

**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.

# Table des matières

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Introduction</b>                               | <b>1</b> |
| <b>2</b> | <b>Régression linéaire simple</b>                 | <b>2</b> |
| 2.1      | Introduction . . . . .                            | 2        |
| 2.1.1    | Régression linéaire simple . . . . .              | 3        |
| 2.1.2    | Régression linéaire multiple . . . . .            | 3        |
| 2.1.3    | Régression exponentielle . . . . .                | 5        |
| 2.1.4    | Régression quadratique . . . . .                  | 5        |
| 2.2      | Le modèle de régression linéaire simple . . . . . | 8        |

## Résumé

*abstrat*

# Chapitre 1

## Introduction

L'établissement de prévisions joue un rôle central dans notre vie de tous les jours (prévisions météorologique, 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 <sup>1</sup> | Input  |
| $\beta_0, \beta_1, \dots, \beta_n$ | Les paramètres à estimer   |        |

Voici une illustration du concept de régression linéaire

#### Étape 1

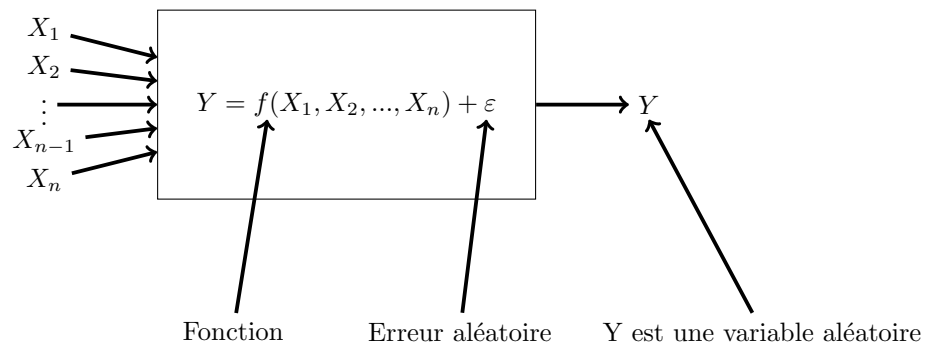
Observation  
des  $X_i$

#### Étape 2

Modèle de ré-  
gression

#### Étape 3

Prévision de  $Y$



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

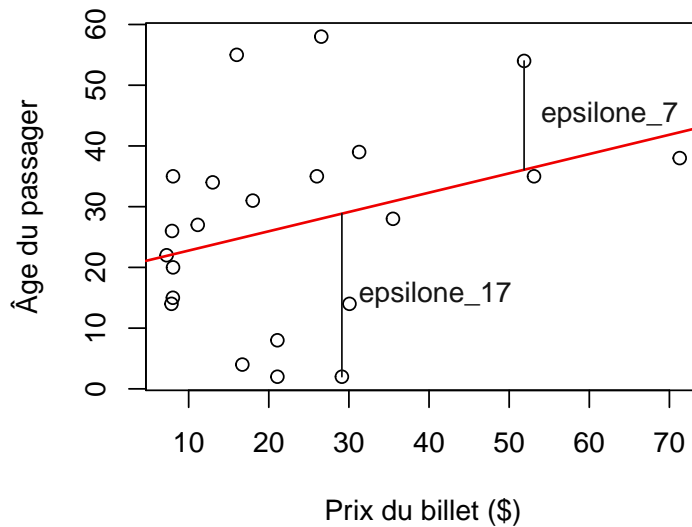
### 2.1.1 Regression linéaire simple

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 passa-      Prix du billet      Erreur aléa-  
ger                      toire

**Âge prédit des passagers du Titanic**



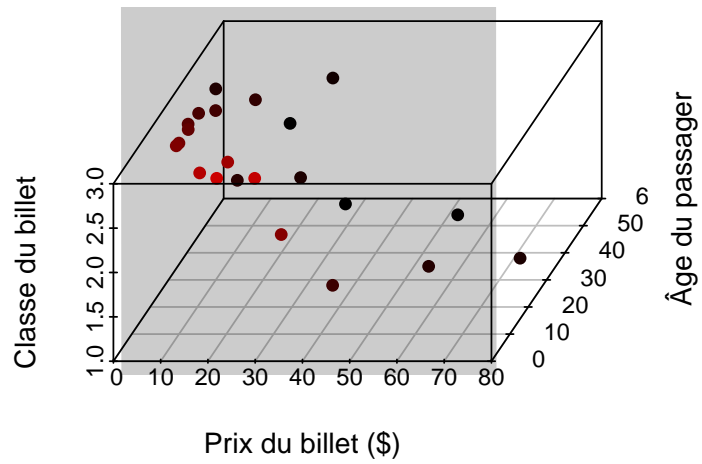
### 2.1.2 Regression linéaire multiple

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,

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

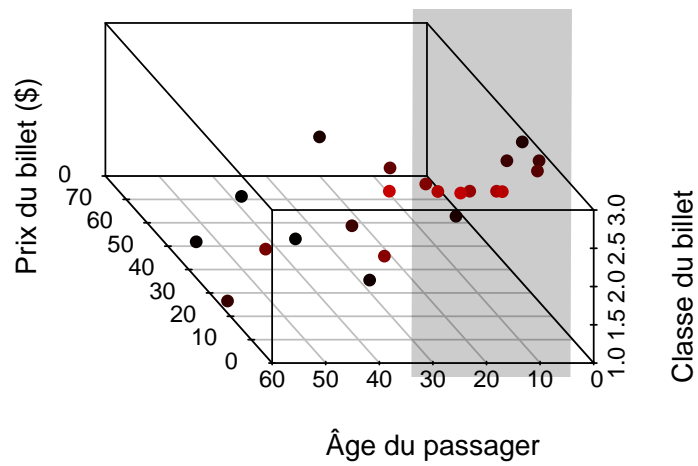
↑                      ↑                      ↑                      ↖  
Âge du passa-      Prix du billet      Sexe du passa-      Erreur aléa-  
ger                      ger                      toire

### Âge predict des passagers du Titanic



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

### Âge predict des passagers du Titanic





### 2.1.3 Régression exponentielle

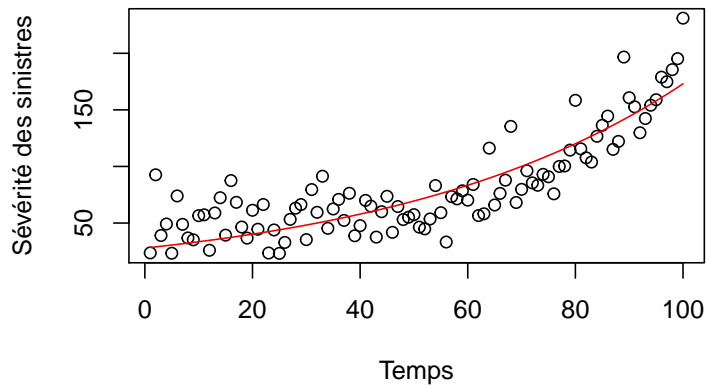
On cherche à prédire la sévérité d'un sinistre automobile en fonction du temps à l'aide du modèle exponentielle suivant,

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

↑
↗
↖

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.

$$\begin{aligned}\ln(Y) &= \ln(\beta_0) + \beta_1 \times X + \ln(\varepsilon) \\ Y^* &= \beta_0^* + \beta_1 \times X + \varepsilon^*\end{aligned}$$

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

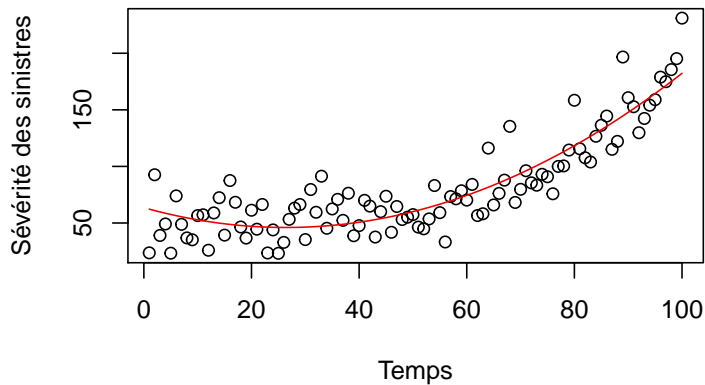
### 2.1.4 Régression quadratique

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 car,

- 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, tel que l'assurance.
- Se généralisent facilement, tel que les *Generalized Linear Models*.

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

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

soit minimiser.

Il existe plusieurs méthode pour calcul l'erreur. Soit les erreurs suivants :

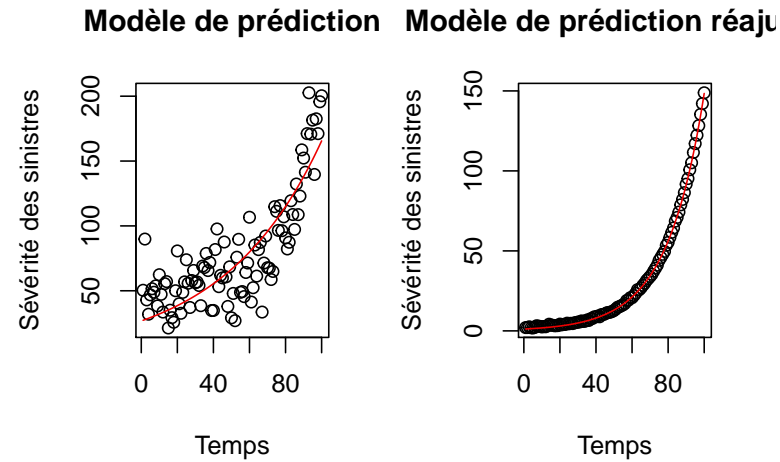
- Erreur totale
- Erreur absolue
- Erreur quadratique

Quel type d'erreur est suffisante pour déterminer  $\varepsilon$  ?

#### 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
- Pas fiable 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 compliqué 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 quadratique.
- Donne beaucoup de poids aux grandes erreurs

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

## 2.2 Le modèle de régression linéaire simple

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

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  $\beta_0$  est le paramètre associé à l'ordonnée à l'origine du modèle ;  $\beta_1$  est le paramètre associé à la pente de la droite ; et  $\varepsilon$  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 tiré d'une variable aléatoire ;
- Les observations de  $X_i$  sont considérées comme des valeurs connues et non aléatoires ;
- Les paramètres  $\beta_0$  et  $\beta_1$  sont inconnus au départ et doivent être estimer ;
- $\varepsilon_i$  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  $t$

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'année de scolarité est *en moyenne* de  $\beta_0$ . Par exemple,  $\beta_0$  serait le salaire moyen d'un stagiaire.

---

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

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

- Par la suite, pour chaque année additionnelle de scolarité, le salaire augmente *en moyenne* de  $\beta_1$  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  $\varepsilon_t$  et est tel 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.

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

Taux de croissance du salaire

Salaire

Ordonnées à l'origine

Nombre d'années de scolarité

Erreur aléatoire

On doit maintenant trouver les paramètres  $\beta_0$  et  $\beta_1$  de manière à minimiser l'erreur  $\varepsilon_t$ .

Si  $\varepsilon_t$  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  $\beta_i$ .

#### 2.2.0.1 Coefficients de régression

Les paramètres  $\beta_0$  et  $\beta_1$  sont déterminés en minimisant l'erreur quadratique à l'aide de la méthode des moindres carrées.

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

Où  $S(\psi)$  peut être considéré comme une mesure de la *distance* entre les données observées et le modèle théorique qui prédit ces données<sup>4</sup>.

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

### Minimisation de $\beta_0$

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

$$\sum_{t=1}^n Y_t - n \times \hat{\beta}_0 - \hat{\beta}_1 \sum_{t=1}^n X_t = 0 \tag{2.6}$$

---

4. Pour de plus ample information sur la méthode des moindres carrées et la fonction de *distance*, la page [Wikipédia](#) contient des bonnes explications sur le sujet.

### Minimisation de $\beta_1$

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

À l'aide des équations 2.6 et 2.7, on peut trouver les deux inconnus  $\beta_0$  et  $\beta_1$ .  
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 2.10 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

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

## Remarques