dépendante une variable **endogène**. Qui s'interprète comme étant une variable qui est due à une cause interne.

3. On appelle parfois les variables dépendantes des variables **exogène**. Qui s'interprète comme étant extérieur à un système.

Comment résoudre le modèle pour prédire les salaires des actuaires en fonction du nombre d'années de scolarité ?

Raisonnement :

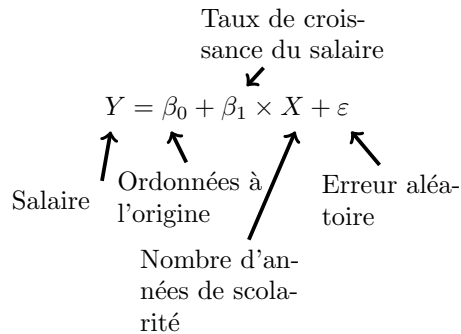
- Pour $X_t = 0$; on a $Y_t = \beta_0$. Autrement dit, le salaire avec un nombre d'années nulle de scolarité est *en moyenne* de β_0 . Par exemple, β_0 serait le salaire moyen d'un stagiaire.
 - Par la suite, pour chaque année additionnelle de scolarité, le salaire augmente *en moyenne* de β_1 unités.
- Ainsi, *en moyenne* on a

$$E[Y_t|X_t] = \beta_0 + \beta_1 \times X_t$$

Habituellement, la relation n'est pas parfaitement exacte dans la réalité. On se retrouve ainsi avec une *différence* dans notre variable exogène prédite. L'erreur est notée ε_t et est telle que mentionnée plus tôt, assumée aléatoire.

$$\begin{aligned}\varepsilon_t &= Y_t - E[Y_t|X_t] \\ &= Y_t - (\beta_0 + \beta_1 \times X_t)\end{aligned}$$

En réorganisant, on retrouve l'équation 2.5.



On doit maintenant trouver les paramètres β_0 et β_1 de manière à minimiser l'erreur ε_t .

Si ε_t est minimal, cela veut dire que $Y_t \approx \beta_0 + \beta_1 \times X_t$. Ce qui signifie que la droite de régression est une bonne approximation de Y_t .



En résumé

En résumé, on cherche à minimiser nos résidus en optimisant les paramètres β_i .

2.2.1 Coefficients de régression

Les paramètres β_0 et β_1 sont déterminés en minimisant l'erreur quadratique à l'aide de la méthode des moindres carrés.

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{t=1}^n \varepsilon_t^2 \\ &= \sum_{t=1}^n (Y_t - (\beta_0 + \beta_1 \times X_t))^2 \\ &= \sum_{t=1}^n (Y_t - \beta_0 - \beta_1 \times X_t)^2 \end{aligned}$$

Où $S(\psi)$ peut être considérée comme une mesure de la *distance* entre les données observées et le modèle théorique qui prédit ces données⁴.

Afin de minimiser la fonction $S(\beta_0, \beta_1)$, on dérive la fonction partiellement en fonction de chacun des paramètres.

Minimisation de β_0

$$\begin{aligned} \frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} &= 0 \\ \frac{\partial}{\partial \beta_0} \sum_{t=1}^n (Y_t - \hat{\beta}_0 - \hat{\beta}_1 \times X_t)^2 &= 0 \\ -2 \sum_{t=1}^n (Y_t - \hat{\beta}_0 - \hat{\beta}_1 \times X_t) &= 0 \\ \sum_{t=1}^n Y_t - n \times \hat{\beta}_0 - \hat{\beta}_1 \sum_{t=1}^n X_t &= 0 \end{aligned} \tag{2.6}$$

4. Pour de plus amples informations sur la méthode des moindres carrés et la fonction de *distance*, la page [Wikipédia](#) contient une bonne explication sur le sujet.

Minimisation de β_1

$$\begin{aligned}
\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} &= 0 \\
\frac{\partial}{\partial \beta_1} \sum_{t=1}^n (Y_t - \hat{\beta}_0 - \hat{\beta}_1 \times X_t)^2 &= 0 \\
-2 \sum_{t=1}^n (Y_t - \hat{\beta}_0 - \hat{\beta}_1 \times X_t) \times X_t &= 0 \\
\sum_{t=1}^n Y_t \times X_t - \hat{\beta}_0 \sum_{t=1}^n X_t - \hat{\beta}_1 \sum_{t=1}^n X_t^2 &= 0
\end{aligned} \tag{2.7}$$

À l'aide des équations 2.6 et 2.7, on peut trouver les deux inconnus β_0 et β_1 .
À partir de 2.6 :

$$\begin{aligned}
\sum_{t=1}^n Y_t - n \times \hat{\beta}_0 - \hat{\beta}_1 \sum_{t=1}^n X_t &= 0 \\
\sum_{t=1}^n Y_t - \hat{\beta}_1 \sum_{t=1}^n X_t &= n \times \hat{\beta}_0 \\
\frac{\sum_{t=1}^n Y_t}{n} - \hat{\beta}_1 \frac{\sum_{t=1}^n X_t}{n} &= \hat{\beta}_0 \\
\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}} &
\end{aligned} \tag{2.8}$$

Et à partir de 2.7 :

$$\begin{aligned}
\sum_{t=1}^n Y_t \times X_t - \hat{\beta}_0 \sum_{t=1}^n X_t - \hat{\beta}_1 \sum_{t=1}^n X_t^2 &= 0 \\
\sum_{t=1}^n Y_t \times X_t - \hat{\beta}_0 \sum_{t=1}^n X_t &= \hat{\beta}_1 \sum_{t=1}^n X_t^2 \\
\hat{\beta}_1 &= \frac{\sum_{t=1}^n Y_t \times X_t - \hat{\beta}_0 \sum_{t=1}^n X_t}{\sum_{t=1}^n X_t^2}
\end{aligned} \tag{2.9}$$

On utilise l'équation 2.8 de $\hat{\beta}_0$ avec l'équation 2.9 de $\hat{\beta}_1$, on développe l'équation résultante afin d'isoler $\hat{\beta}_1$.

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^n Y_t \times X_t - (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{t=1}^n X_t}{\sum_{t=1}^n X_t^2} \\ &= \frac{\sum_{t=1}^n Y_t \times X_t - (\bar{Y} - \hat{\beta}_1 \bar{X}) \times n\bar{X}}{\sum_{t=1}^n X_t^2} \\ &= \frac{\sum_{t=1}^n Y_t X_t - n\bar{Y}\bar{X} + \hat{\beta}_1 \times \bar{X}^2 \times n}{\sum_{t=1}^n X_t^2}\end{aligned}$$

En isolant $\hat{\beta}_1$, on obtient la définition suivante

$$\hat{\beta}_1 = \frac{\sum_{t=1}^n Y_t X_t - n\bar{Y}\bar{X}}{\sum_{t=1}^n X_t^2 - n\bar{X}^2} \quad (2.10)$$

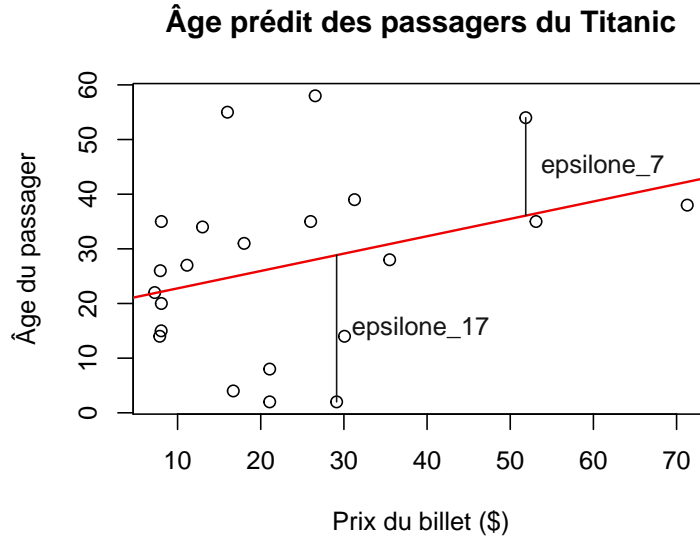
Remarques

1. On note $\hat{\varepsilon}_t$ les résidus générés par le modèle estimé :

$$\begin{aligned}\hat{\varepsilon}_t &= Y_t - \hat{Y}_t \\ \hat{\varepsilon}_t &= Y_t - (\hat{\beta}_0 - \hat{\beta}_1 X_t); \text{ pour } t = 1, 2, \dots, n\end{aligned}$$

Si on illustre graphiquement les résidus, il s'agit du segment le plus court entre la droite de régression et la donnée observée.

Si on reprend le graphique de la section 2.1.1, on observe facilement les résidus sur cette représentation graphique :



2. Le *centre de gravité*⁵ des données (\bar{X}, \bar{Y}) se trouvent exactement sur la droite de régression.

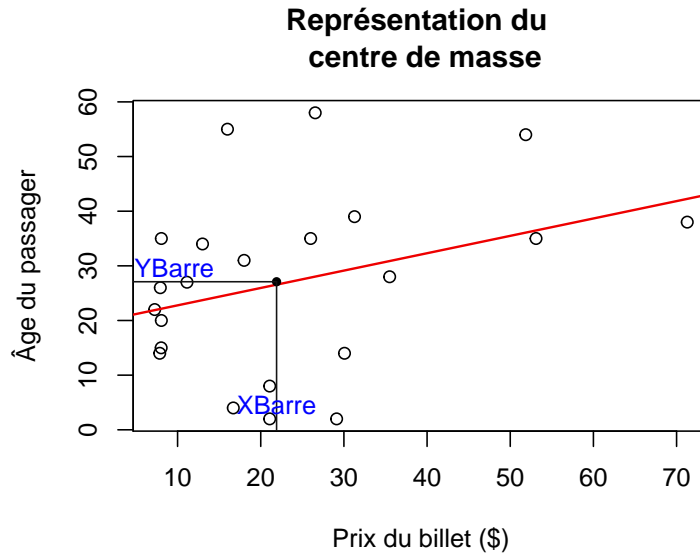
On peut facilement effectuer cette preuve à partir de l'équation 2.8,

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \bar{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + 0\end{aligned}$$

On note ainsi une absence de résidus pour le centre de masse.

Si on reprend (encore) le graphique de la section 2.1.1, on observe facilement le centre de masse sur le graphique.

5. Qu'on appelle parfois le centre de masse.



3. La somme des résidus de tout modèle de régression linéaire est nulle.

$$\begin{aligned}
 \sum_{t=1}^n \hat{\varepsilon}_t &= \sum_{t=1}^n (Y_t - (\hat{\beta}_0 + \hat{\beta}_1 X_t)) \\
 &\stackrel{2.8}{=} \sum_{t=1}^n (Y_t - (\bar{Y} - \hat{\beta}_1 \bar{X})) \\
 &= \sum_{t=1}^n Y_t - \sum_{t=1}^n \bar{Y} + \hat{\beta}_1 \sum_{t=1}^n \bar{X} - \hat{\beta}_1 \sum_{t=1}^n X_t \\
 &= n\bar{Y} - n\bar{Y} + \hat{\beta}_1 + n\bar{X} - \hat{\beta}_1 + n\bar{X} \\
 &= 0
 \end{aligned}$$

Notation

Afin de faciliter l'écriture, on intègre la notation suivante ; S_{xx} et S_{xy} . Les expressions précédentes sont appelées respectivement : la somme des carrés corrigée de x et la somme des produits croisés corrigée de x et de y . Voici le développement pour

S_{xx} ,

$$\begin{aligned}
S_{xx} &= \sum_{t=1}^n (X_t - \bar{X})^2 \\
&= \sum_{t=1}^n (X_t^2 - 2X_t\bar{X} + \bar{X}^2) \\
&= \sum_{t=1}^n X_t^2 - 2\bar{X} \sum_{t=1}^n X_t + n\bar{X}^2 \\
&= \sum_{t=1}^n X_t^2 - 2\bar{X}n\bar{X} + n\bar{X}^2 \\
&= \sum_{t=1}^n X_t^2 - n\bar{X}^2
\end{aligned}$$

On effectue le même type de développement pour S_{xy} ,

$$\begin{aligned}
S_{xy} &= \sum_{t=1}^n (X_t - \bar{X})(Y_t - \bar{Y}) \\
&\vdots \\
&= \sum_{t=1}^n X_t Y_t - n\bar{X}\bar{Y}
\end{aligned}$$

À l'aide des sommes de carrés corrigés, on peut réécrire la définition de $\hat{\beta}_1$

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}} \tag{2.11}$$

Exemple

On poursuit avec un exemple pour assimiler l'information.

- On dispose des cinq observations suivantes du couple (X_t, Y_t) dans le tableau de gauche ainsi que les éléments calculés nécessaires pour trouver les paramètres dans le tableau de droite.

t	X_t	Y_t
1	2	2
2	3	5
3	6	3
4	9	6
5	12	5
Totaux :	32	21

t	X_t^2	$X_t Y_t$
1	4	4
2	9	15
3	36	18
4	81	54
5	144	60
Totaux :	274	151

À partir des définitions 2.8 et 2.10, on trouve facilement la valeur de $\hat{\beta}_0$ et de $\hat{\beta}_1$.


$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{t=1}^n Y_t X_t - n \bar{Y} \bar{X}}{\sum_{t=1}^n X_t^2 - n \bar{X}^2} \\
&= \frac{151 - (5)(\frac{21}{5})(\frac{32}{5})}{274 - (5)(\frac{32}{5})^2} \\
&= \frac{83}{346} \\
&\approx 0.2399
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\
&= \frac{21}{5} - \left(\frac{83}{346}\right) \times \left(\frac{32}{5}\right) \\
&\approx 2.6647
\end{aligned}$$

On obtient ainsi le modèle de régression suivant :

$$Y_t = 2.6647 + 0.2399X_t + \varepsilon_t$$

t	$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$	$\hat{\varepsilon}_t$
1	3.1445	-1.1445
2	3.3844	1.6156
3	4.1041	-1.1041
4	4.8238	1.1762
5	5.5435	-0.5435

$$\sum_{t=1}^5 \varepsilon_t \approx -0.0003$$


Exécution en R

```
3 > # Dataset
4 > x <- c(2,3,6,9,12); y <- c(2,5,3,6,5)
5 > # Estimations des parametres
6 > reg <- lm(y ~ x)
7 > # Resume de l'estimation
8 > summary(reg)
9 > # Valeurs de Yt
10 > fitted(reg)
11 > # Residus
12 > residuals(reg)
```

Listing 2.1 – Code source en R pour l'exemple



Astuce calculatrice

La calculatrice TI-30XS Multiview permet de créer un tableau de donnée et de sortir rapidement et facilement différentes informations sur une régression à partir des données.

Tel que :

- \bar{X} et \bar{Y} ;
- $\sum_{t=1}^n X_t$, $\sum_{t=1}^n X_t^2$, $\sum_{t=1}^n Y_t$, $\sum_{t=1}^n Y_t^2$ et $\sum_{t=1}^n X_t Y_t$;
- $\hat{\beta}_0$ et $\hat{\beta}_1$

Pour de plus ample information, consulter le [guide](#) sur les calculatrices.

2.2.2 Caractéristiques du terme d'erreur

On rappelle que l'équation du modèle de régression correspond à

$$Y_t = \beta_0 + \beta_1 \times X_t + \varepsilon_t \quad (2.5)$$

De plus, on sait qu'il s'agit des valeurs moyennes de Y_t en sachant X_t , soit

$$Y_t = E[Y_t|X_t] + \varepsilon_t$$

On peut ainsi formuler les trois postulats⁶ suivants,

6. Le [postulat](#) est un principe non démontré, mais utilisé dans la construction d'une théorie mathématique.

1. $E[\varepsilon_t] = 0$, par définition pour que $E[Y_t] = E[Y_t|X_t]$. Il s'agit de l'hypothèse de linéarité de la variable explicative. On dit qu'elle est exogène si elle n'est pas corrélée au terme d'erreur.
2. $Var(\varepsilon_t) = \sigma^2$, la variance des termes d'erreurs est supposée constante. Il s'agit de l'hypothèse d'homoscédasticité.
3. $Cov(\varepsilon_t, \varepsilon_s) = 0$, pour $t \neq s$, il n'y a pas de corrélation entre les termes d'erreurs. Il s'agit de l'hypothèse d'indépendance des erreurs.

i

Quatrième postulat

Les hypothèses de linéarité et d'homoscédasticité sont très intéressantes, si on observe leurs définitions ensemble on remarque qu'il s'agit d'une distribution avec une espérance nulle et une variabilité supposée constante. Ce qui nous amène à une quatrième hypothèse, les résidus sont distribués selon une loi normale.

$$\hat{\varepsilon}_t|x_i \sim N(0, \sigma^2)$$

2.3 Propriétés de l'estimateur des moindres carrés (EMC)

2.3.1 Estimateur sans biais

On rappelle qu'un estimateur est dit sans biais lorsque son espérance est égale à la valeur vraie du paramètre, soit $E[\hat{\theta}] = \theta \Leftrightarrow b(\hat{\theta}) = 0$ ⁷.

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\frac{\sum_{t=1}^n (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^n (X_t - \bar{X})^2}\right] \\ &= \frac{\sum_{t=1}^n (X_t - \bar{X})E[Y_t - \bar{Y}]}{\sum_{t=1}^n (X_t - \bar{X})^2} \\ &= \frac{\sum_{t=1}^n (X_t - \bar{X})(E[Y_t] - E[\bar{Y}])}{\sum_{t=1}^n (X_t - \bar{X})^2} \end{aligned}$$

7. Notes de cours ACT-2000, chapitre 3, Thomas Landry, Hiver 2017.

De l'équation 2.5, et avec le postulat 1, on sait que,

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 \times X_t + \varepsilon_t \\ E[Y_t] &= E[\beta_0 + \beta_1 \times X_t] + E[\varepsilon_t] \\ &\stackrel{1}{=} \beta_0 + \beta_1 \times X_t + 0 \end{aligned}$$

On applique le même raisonnement pour l'espérance de \bar{Y} .

$$\begin{aligned} E[\hat{\beta}_1] &= \frac{\sum_{t=1}^n (X_t - \bar{X})(E[Y_t] - E[\bar{Y}])}{\sum_{t=1}^n (X_t - \bar{X})^2} \\ &= \frac{\sum_{t=1}^n (X_t - \bar{X})(\beta_0 + \beta_1 \times X_t - \beta_0 - \beta_1 \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} \\ &= \frac{\sum_{t=1}^n (X_t - \bar{X})\beta_1(X_t - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} \\ &= \beta_1 \frac{\sum_{t=1}^n (X_t - \bar{X})^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \\ E[\hat{\beta}_1] &= \beta_1 \end{aligned}$$

Par conséquent,

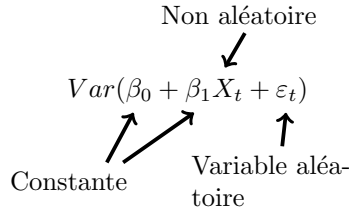
$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{X}] \\ &= E[\bar{Y}] - \bar{X}E[\hat{\beta}_1] \\ &= \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_1 \\ E[\hat{\beta}_0] &= \beta_0 \end{aligned}$$

On peut ainsi conclure que les deux estimateurs des paramètres sont sans biais.

2.3.2 Variances et covariances des estimateurs

On s'intéresse aux variances et aux covariances des estimateurs, cette deuxième propriété ainsi que la première nous permettra de déduire une conclusion en lien avec le quatrième postulat.

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{t=1}^n (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^n (X_t - \bar{X})^2}\right) \\
&= \frac{Var\left(\sum_{t=1}^n (X_t - \bar{X})Y_t - \sum_{t=1}^n (X_t - \bar{X})\bar{Y}\right)}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2} \\
&= \frac{Var\left(\sum_{t=1}^n (X_t - \bar{X})Y_t\right) + Var\left(\sum_{t=1}^n (X_t - \bar{X})\bar{Y}\right)}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2} \\
&= \frac{\sum_{t=1}^n (X_t - \bar{X})^2 Var(Y_t) + Var(\bar{Y}(n\bar{X} - n\bar{X}))}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2} \\
&= \frac{\sum_{t=1}^n (X_t - \bar{X})^2 Var(\beta_0 + \beta_1 X_t + \varepsilon_t) + 0}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2}
\end{aligned}$$



$$\begin{aligned}
&= \frac{\sum_{t=1}^n (X_t - \bar{X})^2 Var(\varepsilon_t)}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2} \\
&\stackrel{2}{=} \frac{\sum_{t=1}^n (X_t - \bar{X})^2 \sigma^2}{\left(\sum_{t=1}^n (X_t - \bar{X})^2\right)^2}
\end{aligned}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

(2.12)

$$\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1 \bar{X}) \\
&= Var(\bar{Y}) + Var(\hat{\beta}_1 \bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X}) \\
&= Var\left(\frac{\sum_{t=1}^n Y_t}{n}\right) + \bar{X}^2 Var(\hat{\beta}_1) - 2\bar{X} Cov(\bar{Y}, \hat{\beta}_1) \\
&= \frac{n \times Var(Y_t)}{n^2} + \bar{X}^2 \left(\frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right) - 2\bar{X} Cov(\bar{Y}, \hat{\beta}_1)
\end{aligned}$$

$$\begin{aligned}
Cov(\bar{Y}, \hat{\beta}_1) &= Cov\left(\frac{\sum_{t=1}^n Y_t}{n}, \frac{\sum_{s=1}^n (X_s - \bar{X})(Y_s - \bar{Y})}{\sum_{s=1}^n (X_s - \bar{X})^2}\right) \\
&= \frac{1}{n} \frac{1}{\sum_{s=1}^n (X_s - \bar{X})^2} Cov\left(\sum_{t=1}^n Y_t, \sum_{s=1}^n (X_s - \bar{X}) Y_s - \bar{Y} \sum_{s=1}^n (X_s - \bar{X})\right) \\
&= \frac{1}{n} \frac{1}{\sum_{s=1}^n (X_s - \bar{X})^2} \sum_{t=1}^n \sum_{s=1}^n (X_s - \bar{X}) Cov(Y_t, Y_s) \\
&= \frac{1}{n} \frac{1}{\sum_{s=1}^n (X_s - \bar{X})^2} \left(\sum_{\substack{t=1 \\ :t \neq s}}^n \sum_{s=1}^n (X_s - \bar{X}) \times 0 + \sum_{\substack{t=1 \\ :t=s}}^n \sum_{s=1}^n (X_s - \bar{X}) \sigma^2 \right) \\
&= \frac{1}{n} \frac{1}{\sum_{s=1}^n (X_s - \bar{X})^2} \sigma^2 \left(\sum_{t=1}^n (X_t - \bar{X}) \right) \\
&= \frac{1}{n} \frac{1}{\sum_{s=1}^n (X_s - \bar{X})^2} \sigma^2 (n\bar{X} - n\bar{X}) \\
&= 0
\end{aligned}$$

$$\boxed{Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{X}^2 \left(\frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)} \quad (2.13)$$

Finalement, pour la covariance entre $\hat{\beta}_0$ et $\hat{\beta}_1$

$$\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{Y} - \hat{\beta}_1 \bar{X}, \hat{\beta}_1) \\
&= Cov(\bar{Y}, \hat{\beta}_1) - \bar{X} Var(\hat{\beta}_1) \\
&= 0 - \bar{X} \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2}
\end{aligned}$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{X} \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \quad (2.14)$$

i

Résumé des propriétés des estimateurs

Les équations 2.13 et 2.12 ainsi que le postulat 4 à la section 2.2.2 nous permettent de conclure que

$$\begin{aligned} \hat{\beta}_0 &\sim N\left(\beta_0, \frac{\sigma^2}{n} + \bar{X}^2 \left(\frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)\right) \\ \hat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2}\right) \end{aligned}$$

2.3.3 Optimalité

Le théorème de [Gauss-Markor](#) nous permet d'établir que l'estimateur des moindres carrés est l'estimateur non biaisé à variance minimale.

Notions importantes à retenir du théorème :

1. Considérer l'estimateur $\Theta^* = \sum_{t=1}^n C_t \times Y_t$
2. Minimiser $\text{Var}(\Theta^*)$ sous la contrainte que $E[\Theta^*] = \beta$; où

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

2.4 Régression passant par l'origine

Dans certaines situations, il est possible que l'on souhaite forcer la droite de régression à passer par l'origine. Voici un exemple de situation où il est plus logique de forcer le modèle,

X_t : Nombre de Km parcourut

Y_t : Consommation d'essence en L d'une voituret

Il est plus logique d'avoir une consommation de 0 L pour une distance de 0 Km.

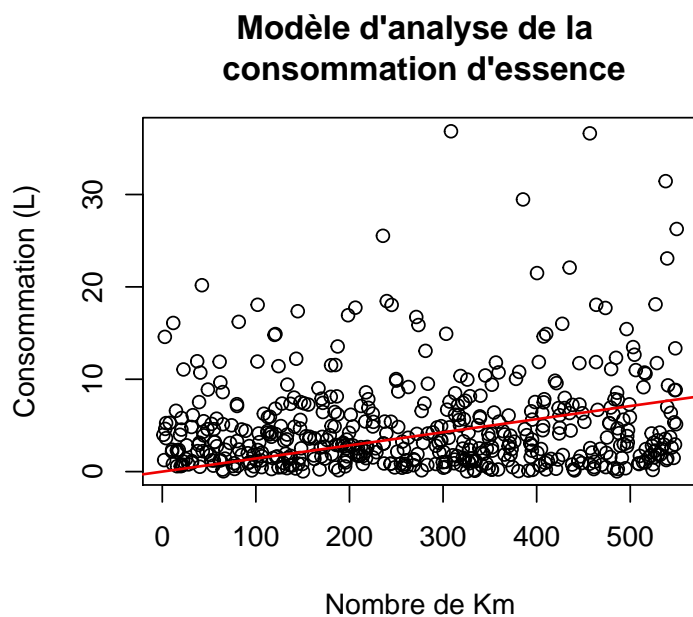
Dans ce cas, on peut postuler le modèle suivant :

$$Y_t = \beta \times X_t + \varepsilon_t \quad (2.15)$$

On peut démontrer par le même raisonnement qu'à la section 2.2.1 que de minimisation du paramètre β correspond à :

$$\hat{\beta} = \frac{\sum_{t=1}^n X_t Y_t}{\sum_{t=1}^n X_t^2} \quad (2.16)$$

On reprend l'exemple énoncé plus haut, voici le modèle représenté graphiquement :



Code R

Voici le code R permettant de créer un modèle linéaire simple avec une droite passant par l'origine.

```
3 > # dataset
4 > # X Km parcourus
5 > # Y consommation essence en L
6 > simul <- 500
7 > alpha <- 1
8 > beta <- alpha/5.1
9 > y <- rgamma(simul, alpha, beta)
10 > x <- runif(simul, 0, 550)
```

```

11 > # Estimation de beta
12 > reg <- lm(y ~ x - 1)
13 > plot(x, y, xlab = "Nombre de Km", ylab = "Consommation (L)",
14 >      main= "Modele d'analyse de la \n consommation d'essence")
14 > abline(reg, col="red2", lwd = 1.5)

```

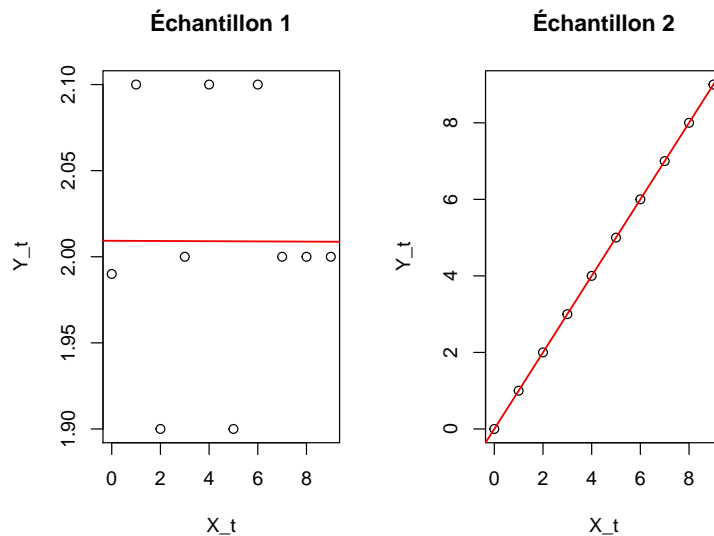
Listing 2.2 – Code source en R pour l'exemple

2.5 Analyse de la variance

Un tableau d'analyse de la variance permet d'évaluer la qualité de l'ajustement du modèle aux observations.

Idée

1. Si on décide de modéliser Y_t sans la régression, autrement dit de l'analyse statistique⁸, alors Y est vue comme une variable aléatoire avec une certaine variance, soit $Var(y)$.
2. En utilisant la régression pour modéliser Y_t en fonction de X_t une partie de la variance de Y_t est *expliquée* par la variance de X_t , alors que l'autre partie reste *inexpliquée*.
3. L'utilité de la régression est de trouver la proportion de la variance de Y_t qui est expliquée par la variance de X_t .



8. Cours ACT-2000


On voit que les résidus de l'échantillon 1 sont très mal expliqués par notre modèle, les résidus sont très élevés. Tandis que les résidus de l'échantillon 2 sont parfaitement expliqués par notre modèle.


```
$residusMauvaisFit
      1      2      3      4      5      6      7      8
2.009273 2.009212 2.009152 2.009091 2.009030 2.008970 2.008909 2.008848
      9     10
2.008788 2.008727

$residusBonFit
      1      2      3      4      5
-2.580003e-16 1.000000e+00 2.000000e+00 3.000000e+00 4.000000e+00
      6      7      8      9     10
 5.000000e+00 6.000000e+00 7.000000e+00 8.000000e+00 9.000000e+00
```

Il y a peu d'intérêt de construire un modèle avec les données de l'échantillon 1 car,

$$Var(Y_t) \approx 0\% \times Var(X_t) + 100\% \times Var(\varepsilon_t)$$


 Expliquée


 Inexpliquée

Il est préférable dans ce cas-ci d'utiliser les modèles statistiques vus dans le cours ACT-2000.

Par contre, il y a un intérêt à utiliser un modèle avec les données de l'échantillon 2 car,

$$Var(Y_t) = Var(X_t)$$

Autrement dit, la variable X explique bien la variable Y.

Note

Noter que les modèles précédents ont été ajustés pour mieux représenter le concept, un modèle avec un fit parfait n'est pas réaliste dans la réalité.

2.5.1 Notions préliminaires : Somme des carrés

La variance totale de Y_t est décomposable sous le modèle de régression linéaire, cette décomposition permet d'analyser l'ajustement du modèle. On la représente ainsi :

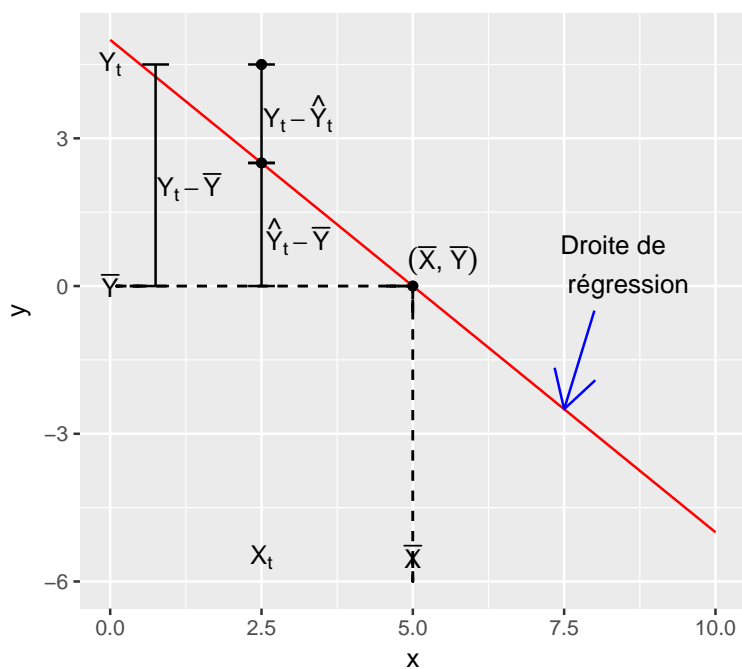
$$SST = \sum_{t=1}^n (Y_t - \bar{Y})^2$$

Décomposition

$$(Y_t - \bar{Y}) = Y_t - \hat{Y}_t + \hat{Y}_t - \bar{Y}$$

$$(Y_t - \bar{Y}) = (Y_t - \hat{Y}_t) + (\hat{Y}_t - \bar{Y})$$

$\underbrace{(Y_t - \bar{Y}) =}$	$\underbrace{(Y_t - \hat{Y}_t) +}$	$\underbrace{(\hat{Y}_t - \bar{Y})}$
Variation totale de Y_t	Variation de Y_t Variation expliquée par la régression	Résidu Variation inexpliquée par la régression



Par conséquent, on a que

$$\begin{aligned}
 SST &= \sum_{t=1}^n \left[(\hat{Y}_t - \bar{Y}) + (Y_t - \hat{Y}_t) \right]^2 \\
 &= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 + \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 + 2 \sum_{t=1}^n (\hat{Y}_t - \bar{Y})(Y_t - \hat{Y}_t)
 \end{aligned}$$

$$= \underbrace{\sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2}_{\substack{\text{SSR} \\ \text{Régression}}} + \underbrace{\sum_{t=1}^n (Y_t - \hat{Y}_t)^2}_{\substack{\text{SSE} \\ \text{Erreur}}} + \underbrace{2 \sum_{t=1}^n (\hat{Y}_t - \bar{Y})(Y_t - \hat{Y}_t)}_{\psi}$$

Développement de ψ

$$\begin{aligned} 2 \sum_{t=1}^n (\hat{Y}_t - \bar{Y})(Y_t - \hat{Y}_t) &\Rightarrow 2 \sum_{t=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_t - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})(Y_t - \bar{Y} + \bar{Y} - \hat{Y}_t) \\ &= 2 \sum_{t=1}^n \hat{\beta}_1 (\hat{X}_t - \bar{X})(Y_t - \bar{Y} + \hat{\beta}_0 + \hat{\beta}_1 \bar{X} - \hat{\beta}_0 - \hat{\beta}_1 X_t) \\ &= 2 \sum_{t=1}^n \hat{\beta}_1 (\hat{X}_t - \bar{X}) \left((Y_t - \bar{Y}) - \hat{\beta}_1 (X_t - \bar{X}) \right) \\ &= 2 \hat{\beta}_1 \sum_{t=1}^n (\hat{X}_t - \bar{X})(Y_t - \bar{Y}) - 2 \hat{\beta}_1^2 \sum_{t=1}^n (X_t - \bar{X})^2 \\ &= 2 \hat{\beta}_1 (S_{xy} - \hat{\beta}_1 S_{xx}) \\ &\stackrel{2.11}{=} 2 \hat{\beta}_1 \left(S_{xy} - \frac{S_{xy}}{S_{xx}} S_{xx} \right) \\ &= 2 \hat{\beta}_1 (S_{xy} - S_{xy}) \\ &= 0 \end{aligned}$$

Ainsi,

$$\boxed{\text{SST} = \text{SSR} + \text{SSE}} \quad (2.17)$$

Où SSR est la variation expliquée par le modèle de régression linéaire et SSE signifie la variation inexpliquée, ou résiduelle du modèle de régression linéaire.

Intuitivement,

- Dans un bon modèle de régression, on aimerait que
 - $\text{SST} \approx \text{SSR}$, soit que $\text{Var}(Y_t) \approx \text{Var}(\hat{Y}_t)$
- ou
- $\text{SSE} \approx 0$, soit que la variation résiduelle soit très faible
- On définit le coefficient de détermination par

$$\boxed{R^2 = \text{Corr}^2(Y, \hat{Y}) = \frac{\text{SSR}}{\text{SST}} \Leftrightarrow 1 - \frac{\text{SSE}}{\text{SST}}} \quad (2.18)$$

Par rapport au ratio, $\frac{SSR}{SST}$ signifie le pourcentage de la variance dans Y_t expliqué par la régression et $1 - \frac{SSR}{SST}$ signifie le pourcentage de la variance dans Y_t qui n'est pas expliquée par la régression.

- $R^2 \in [0, 1]$
- Si $R^2 = 100\%$, la régression est parfaite et utile ; si $R^2 = 0\%$, la régression n'est pas parfaite et est inutile.

2.5.2 Notions préliminaires : Degrés de liberté

Le nombre de *degrés de liberté*⁹ d'une *somme de carrés* est :

- Le nombre de composants *indépendants* dans la somme ;

ou

- Le nombre minimal de fonctions de Y_1, \dots, Y_n qu'il faut connaître pour obtenir la somme ;

ou

- **Pour SST et SSE seulement**

$$d.l. = (\text{Nombre de termes dans la somme}) - (\text{Nombre de paramètres estimés dans cette somme})$$

Ainsi,

- $SST = \sum_{t=1}^n (Y_t - \bar{Y})^2 \rightarrow n \text{ termes} - (1 \text{ paramètre estimé}^{10}) = \boxed{(n-1)d.l.}$
- $SSE = \sum_{t=1}^n (Y_t - \hat{Y}_t)^2$

$$\sum_{t=1}^n (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^n (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t)^2 \rightarrow n \text{ termes} - (2 \text{ paramètres estimés}^{11}) = \boxed{(n-2)d.l.}$$

- $SSR = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2$

$$\begin{aligned} \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 &= \sum_{t=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_t - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 \\ &= \underbrace{\hat{\beta}_1^2}_{f(y_1, \dots, y_n)} \times \underbrace{\sum_{t=1}^n (X_t - \bar{X})^2}_{f(x_1, \dots, x_n)} \end{aligned}$$

Soit une seule fonction des Y_1, \dots, Y_n doit être connue pour obtenir $SSR \rightarrow \boxed{1 d.l.}$

9. Couramment l'abréviation *d.l.* sera utiliser pour signifié *degrés de liberté*.

10. \bar{Y}

11. $\hat{\beta}_0$ et $\hat{\beta}_1$

Remarque

On sait que :

$$SST = SSE + SSR$$

On note aussi que

$$\begin{aligned} d.l.(SST) &= d.l.(SSE) + d.l.(SSR) \\ (n-1) &= (n-2) + (1) \end{aligned}$$

On aurait donc pu retrouver $d.l.(SST) = d.l.(SSE) + d.l.(SSR)$

2.5.3 Tableau d'analyse de la variance

On appelle couramment le tableau d'analyse de la variance le tableau ANOVA. Ce type de tableau est utilisé dans tous les logiciels de régression pour évaluer la qualité d'un modèle.

Source de la variance	Somme des carrés (SS)	Degrés de liberté ($d.l.$)	Carrés moyens (MS)	Ratio de Fisher (F)
Régression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Erreur	SSE	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n - 1$		

Exemple

On poursuit avec un exemple pour assimiler l'information, on reprend l'exemple de la section 2.2.1.

t	X_t	Y_t	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$	$\hat{\varepsilon}_t$
1	2	2	3.1445	-1.1445
2	3	5	3.3844	1.6156
3	6	3	4.1041	-1.1041
4	9	6	4.8238	1.1762
5	12	5	5.5435	-0.5435
Totaux :	32	21		

$$SSE = \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^n \varepsilon_t^2 = 6.8179$$

$$SSR = \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 = 3.9821$$

$$SST = SSE + SSR = 6.8179 + 3.9821 = 10.8000$$

ANOVA

Source	<i>SS</i>	<i>d.l.</i>	<i>MS</i>	<i>F</i>
Régression	3.9821	1	3.9821	1.7522
Erreur	6.8179	3	2.2726	
Totaux	10.8000	4		

R^2

$$R^2 = \frac{SSR}{SST} = \frac{3.9821}{10.8000} = 36.87\%$$
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{6.8179}{10.8000} = 36.87\%$$

Autrement dit, seulement 36.87 % de la variabilité des Y_t est expliquée par la variabilité des X_t . La régression n'est pas très efficace et utile.

Code R

Voici le code R permettant de créer un modèle linéaire simple avec une droite passant par l'origine.

```
3 > # Dataset
4 > y <- c(2, 5, 3, 6, 5); x <- c(2, 3, 6, 9, 12)
5 > # Estimation des betas
6 > reg <- lm(y ~ x)
7 > anova(reg)
```

Listing 2.3 – Code source en R pour l'exemple

2.6 Intervalles de confiance (I.C.) et test d'hypothèses

On poursuit l'objectif des sections 2.3 et 2.5m soit de valider la qualité du modèle de régression.

2.6.1 Distribution des variables aléatoires

On rappelle qu'avec le postulat 4 (2.2.2), on suppose que les résidus suivent une loi normale d'espérance nulle et de variance de σ^2 .

$$\hat{\varepsilon}_t | x_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

Les conséquences de ce postulat sont les suivantes :

1. $(Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t) \sim N(\beta_0 + \beta_1 X_t, \sigma^2)$ (Postulat 1)
2. Les propriétés de l'estimateur des moindres carrés avaient permis de démontrer que (section 2.3) :

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \bar{X}^2 \left(\frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2}\right)$$



Alternative

On peut tirer la même conclusion à partir de la propriété des fonctions linéaires de $\hat{\beta}_0$ et $\hat{\beta}_1$.

3. L'estimateur sans biais pour σ^2 est :

$$\sigma^2 = S^2 = MSE$$

$$MSE = \frac{SSE}{d.l.(SSE)}$$

$$\frac{SSE}{d.l.(SSE)} = \frac{\sum_{t=1}^n (Y_t - \hat{Y}_t)^2}{n - 2}$$

$$\sigma^2 = \frac{\sum_{t=1}^n \varepsilon_t^2}{n - 2} \quad (2.19)$$

4. On peut montrer que

$$\left(\frac{SSE}{\sigma^2} \right) \sim \chi^2(n - 2) \quad (2.20)$$

2.6.2 Intervalle de confiance pour β_1

Attention de ne pas confondre avec $\hat{\beta}_1$. Puisque $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$, on a que

$$\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} \right) \sim N(0, 1)$$

Si σ^2 était connu, l'intervalle de confiance serait de la forme suivante

$$\left[\hat{\beta}_1 \pm Z_{\alpha/2} \times \sqrt{\text{Var}(\hat{\beta}_1)} \right]$$

Par contre, σ^2 n'est souvent pas connu et il est nécessaire de l'estimer. Tel que mentionné plus haut, l'estimateur non biaisé correspond à l'équation 2.19. Mais cet estimateur ne suit pas une distribution normale. À l'aide des notions acquises en ACT-2000, il est possible de démontrer que si on utilise l'estimateur de σ^2 , soit S^2 , dans la formule de $\text{Var}(\hat{\beta}_1)$, c'est-à-dire :

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{S^2}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

Alors, on peut conclure que :

$$\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \right) \sim t(n-2)$$

On obtient ainsi l'intervalle de confiance suivant au niveau $100 \times (1 - \alpha)\%$ pour β_1 ,

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\frac{S^2}{\sum_{t=1}^n (X_t - \bar{X})^2}} \quad (2.21)$$

2.6.3 Intervalle de confiance pour β_0

De manière similaire, un intervalle de confiance au niveau $100 \times (1 - \alpha)\%$ pour β_0 est,

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\frac{S^2}{n} + \frac{S^2 \bar{X}^2}{\sum_{t=1}^n (X_t - \bar{X})^2}} \quad (2.22)$$

2.6.4 Test d'hypothèses sur les paramètres

Principales questions auxquelles on aimerait répondre :

1. L'ordonnée à l'origine (β_0) est-elle significativement différente de 0 ?
Sinon, on considère le modèle $Y_t = \beta_1 \times X_t + \varepsilon_t$.
2. La pente (β_1) est-elle significativement différente de 0 ?
Sinon, on considère le modèle $Y_t = \beta_0 + \varepsilon_t$.

Pour tester la question 1 :

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

On utilise la statistique suivante,

$$t = \frac{\hat{\beta}_0 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}}$$

Pour tester la question 2 :

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

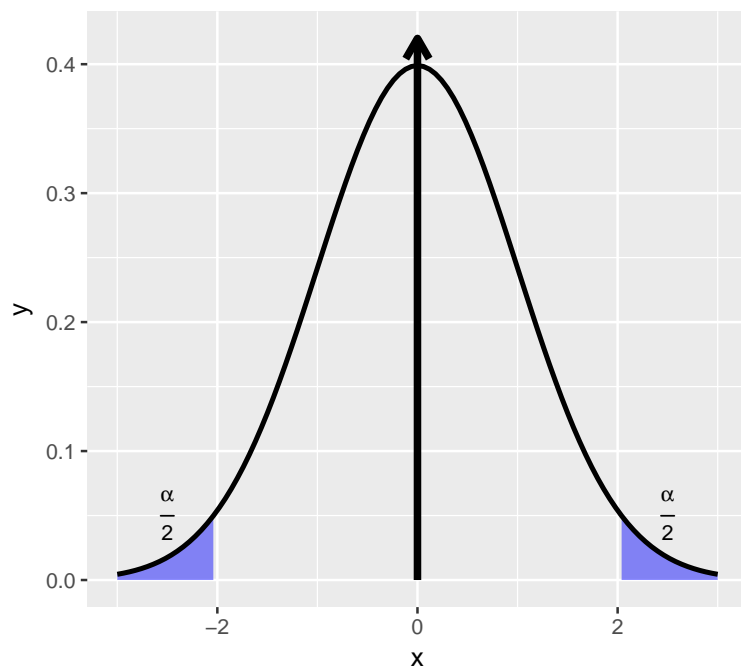
On utilise la statistique suivante,

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$$

On rejette H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ pour β_0 si :

$$|t| > t_{\frac{\alpha}{2}(n-2)}$$

Voici une représentation graphique de la zone de rejet bilatéral :



Qui correspond à la probabilité de *se tromper* en rejetant H_0 .

Remarques

De manière générale, on utilise plutôt les tests d'hypothèses suivants pour nos deux questions :

Pour tester la question 1 :

$$H_0 : \beta_0 = \beta_0^*$$

$$H_1 : \beta_0 \neq \beta_0^*$$

On utilise la statistique suivante,

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\widehat{Var}(\hat{\beta}_0)}}$$

Pour tester la question 2 :

$$H_0 : \beta_1 = \beta_1^*$$

$$H_1 : \beta_1 \neq \beta_1^*$$

On utilise la statistique suivante,

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$$

On rejette H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ pour β_0 si :

$$|t| > t_{\frac{\alpha}{2}(n-2)}$$

On poursuit avec un exemple pour assimiler l'information.

Exemple

Dans une régression sur un ensemble de 14 observations, on a obtenu :

$$\hat{Y}_t = 68.494 - 0.468X_t$$

ainsi que

$$\begin{aligned}\widehat{Var}(\hat{\beta}) &= \widehat{Var}\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right) \\ &= \begin{bmatrix} \widehat{Var}(\hat{\beta}_0) & \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{Var}(\hat{\beta}_1) \end{bmatrix} \\ &= \begin{bmatrix} 66.8511 & 1.2544 \\ 1.2544 & 0.0237 \end{bmatrix}\end{aligned}$$

Question 1

Tester si β_0 est significativement différent de 0 à un taux de confiance de 95 %.

$$H_0 : \beta_0 = 0 \text{ Hypothèse nulle}$$

$$H_1 : \beta_0 \neq 0$$

On utilise la statistique suivante,

$$\begin{aligned} t &= \frac{\hat{\beta}_0 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \\ &= \frac{68.494 - 0}{\sqrt{66.8511}} \\ &= 8.38 \\ t_{\frac{0.05}{2}(14-2)} &= 2.18 \end{aligned}$$

Étant donné que $|8.38| > 2.18$, on rejette H_0 au niveau de confiance de 95 %. Autrement dit, l'ordonnée à l'origine est significative.

Question 2

Tester si β_1 est significativement différent de 0 à un taux de confiance de 95 %.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

On utilise la statistique suivante,

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \\ &= \frac{-0.468 - 0}{\sqrt{0.0237}} \\ &= -3.040 \\ t_{\frac{0.05}{2}(14-2)} &= 2.18 \end{aligned}$$

Étant donné que $|-3.040| > 2.18$, on rejette H_0 au niveau de confiance de 95 %. Autrement dit, il y a 96 % de chance que la régression soit utile.

Question 2

Tester si β_1 est significativement différent de 0 à un taux de confiance de 95 %.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

On utilise la statistique suivante,

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \\ &= \frac{-0.468 - 0}{\sqrt{0.0237}} \\ &= -3.040 \\ t_{\frac{0.05}{2}(14-2)} &= 2.18 \end{aligned}$$

Étant donné que $|-3.040| > 2.18$, on rejette H_0 au niveau de confiance de 95 %. Autrement dit, il y a 95 % de chance que la régression soit utile.

Question 3

Tester si β_1 est significativement négatif à un taux de confiance de 95 %.

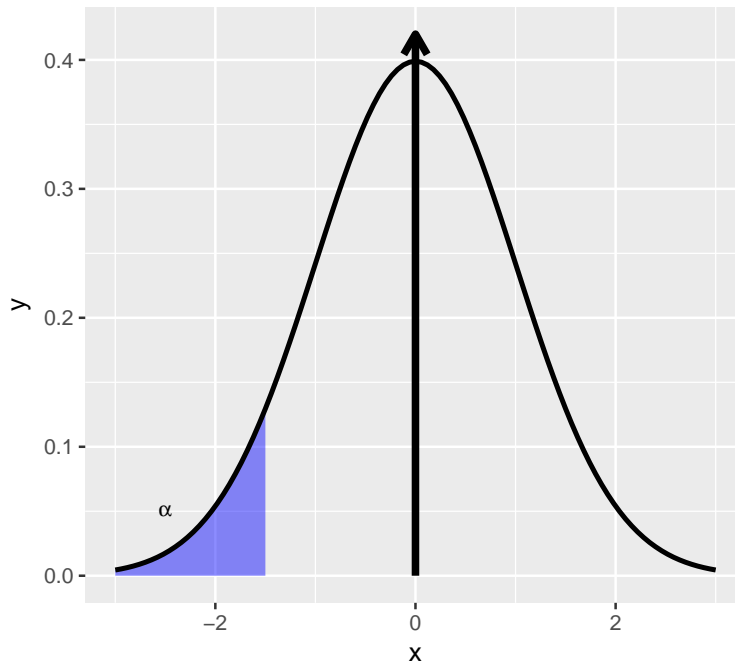
$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 < 0$$

On utilise la statistique suivante,

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \\ &= \frac{-0.468 - 0}{\sqrt{0.0237}} \\ &= -3.040 \\ -t_{\frac{0.05}{2}(14-2)} &= -1.78 \end{aligned}$$

Il s'agit d'un test unilatéral, la zone de rejet est la suivante



Étant donné que $|-3.040| < -1.78$, on rejette H_0 au niveau de confiance de 95 %. Autrement dit, la pente de la droite est significativement négative.

Question 4

Obtenir un I.C. au niveau de confiance de 95 % pour β_0 .

$$\begin{aligned}\beta_0 &\in \hat{\beta}_0 \pm t_{\frac{0.95}{2}}(14-2)\sqrt{\widehat{Var}(\hat{\beta}_0)} \\ &\in 68.494 \pm 2.18 \times \sqrt{66.8511} \\ &\in]50.670, 86.318[\end{aligned}$$

L' I.C. permet de valider le test d'hypothèse de la question 1, car il ne comprend pas la valeur zéro.

Question 5

Obtenir un I.C. au niveau de confiance de 95 % pour β_1 .

$$\begin{aligned}\beta_1 &\in \hat{\beta}_1 \pm t_{\frac{0.95}{2}}(14-2)\sqrt{\widehat{Var}(\hat{\beta}_1)} \\ &\in -0.468 \pm 2.18 \times \sqrt{0.0237} \\ &\in]-0.804, -0.132[\end{aligned}$$

L'I.C. permet de valider le test d'hypothèse de la question 2 et 3, il ne comprend pas la valeur zéro et est strictement négatif.

2.6.5 Test de la validité *globale* de la régression

Une régression linéaire simple est valide, ou significative si $\beta_1 \neq 0$. Le tableau ANOVA obtenue en 2.5.3 peut être utilisé pour tester les hypothèses :

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 < 0$$

avec la statistique de Fisher,

$$\begin{aligned} F &= \frac{MSR}{MSE} \\ &= \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}} \end{aligned}$$

Sous H_0 , on a que $F \sim F(1, n-2)$.

On rejette donc H_0 au niveau $100 \times (1 - \alpha)\%$ si

$$F > F_{\alpha}(1, n-2) \quad (2.23)$$

i

Équivalent

En régression linéaire simple **seulement**, le test F est équivalent au test t pour $\beta_1 = 0$

$$\begin{aligned} F &= \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}} = \frac{SSR}{\sigma^2} = \frac{SSR}{S^2} = \frac{\sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2}{S^2} \\ &= \frac{\sum_{t=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_t - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2}{S^2} = \frac{\hat{\beta}_1^2 \times \sum_{t=1}^n (X_t - \bar{X})^2}{S^2} \\ &= \frac{\hat{\beta}_1^2}{\frac{S^2}{\sum_{t=1}^n (X_t - \bar{X})^2}} \\ &= \frac{(\hat{\beta}_1 - 0)^2}{\widehat{Var}(\hat{\beta}_1)} \\ &= t^2 \end{aligned}$$

On poursuit avec un exemple pour assimiler l'information.

Exemple

Soit le tableau ANOVA suivant :

Source	SS	$d.l.$	MS	F
Régression	48.845	1	48.845	9.249
Erreur	63.374	12	5.281	
Total	112.219	13		

On cherche à vérifier la validité de la régression à l'aide du test F .

On a que $F = 9.249$, par contre $F_{0.05}(1, 12) = 4.75$

Puisque $F > F_{0.05}(1, 12)$; on rejette H_0 . La régression est significative au niveau de confiance de 95 %.

2.7 Prévisions et intervalles de confiance

On peut utiliser la droite de régression pour faire des types de prévisions de Y^* en sachant X^* :

Type 1

Prévision pour la *valeur moyenne* de Y^*

$$E[Y^*] = \beta_0 + \beta_1 X^*$$

Type 2

Prévision pour la *vraie valeur* de Y^*

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

Remarques

1. Dans les deux types, la prévision est le point sur la droite de régression

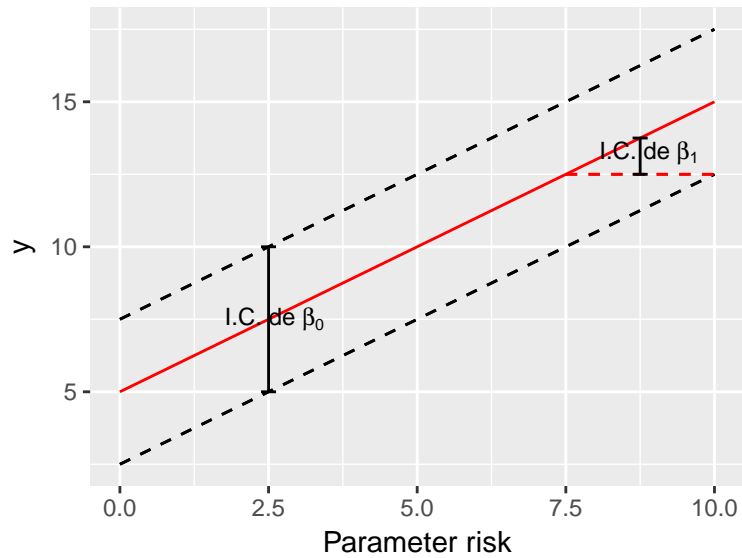
$$\begin{aligned}\widehat{E}[Y^*] &= \hat{Y}^* \\ \hat{Y}^* &= \hat{\beta}_0 + \hat{\beta}_1 X^*\end{aligned}$$

2. La prévision est sans biais

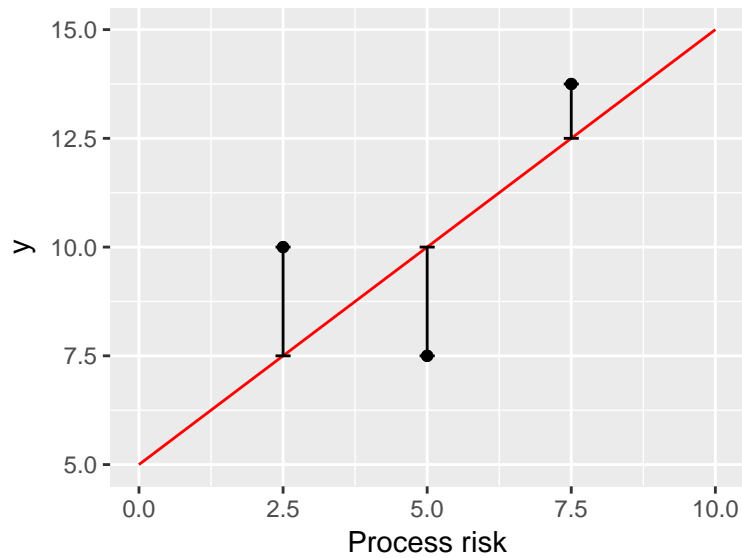
$$\begin{aligned}E[\hat{\beta}_0 + \hat{\beta}_1 X^*] &= E[\hat{\beta}_0] + E[\hat{\beta}_1] X^* \\ &= \beta_0 + \beta_1 X^*\end{aligned}$$

3. Il y a deux sources d'erreur dans les prévisions,

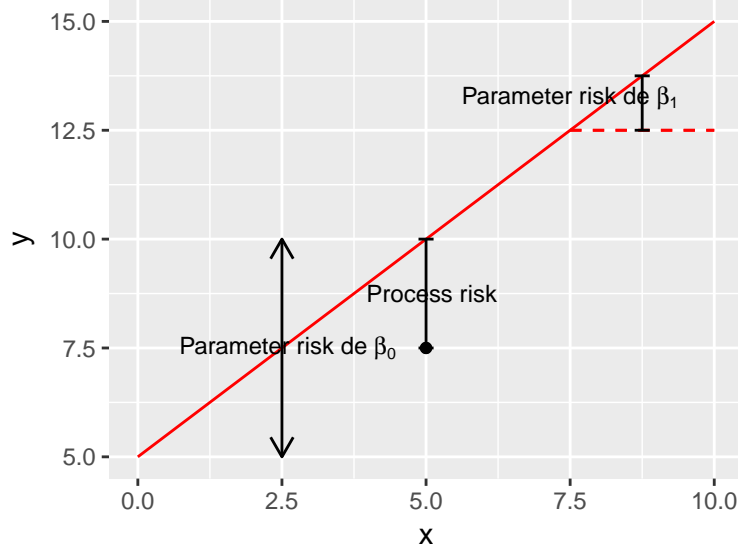
- Parameter risk : Incertitude sur les estimateurs. Autrement dit, la variance des estimateurs des paramètres.



- Process risk : Fluctuations autour de la droite de régression. Autrement dit, la variance des résidus.



Effet combiner des deux sources d'erreur dans les prévisions,



2.7.1 I.C. pour la prévision de type I (Valeur moyenne)

Aussi appelé intervalle de confiance pour la droite de régression.

Tel que vue à la section 2.6.1, on a que

$$(\hat{E}[Y^*] = \hat{\beta}_0 + \hat{\beta}_1 X^*) \sim N(\beta_0 + \beta_1 X^*; Var(\hat{\beta}_0 + \hat{\beta}_1 X^*))$$

Par conséquent,

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 X^*) - (\beta_0 + \beta_1 X^*)}{\sqrt{Var(\hat{\beta}_0 + \hat{\beta}_1 X^*)}} \sim N(0, 1)$$

En substituant σ^2 par S^2 dans la $Var(\hat{\beta}_0 + \hat{\beta}_1 X^*)$; on a

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 X^*) - (\beta_0 + \beta_1 X^*)}{\sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 X^*)}} \sim t(n - 2)$$

Ainsi, un I.C. au niveau $100 \times (1 - \alpha)\%$ pour la valeur moyenne est

$$(\hat{\beta}_0 + \hat{\beta}_1 X^*) \pm t_{\frac{\alpha}{2}}(n - 2) \times \sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 X^*)} \quad (2.24)$$

On rappelle que comme σ^2 n'est souvent pas connu, il est nécessaire d'utiliser son estimateur S^2 .

Or,

$$\begin{aligned}
Var(\hat{\beta}_0 + \hat{\beta}_1 X^*) &= Var(\bar{Y} - \bar{Y} + \hat{\beta}_0 + \hat{\beta}_1 X^*) \\
&= Var(\bar{Y} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) + \hat{\beta}_0 + \hat{\beta}_1 X^*) \\
&= Var(\bar{Y} + \hat{\beta}_1(X^* - \bar{X})) \\
&= Var(\bar{Y}) + Var(\hat{\beta}_1)(X^* - \bar{X})^2 \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} (X^* - \bar{X})^2 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)
\end{aligned}$$

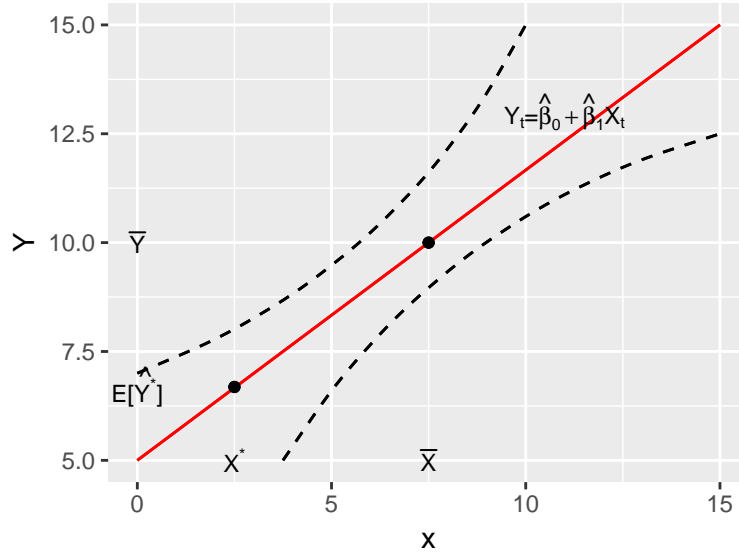
Et ainsi, on obtient,

$$\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 X^*) = S^2 \left(\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right) \quad (2.25)$$

L'I.C. est donc,

$$(\hat{\beta}_0 + \hat{\beta}_1 X^*) \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{S^2 \left(\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{t=1}^n (X_t - \bar{X})^2} \right)} \quad (2.26)$$

Remarque



1. Plus X^* s'éloigne de \bar{X} , plus l'I.C. est large, parce que l'incertitude augmente.
2. Les limites de l'intervalle sont des *hyperboles* centrées en (\bar{X}, \bar{Y})
3. Cet I.C. peut être appelé :
 - I.C. pour la valeur moyenne ;
 - I.C. pour la droite de régression ;
 - I.C. pour la tendance.
4. Dans ce type d' I.C., on tient seulement compte du **risque de paramètre**.

2.7.2 I.C. pour la prévision de type II (Vraie valeur)

Aussi appelé I.C. pour les points de Y^* . Pour obtenir un I.C. pour la vraie valeur de

Y^* , il faut tenir compte du parameter risk ($Var(\hat{\beta}_i)$) ET du process risk ($Var(\varepsilon_t)$). On considère donc de manière équivalente à la section 2.7.1,

$$\frac{Y^* - \hat{Y}^*}{\sqrt{Var(Y^* - \hat{Y}^*)}} \sim N(0, 1)$$

En substituant σ^2 par S^2 dans $Var(Y^* - \hat{Y}^*)$, on a

$$\frac{Y^* - \hat{Y}^*}{\sqrt{\widehat{Var}(Y^* - \hat{Y}^*)}} \sim t(n - 2)$$

Ainsi, un I.C. au niveau $100 \times (1 - \alpha)\%$ pour β_1 pour la vraie valeur de Y^* est,

$$\hat{Y}^* \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\widehat{Var}(Y^* - \hat{Y}^*)}$$

Or par hypothèse on a

$$\begin{aligned} Var(Y^* - \hat{Y}^*) &= Var(Y^*) + Var(\hat{Y}^*) \\ &= \underbrace{\sigma^2}_{\text{Process risk}} + \underbrace{\sigma^2 \left(\frac{1}{n} + \frac{(X^* - \bar{X}^*)^2}{\sum_{t=1}^n (X_t - \bar{X}^*)^2} \right)}_{\text{Parameter risk}} \end{aligned}$$

D'où

$$\widehat{Var}(Y^* - \hat{Y}^*) = S^2 \left(1 + \frac{1}{n} + \frac{(X^* - \bar{X}^*)^2}{\sum_{t=1}^n (X_t - \bar{X}^*)^2} \right) \quad (2.27)$$

L'I.C. est donc,

$$(\hat{\beta}_0 + \hat{\beta}_1 X^*) \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{S^2 \left(1 + \frac{1}{n} + \frac{(X^* - \bar{X}^*)^2}{\sum_{t=1}^n (X_t - \bar{X}^*)^2} \right)} \quad (2.28)$$

Exemple en R

Il est possible d'obtenir le résultat des formules des sections 2.7.1 et 2.7.2.

I.C. de type I pour tous les X dans les observations

	fit	lwr	upr
1	2.6439086	1.6893246	3.5984925
2	0.9232698	0.4385055	1.4080340
3	-0.7324702	-1.2330250	-0.2319153
4	0.8762661	0.4004240	1.3521083
5	1.8000986	1.0990848	2.5011124

I.C. de type I pour un vecteur X^*

	fit	lwr	upr
1	0.04295184	-0.35943041	0.4453341
2	0.24265089	-0.16020308	0.6455049
3	0.44234993	0.02740751	0.8572924
4	0.64204898	0.20436284	1.0797351
5	0.84174803	0.37220874	1.3112873

I.C. de type II pour un vecteur X^*

	fit	lwr	upr
1	0.04295184	-1.5618830	1.647787
2	0.24265089	-1.3623022	1.847604
3	0.44234993	-1.1656801	2.050380
4	0.64204898	-0.9719995	2.256097
5	0.84174803	-0.7812278	2.464724

```
3 > # dataset
4 > x <- rnorm(15)
5 > y <- x + rnorm(15)
6 > xStar <- data.frame(x = seq(0, 2, by = 0.2))
7 > # Modele de regression
8 > fit <- lm(y ~x)
9 > # I.C. de type 1
10 > predict(fit, interval = "confidence") # I.C. pour tous les X
    dans les observations
11 > predict(fit, interval = "confidence", newdata = xStar) # I.C.
    pour un vecteur de X^*
12 > # I.C. de type 2
13 > predict(fit, interval = "prediction", newdata = xStar) # I.C.
    pour un vecteur de x^*
```

Listing 2.4 – Code source en R pour l'exemple

Chapitre 3

Régression multiple

Il n'est pas rare que plus d'une variable soit nécessaire pour expliquer un phénomène. Tel que vue à la section 2.1.2, voici un exemple de modèle de régression multiple :

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \varepsilon$$

Diagram illustrating the variables in the regression model:

- β_0 : Intercept
- β_1 : Coefficient for X_1 (Age du passager)
- β_2 : Coefficient for X_2 (Prix du billet)
- ε : Error term (Erreur aléatoire)

De manière générale, la régression multiple considère le modèle général suivant :

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \dots + \beta_p X_{t,p} + \varepsilon_t, \text{ pour } t = 1, \dots, n$$

- n observations
- p variables exogènes (X_1, \dots, X_p)
- $(p + 1)$ paramètres à estimer ($\beta_0, \beta_1, \dots, \beta_p$)

Quelques éléments d'algèbre matricielle pour les vecteurs et matrices aléatoires

Soient X_1, \dots, X_n des variables aléatoires, on définit le vecteur aléatoire X suivant

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}_{n \times 1}$$

On définit le vecteur espérance de la façon suivante

$$E[X] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix}_{n \times 1}$$

et la matrice de variance-covariance

$$Var(X) = \underbrace{E[(X - E[X])(X - E[X])^T]}_{\text{Produit matriciel}} = \begin{bmatrix} Var(X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \\ Cov(X_n, X_1) & \cdots & Var(X_n) \end{bmatrix}_{n \times 1}$$

Théorème

Soit \mathbb{X} , un vecteur aléatoire et \mathbb{A} une matrice de constantes telle que :

$$\mathbb{X} = \mathbb{X}_{n \times 1} \text{ et } \mathbb{A} = \mathbb{A}_{p \times n}$$

Alors,

$$\begin{aligned} E[\mathbb{A}\mathbb{X}] &= \mathbb{A}E[\mathbb{X}] \\ Var(\mathbb{A}\mathbb{X}) &= \mathbb{A}Var(\mathbb{X})\mathbb{A}^T \end{aligned}$$

Exemple

$$\begin{aligned} \mathbb{A} &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}_{2 \times 1}^T = [1 \quad 1]_{1 \times 2} \\ \mathbb{X} &= \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}_{2 \times 1} \end{aligned}$$

Intuitivement,

$$\begin{aligned} \mathbb{A}\mathbb{X} &= X_1 + X_2 \\ \Rightarrow E[\mathbb{A}\mathbb{X}] &= E[X_1 + X_2] = E[X_1] + E[X_2] \\ \Rightarrow Var(\mathbb{A}\mathbb{X}) &= Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2cov(X_1, X_2) \end{aligned}$$

En calcul matriciel,

$$\begin{aligned}
E[\mathbb{A}\mathbb{X}] &= \mathbb{A}E[\mathbb{X}] \\
&= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} E[X_1] \\ E[X_2] \end{bmatrix} \\
&= E[X_1] + E[X_2] \\
Var(\mathbb{A}\mathbb{X}) &= \mathbb{A}Var(\mathbb{X})\mathbb{A}^\top \\
&= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} (Var(X_1) + Cov(X_1, X_2)) & (Cov(X_1, X_2) + Var(X_2)) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= Var(X_1) + Var(X_2) + 2cov(X_1, X_2)
\end{aligned}$$

3.1 Le modèle sous forme matricielle

À partir du modèle général suivant,

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \dots + \beta_p X_{t,p} + \varepsilon_t, t = 1, \dots, n$$

On représente les n formules suivantes

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{1,1} + \dots + \beta_p X_{1,p} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{2,1} + \dots + \beta_p X_{2,p} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{n,1} + \dots + \beta_p X_{n,p} + \varepsilon_n
\end{aligned}$$

Qu'il est possible de réécrire sous forme matricielle,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} \beta_0 + \beta_1 X_{1,1} + \dots + \beta_p X_{1,p} \\ \beta_0 + \beta_1 X_{2,1} + \dots + \beta_p X_{2,p} \\ \vdots \\ \beta_0 + \beta_1 X_{n,1} + \dots + \beta_p X_{n,p} \end{bmatrix}_{n \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

Ou encore de la façon suivante,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix}_{n \times (p+1)} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

De manière plus compacte, on utilise la notation suivante

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

, avec :

- Y est un vecteur de dimension $n \times 1$ des variables réponses.
- X est une matrice schéma de dimension $n \times (p + 1)$ qui correspond aux variables explicatives.
- β est un vecteur de dimension $(p + 1) \times 1$ des coefficients à estimer.
- ε est un vecteur de dimension $n \times 1$ des erreurs de telle sorte que
 - $E[\varepsilon] = \mathbb{O}_{n \times 1}$, où \mathbb{O} correspond à une matrice nulle.
 - $Var(\varepsilon) = \sigma^2 \mathbb{I}_{n \times n}$, où \mathbb{I} correspond à une matrice identité.

Remarques

1. On suppose que $(\mathbb{X}^\top \mathbb{X})^{-1}$ existe, que \mathbb{X} est de rang complet et que $\left((\mathbb{X}^\top \mathbb{X})^{-1}\right)^\top = (\mathbb{X}^\top \mathbb{X})^{-1}$
2. Pour un modèle de régression linéaire simple, il suffit de définir la matrice schéma de la façon suivante :

$$\mathbb{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}_{n \times 2}$$

3. Pour un modèle passant par l'origine, il n'y a pas de colonne de 1 dans la matrice schéma :

$$\mathbb{X} = \begin{bmatrix} X_{1,1} & \dots & X_{1,p} \\ X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \vdots \\ X_{n,1} & \dots & X_{n,p} \end{bmatrix}_{n \times (p)}$$

4. Pour un modèle du type $Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t$, il ne suffit que de définir la matrice schéma telle que :

$$\mathbb{X} = \begin{bmatrix} 1 & X_1 & \dots & X_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & \dots & X_n^2 \end{bmatrix}_{n \times (p)}$$

$$X_{t,1} = X_t$$

$$X_{t,2} = X_t^2$$

3.1.1 Estimateur des moindres carrés (EMC)

On peut démontrer que l'estimateur $\hat{\beta}$ de β qui minimise la somme résiduelle des carrés correspond à l'équation suivante :

$$\begin{aligned} S(\beta) &= \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 \\ &= \sum_{t=1}^n \varepsilon_t^2 \\ &= \varepsilon^\top \varepsilon \\ &= (\mathbb{Y} - \mathbb{X}\beta)^\top (\mathbb{Y} - \mathbb{X}\beta) \end{aligned}$$

est donné par

$$\boxed{\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}} \quad (3.1)$$

Exemple

On poursuit avec un exemple en régression linéaire simple pour assimiler l'information.

À l'aide des matrices suivantes, déterminer les paramètres de la droite de régression.

$$\mathbb{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \mathbb{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}; \mathbb{X}^\top = \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix}$$

$$\begin{aligned} \mathbb{X}^\top \mathbb{X} &= \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{t=1}^n X_t \\ \sum_{t=1}^n X_t & \sum_{t=1}^n X_t^2 \end{bmatrix} \end{aligned}$$

$$(\mathbb{X}^\top \mathbb{X})^{-1} = \frac{1}{n \sum_{t=1}^n X_t^2 - (n \bar{X})^2} \begin{bmatrix} \sum_{t=1}^n X_t^2 & n \bar{X} \\ n \bar{X} & n \end{bmatrix}$$

$$\begin{aligned}
\mathbb{X}^\top \mathbb{Y} &= \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{t=1}^n Y_t \\ \sum_{t=1}^n X_t Y_t \end{bmatrix} \\
&= \begin{bmatrix} n\bar{Y} \\ \sum_{t=1}^n X_t Y_t \end{bmatrix}
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\hat{\beta} &= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \\
&= (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y} \\
&= \begin{bmatrix} \frac{n\bar{Y} \sum_{t=1}^n X_t^2 - n\bar{X} \sum_{t=1}^n X_t Y_t}{n \sum_{t=1}^n X_t^2 - (n\bar{X})^2} \\ \frac{n \sum_{t=1}^n X_t Y_t - (n\bar{Y})(n\bar{X})}{n \sum_{t=1}^n X_t^2 - (n\bar{X})^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\bar{Y} \sum_{t=1}^n X_t^2 - \bar{X} \sum_{t=1}^n X_t Y_t + n\bar{X}^2 \bar{Y} - n\bar{X}^2 \bar{Y}}{n \sum_{t=1}^n X_t^2 - (n\bar{X})^2} \\ \frac{\sum_{t=1}^n X_t Y_t - n\bar{Y}\bar{X}}{\sum_{t=1}^n X_t^2 - n\bar{X}^2} \end{bmatrix} \\
&= \begin{bmatrix} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 \end{bmatrix}
\end{aligned}$$

Qui corresponde bien aux estimateurs de $\hat{\beta}_0$ (2.8) et $\hat{\beta}_1$ (2.10) trouver précédemment.

Propriétés des estimateurs

1. Sans biais

$$\begin{aligned}
E[\hat{\beta}] &= E[(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}] \\
&\stackrel{3}{=} (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top E[\mathbb{Y}] \\
&= (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top (\mathbb{X}\beta) \\
&= (\mathbb{X}^\top \mathbb{X})^{-1} (\mathbb{X}^\top \mathbb{X})\beta \\
&= \mathbb{I}\beta \\
&= \beta
\end{aligned}$$

2. Variance-covariance

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{Var}((\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}) \\
&\stackrel{3}{=} (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \text{Var}(\mathbb{Y}) [(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top]^\top \\
&= (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \sigma^2 \mathbb{I} \left[\mathbb{X} [(\mathbb{X}^\top \mathbb{X})^{-1}]^\top \right] \\
&\stackrel{1}{=} \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \left[\mathbb{X} [(\mathbb{X}^\top \mathbb{X})^{-1}] \right] \\
&= \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1} (\mathbb{X}^\top \mathbb{X}) (\mathbb{X}^\top \mathbb{X})^{-1} \\
&= \sigma^2 \mathbb{I} (\mathbb{X}^\top \mathbb{X})^{-1} \\
&= \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}
\end{aligned}$$

3.1.2 Résidus et tableau ANOVA

On définit les résidus comme ceci,

$$\begin{aligned}
\varepsilon_{n \times 1} &= \mathbb{Y} - \hat{\mathbb{Y}} \\
&= \mathbb{Y} - \mathbb{X} \hat{\beta} \\
&= \mathbb{Y} - \mathbb{X} (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y} \\
&= \mathbb{Y} (\mathbb{I} - \mathbb{X} (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top) \\
&= \mathbb{Y} (\mathbb{I} - \mathbb{H})
\end{aligned}$$

Où \mathbb{H} correspond à la matrice de projection (Hat matrix).

Les sommes des carrés du tableau ANOVA sont données par les expressions suivantes :

$$\begin{aligned}
\bullet SST &= \sum_{t=1}^n (Y_t - \bar{Y})^2 \\
&= \sum_{t=1}^n Y_t^2 - n \bar{Y}^2 \\
&= \mathbb{Y}^\top \mathbb{Y} - n \bar{Y}^2
\end{aligned}$$

Avec $(n - 1)$ degré de liberté

$$\begin{aligned}
\bullet SSE &= \sum_{t=1}^n (Y_t - \hat{Y})^2 \\
&= \sum_{t=1}^n \varepsilon_t^2 \\
&= \varepsilon^T \varepsilon
\end{aligned}$$

Avec $(n - (p + 1))$ degré de liberté

$$\begin{aligned}
\bullet SSR &= \sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2 \\
&= \sum_{t=1}^n \hat{Y}_t^2 - n\bar{Y}^2 \\
&= \hat{Y}^T \hat{Y} - n\bar{Y}^2
\end{aligned}$$

Avec (p) degré de liberté

Dans le cas de régression multiple, le tableau ANOVA est le suivant :

Source de la variance	Somme des carrés (SS)	Degrés de liberté ($d.l.$)	Carrés moyens (MS)	Ratio de Fisher (F)
Régression	SSR	p	$\frac{SSR}{p}$	$\frac{MSR}{MSE}$
Erreur	SSE	n - (p+1)	$\frac{SSE}{n-(p+1)}$	
Total	SST	n - 1		

3.1.3 Estimateur de σ^2

Dans le cas de la régression multiple, on peut démontrer qu'un bon estimateur sans biais de σ^2 est S^2 sous la forme suivante :

$$\begin{aligned}
S^2 &= MSE \\
&= \frac{SSE}{n - (p + 1)}
\end{aligned}$$

3.1.4 Intervalles de confiance et tests d'hypothèses

Essentiellement, on a la même chose qu'au chapitre 2 pour les tests t et F , sauf qu'il faut adapter les degrés de liberté.

On rappelle qu'avec le postulat 4 (2.2.2), on suppose que les résidus suivent une loi

normale d'espérance nulle et de variance de σ^2 .

$$\begin{aligned}\varepsilon_t &\stackrel{i.i.d.}{\sim} N(0, \sigma^2) \\ \underset{n \times 1}{\varepsilon} &\sim N(\mathbb{O}, \sigma^2 \mathbb{I}_{n \times n})\end{aligned}$$

Ainsi on a que $(\mathbb{Y} = \mathbb{X}\beta + \varepsilon) \sim N_n(\mathbb{X}\beta; \sigma^2 \mathbb{I}_{n \times n})$
et que $\hat{\beta} \sim N_n(\beta; (\mathbb{X}^\top \mathbb{X})^{-1} \sigma^2)$.

3.1.5 Test de Student sur un seul paramètre

On effectue le test suivant,

$$H_0 : \beta_i = \beta_i^*$$

$$H_1 : \beta_i \neq \beta_i^*$$

Où β_i^* est une constante.

On teste l'hypothèse à l'aide de la statistique suivante,

$$t = \frac{\hat{\beta}_i - \beta_i^*}{\sqrt{[Var(\hat{\beta})]_{i+1 \times i+1}}} \sim N(0, 1)$$

et en remplaçant σ^2 par S^2 dans la matrice de la variance, on obtient

$$t = \frac{\hat{\beta}_i - \beta_i^*}{\sqrt{[\widehat{Var}(\hat{\beta})]_{i+1 \times i+1}}} \sim t(n - (p + 1))$$

On rejette H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ pour si :

$$|t| > t_{\frac{\alpha}{2}(n-(p+1))}$$

Or, on a que $Var((\hat{\beta})) = (\mathbb{X}^\top \mathbb{X})^{-1} \sigma^2$

Ainsi

$$\widehat{Var}((\hat{\beta})) = (\mathbb{X}^\top \mathbb{X})^{-1} S^2 \quad (3.2)$$

Avec un peu d'algèbre, on transforme ce test d'hypothèse en un intervalle de confiance pour β_i . L'I.C. marginal est donc le suivant :

$$\hat{\beta}_i \pm t_{\frac{\alpha}{n}}(n - (p + 1)) \times \sqrt{[(\mathbb{X}^\top \mathbb{X})^{-1} S^2]_{i+1 \times i+1}} \quad (3.3)$$

3.1.6 Test de Fisher pour la validité globale de la régression

Dans le cas de la régression multiple, on teste

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{Au moins un coefficient parmi } \beta_1, \dots, \beta_p \text{ est } \neq 0.$$

On teste l'hypothèse à l'aide de la statistique suivante,

$$F = \frac{MSR}{MSE} \sim F(\text{d.l. de SSR, d.l. de SSE})$$

$$F = \frac{MSR}{MSE} \sim F(p, n - (p + 1))$$

On rejette H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ pour si :

$$F > F_\alpha(p, n - (p + 1))$$

Remarque importante

De manière générale, avec p variables explicatives, on a que :

$$F \neq t^2$$

L'égalité ne survient que lorsque $p = 1$ (Voir la section 2.6.5).

3.1.7 Test de Fisher partiel

À la section 3.1.6 on a testé si **tous les** $\beta_1, \beta_2, \dots, \beta_p$ étaient nuls.

Dans cette section, on teste simultanément si certains β_i parmi $\beta_1, \beta_2, \dots, \beta_p$ sont nuls.

On teste donc :

H_0 : Un modèle *réduit*, noté M_0 dont certains $\beta_i = 0$ parmi $\beta_1, \beta_2, \dots, \beta_p$ est acceptable.

H_1 : On doit utiliser le modèle *complet*, noté M_1 avec les p variables.

On utilise la statistique de Fisher partielle suivante,

$$F^* = \frac{\frac{[SSE(M_0) - SSE(M_1)]}{[d.l.(SSE(M_0)) - d.l.(SSE(M_1))]}{\frac{SSE(M_1)}{d.l.(SSE(M_1))}} \quad (3.4)$$

On rejette H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ si :

$$F^* > F_\alpha(d.l.(SSE(M_0)) - d.l.(SSE(M_1)); d.l.(SSE(M_1)))$$

Remarque

Si le modèle réduit de H_0 ne consiste qu'à $\beta_i = 0$, autrement dit un seul paramètre, alors on aura que $F^* = t^2$. Dans ce cas **seulement**, le test Fisher partiel est équivalent au test de Student.

Exemple

On poursuit avec un exemple pour assimiler l'information.

Soit le modèle de régression multiple suivant :

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \beta_3 X_{t,3} + \beta_4 X_{t,4} + \varepsilon_t$$

On teste donc :

$$\begin{aligned} H_0 : \beta_2 = \beta_3 = 0 \\ H_1 : \beta_2 \neq 0 \text{ et/ou } \beta_3 \neq 0 \end{aligned}$$

Afin d'effectuer le test, on effectue les étapes suivantes :

Étape 1 Obtenir le tableau ANOVA pour le modèle sous le modèle complet M_0 .

Extraire $SSE(M_0)$ et $d.l.(SSE(M_0)) \Rightarrow n - 3$

Étape 2 Obtenir le tableau ANOVA pour le modèle sous le modèle complet M_1 .

Extraire $SSE(M_1)$ et $d.l.(SSE(M_1)) \Rightarrow n - 5$

Étape 3 Calculer la valeur de la statistique de Fisher partielle.

$$F^* = \frac{\frac{[SSE(M_0) - SSE(M_1)]}{[(n-3) - (n-5)]}}{\frac{SSE(M_1)}{(n-5)}}$$

Puis rejeter H_0 au niveau de confiance $100 \times (1 - \alpha)\%$ si :

$$F^* > F_\alpha(2; n - 5)$$

3.2 Sélection d'un modèle optimal

Lorsque l'on dispose de plusieurs variables explicatives (X_1, X_2, \dots, X_p) , un modèle optimal est tel que :

1. Pouvoir prédictif **maximal**
2. Avec un nombre de variables **minimal**

En régression, il existe plusieurs algorithmes pour obtenir un modèle optimal.

3.2.1 Technique 1 : Essai de tous les modèles

La stratégie la plus simple consiste à examiner tous les modèles possibles, soit les 2^p combinaisons existantes.

On choisit le modèle ayant le plus grand R_{adj}^2 , qui correspond à l'une des expressions suivantes :

$$R_{adj}^2 = 1 - \frac{\frac{SSE}{(n-p-1)}}{\frac{SST}{(n-1)}} \quad (3.5)$$

$$R_{adj}^2 = 1 - (1 - R^2) \left(\frac{n-1}{n-p-1} \right) \quad (3.6)$$

On note que contrairement au R^2 , le R_{adj}^2 pénalise pour l'ajout de variables dans le modèle.

Exemple

Si on dispose de X_1, X_2 et X_3 , on ajuste les 2^3 modèles possibles :

1. $Y = \beta_0 + \varepsilon$
2. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
3. $Y = \beta_0 + \beta_1 X_2 + \varepsilon$
4. $Y = \beta_0 + \beta_1 X_3 + \varepsilon$
5. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
6. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$
7. $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$
8. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

On fait le calcul du R_{adj}^2 pour chaque modèle et on choisit le modèle avec le plus grand R_{adj}^2 .

Remarque

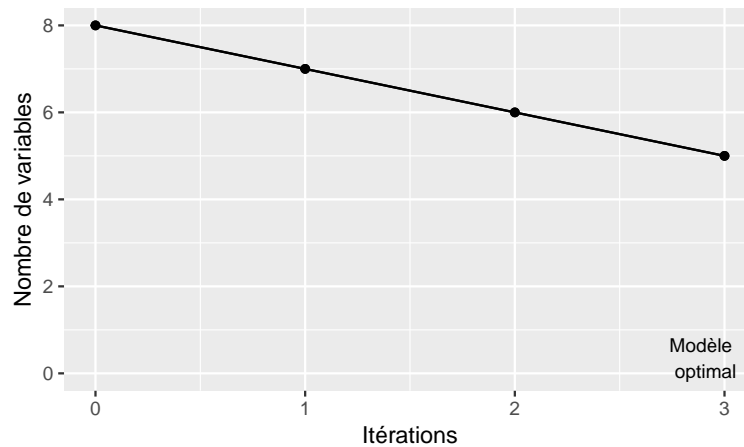
En pratique cette méthode n'est pas *efficace*, car le temps d'exécution devient énorme lorsque p augmente :

p	2^p
1	2
2	4
3	8
\vdots	
10	1024
\vdots	
25	33554432
\vdots	
100	1.26×10^{30}

3.2.2 Technique 2 : Élimination régressive (*Backward elimination*)

- Étape 1 Débuter avec toutes les variables disponibles dans le modèle.
Étape 2 Chercher la variable qui génère la plus faible augmentation de SSE lorsqu'exclue du modèle. Autrement dit, la pire variable.
Étape 3 Utiliser les test Fisher partiels (3.1.7) pour tester s'il est possible d'exclure la variable de l'étape 2.
Étape 4 Continuer les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de variables à éliminer selon les tests Fisher partiels.

Voici une illustrations de l'élimination régressive.



Remarque

Le principal inconvénient de cette technique est qu'une variable éliminée ne peut jamais être réintégrée.

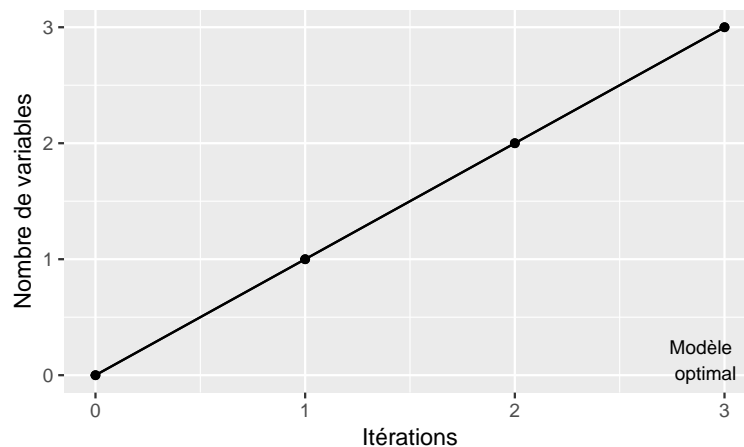
3.2.3 Technique 3 : Sélection progressive (*forward selection*)

Étape 1 Débuter avec le modèle $Y = \beta_0 + \varepsilon$

Étape 2 Chercher la variable qui génère la plus grande diminution de SSE lorsqu'incluse dans le modèle. Autrement dit, la meilleure variable.

Étape 3 Utiliser les test Fisher partiels (3.1.7) pour tester s'il est possible d'inclure la variable de l'étape 2.

Étape 4 Continuer les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de variables à inclure selon les tests Fisher partiels.



Remarque

Le principal inconvénient de cette technique est qu'une variable incluse ne peut jamais être éliminée par la suite.

3.2.4 Technique 4 : Régression pas à pas (*stepwise regression*)

Il s'agit d'une combinaison de l'élimination régressive et de la sélection progressive.

Étape 1 Débuter avec le modèle $Y = \beta_0 + \varepsilon$

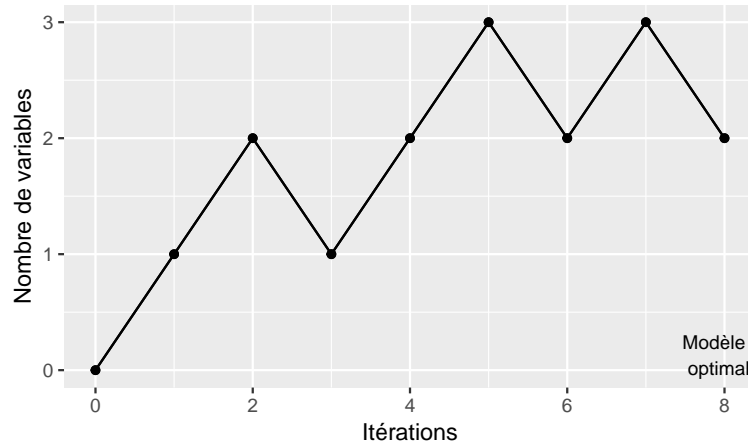
Étape 2 Chercher la variable qui génère la plus grande diminution de SSE si incluse dans le modèle. Autrement dit, la meilleure variable.

Étape 3 Utiliser les test Fisher partiels (3.1.7) pour tester s'il est possible d'inclure la variable de l'étape 2.

Étape 4 Chercher la variable qui génère la plus faible diminution de SSE si incluse dans le modèle. Autrement dit, la pire variable.

Étape 5 Utiliser les test Fisher partiels (3.1.7) pour tester s'il est possible d'exclure la variable de l'étape 4.

Étape 6 Continuer les étapes 2 à 5 jusqu'à ce que l'algorithme élimine la variable qui vient d'entrer.



Exemple

On poursuit avec un exemple pour assimiler l'information.

À partir des informations suivantes on cherche à trouver le meilleur modèle de régression des 20 observations.

Variables dans le modèle	<i>SSE</i>	<i>SSR</i>	<i>SST</i>	R^2_{adj}
ϕ	10	0	10	0 %
X_1	5	5	10	47.2 %
X_2	9	1	10	5 %
X_3	8	2	10	15.5 %
X_1, X_2	4	6	10	55.3 %
X_1, X_3	3.9	6.1	10	56.4 %
X_2, X_3	8.5	1.5	10	5 %
X_1, X_2, X_3	3.8	6.2	10	54.9 %

Technique 1

Avec la technique 1 (3.2.1), on trouve le modèle suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Technique 2

Avec la technique 2 (3.2.2), on débute avec le modèle initial suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

La pire variable est X_2 , on effectue le test de Fisher partiel (3.4) avec modèle M_0 sans la variable X_2 et M_1 avec le modèle complet.

H_0 : Modèle avec X_1 et X_3

H_1 : Modèle avec X_1, X_2 et X_3

$$F = \frac{\frac{3.9-3.8}{1}}{\frac{3.8}{16}}$$

$$= 0.4211$$

$$F_{5\%}(1.16) = 4.49$$

$$0.4211 < 4.49$$

On accepte H_0 et on exclut X_2 .

La prochaine pire variable est la variable X_1 .

H_0 : Modèle avec X_3

H_1 : Modèle avec X_1 et X_3

$$F = \frac{\frac{5-3.9}{1}}{\frac{3.9}{17}}$$

$$= 4.79$$

$$F_{5\%}(1.17) = 4.49$$

$$4.79 > 4.45$$

On rejette H_0 et on n'exclut pas X_3 .

On trouve le modèle suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$$

Exemple en R

À l'aide du jeu de donnée *mtcars*¹ de *R*, construis un modèle pour prédire la consommation en gallon par miles à l'aide de la technique de régression pas à pas.

Start: AIC=65.47

```
mpg ~ wt + drat + disp + qsec + hp
```

1. Voici les [informations](#) sur les données.

	Df	Sum of Sq	RSS	AIC
- disp 1	1	3.974	174.10	64.205
<none>			170.13	65.466
- hp 1	1	11.886	182.01	65.627
- qsec 1	1	12.708	182.84	65.772
- drat 1	1	15.506	185.63	66.258
- wt 1	1	81.394	251.52	75.978

Step: AIC=64.21

mpg ~ wt + drat + qsec + hp

	Df	Sum of Sq	RSS	AIC
- hp 1	1	9.418	183.52	63.891
- qsec 1	1	9.578	183.68	63.919
<none>			174.10	64.205
- drat 1	1	11.956	186.06	64.331
+ disp 1	1	3.974	170.13	65.466
- wt 1	1	113.882	287.99	78.310

Step: AIC=63.89

mpg ~ wt + drat + qsec

	Df	Sum of Sq	RSS	AIC
<none>			183.52	63.891
- drat 1	1	11.942	195.46	63.908
+ hp 1	1	9.418	174.10	64.205
+ disp 1	1	1.506	182.02	65.627
- qsec 1	1	85.720	269.24	74.156
- wt 1	1	275.686	459.21	91.241

Call:

lm(formula = mpg ~ wt + drat + qsec, data = mtcars)

Coefficients:

(Intercept)	wt	drat	qsec
11.3945	-4.3978	1.6561	0.9462

On remarque que la variable *disp* et *hp* n'ont pas été retenues dans le modèle.

```
3 > step(lm(mpg~wt+drat+disp+qsec+hp, data=mtcars), direction="both")
```

Listing 3.1 – Code source en R pour l'exemple



Technique de sélection & R

Toutes les techniques présentées à la section 3.2 sont intégrées dans le système de base de *R*. Cette vignette sur la sélection des variables comprends les différentes méthodes ainsi que des exemples utiles.

3.3 Régression avec variables indicatrices

Permettent de traiter des variables explicatives *catégoriques* dans les modèles.

Exemples

- Couleur des yeux (bleu, brun, vert et autres)
- Type de véhicule (sport et autres)
- emploi (ACT, ETUm RTR, GOU et autres)

Pour inclure une variable catégorique ayant r valeurs possibles, on doit créer $(r - 1)$ variables indicatrices.

Exemple

- Couleur des yeux :

$$X_{t,1} = 1_{\{Couleur_t=Bleu\}}$$

$$X_{t,2} = 1_{\{Couleur_t=Brun\}}$$

$$X_{t,3} = 1_{\{Couleur_t=Vert\}}$$

- Type de véhicule :

$$X_{t,4} = 1_{\{Type_t=Sport\}}$$

- Emploi :

$$X_{t,5} = 1_{\{Emploi_t=ACT\}}$$

$$X_{t,6} = 1_{\{Emploi_t=ETU\}}$$

$$X_{t,7} = 1_{\{Emploi_t=RTR\}}$$

$$X_{t,8} = 1_{\{Emploi_t=GOU\}}$$

; où

$$1_{\{A\}} = \begin{cases} 1 & , \text{si } A \text{ vrai} \\ 0 & , \text{sinon} \end{cases}$$

Exemple

À partir des 5 observations suivantes, définir la matrice des variables réponses et la matrice schéma.

Y_t	$Couleur_t$	$Type_t$	$Emploi_t$
70	Bleu	Autres	ETU
75	Brun	Sport	GOU
50	Vert	Autres	Autres
55	Autres	Autres	Autres
85	Brun	Sport	ACT

On utilise le modèle de régression multiple à partir du modèle d'indicatrice précédent. On obtient le modèle de régression suivant :

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \beta_3 X_{t,3} + \beta_4 X_{t,4} + \beta_5 X_{t,5} + \beta_6 X_{t,6} + \beta_7 X_{t,7} + \beta_8 X_{t,8} + \varepsilon_t$$

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

La matrice des variables réponses correspond à :

$$\mathbb{Y} = \begin{bmatrix} 70 \\ 75 \\ 50 \\ 55 \\ 85 \end{bmatrix}$$

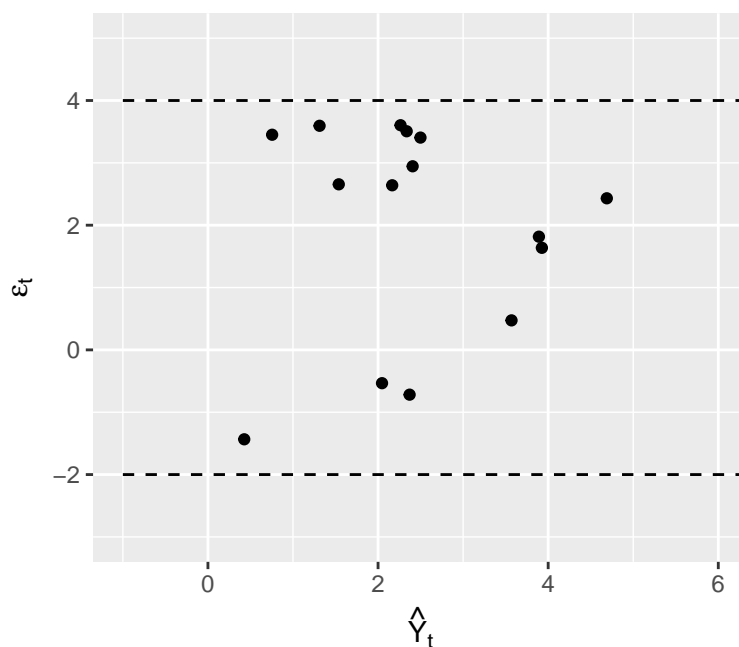
La matrice schéma correspond à :

$$\mathbb{X} = \begin{matrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

3.4 Analyse qualitative des résidus

Même si les tests t et F sont concluents, le modèle choisi peut ne pas être adéquat. En effet, l'analyse qualitative des résidus est la principale façon de valider un modèle sélectionné.

Distribution *uniforme*



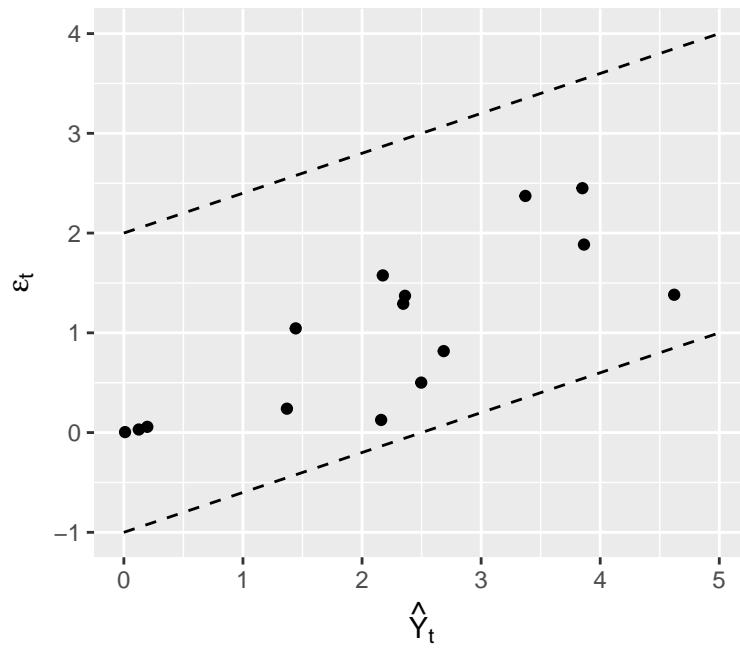
On remarque que les résidus sont uniformément distribués autour de l'axe des x . Il s'agit d'une situation idéale.

3.4.1 Problèmes possibles dans la distribution des résidus

Plusieurs problèmes de distribution des résidus peuvent être observés, voici leurs représentations graphiques et leurs possibles significations :

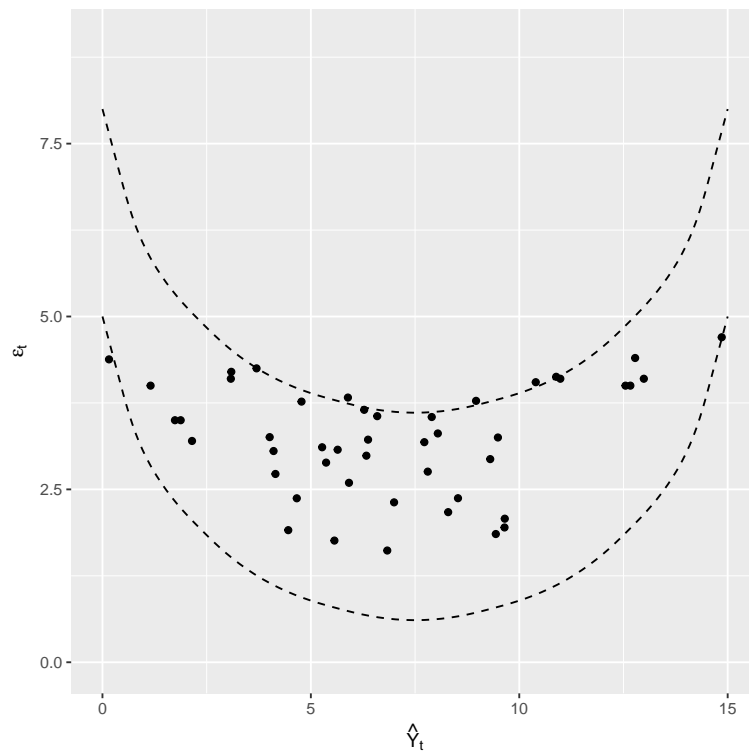
3.4.1.1 Distribution *uniforme* avec rotation

Cette distribution des résidus est très similaire à la distribution *uniforme*, par contre les résidus ne sont pas distribués autour de l'axe des x . La distribution semble avoir effectué une rotation. Il manque probablement un terme linéaire dans X .



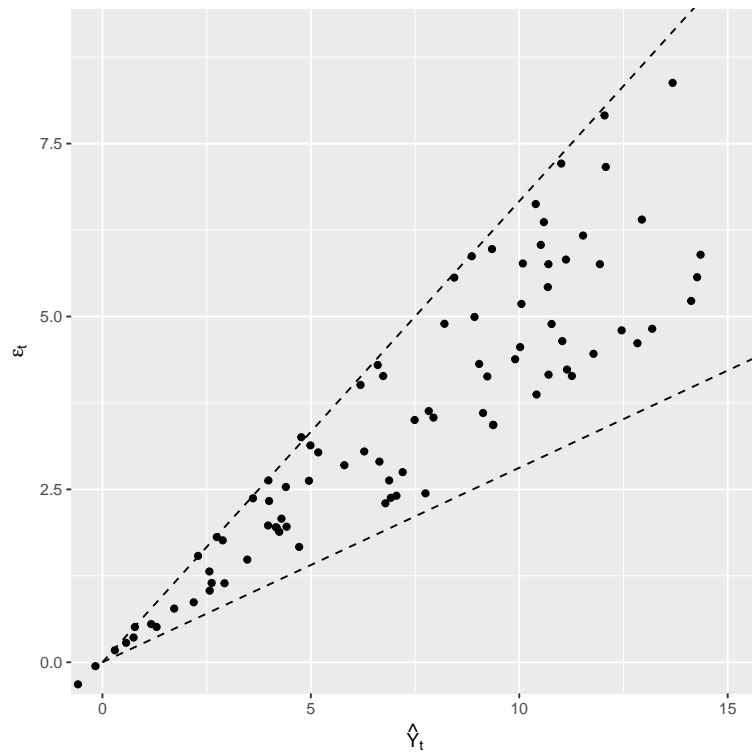
3.4.1.2 Distribution *quadratique*

Cette distribution des résidus semble suivre une distribution quadratique. Il manque probablement une variable quadratique dans X .



3.4.1.3 Distribution *conique*

Cette distribution des résidus semble suivre être distribuer dans un cone. La variance n'est probablement pas constante. Il y a violation de l'hypothèse 2.

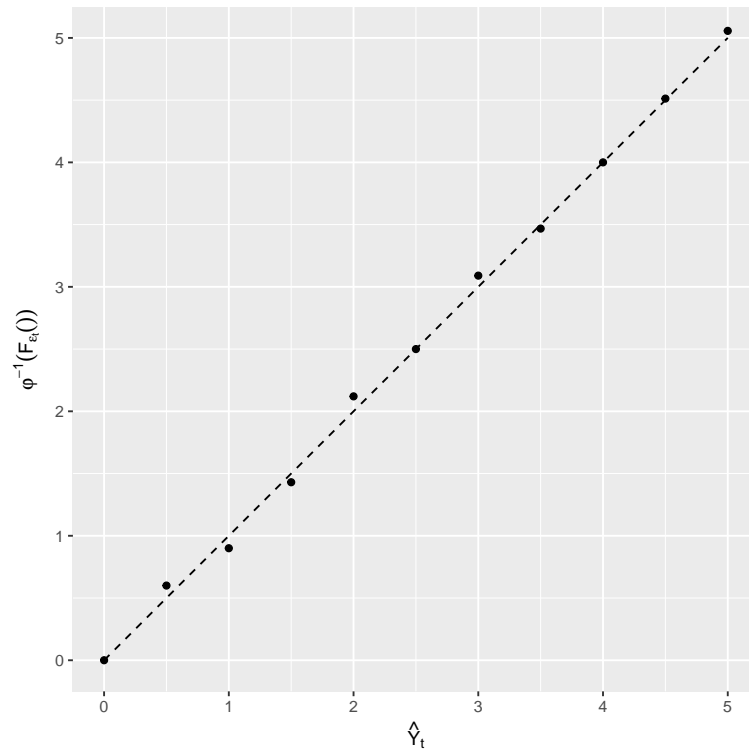


3.4.2 Quantiles normaux

On appelle parfois le diagramme Quantile-Quantile ou Q-Q plot. Il s'agit d'un outil permettant d'évaluer la pertinence de l'ajustement d'une distribution.

On ajuste selon une loi normale la fonction empirique des résidus par rapport aux résidus.

On cherche à avoir une droite à 45° . Dans cette situation, cela signifie $\varepsilon_t \sim N(0, 1)$.



3.4.3 Exemple complet

On poursuit avec un exemple complet pour synthétiser l'information du chapitre 3.

On reprend le scénario du Titanic de la section 2.1.2, cette fois on va utiliser un modèle complet avec des catégories et déterminer le meilleur modèle avec la technique de régression pas à pas².

Tout d'abord, voici la signification des variables :

2. Le code source complet de l'exemple est disponible à l'annexe [A](#)

Variable	Définition
Survival	Survie du passager au naufrage du Titanic
Pclass	Catégorie du billet
Sex	Sexe
Age	Âge
sibsp	Nombre se frères et sœurs / époux à bord du Titanic
parch	Nombre de parents / enfants à bord du Titanic
ticket	Numéro du billet
fare	Prix du billet
cabin	Numéro de la cabine
embarked	Port d'embarquement

L'étape suivante consiste à analyser notre jeu de donnée et de retirer les variables inutiles :

PassengerId	Survived	Pclass
Min. : 1.0	Min. :0.0000	Min. :1.000
1st Qu.:223.5	1st Qu.:0.0000	1st Qu.:2.000
Median :446.0	Median :0.0000	Median :3.000
Mean :446.0	Mean :0.3838	Mean :2.309
3rd Qu.:668.5	3rd Qu.:1.0000	3rd Qu.:3.000
Max. :891.0	Max. :1.0000	Max. :3.000

	Name	Sex	Age
Abbing, Mr. Anthony	: 1	female:314	Min. : 0.42
Abbott, Mr. Rossmore Edward	: 1	male :577	1st Qu.:20.12
Abbott, Mrs. Stanton (Rosa Hunt)	: 1		Median :28.00
Abelson, Mr. Samuel	: 1		Mean :29.70
Abelson, Mrs. Samuel (Hannah Wozosky)	: 1		3rd Qu.:38.00
Adahl, Mr. Mauritz Nils Martin	: 1		Max. :80.00
(Other)	:885		NA's :177

SibSp	Parch	Ticket	Fare
Min. :0.000	Min. :0.0000	1601 : 7	Min. : 0.00
1st Qu.:0.000	1st Qu.:0.0000	347082 : 7	1st Qu.: 7.91
Median :0.000	Median :0.0000	CA. 2343: 7	Median : 14.45
Mean :0.523	Mean :0.3816	3101295 : 6	Mean : 32.20
3rd Qu.:1.000	3rd Qu.:0.0000	347088 : 6	3rd Qu.: 31.00
Max. :8.000	Max. :6.0000	CA 2144 : 6	Max. :512.33
		(Other) :852	

Cabin	Embarked
:687	: 2
B96 B98 : 4	C:168
C23 C25 C27: 4	Q: 77

```
G6      : 4   S:644
C22 C26 : 3
D       : 3
(Other) :186
```

Certaines variables ne sont d'aucun intérêt pour estimer l'âge d'un passager, tel que :

- *PassengerId*, car il s'agit d'un numéro unique pour chaque passager. (Max. : 891 pour 891 observations)
- *Name*, dans son état actuel le nom du passager n'est pas très utile car il s'agit d'observation unique.
- *Ticket*, car il s'agit d'un numéro unique pour chaque passager.
- *SibSP*, *Parch* et *Cabin* sont retirés pour des fins de simplification.

On cherche maintenant à transformer une donnée pour en tirer de l'information.

On observe que le nom du passager est unique dans son format actuel, mais avec un peu de manipulation de donnée, il est très facile d'extraire son nom de famille.

```
> # Visualiser les 6 premières observations par catégorie
> head(summary(data$Surname))
```

```
Andersson      Sage      Carter      Goodwin      Johnson      Panula
           9           7           6           6           6           6
```

On obtient ainsi le modèle complet suivant :

$$\hat{\text{Age}}_t = \beta_0 + \beta_1 \times \text{Survived}_t + \beta_2 \times \text{Pclass}_t + \beta_3 \times \text{Sex}_t + \beta_4 \times \text{Fare}_t + \beta_5 \times \text{Embarked}_t + \beta_6 \times \text{Surname}_t + \varepsilon_t$$

On peut maintenant trouver le meilleur modèle,

```
> fit <- step(lm(Age ~ Survived + Pclass + Sex + Fare + Embarked + Surname, data),
+             direction = "both")
```

Start: AIC=3796.05

Age ~ Survived + Pclass + Sex + Fare + Embarked + Surname

	Df	Sum of Sq	RSS	AIC
- Surname	533	87709	119664	3672.8
- Pclass	1	3	31958	3794.1
- Fare	1	43	31997	3795.0
<none>			31955	3796.0
- Embarked	2	336	32291	3799.5
- Sex	1	379	32334	3802.5
- Survived	1	2616	34571	3850.2

Step: AIC=3672.79

Age ~ Survived + Pclass + Sex + Fare + Embarked

	Df	Sum of Sq	RSS	AIC
- Embarked	3	664	120328	3670.7
- Sex	1	187	119851	3671.9
<none>			119664	3672.8
- Fare	1	1574	121238	3680.1
- Survived	1	4062	123726	3694.6
+ Surname	533	87709	31955	3796.0
- Pclass	1	25962	145626	3811.0

Step: AIC=3670.74

Age ~ Survived + Pclass + Sex + Fare

	Df	Sum of Sq	RSS	AIC
- Sex	1	145	120473	3669.6
<none>			120328	3670.7
+ Embarked	3	664	119664	3672.8
- Fare	1	1747	122075	3679.0
- Survived	1	4200	124528	3693.2
+ Surname	534	88037	32291	3799.5
- Pclass	1	26337	146664	3810.1

Step: AIC=3669.61

Age ~ Survived + Pclass + Fare

	Df	Sum of Sq	RSS	AIC
<none>			120473	3669.6
+ Sex	1	145	120328	3670.7
+ Embarked	3	623	119851	3671.9
- Fare	1	1850	122323	3678.5
- Survived	1	6913	127386	3707.4
+ Surname	534	87812	32662	3805.7
- Pclass	1	26875	147349	3811.4

> fit

Call:

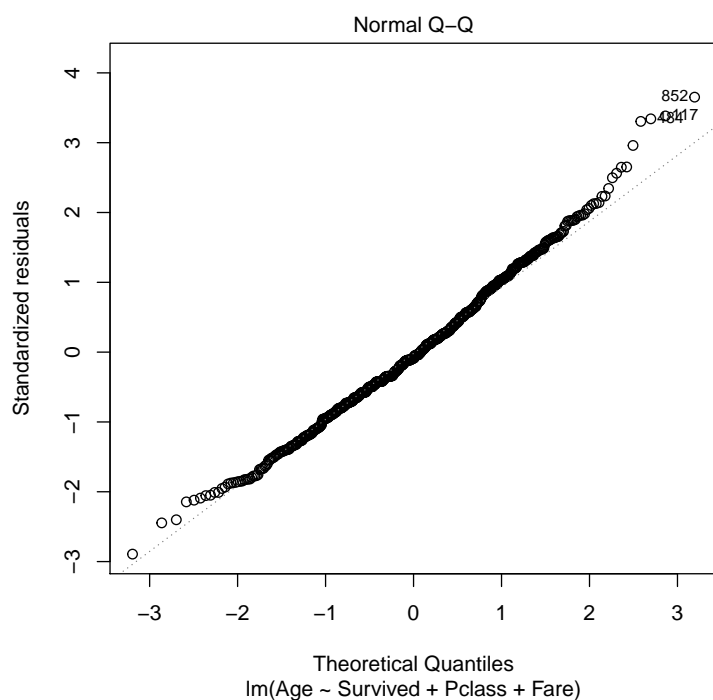
lm(formula = Age ~ Survived + Pclass + Fare, data = data)

Coefficients:

(Intercept)	Survived	Pclass	Fare
54.14124	-6.81709	-9.12040	-0.03671

On observe que le Q-Q plot suivant pour le modèle

```
> # Q-Q plot
> plot(fit, which=2)
```



Notre modèle prédit assez bien l'âge, par contre pour des valeurs extrêmes notre prédiction est moins précise.

On va maintenant tester une variable modifier, le log-Fare. En appliquant un log à une valeur numérique, on réduit l'échelle d'écart entre les variables et certaines peuvent mieux prédire après avoir être transformer.

Start: AIC=3796.92

Age ~ Survived + Pclass + Sex + Fare + Embarked + Surname + LogFare

	Df	Sum of Sq	RSS	AIC
- Surname	533	84253	116157	3653.6
- Pclass	1	6	31910	3795.0
- LogFare	1	51	31955	3796.0
<none>			31904	3796.9
- Fare	1	93	31997	3797.0
- Embarked	2	374	32278	3801.2

```
- Sex      1      409  32313 3804.0
- Survived 1      2614  34518 3851.1
```

Step: AIC=3653.56

Age ~ Survived + Pclass + Sex + Fare + Embarked + LogFare

```
      Df Sum of Sq    RSS    AIC
- Sex      1         6 116163 3651.6
- Embarked  3        682 116839 3651.7
- Fare      1       116 116274 3652.3
<none>                        116157 3653.6
- LogFare   1      3507 119664 3672.8
- Survived  1      4214 120372 3677.0
+ Surname  533     84253  31904 3796.9
- Pclass    1     28085 144242 3806.2
```

Step: AIC=3651.59

Age ~ Survived + Pclass + Fare + Embarked + LogFare

```
      Df Sum of Sq    RSS    AIC
- Embarked  3        677 116840 3649.7
- Fare      1       120 116283 3650.3
<none>                        116163 3651.6
+ Sex      1         6 116157 3653.6
- LogFare   1      3687 119851 3671.9
- Survived  1      5931 122094 3685.1
+ Surname  533     83850  32313 3804.0
- Pclass    1     29036 145199 3808.9
```

Step: AIC=3649.74

Age ~ Survived + Pclass + Fare + LogFare

```
      Df Sum of Sq    RSS    AIC
- Fare      1        99 116939 3648.3
<none>                        116840 3649.7
+ Embarked  3        677 116163 3651.6
+ Sex      1         1 116839 3651.7
- LogFare   1      3633 120473 3669.6
- Survived  1      5953 122794 3683.2
- Pclass    1     29179 146019 3806.9
+ Surname  534     84180  32660 3807.6
```

Step: AIC=3648.35

Age ~ Survived + Pclass + LogFare

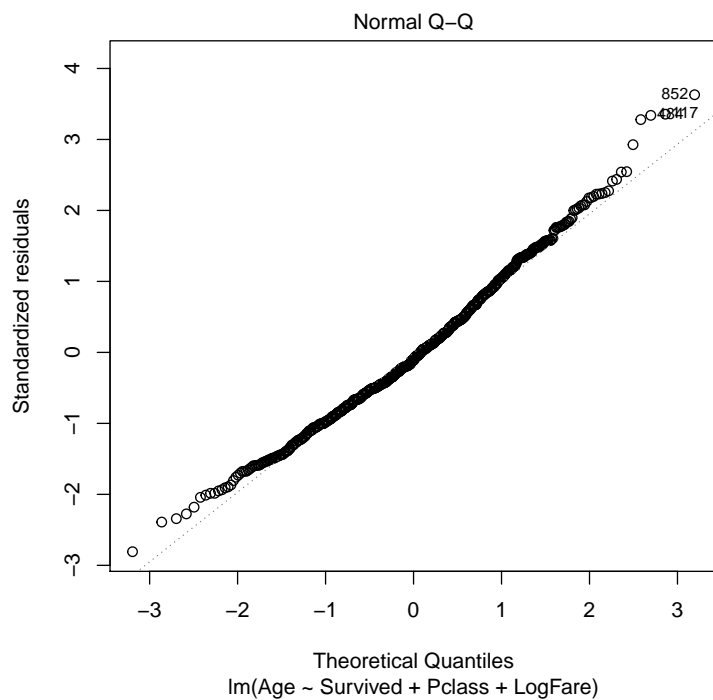
	Df	Sum of Sq	RSS	AIC
<none>			116939	3648.3
+ Fare	1	99	116840	3649.7
+ Embarked	3	656	116283	3650.3
+ Sex	1	2	116937	3650.3
- LogFare	1	5384	122323	3678.5
- Survived	1	5953	122892	3681.8
- Pclass	1	29081	146021	3804.9
+ Surname	534	84275	32664	3805.7

Call:

```
lm(formula = Age ~ Survived + Pclass + LogFare, data = data)
```

Coefficients:

(Intercept)	Survived	Pclass	LogFare
69.419	-6.354	-11.201	-4.060



On obtient une légère amélioration de notre modèle et on retient ce modèle.

On cherche maintenant à prédire à partir d'un autre échantillon des données.

```
> head(predict(fitL, dataTest))
```

1	2	3	4	5	6
21.10578	21.56033	37.79596	27.04870	19.27570	20.43965

i

Segmentation des données

En analyse des données, il est primordial de fragmenter les données. On utilise habituellement réduit l'algorithme suivant :

1. 80 % des données pour l'entraînement (training) et 20 % pour le test (testing) ;
2. 80 % des données de test pour l'entraînement (training) et 20 % pour la validation (validation).

On segmente les données afin d'éviter le surapprentissage (overfitting). Il s'agit de la situation ou toutes les situations possibles sont incluses dans le modèle et celui-ci perd de la qualité prédictive. Dans le cadre du cours, seulement l'étape 1 est suffisante.

Des méthodes plus [élaborée](#) et complexe existent pour les données massives et l'apprentissage automatique.

Chapitre 4

Les modèles linéaires généralisés

4.1 Introduction

Le modèle de régression linéaire multiple étudié lors des derniers chapitres peut parfois avoir certaines limitations :

- On suppose une distribution normale. Dans la plupart des contextes, cette distribution est inappropriée car elle permet des valeurs négatives. On comprend qu'en actuariat, cette situation n'est pas désirable.
- Hypothèse contraignante de variance constante.
- Le domaine des variables réponses permet des valeurs entre $-\infty$ et ∞ . Plusieurs contextes ne se retrouvent que dans un domaine non négatifs. De plus, certaines situations pourraient être une variable réponse discrète.

Le modèle linéaire généralisé, parfois appelé GLM pour *Generalized Linear Models*, est une généralisation de la régression linéaire multiple dont l'objectif est de palier aux limitations précédentes.

i

Reformulation

On peut voir le GLM comme une généralisation souple de la régression linéaire. Cette généralisation de la régression linéaire permet au modèle linéaire d'être relié aux variables réponses via une fonction de lien. De plus, le GLM autorise l'amplitude de la variance de chaque mesure en étant une fonction de sa valeur prévue.

4.2 Notions préliminaires : La famille exponentielle

De manière générale, une variable aléatoire y obéit à une distribution faisant partie de la famille exponentielle si :

$$f_Y(y) = \exp \left\{ \frac{y \times \theta - b(\theta)}{a(\phi)} - c(y, \phi) \right\}$$

; où

- θ : Paramètre canonique
- ϕ : Paramètre de dispersion
- $a(\phi)$, $b(\theta)$ et $c(y, \phi)$: 3 fonctions générales de y , θ et ϕ .

4.2.1 Loi Normale

En posant,

- $\theta = \mu$
 - $\phi = \sigma^2$
 - $b(\theta) = \frac{\theta^2}{2}$
 - $a(\phi) = \phi$
 - $c(y, \phi) = \frac{-1}{2} \left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right)$
- alors,

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2} \times \left(\frac{y-\mu}{\sigma} \right)^2}$$

et

$$Y \sim N(\mu, \sigma^2)$$

Preuve

$$\begin{aligned} f_Y(y) &= \exp \left\{ \frac{y \times \mu - \frac{\mu^2}{2}}{\sigma^2} + \frac{-1}{2} \left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) \right\} \\ &= \exp \left\{ \frac{-y^2}{2\sigma^2} + \frac{y \times \mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right\} \\ &= \exp \left\{ \frac{-1}{2\sigma^2} (y - \mu)^2 - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2} \times \left(\frac{y-\mu}{\sigma} \right)^2} \end{aligned}$$

4.2.2 Loi Poisson

En posant,

- $\theta = \ln(\mu)$
 - $\phi = 1$
 - $b(\theta) = e^\theta$
 - $a(\phi) = \phi$
 - $c(y, \phi) = -\ln(y!)$
- alors,

$$f_Y(y) = \frac{e^{-\mu} \mu^y}{y!}$$

et

$$Y \sim \text{Poisson}(\mu)$$

Preuve

$$\begin{aligned} f_Y(y) &= \exp \left\{ \frac{y \ln(\mu) - e^{\ln(\mu)}}{1} + (-1) \ln(y!) \right\} \\ &= \exp \left\{ y \ln(\mu) - \mu + \ln\left(\frac{1}{y!}\right) \right\} \\ &= \frac{e^{-\mu} \mu^y}{y!} \end{aligned}$$

4.2.3 Loi Bernouilli

En posant,

- $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
 - $\phi = 1$
 - $b(\theta) = \ln(1 + e^\theta)$
 - $a(\phi) = \phi$
 - $c(y, \phi) = \emptyset$
- alors,

$$f_Y(y) = \pi^y \times (1 - \pi)^{1-y}$$

et

$$Y \sim \text{Bern}(\pi)$$

Preuve

$$\begin{aligned} f_Y(y) &= \exp \left\{ \frac{y \times \ln \left(\frac{\pi}{(1-\pi)} \right) - \ln \left(1 + e^{\ln \left(\frac{\pi}{(1-\pi)} \right)} \right)}{1} + 0 \right\} \\ &= \exp \left\{ y \times \ln(\pi) - y \ln(1-\pi) + \ln(1-\pi) \right\} \\ &= \exp \left\{ y \times \ln(\pi) + (1-y) \ln(1-\pi) \right\} \\ &= \pi^y \times (1-\pi)^{1-y} \end{aligned}$$

4.2.4 Autres lois

- Loi beta
- Loi binomiale
- Loi gamma
- Loi inverse-Gaussienne
- Loi binomiale négative
- Loi pareto
- Loi weibull

sont aussi des distributions qui appartiennent à la famille exponentielle.

4.3 Généralités sur les modèles de régression avec la famille exponentielle

4.3.1 Contexte

Le contexte est très similaire à celui de la régression multiple :

4.3.2 Autres lois

- $\mathbb{Y}_{n \times 1}$: Vecteur des observations de Y_i , où $i = 1, \dots, n$.
- $\mathbb{X}_{n \times (p+1)}$: Matrice schéma contenant n lignes d'observations et $(p+1)$ colonnes de variables explicatives.
- $\beta_{(p+1) \times 1}$: Vecteur des $(p+1)$ paramètres $\beta_0, \beta_1, \dots, \beta_p$ à estimer.

4.3.3 Structure du modèle

On suppose maintenant que

$$(Y_i | X_i) \sim \text{Famille exponentielle}$$

et que

$$\begin{aligned}g(E[Y_i|X_i]) &= X_i\beta \\ &= \beta_0 + \beta_1 X_{i,1} + \dots + \beta_p X_{i,p}\end{aligned}$$

où $g(\bullet)$ est une fonction continue appelée une **fonction de lien**.

4.3.4 Propriété de la fonction de lien

4.3.4.1 Domaine de $g(\bullet)$

Domaine des valeurs possible de $\mu = \varphi(\mu)$

Voici quelques exemple de domaine de la fonction de lien :

- Loi gamma $\Rightarrow \varphi(\mu) = [0, \infty[$
- Loi bernouilli $\Rightarrow \varphi(\mu) = [0, 1]$

4.3.4.2 Image de $g(\bullet)$

L'image de la fonction de lien est $\mathbb{R} =] - \infty, \infty[$

4.3.4.3 Conclusion

Le but de la fonction de lien est d'obtenir des valeurs de μ qui correspond bien au contexte du modèle à parti du prédicteur linéaire $X_i\beta$ qui peut prendre des valeurs dans \mathbb{R} . Autrement dit, on défini un domaine cohérent pour les valeurs réponses et on ne restreint pas les valeurs possibles des observations.

Ainsi, on obtient μ en inversant $g(\bullet)$:

$$g(E[Y_i|X_i]) = X_i\beta$$

$$\boxed{E[Y_i|X_i] = g^{-1}(X_i\beta)} \tag{4.1}$$

Il est donc nécessaire de choisir une fonction inversible pour la fonction de lien.

4.4 Approche générale

4.4.1 Procédure avec les GLM

- Choisir une distribution pour Y dans la famille exponentielle.
- Choisir une fonction de lien $g(\bullet)$.
- Estimer les paramètres β et ϕ .
- Valider le modèle.

4.4.2 Estimation des paramètres

Dans le cas des GLM, on estime les paramètres en utilisant la méthode du maximum de vraisemblance.

On souhaite choisir $\underline{\beta}$ qui maximise la fonction de vraisemblance suivante qui sera notre nouvelle métrique de distance :

$$l(\underline{\beta}) = \sum_{i=1}^n \ln(f_Y(y_i))$$

En pratique, il est d'usage d'utiliser la méthode de Newton-Raphson pour maximiser numériquement $l(\underline{\beta})$.

Pour ce faire, on pose $\hat{\underline{\beta}}^{(k)}$, le vecteur contenant les valeurs estimées pour $\underline{\beta}$ après la k^e itération de l'algorithme.

En supposant que l'algorithme ait convergé après l'itération $(i+1)$, alors on a que :

$$\frac{\partial}{\partial \underline{\beta}} l(\underline{\beta}^{(i+1)}) = 0$$

ou encore que

$$S(\underline{\beta}^{(i+1)}) = 0 \quad (4.2)$$

où $S(\bullet)$ est appelé le vecteur score et est défini ainsi :

$$S(\underline{\beta}^{(i+1)}) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} l(\underline{\beta}) \\ \frac{\partial}{\partial \beta_1} l(\underline{\beta}) \\ \vdots \\ \frac{\partial}{\partial \beta_p} l(\underline{\beta}) \end{bmatrix}_{(p+1) \times 1}$$

Pour obtenir un algorithme permettant d'obtenir un estimateur récursivement, à partir de $\hat{\underline{\beta}}^{(i)}$, il faut développer $S(\underline{\beta}^{(i+1)})$ autour de $\hat{\underline{\beta}}^{(i)}$ à l'aide d'une série de Taylor :

$$S(\underline{\beta}^{(i+1)}) = S(\underline{\beta}^{(i)}) + \left[\frac{\partial}{\partial \underline{\beta}} (S(\underline{\beta}^{(i)})) \right] \left(\hat{\underline{\beta}}^{(i+1)} - \hat{\underline{\beta}}^{(i)} \right)$$

En substituant cette expression dans l'équation 4.2, on obtient le développement suivant :

$$S(\underline{\beta}^{(i)}) + \frac{\partial}{\partial \underline{\beta}} (S(\underline{\beta}^{(i)})) \left(\hat{\underline{\beta}}^{(i+1)} - \hat{\underline{\beta}}^{(i)} \right) = 0$$

ou encore,

$$S(\hat{\beta}_{\sim}^{(i)}) + (-I(\hat{\beta}_{\sim}^{(i)}))(\hat{\beta}_{\sim}^{(i+1)} - \hat{\beta}_{\sim}^{(i)}) = 0$$

où $I(\beta)$ correspond à la matrice d'information de Fisher et est donnée par :

$$I(\beta) = \begin{bmatrix} \frac{\partial^2}{\partial \beta_0^2} l(\beta) & \frac{\partial^2}{\partial \beta_0 \partial \beta_1} l(\beta) & \cdots & \frac{\partial^2}{\partial \beta_0 \partial \beta_p} l(\beta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial \beta_0 \partial \beta_p} l(\beta) & \frac{\partial^2}{\partial \beta_1 \partial \beta_p} l(\beta) & \cdots & \frac{\partial^2}{\partial \beta_p^2} l(\beta) \end{bmatrix}_{(p+1) \times (p+1)}$$

Ainsi, on peut réécrire l'expression de la façon suivante :

$$\begin{aligned} S(\hat{\beta}_{\sim}^{(i)}) &= I(\hat{\beta}_{\sim}^{(i)})(\hat{\beta}_{\sim}^{(i+1)} - \hat{\beta}_{\sim}^{(i)}) \\ I^{-1}(\hat{\beta}_{\sim}^{(i)})S(\hat{\beta}_{\sim}^{(i)}) &= \hat{\beta}_{\sim}^{(i+1)} - \hat{\beta}_{\sim}^{(i)} \end{aligned}$$

$$\boxed{\hat{\beta}_{\sim}^{(i+1)} = \hat{\beta}_{\sim}^{(i)} + I^{-1}(\hat{\beta}_{\sim}^{(i)})S(\hat{\beta}_{\sim}^{(i)})} \quad (4.3)$$

Cette équation correspond à

Vecteur de paramètre mis à jour = Ancien vecteur de paramètre

+ Produit matriciel entre l'inverse de la matrice Fisher et le vecteur score

4.4.3 Validation *global* du modèle avec la *déviante*

En réécrivant la fonction de vraisemblance

Annexe A

Code source de l'exemple chapitre 3

(Section 3.4.3)

```
3 > # Import data
4 > data <- read.csv('data/Titanic/train.csv', stringsAsFactors =
  T)
5 > summary(data)
6 > # Ajout de la variable nom de famille
7 > data$Surname <- as.factor(sapply(as.character(data$Name),
8 +                             function(x) strsplit(x, split = '[,]')
  [[1]][1]))
9 > # Modele de regression stepwise
10 > fit <- step(lm(Age ~ Survived + Pclass + Sex + Fare + Embarked
  + Surname, data), direction = "both")
11 > fit
12 > # Q-Q plot
13 > plot(fit, which=2)
14 > # Ajout de la variable Log - Fare
15 > data$LogFare <- log(data$Fare)
16 > data$LogFare[data$LogFare == -Inf] <- 0
17 > fitL <- step(lm(Age ~ Survived + Pclass + Sex + Fare +
  Embarked + Surname + LogFare, data), direction = "both")
18 > fitL
19 > # Import data
20 > dataTest <- read.csv('data/Titanic/test.csv', stringsAsFactors
  = T)
21 > head(predict(fitL, dataTest))
```

Listing A.1 – Code source en R pour l'exemple