Guide de survie en Actuariat

Guide de survie en Actuariat

Gabriel Crépeault-Cauchon
Nicholas Langevin

Dépôt officiel de ce document
Dernière mise à jour : 13 octobre 2019

Introduction

On explique ici les motivations du document

Liste des collaborateurs

Voici la liste des personnes ayant collaboré à ce document :

- Gabriel Crépeault-Cauchon
- Nicholas Langeevin

Table des matières

Première partie

Fondements mathématiques utiles

Chapitre 1

Notions de calcul différentiel et calcul intégral

1.1 Règles de dérivation

Dans le tableau, on utilise $k \in \mathbb{R}$ ou n pour parler d'une constante, u, v ou w pour parler d'une fonction.

Fonction	Dérivée
$f(x) = k$	$f'(x) = 0$
$f(x) = kx$	$f'(x) = k$
$f(x) = x^n$	$f'(x) = nx^{n-1}$
$f(x) = kg(x)$	$f'(x) = kg'(x)$
$f(x) = g(x) \pm h(x)$	$f'(x) = g'(x) \pm h'(x)$
$f(x) = g(x) \cdot h(x)$	$f'(x) = g'(x) \cdot h(x) + g(x) \cdot h'(x)$
$f(x) = \frac{g(x)}{h(x)}$	$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$
$f(x) = g(x)^n$	$f'(x) = n \cdot g(x)^{n-1} \cdot g'(x)$
$f(x) = k^{g(x)}$	$f'(x) = k^{g(x)} \ln k \cdot g'(x)$
$f(x) = e^{g(x)}$	$f'(x) = e^{g(x)} \cdot g'(x)$
$f(x) = \ln(g(x))$	$f'(x) = \frac{g'(x)}{g(x)}$

Chapitre 2

Algèbre linéaire

2.1 Définition d'un vecteur et une matrice

Vecteur ligne Un vecteur ligne $\boldsymbol{x}$est un vecteur de dimension $p \times 1$, tel que

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_p \end{bmatrix}$$

Matrice Une matrice $\mathbf{A} = [a_{ij}]_{m \times n}$ de dimension m lignes par n colonnes , définie telle que

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & \ddots & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{bmatrix} \tag{2.1}$$

Matrice carrée Une matrice carrée $\mathbf{A}$ de dimensions $m \times m$ a autant de lignes que de colonnes.

non-négative $\mathbf{A}$ est définie comme *non-négative* si $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$, $\forall \mathbf{x} \in \mathbb{R}^n$.

positive $\mathbf{A}$ est définie comme *positive* si $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$, $\forall \mathbf{x} \neq 0$.

semi-positive $\mathbf{A}$ est définie comme non-négative, mais elle n'est pas définie positive.

Orthogonale $\mathbf{A}$ est *orthogonale* si elle est non-singulière et $\mathbf{A}^{-1} = \mathbf{A}^{\top}$ (voir sous-section 2.4.3 pour définition de $\mathbf{A}^{-1}$)

Matrice symétrique La mactrice $\mathbf{A}$ est symétrique si $a_{ij} = a_{ji}\ \forall i, j$, i.e

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 1 \end{bmatrix}$$

Matrice triangulaire (inférieure ou supérieure) Une matrice inférieure $\mathbf{L}$ est constituée de 0 en dessous de la diagonale :

$$\mathbf{L} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$

À l'inverse, on peut aussi avoir une matrice triangulaire supérieure $\mathbf{U}$, où les éléments en haut de la diagonale sont tous égaux à 0 :

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & 0 \\ 8 & 3 & 0 \\ 2 & 3 & 9 \end{bmatrix}$$

Matrice diagonale Une matrice diagonale $\mathbf{D}$ a des éléments $d_{ii} > 0$ sur sa diagonale seulement. Cette matrice est à la fois triangulaire inférieure et supérieure. i.e.

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Un cas spécial de la matrice diagonale est la matrice identité $\mathbf{I}$, où $\mathbf{I}_{ii} = 1,\ \forall i$, i.e

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.2}$$

Matrice diagonalisable Une matrice $\mathbf{A}_{n\times n}$ est dite *diagonalisable* s'il existe une matrice carrée $\mathbf{Q}_{n\times n}$ inversible (ou non-singulière) et une matrice $\mathbf{D}$ diagonale telle que

$$\mathbf{Q^{-1}AQ} = \mathbf{D} \leftrightarrow \mathbf{A} = \mathbf{QDQ^{-1}} \tag{2.3}$$

(Théorème sur les matrices symétriques) : Toute matrice carrée symétrique est diagonalisable apr uen matrice orthogonale $\mathbf{Q}$.

2.2 Matrice transposée

Soit la matrice $\mathbf{A}$ définie en (2.1). On peut trouver la matrice transposée $\mathbf{A}^\top$, où $[a_{ij}] = [a_{ji]}$. **En d'autres mots, les lignes deviennent des colonnes.** Voici quelques propriétés intéressantes avec les matrices transposées :

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
- $(\mathbf{kA})^\top = \mathbf{kA}^\top$
- $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$
- $\mathbf{A}^\top\mathbf{A}$ et $\mathbf{AA}^\top$ sont symétriques.

2.3 Opérations matricielles

Voici une liste non-exhaustive des opérations matricielles possibles. Côté notation, A et B représente des matrices, c re présente une constante

- $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$
- $\mathbf{A} - \mathbf{B} = [a_{ij} - b_{ij}]$
- $c\mathbf{A} = [ca_{ij}]$
- Produit matriciel :

$$\mathbf{AB} = \left[\sum_{k=1}^{p} a_{ik}b_{kj}\right]_{i\times j} \tag{2.4}$$

  , avec $\mathbf{A} = [a_{ip}]$ et $B = [b_{pj}]$
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $\mathbf{A^{-1}A} = \mathbf{I} = \mathbf{AA^{-1}}$, où I est la matrice identité (voir Équation 2.2) et A^{-1} est la matrice inverse de A (voir sous-section 2.4.3 au besoin)
- $(AB)^{-1} = B^{-1}A^{-1}$

2.4 Trace, déterminant et matrice inverse

2.4.1 Trace d'une matrice

Soit la matrice carrée $\mathbf{A}$. On peut trouver la trace de cette matrice en sommant les éléments de sa diagonale, i.e.

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \tag{2.5}$$

Propriétés de la trace d'une matrice

- $\mathrm{Tr}(\mathbf{A} + \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B})$
- $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ et $\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA})$

2.4.2 Déterminant d'une matrice

Soit la matrice carrée $\mathbf{A}$. On peut trouver le déterminant de $\mathbf{A}$, noté $\det(\mathbf{A})$ ou $|\mathbf{A}|$, avec

$$\det(\mathbf{A}) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \tag{2.6}$$

De façon générale, lorsque les dimensions de la matrice carrée sont supérieures à 2, on a

$$\det(A) = \sum_{j=1}^{n} a_{ij} C_{ij} \tag{2.7}$$

avec $1 \leq i \leq n$ où $C_{ij} = (-1)^{i+j} M_{ij}$ et M_{ij} est le déterminant de la nouvelle matrice en enlevant la ligne i et la colonne j.
Si la matrice $\mathbf{A}$ est inversible (ou non-singulière, voir la sous-section 2.4.3), alors le déterminant aura les propriétés suivantes :

- $\det(A^\top) = \det(A)$
- $\det(kA) = k^n \det(A)$
- $\det(A + B) \neq \det(A) + \det(B)$
- $\det(AB) = \det(A) \det(B)$
- $\det(A^{-1}) = \frac{1}{\det(AB)} = \det(A)^{-1}$

2.4.3 Matrice inverse

Soit la matrice carrée $\mathbf{A}$. On peut trouver la matrice inverse $\mathbf{A}^{-1}$ telle que

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \operatorname{Adj}(\mathbf{A}) \tag{2.8}$$

où $\operatorname{Adj}(\mathbf{A}) = [C_{ij}]^T_{m\times n}$ et $C_{ij} = (-1)^{i+j} M_{ij}$.

2.5 Décomposition LDU de Choleski

Soit $\mathbf{A}$ une matrice carrée symétrique définie positive. Alors, il existe une décomposition unique telle que

$$\mathbf{A} = \mathbf{LDU} \tag{2.9}$$

où $\mathbf{L}, \mathbf{D}, \mathbf{U}$ sont respectivement des matrices triangulaire inférieure, triangulaire supérieure et diagonale.
Cette décomposition peut être fortement utile en programmation lorsqu'on fait des opérations sur des matrices, afin de limiter le nombre d'opérations.

2.6 Vecteurs et valeurs propres

2.6.1 Définition

Soit $\mathbf{A}$ une matrice carrée. On dit que λ est une *valeur propre* de $\mathbf{A}$ s'il existe un vecteur $\mathbf{x} \neq 0$ tel que

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{2.10}$$

On appelle le vecteur $\mathbf{x}$ un *vecteur propre* correspondant à la valeur propre λ. De plus, l'ensemble des nombres réels λ satisfaisant l'Équation 2.10 est appelé *spectre* de la matrice $\mathbf{A}$.

2.6.2 Propriétés intéressantes

Les vecteurs propres et valeurs propres permettent d'avoir plusieurs propriétés appréciables, notamment :

- Si $\mathbf{x}$ est un vecteur propre de $\mathbf{A}$ correspondant à la valeur propre λ, alors $c\mathbf{x}$ sera également un vecteur propre de $\mathbf{A}$ correspondant à λ.

- Si $\mathbf{A}$ est symétrique et $\mathbf{x_1}$ et $\mathbf{x_2}$ sont des vecteurs propres correspondant à des valeurs propres différentes de $\mathbf{A}$, alors $\mathbf{x_1}$ et $\mathbf{x_2}$ sont des vecteurs ortogonaux, i.e. $\mathbf{x_1}^\top \mathbf{x_2} = 0$.
- Si $\mathbf{A}$ a les valeurs propres (pas nécessairement distinctes) $\lambda_1, \ldots, \lambda_n$, alors $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$ et $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$.

2.6.3 Décomposition spectrale

Soit $\mathbf{A}_{n\times n}$ une matrice symétrique avec les n valeurs propres $\lambda_1, \ldots, \lambda_n$. Il existe une matrice orthogonale $\mathbf{Q}$ telle que

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top \tag{2.11}$$

avec $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$. Cette décomposition est fort utile lorsqu'on veut faire des produits matriciels successifs de la même matrice (appliqué directement dans les chaînes de Markov, voir section 5.1) :

$$\begin{aligned}\mathbf{A}\mathbf{A} &= \mathbf{Q}\mathbf{\Lambda}\underbrace{Q^\top Q}_{=\mathbf{I}}\mathbf{\Lambda}\mathbf{Q}^\top \\ &= \mathbf{Q}\mathbf{\Lambda}^\mathbf{2}\mathbf{Q}^\top\end{aligned}$$

2.7 Dérivées de matrice ou vecteurs

Voici quelques entités pratiques :

$$\frac{\partial}{\partial \mathbf{v}}\mathbf{w}^\top\mathbf{v} = w$$

$$\frac{\partial}{\partial \mathbf{v}} = \mathbf{v}^\top\mathbf{A}\mathbf{v} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{v}$$

Deuxième partie

Matière vue dans le baccalauréat en actuariat

Chapitre 3

Probabilités et statistiques

3.1 Concepts de probabilité de base

3.1.1 Probabilité conditionnelle

Définition de base

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{3.1}$$

Loi des probabilités totales Soit E_i le *outcome* i parmi l'ensemble des n *outcome* possibles de l'évènement E, alors, on peut représenter la probabilité que l'évènement A survienne comme

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A|E_i) \Pr(E_i) \tag{3.2}$$

avec $\sum_{i=1}^{n} \Pr(E_i) = 1$.

Relation importante de l'Équation 3.1, on peut représenter $\Pr(A|B)$ comme

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)} \tag{3.3}$$

3.1.2 Théorème de Bayes

En combinant l'Équation 3.3 et la loi des probabilités totales (l'Équation 3.2), on obtient le théorème de Bayes :

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\sum_{i=1}^{n}\Pr(B|A_i)\Pr(A_i)} \tag{3.4}$$

3.2 Définition d'une variable aléatoire

3.3 Distribution d'une variable aléatoire

Fonction de densité, répartition, survie, hazard rate, etc.

3.4 Moments et quantités importantes

Espérance, variance, covariance, coefficient de variation, corrélation

Espérance Soit une v.a. X (continue ou discrète). Son espérance est définie telle que

$$\mathrm{E}[X] = \mu = \sum_{x=0}^{\infty} x\Pr(X=x) = \int_0^{\infty} x f_X(x)dx \tag{3.5}$$

L'espérance d'une fonction de la v.a X est

$$\mathrm{E}[g(X)] = \sum_{x=0}^{\infty} g(x)\Pr(X=x) = \int_0^{\infty} g(x) f_X(x)dx \tag{3.6}$$

Variance

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \tag{3.7}$$

quelques propriétés à savoir :

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$
$$\mathrm{Var}(X+b) = \mathrm{Var}(X)$$

Covariance

$$\mathrm{Cov}\,(X,Y) = \sigma_{X,Y} = \mathrm{E}\,[(X - \mathrm{E}\,[X])(Y - \mathrm{E}\,[Y])] = \mathrm{E}\,[XY] - \mathrm{E}\,[X]\,\mathrm{E}\,[Y] \tag{3.8}$$

3.5 Distribution de probabilité qui reviennent souvent

Un tableau récapitulatif des différentes distribution de probabilité est disponible à l'

Chapitre 4

Mathématiques financières

to-do

Chapitre 5

Processus aléatoires

to-do

5.1 Chaîne de Markov

Chapitre 6

Théorie du risque

6.1 Modèle pour les risques et méthodes d'estimation

Introduction Dans le cours *Introduction à l'actuariat II*, on a vu comment produire des réalisation $x^{(j)}$ de x du modèle fréquence-sévérité, dans le cas ou $B \sim$ Gamma. Dans ce cours, on développe des technique récursive pour évaluer la convolution.

6.1.1 Méthode d'estimation

Context #1

(1) Pour chaques contrats (j), on dispose du nombre de sinistres (n_j) et des montants de chaques sinistres $(y_1, ..., y_{n_j})$.

(2) On définie

$$X_j = \sum_{k=1}^{n_j} y_{j,k}$$

où $X_j = 0$ si $n_j = 0$.

(3) On pose $\underline{\theta}^N$ et $\underline{\theta}^Y$, les paramètres à estimer. On utilise la méthode du

maximum de vraisemblance pour estimer ses paramètres.

$$\begin{aligned}\mathcal{L}(\underline{\theta}^N, \underline{\theta}^Y) &= \prod_{j=1}^{m} \left\{ f_n(n_j|\underline{\theta}^n) \prod_{k=1}^{n_j} f_y(y_{j,k}|\underline{\theta}^y) \right\} \\ &= \left(\prod_{j=1}^{m} f_n(n_j|\underline{\theta}^n) \right) \left(\prod_{j=1}^{m} \prod_{k=1}^{n_j} f_y(y_{j,k}|\underline{\theta}^y) \mathbb{1}_{\{n_j>0\}} \right) \\ &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \mathcal{L}(\underline{\theta}^N)\mathcal{L}(\underline{\theta}^Y)\end{aligned}$$

(4) Remarque :

- Le résultat découle de l'indépendence entre N et $\underline{Y}$.
- Ce résultat facilite l'estimation.
- On peut estimer des paramètres avec des lois de fréquence et sévérité séparément.

Context #2

(1) Pour chaque contrat (j), on dispose du nombre de sinistres (n_j) et si le nombre de sinistre est non null, on connait le montant **total** des sinistres. On ne connait pas les montants de chaques sinistre.

(2) On pose $\underline{\theta}^N$ et $\underline{\theta}^Y$, les paramètres à estimer. On utilise la méthode du maximum de vraisemblance pour estimer ses paramètres.

$$\begin{aligned}\mathcal{L}(\underline{\theta}^N, \underline{\theta}^Y) &= \prod_{j=1}^{m} f_n(n_j|\underline{\theta}^n) f_{Y_1+...+Y_{n_j}}(x_j|\underline{\theta}^y) \mathbb{1}_{\{n_j>0\}} \\ &= \mathcal{L}(\underline{\theta}^N)\mathcal{L}(\underline{\theta}^{Y_1+...+Y_{n_j}})\end{aligned}$$

(3) Remarque :

- Le résultat découle de l'indépendence entre N et $\underline{Y}$.
- Ce résultat facilite l'estimation.
- On peut estimer des paramètres avec des lois de fréquence et sévérité séparément si on connait la loi de $Y_1 + ... + Y_{n_j}$.

Context #3

(1) Pour chaque contrat (j), on connait uniquement les coûts totaux, null ou non null.

(2) On pose $\underline{\theta}^N$ et $\underline{\theta}^Y$, les paramètres à estimer. On utilise la méthode du maximum de vraisemblance pour estimer ses paramètres.

(3) Possibilités #1, modèle forfaitaire, $X_j = C \cdot \mathbb{1}_{\{I=1\}}$.

$$\mathcal{L}(\underline{\theta}^N, \underline{\theta}^Y) = \prod_{j=1, x_j=0}^{m} f_I(0|\underline{\theta}^I) \prod_{j=1, x_j>0}^{m} f_I(1|\underline{\theta}^I) \cdot f_C(x_j|\underline{\theta}^C)$$

(4) Possibilités #2, modèle fréquence-sévérité. Si on connait la loi de la somme $(Y_1 + ... + Y_k)$. La distribution est donc mixte avec masse de probabilité à 0. Ainsi on a

$$f_X(0|\underline{\theta}^N, \underline{\theta}^Y) = \Pr\left(N = 0|\underline{\theta}^N\right)$$

$$f_X(x_j|\underline{\theta}^N, \underline{\theta}^Y) = \sum_{k=1}^{\infty} \Pr\left(N = k|\underline{\theta}^N\right) f_{Y_1+...+Y_k}(x_j|\underline{\theta}^Y)$$

$$\mathcal{L}(\underline{\theta}^N, \underline{\theta}^Y) = \prod_{j=1, x_j=0}^{m} \Pr\left(N = 0|\underline{\theta}^N\right) \prod_{j=1, x_j>0}^{m} f_X(x_j|\underline{\theta}^N, \underline{\theta}^Y)$$

(5) Remarque : Contrairement au deux autre context, on ne peut pas estimé séparément $\underline{\theta}^N$ et $\underline{\theta}^Y$.

6.2 Processus Stochastique

6.2.1 Processus de poisson homogène

Definition

Le processus de comptage $N = \{N(t), t \geq 0\}$ est un processus de Poisson si les conditions suivantes sont satisfaites :

(1) $N(t) = 0$

(2) un accroissement sur un intervalle de temps de longueur t obéit à une loi Poisson de paramètre $\lambda t (t > 0)$:
$\cdot N(t) \sim$ Pois (λt) :
$\cdot N(s, s+t] \sim$ Pois (λt) :

(3) I des accroissements indépendants :
pour $0 \leq s_1 < s_2 \leq t_1 < t_2 < \infty, N(s_1, s_2]$ et $N(t_1, t_2]$ sont indépendants ;
c. - $\ddot{a} - d$. I les accroissements sur deux intervalles disjoints de temps sont indépendants

(4) $\underline{N}$ a des accroissements stationnaires : $N(t) \sim N(s, s+t]$

Proposition 1

Soit les processus de Poisson indépendants $\underline{N}_1 = \{N_1(t), t \geq 0\}$ et $\underline{N}_2 = \{N_2(t), t \geq 0\}$ avec des taux λ_1 et λ_2 . Alors, le processus defini par $\underline{M} = \{M(t), t \geq 0\}$
ou

$$M(t) = N_1(t) + N_2(t)$$

est aussi un processus de Poisson process avec un taux $\lambda_1 + \lambda_2$

Algorithme PP1

1. On fixe $T_0^{(j)} = 0$
2. Pour $i = 1, ..., n$ on a,
   2.1 On simule $W_i^{(j)}$
   2.2 On calcule $T_i^{(j)} = T_{i-1}^{(j)} + W_i^{(j)}$

```
PP1 <- function(time, lam, giveN = FALSE){
    T_i <- 0
    while(tail(T_i, 1) <= time){
        u <- runif(1)
        w <- qexp(u, lam)
        T_i <- c(T_i, tail(T_i, 1) + w)
    }
    if(giveN) length(T_i[-c(1, length(T_i))])
    else T_i[-c(1, length(T_i))]
}
```

Listing 6.1 – Mise en oeuvre de PP1 en R

Algorithme PP2

1. On fixe $T_0^{(j)} = 0$
2. On simule la réalisation $N(t)^{(j)}$ de $N(t)$
3. Sachant $N(t) = N(t)^{(j)} > 0$
   3.1 On simule le vecteur de réalisations $\left(U_1^{(j)}, \ldots, U_{N(t)}^{(j)}\right)$ *de* $\left(U_1, \ldots, U_{N(t)}(t)\right)$ ou les va. $U_i \sim U \sim$ Unif $(0, 1)$:

3.2 On trie les réalisations en [1] et on obtient $\left(U_{[1]}^{(j)}, \ldots, U_{[N(t)}^{(j)}(j)\right]\Big)$ ou $U_{[1]}^{(j)} < \ldots < U_{\left[N(t)^{(j)}\right]}^{(j)}$

3.3 On calcule $T_i^{(j)} = t \times U_{[i]}^{(j)},$ pour $i = 1, \ldots, N(t)^{(j)}$

```
PP2 <- function(t, lam, giveN = FALSE){
    N_t <- rpois(1, t * lam)
    if(N_t == 0) return(0)
    U_i <- runif(N_t)
    T_i <- t * sort(U_i)
    if(giveN) N_i
    else T_i
}
```

Listing 6.2 – Mise en oeuvre de PP2 en R

Propisition 2

Soit un processus de Poisson $N = \{N(t), t \geq 0\}$ avec le taux λ. Soit le vecteur de v.a. continues iid $(Y_1 \ldots, Y_n)$ ou $Y_i \sim Y \sim$ Unif $(0, t)$ avec $f_Y(t) = \frac{1}{t}, t \in (0, t],$ pour $i = 1, 2, \ldots, n$. On definit le vecteur de statistiques d' ordre $\left(Y_{[1]}, \ldots, Y_{[n]}\right)$ à partir de $(Y_1 \ldots, Y_n)$
Alors, on a

$$(T_1, T_2, \ldots, T_n | N(t) = n) \sim \left(Y_{[1]}, Y_{[2]}, \ldots, Y_{[n]}\right)$$

Note

$$\Pr\left(X \in (x, x + dx)\right) = f_X(t)dx$$

Preuve (Proposition 2)

(1) On pose $h_1, ..., h_n$ très très petit. On fixe $0 \leq s_1 \leq s_1 + h_1 \leq ... \leq s_n \leq s_n + h_n$.

$$\Pr\left(T_1 \in (s_1, s_1 + h_1), ..., T_n \in (s_n, s_n + h_n) | N(t) = n\right) \approx f_{T_1, ..., T_n | N(t)}(s_1, ..., s_n | n) h_1 \cdots h_n$$

(2) On veut identifier $f_{T_1,...,T_n|N(t)}$, on développe

$$\begin{aligned}&\Pr\left(T_1 \in (s_1, s_1+h_1), ..., T_n \in (s_n, s_n+h_n) | N(t)=n\right)\\ =&\frac{\Pr\left(T_1 \in (s_1, s_1+h_1), ..., T_n \in (s_n, s_n+h_n), N(t)=n\right)}{\Pr\left(N(t)=n\right)}\end{aligned}$$

(3) On réécrit en fonction des accroissements du processus

$$= \frac{N(0,s_1]=0, N(s_1,s_1+h_1]=1, ..., N(s_n,s_n+h_n]=1, N(s_n+h_n,t]=0}{\Pr\left(N(t)=n\right)}$$

Note : la condition $N(s_n+h_n,t]=0$ permet de respecter $N(t)=n$.

(4) On applique la proposition des accroisements indépendant du processus de Poisson.

$$\begin{aligned}&= \frac{\Pr\left(N(0,s_1]=0\right]\right)\Pr\left(N(s_1,s_1+h_1]=1\right)\cdots\Pr\left(N(s_n,s_n+h_n]=1\right)\Pr\left(N(s_n+h_n,t]=0\right)}{\Pr\left(N(t)=n\right)}\\ &= \frac{\left(e^{-\lambda s_1}\right)\left(\lambda h_1 e^{-\lambda h_1}\right)\cdots\left(\lambda h_n e^{-\lambda h_n}\right)\left(e^{-\lambda(t-s_n-h_n)}\right)}{\frac{(\lambda t)^n e^{-\lambda t}}{n!}}\\ &= \frac{\lambda^n h_1\cdots h_n e^{s_1+h_1+s_2-s_1-h_1+...+h_n+t-s_n-h_n}}{\frac{(\lambda t)^n e^{-\lambda t}}{n!}}\\ &= n!\frac{h_1}{t}\cdots\frac{h_n}{t}\end{aligned}$$

(5) En éliminant les h_i de chaque coté avec l'équation en (1), on obtient

$$f_{T_1,...,T_n|N(t)}(s_1,...,s_n|n) = \frac{n!}{t^n}$$

(6) Lemme :

Soit le vecteur de v.a. continues id $(Y_1\ldots,Y_n)$ oi $Y_i \sim Y$ avec la fonction de densité $f_{Y_i}=f_Y,$ pour $i=1,2,\ldots,n$. On définit le vecteur de statistiques d' ordre $\left(Y_{[1]},\ldots,Y_{[n]}\right)$ a partir de $(Y_1\ldots,Y_n)$. Alors, la fonction de densité de conjointe de $\left(Y_{[1]},\ldots,Y_{[n]}\right)$ est don'ee par $f_{Y_{[1]},X_{[3]},...,X_{[n]}}(y_1,y_2,...,y_n)=n!f_Y(y_1)\times f_Y(y_2)\times...,Xf_Y(y_n),y_1<y_2<...<y_n.$(2)

(7) En effet, on défini

$$Y_1, ..., Y_n \sim Y \sim Unif(0, t)$$

avec fonction de densité $f_Y(s) = \frac{1}{t}$

(8) On conclut que

$$(T_1, ..., T_n | N(t) = n) \sim (Y_{[1]}, ..., Y_{[n]})$$

6.2.2 Processus non-homogene

Definition Le processus de comptage $N = \{N(t), t \geq 0\}$ est dit un processus de Poisson non-homogéne de fonction d'intensite $\lambda(t) \geq 0$ pour $t \geq 0$ si

(1) $N(0) = 0$

(2) $\{N(t), t \geq 0\}$ posséde des accroissements indépendants

(3) $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$

(4) $(N(t+h) - N(t) \geq 2) = o(h)$

Proposition 1

Soit $N = \{N(t), t \geq 0\}$ un processus de Poisson non-homogène de fonction d'intensité $\lambda(t)$. Alors,

$N(t+s) - N(t) \sim \text{Poisson}(\Lambda(t+s) - \Lambda(s)), \forall t, s \geq 0$

ou

$\Lambda(t) = \int_0^t \lambda(y)dy$

est la fonction d'intensité cumulee) du processus. Ainsi,

$P(N(t+s) - N(s) = n) = \frac{[m(t+s)-m(s)]^n e^{-[m(t+s)-m(s)]}}{n!}$

Algorithme PPNH1

1. On fixe $T_0^{(j)} = 0$
2. Pour $i = 1, \ldots, n,$ on a
   - 2.1 On simule les réalisations $\left(Z_1^{(j)}, \ldots, Z_n^{(j)}\right)$ du vecteur de v.a. iid avec $Z_i \sim Z \sim \text{Exp}(1)$
   - 2.2 On simule $W_i^{(j)} = \Lambda_{T^{-1}}^{-1}(Z_i)$
   - 2.3 On calcule $T_i^{(j)} = T_{i-1}^{(j)} + W_i^{(j)}$

```
PPNH1 <- function(t, inverseFUN, giveN = FALSE){
    T_i <- 0
    while(tail(T_i, 1) <= t){
        u <- runif(1)
        z <- qexp(u, 1)
        w <- inverseFUN(z, tail(T_i, 1))
        T_i <- c(T_i, tail(T_i, 1) + w)
    }
    if(giveN) length(T_i[-c(1, length(T_i))])
    else T_i[-c(1, length(T_i))]
}
```

Listing 6.3 – Mise en oeuvre de PPNH1 en R

Algorithme PPNH2

1. On fixe $T_0^{(j)} = 0$
2. On simule la réalisation $N(t)^{(j)}$ de $N(t) \sim$ Poisson $(\Lambda(t))$.
3. Sachant $N(t) = N(t)^{(j)} > 0$
   - 3.1 On simule le vecteur de réalisations $\left(V_1^{(j)}, \ldots, V_{N(t)}^{(j)}\right)$ du vecteur de v.a. ind $\left(V_1, \ldots, V_{N(t)}\right)$ ou $V_i \sim V$ avec $f_V(x) = \frac{\lambda(x)}{\Lambda(t)}, 0 < x < t\,(i = 1, 2, \ldots, N(t)^j)$
   - 3.2 On trie les réalisations en [3.1] et on obtient $\left(V_{[1]}^{(j)}, \ldots, V_{\left[N(t)^{(j)}\right]}^{(j)}\right)$ ou $V_{[1]}^{(j)} < \ldots < V_{\left[N(t)^{(j)}\right]}^{(j)}$
   - 3.3 On calcule $T_i^{(j)} = V_{[i]}^{(j)},$ pour $i = 1, \ldots, N(t)^{(j)}$

```
to-do
```

Listing 6.4 – Mise en oeuvre de PPNH2 en R

6.2.3 Processus Homogène Composée

Definition

$$S(t) = \sum_{k=1}^{N(t)} X_k$$

Fonction de répartition

$$F_{S(t)}(x) = \Pr(N(t) = 0) + \sum_{k=1}^{\infty} \Pr(N(t) = k) * F_{X_1+...+X_k}(x)$$

```
F_s <- function(x, t){
    dpois(0, lambda * t) + sum(sapply(1:k0, function
       (k) dpois(k, lambda * t) * pgamma(x, alpha *
       k, beta)
}
```

Listing 6.5 – Exemple Pois-Gamma

Value at risk

$$\mathrm{VaR}_k(S(t)) = F^{-1}_{S(t)}(k)$$

```
VaR_s <- function(kappa, t){
    if(kappa <= dpois(0, lambda * t)
        return(0)
    uniroot(function(x) F_s(x, t) - kappa, c(0,
       10000))$root
}
```

Listing 6.6 – Exemple Pois-Gamma

Tail Values at Risk

$$\mathrm{TVaR}_k(S(t)) = \sum_{k=0}^{\infty} \Pr(N(t) = k) \cdot \mathrm{TVaR}_k(X_1 + ... + X_k)$$

```
TvaR_S <- function(kappa, t){
    sum(sapply(1:k0, function(k) dpois(k, 1.8 * t) *
        alp    ha * k / beta * (1 - pgamma(VaR_s(
       kappa, t), (alpha*k)+1, beta)))) / (1 - kappa
       )
}
```

Listing 6.7 – Exemple Pois-Gamma

6.2.4 Processus Poisson Mixte

Definition Soit Λ une variable aléatoire positive (continue ou discrète). Si le processus de comptage $\underline{N} = \{N(t); t \geq 0\}$ étant donné que $\Lambda = \lambda$ est un processus de Poisson de taux Λ alors $\underline{N} = \{N(t); t \geq 0\}$ est appelé un processus de Poisson mixte.

Les accroissements du processus de Poisson mixte $\underline{N}$ sont **indépendant** et **stationnaire**.

Preuve (stationnaire)

$$\begin{aligned} M_{N(t,t+s]}(r) &= \mathrm{E}\left[e^{rN(t,t+s}\right] \\ &= \mathrm{E}_\Lambda\left[\mathrm{E}\left[e^{rN(t,t+s]}|\Lambda\right]\right] \\ &= \mathrm{E}_\Lambda\left[e^{\Lambda t(e^r-1)}\right] \\ &= M_\Lambda(t(e^r-1)) \\ &= M_{N(t)}(r) \\ &= \text{Stationnaire car fonction de } t \text{ seulement} \end{aligned}$$

Preuve (indépendance) A faire

Note

Les temps-inter siniste sont échangeable, mais ne sont pas indépendant. Par contre, les temps-inter siniste $(W_1|\Lambda)$ et $(W_2|\Lambda)$ sont conditionnelement indépendant et $(w|\Lambda) \sim \mathrm{Exp}(\lambda)$.

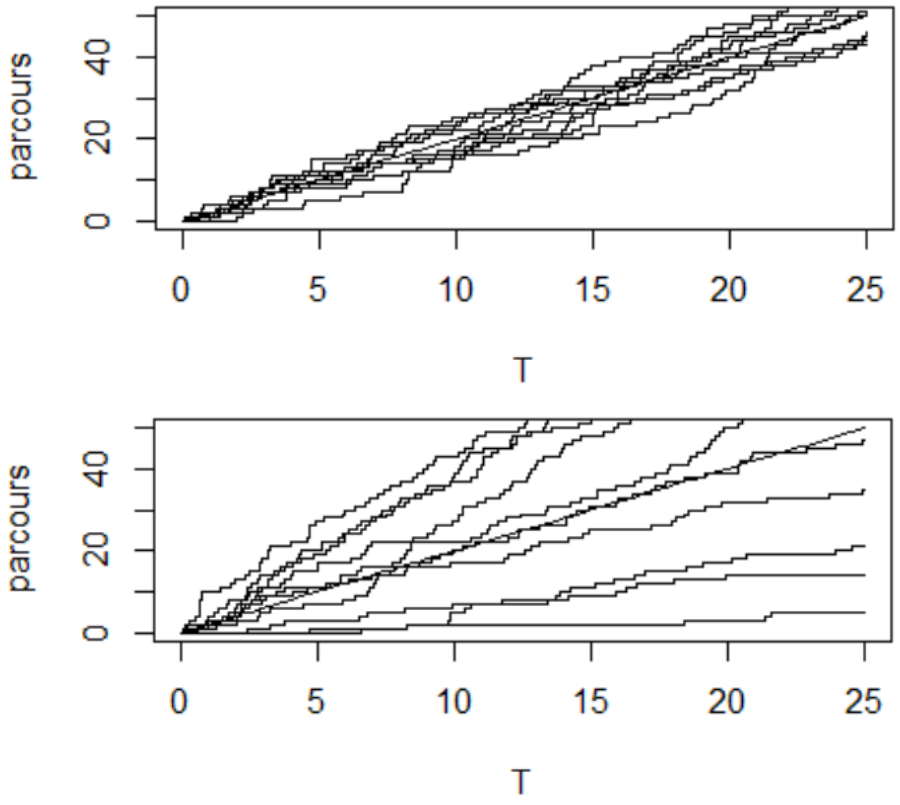

FIGURE 6.1 – Comparaison entre un processus de Poisson homogène et un processus de poisson mixte. Le graphique du bas représente le processus mixte.

$$\begin{aligned}
\Pr\left(N(t)=n\right) &= \int_{-\infty}^{\infty} \Pr\left(N(t)=n|\Lambda\right)\cdot f_{\Lambda}(\lambda)\,d\lambda \\
&\quad \text{si } \Lambda \sim \Gamma(\alpha,\beta) \\
&= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)k!}\left(\frac{\beta}{\beta+t}\right)^{\alpha}\left(\frac{1}{\beta+t}\right)^{n} \sim \text{BinNeg}(\alpha,\frac{\beta}{\beta+t}) \\
\Pr\left(N(t,t+s]=n\right) &= \int_{-\infty}^{\infty} \Pr\left(N(t,t+s]=n|\Lambda\right)\cdot f_{\Lambda}(\lambda)\,d\lambda \\
&= \int_{-\infty}^{\infty} \Pr\left(N(s)=n|\Lambda\right)\cdot f_{\Lambda}(\lambda)\,d\lambda
\end{aligned}$$

$$\begin{aligned}
\Pr\left(N(t)=k_1, N(t,t+s]=k_2\right) &= \int_{-\infty}^{\infty} \Pr\left(N(t)=k_1, N(t,t+s]=k_2|\Lambda\right)\cdot f_{\Lambda}(\lambda)\,d\lambda \\
&= \int_{-\infty}^{\infty} \Pr\left(N(t)=k_1|\Lambda\right)\Pr\left(N(s)=k_2|\Lambda\right)\cdot f_{\Lambda}(\lambda)\,d\lambda
\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}\left[N(t+s)|N(t)=k_1\right] &= \mathrm{E}\left[N(t)+N(t,t+s]|N(t)=k_1\right] \\
&= k_1 + \mathrm{E}_\Lambda\left[\mathrm{E}\left[N(t,t+s]|N(t)=k_1,\Lambda\right]|N(t)=k_1\right] \\
&= k_1 + \mathrm{E}_\Lambda\left[\lambda t|N(t)=k_1\right] \\
&= k_1 + \int_{-\infty}^{\infty} \lambda t \frac{f_{\Lambda,N(t)}(\lambda,k_1)}{\Pr\left(N(t)=k_1\right)}\, d\lambda \\
&= k_1 + \frac{\int_{-\infty}^{\infty} \lambda t \Pr\left(N(t)=k_1|\Lambda\right) f_\Lambda(\lambda)\, d\lambda}{\int_{-\infty}^{\infty} \Pr\left(N(t)=k_1|\Lambda\right) f_\Lambda(\lambda)\, d\lambda} \\
&\quad \text{si } \Lambda \sim \Gamma(\alpha,\beta) \\
&= k_1 + \frac{\alpha+n}{\beta+t}
\end{aligned}
$$

$$
\begin{aligned}
F_{W_1}(t) &= \int_{-\infty}^{\infty} F_{w_1|\Lambda}(t) f_\Lambda(\lambda)\, d\lambda \\
&\quad \text{si } \Lambda \sim \Gamma(\alpha,\beta) \\
&= \left(\frac{\beta}{\beta+t}\right)^\alpha \sim \text{Pareto}(\alpha,\beta)
\end{aligned}
$$

$$
\begin{aligned}
F_{W_1,W_2}(t_1,t_2) &= \int_{-\infty}^{\infty} F_{W_1}(t_1) F_{W_2}(t_2) f_\Lambda(\lambda)\, d\lambda \\
&\quad \text{si } \Lambda \sim \Gamma(\alpha,\beta) \\
&= \left(\frac{\beta}{\beta+t_1+t_2}\right)^\alpha \sim \text{Pareto Mulivarié}
\end{aligned}
$$

6.2.5 Processus de renouvellement

Definition Un processus de renouvellement $\underline{N} = \{N(t), t \geq 0\}$ est un exemple de processus de dénombrement (comptage). IL est une généralisation du processus de Poisson. La généralisation se fait via les v.a. temps inter-sinistres. On définit les temps inter-sinistres associes à N par la suite de v.a. $\underline{W} = \{W_j, j = 1, 2, \ldots\}$ On définit les temps d'occurrence des sinistres par la suite de v.a. $\underline{T} = \{T_j, j \in \mathbb{N}\}$, ou $T_0 = 0$ et $T_j = \sum_{l=1}^{j} W_l, j = 1, 2, \ldots$

Relation fondamentale

$$
\{N(t) \geq k\} = \{T_k \leq t\}\,, \text{ pour } t > 0 \text{ et } k \in \mathbb{N}
$$

Fonction de masse de probabilité

$$\begin{aligned}\Pr\left(N(t)=k\right) &= \Pr\left(\#\text{ de sinistres sur } (0,t] \text{ soit égal } \dot{a}k\right)\\ &= \Pr\left(N(t)\geq k\right) - \Pr\left(N(t)\geq k+1\right)\\ &= F_{T_k}(t) - F_{T_{k+1}}(t)\end{aligned}$$

Espérence de N(t)

$$m(t) = E[N(t)] = \sum_{k=1}^{\infty} E\left[1_{\{T_k\leq t\}}\right] = \sum_{k=1}^{\infty} F_{T_k}(t)$$

pour $t > 0$

Remarques

- $m(t) = E[N(t)]$ = nombre espéré de sinistres sur l'intervalle de temps $(0, t]$
- La fonction $m(t)$ est appelée la fonction de renouvellement.

Algorithme de simulation

1. On fixe $T_0^{(j)} = 0$
2. Pour $i = 1, ..., n$ on a,
   2.1 On simule $W_i^{(j)}$
   2.2 On calcule $T_i^{(j)} = T_{i-1}^{(j)} + W_i^{(j)}$

Troisième partie

Matière pour les examens professionnels

Chapitre 7

Time Series

This section resume the Chapters 1-5 (excluding Sections 3.3 and 3.4), 6, 7 (Sections 7.1, 7.2 and 7.3) of *An Introductory Time Series with R*
A mettre dans la bio : Cowpertwait, P. and Metcalfe, A., Introductory Time Series with R, Springer, 2009.

Trend In general, a systematic change in a time series that does not appear to be periodic is known as a trend.

Seasonal Variation A repeating pattern within any fixed period.

Notation We represent a time series of length n by $x_t : t = 1, ..., n = x_1, x_2, ..., x_n$

Forecast A forecast $\hat{x}_{t+k|t}$ is a predicted future value, and the number of time steps into the future is the **lead time (k)**

7.1 Base Models

Additive Decomposition The additive decomposition model is given by

$$x_t = m_t + s_t + z_t$$

where t is the time, x_t the observed series, m_t the trend, s_t the seasonal effect and z_t the error term.

Multiplication Model If the seasonal effect tends to increase as the trend inscrease, we use a multiplication model define as

$$x_t = m_t \cdot s_t + z_t$$

where t is the time, x_t the observed series, m_t the trend, s_t the seasonal effect and z_t the error term.

7.1.1 Estimating Trends

Centered Moving Average A moving average is an average of a specified number of time series values around each value in the time series, with the exception of the first few and last few terms. The length of the moving average is chosen to average out the seasonal effects, which can be estimated

$$\hat{m}_t = \frac{0.5m_{t-k} + m_{t-k-1} + ... + m_t + ... + m_{t+k-1} + 0.5m_{t+k}}{2k}$$

7.1.2 Estimating Seasonal Effect

Additive Effect An estimate of the monthly additive effect (s_t) at time t is define by

$$\hat{s}_t = x_t - \hat{m}_t$$

It is usual to adjust those estimate in order that the sum of one period of the time serie equal zero. Let c be that adjustment in order to solve this expression.

$$\sum(s_t + c) = 0$$

Multiplicative Effect An estimate of the monthly multiplicative effect (s_t) at time t is define by

$$\hat{s}_t = \frac{x_t}{\hat{m}_t}$$

And we found the ajustment c in order to solve this expression.

$$\sum \frac{(\hat{s}_t + c)}{n} = 1$$

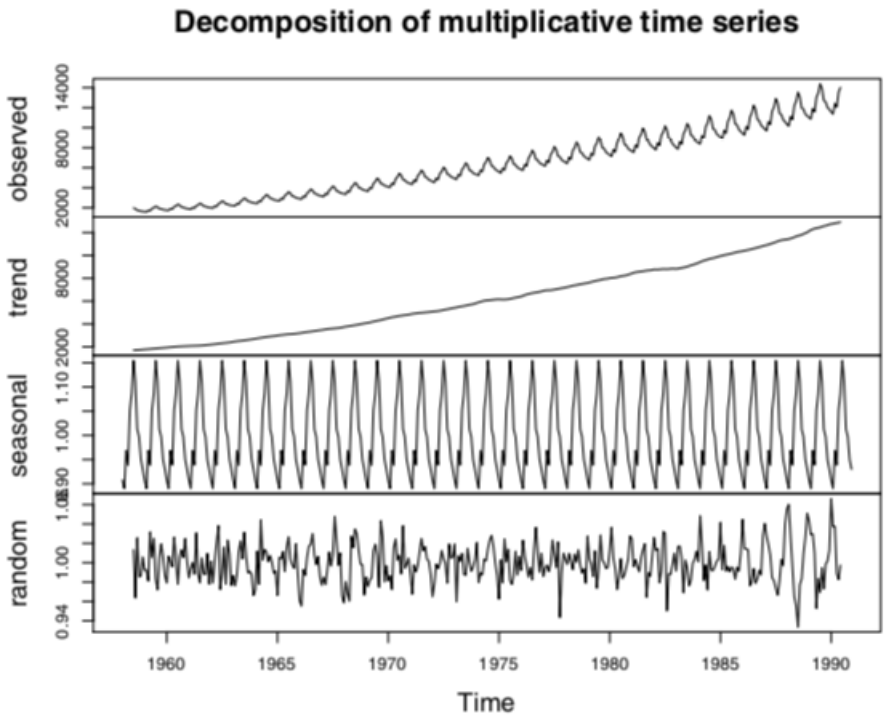

FIGURE 7.1 – In this example, the multiplicative model would seem more appropriate than the additive model because the variance of the original series and trend increase with time

7.1.3 Smoothing Procedure

Smoothing procedures can, and usually do, use points before and after the time at which the smoothed estimate is to be calculated. A consequence is that the smoothed series will have some points missing at the beginning and the end unless the smoothing algorithm is adapted for the end points.

Ex.1 The centering moving average is an exemple of a *smoothing* procedure.

Ex.2 The *loess* technique is also a *smoothing* that use a locally weighted regression.

7.2 Correlation

Mean Function The mean function is, in general, a function of t and it define as

$$\mu(t) = \mathrm{E}\left[x_t\right]$$

Sample Mean The sample mains is define as

$$\bar{x} = \sum \frac{x_i}{n}$$

Variance Function The variance function of a time series model that is stationary in the mean is

$$\sigma^2(t) = \mathrm{E}\left[(x_t - \mu^2)\right]$$

Sample Variance If the model is stationary in the variance, we can estimate the variance with the sample variance define as

$$\mathrm{Var}\,(x) = \frac{\sum (x_t - \bar{x}}{n-1}$$

Stationarity If the mean function is constant, we say that the time series model is stationary in the mean. The time serie can also be stationary in the variance, if the variance function is constant σ^2.

Ergodic Serie A time series model that is stationary in the mean is ergodic in the mean if the time average for a single time series tends to the ensemble mean as the length of the time series increases.

$$\lim_{n\to\infty} \bar{x} = \mu$$

Covariance The covariance measure the *linear association* between two random variables. The covariance is define as

$$\gamma(x, y) = \mathrm{E}\left[(x - \mu_x)(y - \mu_y)\right]$$

Sample Covariance We can estimate the covariance with the sample covariance define as

$$\mathrm{Cov}\,(x, y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation Correlation is a dimensionless measure of the linear association between a pair of variables (x,y) and is obtained by standardising the covariance.

$$\rho(x,y) = \frac{\mathrm{E}\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x\sigma_y} = \frac{\gamma(x,y)}{\sigma_x\sigma_y}$$

Sample Correlation We can estimate the covariance with the sample covariance define as

$$\mathrm{Cor}(x,y) = \frac{\mathrm{Cov}\,(x,y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

Autocovariance (acvf) If a time series model is second-order stationary, we can define an autocovariance function (acvf), γ_k , as a function of the lag k.

$$\gamma_k = \mathrm{E}\left[(x_t-\mu)(x_{t+k}-\mu)\right]$$

Sample Autocovariance The acvf function can be estimated by

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})$$

Note

The sample autocovariance at lag 0, c_0, is the variance calculated with a denominator n.

Autocorrelation (acf) The lag k autocorrelaton function (acf), ρ_k, is defined by

$$\rho_k = \frac{\gamma_k}{\sigma^2}$$

where $\rho_0 = 1$

Sample Autocorrelation The acf function can be estimated by

$$r_k = \frac{c_k}{c_0}$$

Second-order Stationary Suppose a the time serie that is stationary in the mean and the variance. Then, the time serie model is second-order stationary if the correlation between variable depends only on the number of time steps separating them.

7.2.1 The Correlogram

The correlogram is plot of r_k against k. The main use of the correlogram is to detect autocorrelations in the time series after we have removed an estimate of the trend and seasonal variation.

- The x-axis gives the lag (k) and the y-axis gives the autocorrelation (ρ_k) at each lag. The unit of lag is the sampling interval. Correlation is dimensionless, so there is no unit for the y-axis.
- The dotted lines on the correlogram are drawn at

$$-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$$

  If the sample acf(r_k) falls outside these line, we have evidence against the null hypothesis that $p_k = 0$ at the 5% level.
- The lag at 0 is always.

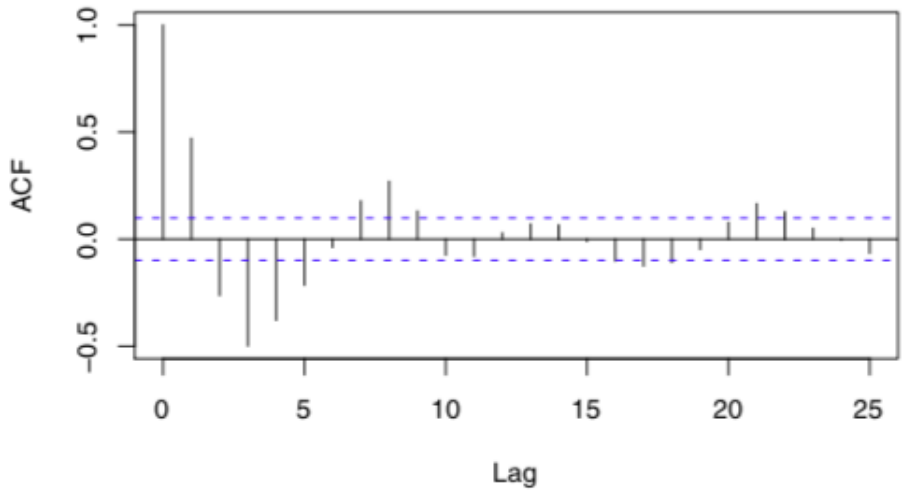

FIGURE 7.2 – Correlogram

7.2.2 Covariance of sums of random variables

Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_m$ be random variables. Then

$$\operatorname{Cov}\left(\sum_{i=1}^{n} x_i, \sum_{j=1}^{m}\right) y_j = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{Cov}(x_i, y_i)$$

The result tells us that the covariance of two sums of variables is the sum of all possible covariance pairs of the variables.

7.3 Forecasting Strategies

7.3.1 Relationships of different time serie

Cross-covariance (ccvf) The cross-correlation (ccvf) between two time series is define as

$$\gamma_k(x, y) = \mathrm{E}\left[(x_{t+k} - \mu_x)(y_t - \mu_y)\right]$$

Note

This is not a symmetric relationship, and the variable x is lagging variable y by k.

$$\gamma_k(x, y) = \gamma_{-k}(y, x)$$

Sample Cross-covariance The ccvf function can be estimed by

$$c_k(x, y) = \frac{1}{n}\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(y_k - \bar{y})$$

Cross-correlation (ccf) The cross-correlation (ccf) between two time serie is define as

$$/rho_k(x, y) = \frac{\gamma_k(x, y)}{\sigma_x \sigma_y}$$

Note

This is not a symmetric relationship, and the variable x is lagging variable y by k.

$$\rho(x, y) = \rho_{-k}(y, x)$$

Sample Cross-correlation The ccf function can be estimed by

$$r_k(x,y) = \frac{c_k(x,y)}{\sqrt{c_0(x,x)x_0(y,y)}}$$

7.4 Basic Stochastic Models

We may consider a **trend** to be **stochastic** when it shows inexplicable changes in direction. Those type of trend can be simulated by the models of this section.

7.4.1 White Noise

Residual Error A residual error is the difference between the observed value and the model predicted value at time t. Then the residual error, x_t, is defined by

$$x_t = y_t - \hat{y}_t$$

where y_t is the observed value and $\hat{y}_t$ the predicted value.
As the residual errors occur in time, they form a time series : $x_1, x_2, ..., x_n$.

Definition A white noise is a time serie $\{w_t : t = 1, 2, ..., n\}$ where each term is independent and with a constant variance σ^2.

$$w_t \sim \mathrm{N}(0, \sigma^2)$$

7.4.2 Random Walks

Definition Let $\{x_t\}$ be a time series. Then $\{x_t\}$ is a random walk if

$$x_t = x_{t-1} + w_t$$

where $\{w_t\}$ is a white noise.

Backward Operator The backward operator(or lag operator) **B** is defined by

$$\mathbf{B}^n = x_{t-n}$$

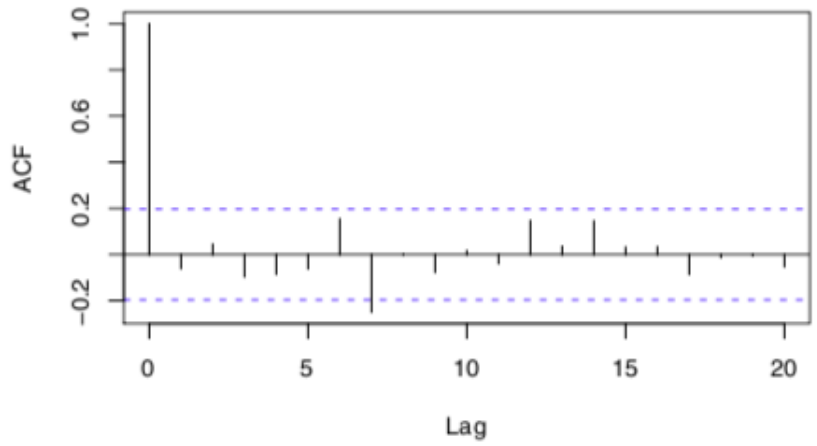

FIGURE 7.3 – Correlogram of a simulated white noise series. The underlying autocorre- lations are all zero (except at lag 0) ; the statistically significant value at lag 7 is due to sampling variation.

Second-order Properties The covariance is in function of time, so a random walk is non-stationary. This model is only suitable for short term predictions.

$$\begin{aligned}\mu(t) &= 0\\ \sigma^2(t) &= t\sigma_w^2\\ \gamma_k(t) &= t\sigma_w^2\\ \rho_k(t) &= \frac{1}{\sqrt{1+\frac{k}{t}}}\end{aligned}$$

The Difference Operator The defference operator is defined by

$$\nabla x_t = x_t - x_{t-1}$$

We can also expresse it with the Backward operator

$$\nabla^n x_t = (1-\mathbf{B})^n x_t$$

Note

Differencing adjacent terms of a series can transform a non-stationary series to a stationary series.

$$x_t - x_{t-1} = w_t$$

Here differencing two non-stationary random walk give a stationary white noise.

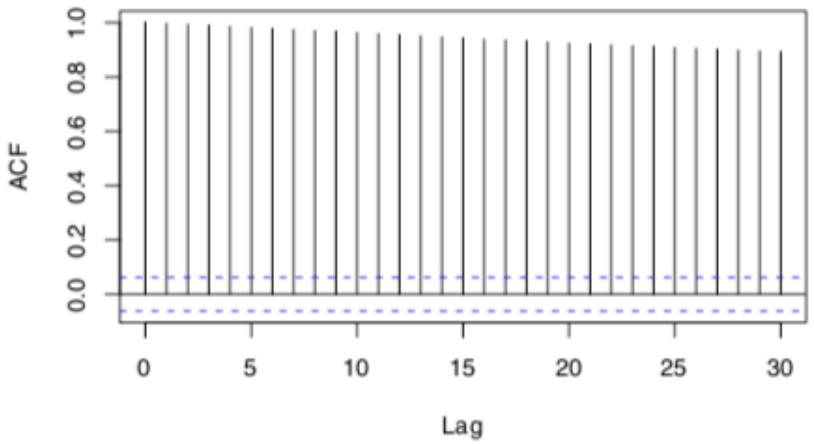

FIGURE 7.4 – The correlogram for the simulated random walk. A gradual decay from a high serial correlation is a notable feature of a random walk series.

Random Walk with Drift A Random walk with drif is a random walk with a mean. it defined as

$$x_t = \delta + x_{x-1} + w_t$$

7.4.3 Autoregressive Models

Definition The series $\{xt\}$ is an autoregressive process of order p, abbreviated to AR(p), if

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{x-2} + ... + \alpha_p x_{t-p} + w_t$$

where α_i are the model parameters with $\alpha_p \neq 0$.
The following points should be noted :

- The random walk is a special case of AR(1) with $\alpha_1 = 1$.
- The model is a regression of x_t on past terms from the same series, that where the name *autoregressive* come.
- A prediction at time t is given by

$$\hat{x}_t = x_t = \alpha_1 x_{t-1} + \alpha_2 x_{x-2} + ... + \alpha_p x_{t-p}$$

- The model parameters can be estimated by minimising the sum of squared errores.

Stationary and non-stationary AR processes We can expressed a AR model as a polynomial of order p in terms of the backward shift operator

$$\theta_p(\mathrm{B}) x_t = (1 - \alpha_1 \mathrm{B} - \alpha_2 \mathrm{B}^2 - ... - \alpha_p \mathrm{B}^p) x_t = w_t$$

Then a AR model is stationary if all root of $\theta_p(\mathrm{B})$ exceed one in absolute values. In other word if

if all $|\mathrm{B_i}| > 1$ then the model is stationary
otherwise, the model is non-stationary

Second-order properties of an AR(1) model

$$\begin{aligned} \mu_k &= 0 \\ \gamma_k &= \frac{\alpha^k \sigma_w^2}{1 - \alpha^2} \\ \rho &= \alpha^k \end{aligned}$$

Partial Autocorrelation An AR(p) process has a correlogram of partial autocorrelation α_k that is zero after lag p.

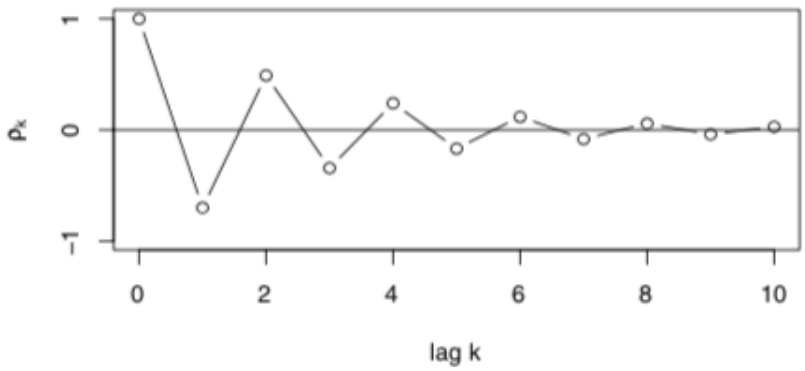

FIGURE 7.5 – For an AR model, the correlograms decays to zero. The correlogram decays to zero more rapidly for small α.

7.5 Regression

When we have some plausible physical explanation for a **trend** we will usually wish to model it in some **deterministic** manner. Therefore, the model of this section can be used.

The difference with **deterministics trend** is the when we make short term forecast, we assume that the trend will change slowly.

Time series regression usually differs from a standart regression analysis because the residual form a time serie and therefore tend to be serially correlated. When this correlation is possitive, the estimated standard errors of the parameter estimates, read from the computer output of a standard regression analysis, will tend to be less than their true value.

7.5.1 Linear Models

Definition A model for a time series $\{x_t : t = 1, ..., n\}$ is linear if it can be expressed as

$$x_t = \alpha_0 + \alpha_1 u_{1,t} + \alpha_2 u_{2,t} + ... + \alpha_m u_{m,t} + z_t$$

where $u_{i,t}$ is the value of the i^{th} explanatory variable, α_i the estimated model parameters, z_t the error at time t.

An example of a linear model is the pth-order polynomial function of t :

$$x_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + ... + \alpha_p t^p + z_t$$

Note

Note that the errors form a time series $\{z_t\}$, with mean 0, that does not have to be Gaussian or white noise.

Stationarity Linear models for time series are non-stationary when they include functions of time.
Differencing can remove both stochastic and deterministic trends from time series. Then for a polynomial of roder m, the mth-order differencing is required to remove the trend.

Autocorrelation variance estimation of the sample mean Let $\{x_t : t = 1, ..., n\}$ be a stationary time serie with mean μ, variance σ^2 and autocovariance $\mathrm{Cov}\,(x_t, x_{t+k})$. Then the variance of the sample mean is given by

$$\mathrm{Var}\,(\bar{x}) = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n-1}(1 - \frac{k}{n})\rho_k\right]$$

Generalised least squares A fitting procedure known as generalised least squares (GLS) can be used to provide better estimates of the standard errors of the regression parameters to account for the autocorrelation in the residual series.

7.5.2 Linear Models with Seasonal Variables

Additive Seasonal Indicator Variables To include seasonal effect, we change the constant α_o depending on the season. Let s be the time serie measured (Ex. for monthly time serie, $s = 12$). Then for each s, we fit a constant term.

$$x_t = s_t + m_t + z_t$$

where s_t is the seasonal constant when t falls in the i$^{\text{th}}$ season, m_t a linear model for the trend and z_t the error term.

Harmonic Seasonal models The advantage of this model is that we can represente the seasonal effect with something that is smoothly. For a time series $\{x_t\}$ with s season, there are $\lfloor s/2 \rfloor$ possible cycles. The harmonic model is defined by

$$x_t = m_t + \sum_{i=1}^{\lfloor s/2 \rfloor} \left\{ s_i \sin(\frac{2\pi i t}{s}) + c_i \cos(\frac{2\pi i t}{s}) \right\} + z_t$$

where m_t is a trend model **that include a constant term** (α_0) and s_i and c_i are unknown parameters.

7.5.3 Forecasting from regression

When we predic a regression time series, we try to predict in the future. The problem is that the trend might change. Therefore, it is better ot think of a forecast from a regression model as an expected value conditional on past trends continuing into the future.

Bias Correction The process of transforming the model introduce some bias in the mean. We need to apply a correction to the mean. Note that this correction doesn't need to be apply in simulation.

$$\hat{x}'_t = \hat{x}_t * \text{Correction}$$

Lognormal Correction

$$e^{\sigma^2/2}$$

Empirical Correction

$$\frac{1}{n} \sum e^{z_t}$$

7.6 Stationary Models

Sometime, the residual will be correlated in time, as this is not accounted in the fitted regression model, we need other model.

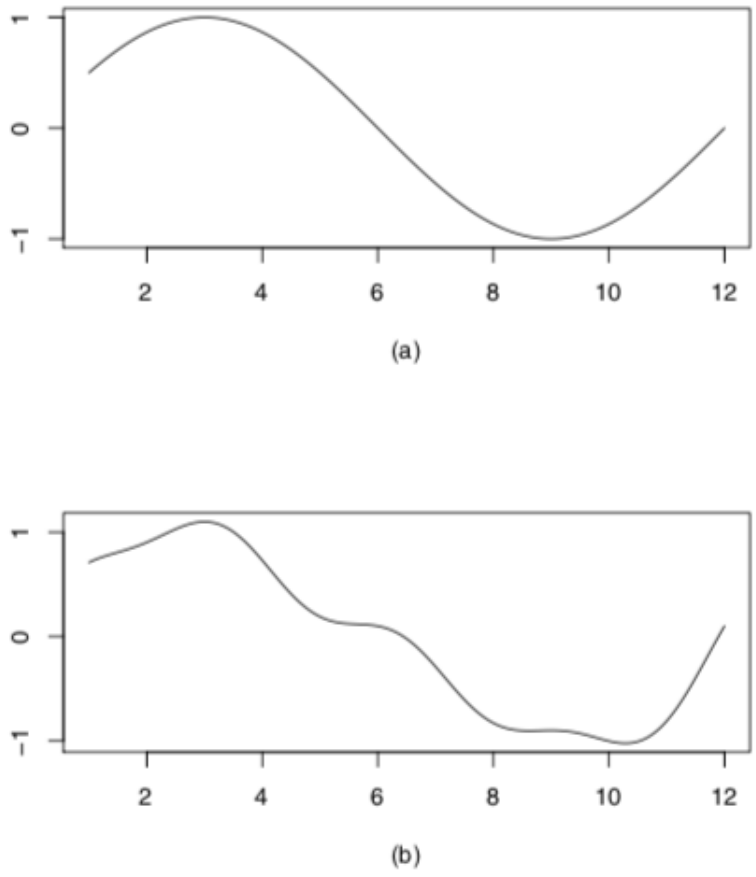

FIGURE 7.6 – Two possible underlying seasonal patterns for monthly series based on the harmonic model (Equation (5.10)). Plot (a) is of the first harmonic over a year and is usually too regular for most practical applications. Plot (b) is of the same wave but with a further two harmonics added. Plot (b) illustrates just one of many ways that an underlying sine wave can be perturbed to produce a less regular, but still dominant, seasonal pattern of period 12 months.

7.6.1 Strictly Stationary Series

A time series model $\{x_t\}$ is *strictly stationry* if the joint statistical distribution $x_{t_1}, ..., x_{t_n}$ is the same as the joint distribution of $x_{t_1+m}, ..., x_{t_n+m}$ for all $t_1, ..., t_n$ and m, so that the distribution is unchanged after an arbitrary time shift.

Note

Note that strict stationarity implies that the mean and variance are constant in time and that the autocovariance $\mathrm{Cov}\,(x_t, x_s) = \gamma_k$ (i.e. only depend on the lag k). If a series is not strictly stationary but the mean and variance are constant in time and the autocovariance only depends on the lag, then the series is called *second-order stationary*.

We focus on the second-order properties in this chapter, but the stochastic processes discussed are strictly stationary.

Stationarity is an idealisation that is a property of models. If we fit a stationary model to data, we assume our data are a realisation of a stationary process. So our first step in an analysis should be to check whether there is any evidence of a trend or seasonal effects and, if there is, remove them. Regression can break down a non-stationary series to a trend, seasonal components, and residual series. It is often reasonable to treat the time series of residuals as a realisation of a stationary error series. Therefore, the models in this chapter are often fitted to residual series arising from regression analyses.

7.6.2 Moving average models

Definition moving average (MA) process of order q is a linear combination of the current white noise term and the q most recent past white noise terms and is defined by

$$x_t = w_t + \beta_1 w_{t-1} + ... + \beta_q w_{t-q} = \phi_q(\mathrm{B}) w_t$$

where ϕ_q is a polynomial of order q. Because MA processes consist of a finite sum of stationary white noise terms, they are stationary and hence have a time-invariant mean and autocovariance.

Second-order properties

$$\begin{aligned}\mu &= 0\\ \sigma^2 &= \sigma_w^2(1+\beta_1^2+...+\beta_q^2)\\ \gamma_k &= \sigma_w^2\sum_{i=0}^{q-k}\beta_i\beta_{i+k}\\ \rho_k &= \frac{\sum_{i=0}^{q-k}\beta_i\beta_{i+k}}{\sum_{i=0}q\beta_i^2}\end{aligned}$$

where $\beta_0 = 1$.

Invertible properties An MA process is invertible if it can be expressed as a stationary AR(∞) process of infinite order without an error term.

$$w_t = (1-\beta \mathrm{B})^{-1}x_t = x_t + \beta x_{t-1} + \beta^2 x_{t-2} + ...$$

provided $|\beta| < 1$, which is required for convergence.

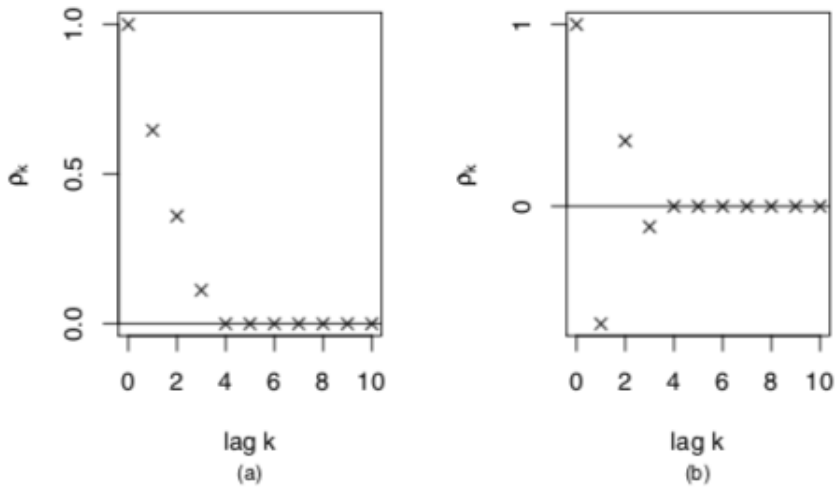

FIGURE 7.7 – Plots of the autocorrelation functions for two MA(3) processes. The autocorrelation for lag $k > q$ are all zero. $(a)\beta_1 = 0.7, \beta_2 = 0.5, \beta_3 = 0.2; (b)\beta_1 = -0.7, \beta_2 = 0.5, \beta_3 = -0.2$.

7.6.3 Mized Models : The ARMA process

Definition The ARMA model is a AR(p) + MA(q) models defined as

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + ... + \alpha_p x_{t-p} + w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} + ... + \beta_q w_{t-q}$$

We can also express it as

$$\theta_p(\mathrm{B}) x_t = \phi_q(\mathrm{B}) w_t$$

- The process is stationary when the roots of θ are all exceed unity in absolute value.
- The process is invertible when the roots of ϕ all exceed unity in absolute value.
- The AR(p) model is the special case ARMA(p, 0).
- The MA(q) model is the special case ARMA(0, q).
- **Parameter parsimony.** When fitting to data, an ARMA model will often be more parameter efficient (i.e., require fewer parameters) than a single MA or AR model.
- **Parameter redundancy.** When θ and ϕ share a common factor, a stationary model can be simplified. For example, the model :

$$(1 - \frac{1}{2}B)(1 - \frac{1}{3}B)x_t = (1 - frac12B)w_t$$
$$(1 - \frac{1}{3}B)x_t = w_t$$

Second-order Properties

$$\sigma^2 = \sigma_w^2 \left(1 + \frac{(\alpha+\beta)^2}{1-\alpha^2}\right)$$
$$\gamma_0 = \sigma_w^2 \left(\frac{1 + 2\alpha\beta + \beta^2}{1-\alpha^2}\right)$$
$$\gamma_k = \sigma_w^2 (\alpha+\beta)\alpha^{k-1} \left(\frac{1+\alpha\beta}{1-\alpha^2}\right)$$
$$\rho_k = \frac{\alpha^{k-1}(\alpha+\beta)(1+\alpha\beta)}{1+\alpha\beta+\beta^2} = \alpha\rho_{k-1}$$

7.7 Non-stationary Models

7.7.1 Differencing

Differencing a serie $\{x_t\}$ can remove trends, whether these trend are stochastic, as in a random walk, or deterministic, as in the case of a linear trend.

Differencing random walk

$$\nabla x_t = x_t - x_{t-1} = w_t$$

which is a stationary white noise.

Differencing linear trend

$$\nabla x_t = x_t - x_{t-1} = b + w_t - w_{t-1}$$

which is a stationary moving average process rather than white noise.

Integrated Model A serie $\{x_t\}$ is integrated of order d, $I(d)$, if the d^{th} difference of $\{x_t\}$ is a white noise

$$(1 - \mathrm{B})^d x_t = w_t$$

Note

The random walk is a special case $I(1)$

7.7.2 Non-Seasonal ARIMA Models

Definition A time series $\{x_t\}$ follows an ARIMA(p, d, q) process id the d^{th} differences of the $\{x_t\}$ series are an ARIMA(p, q) process

$$\theta_p(\mathrm{B}(1 - \mathrm{B})^d x_t = \phi_q(\mathrm{B}) w_t$$

In general :

- ARIMA$(0, d, q) \equiv$ IMA(d, q)
- ARIMA$(p, d, 0) \equiv$ ARI(p, d)

7.7.3 Seasonal ARIMA models

A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term.

The ARIMA$(p, d, q)(P, D, Q)_s$ model can be defined as

$$\Theta_P(\mathrm{B}^s)\theta_p(\mathrm{B})(1-\mathrm{B}^s)^D(1-\mathrm{B})^d x_t = \Phi_Q(\mathrm{B}^s)\phi_q(\mathrm{B})w_t$$

Chapitre 8

Extended Linear Models

Thissection resume the Chapters 5-9 of *An Introduction to Generalized Linear Models.*
Add to biblio :

8.1 Inference

The two tool to do inference are **confidence intervals** and **hypothesis tests**. For GLM, a hypothesis tests can be use to compare two models, but their need to have the same probability function and the same link. Also, the null hypothesis H_0 is a sinpler model and must be a special case of the other.

8.1.1 Sampling distribution for the score statistic

Score Function Let $\ell = \ln f(y)$ be the log-likehood function, then the score function U, is define as

$$U_j = \frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left[\frac{(Y_i - \mu_i)}{\text{Var}\,(Y_i)} x_{i,j} \left(\frac{\partial \mu_i}{\partial \eta} \right) \right]$$

Information matrix The information matrix is define as the variance-covariance matrix of the score function. This information matrix is defined as

$$I(\theta) = \text{Var}\,(U)$$

Score Statistic If there is only one β, the score statistic has the asymptotic sampling distribution

$$\frac{U}{\sqrt{I(\theta)}} \sim N(0,1), \frac{U^2}{I(\theta)} \sim \chi^2(1)$$

If there is a vector of $\underline{\beta}$

$$U^T I(\theta)^{-1} U \sum \chi^2(p)$$

8.2 Normal Linear Models

8.2.1 Basis Result

Maximum likelihood The maximum likelihood estimation of β is given by

$$\mathrm{b} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}$$

The estimator is unbiasedm with the variance-covariance matrix

$$I(\theta)^{-1} = \sigma^2(\mathrm{X}^T\mathrm{X}^{-1})$$

However, the unbiased estimator of σ^2 is given by

$$\hat{\sigma^2} = \frac{1}{N-p}(\mathrm{y} - \mathrm{Xb})^T$$

Least Square Estimation In the case of linear models, we obtain the same result as the maximum likelihood

$$\mathrm{b} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}$$

Deviance The diviance is define by the square of the error ε.

$$\frac{1}{\sigma^2}\varepsilon^T\varepsilon = \frac{1}{\sigma^2}(\mathrm{y} - \mathrm{Xb})^T(\mathrm{y} - \mathrm{Xb})$$

or, in case of simple linear model,

$$\frac{1}{\sigma^2}\sum(Y_i - \hat{Y}_i)^2$$

Chapitre 9

Statistical Learning

reference statistical learning (voir mas1)
This section do not cover all the formulas talk in the book since we already seen it in *Modèle Linéaire en actuariat*. This section talk more about the analysis of some statistical model.

9.1 Statistical Learning

Prediction

$$\begin{aligned} \mathrm{E}\left[Y-\hat{Y}\right] &= \mathrm{E}\left[f(x)+\varepsilon-\hat{f}(x)\right]^2 \\ &= [f(x)-\hat{f}(x)]^2+\mathrm{Var}\left(\varepsilon\right) \\ &= (\text{Réductible})+(\text{Irreductible}) \end{aligned}$$

Inference We are often interested in understanding the way that Y is affected as $X_1, ..., X_p$ is changing. Inference mean that we want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, ..., X_p$

- Which predictors are associated with the response ?
- What is the relationship between the response and each predictor ?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more compli cated ?

9.1.1 How Do We Estimate f ?

Parametric Methods Parametric methods involve a two-step model-based approach.

1. First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X :

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

2. After a model has been selected, we need a procedure that uses the training data to fit or train the model.enumarate

The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f, then our estimate will be poor. We can try to address this problem by choosing flexible models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as overfitting the data.

Non-parametric Methods No assumption about the form of f is made.

- **Advantage :** Fit the model more closly to the data points.
- **Disadvantage :** Since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

9.1.2 Measuring the Quality of Fit

In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{f}(x_i))^2$$

The MSE will be small if the predicted responses are very close to the true responses.
Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE.

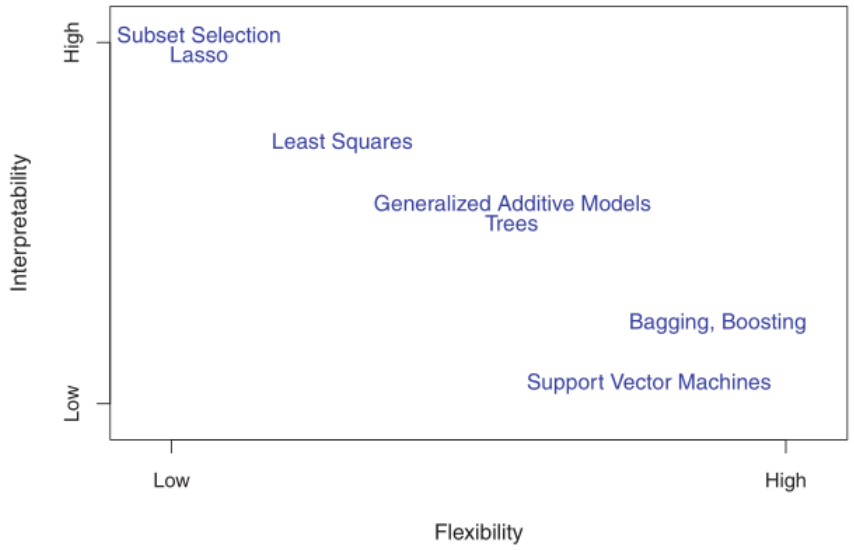

FIGURE 9.1 – A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

Overfitting Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

The Bias-Variance Trade-Off The equation below tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias. Note that variance is inherently a nonnegative quantity, and squared bias is also nonnegative. Hence, we see that the expected test MSE can never lie below Var(?), the irreducible error

$$\mathrm{E}\left[y_0 - \hat{f}(x_0)\right] \mathrm{Var}\left(f(\hat{x}_0)\right) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}\left(\varepsilon\right)$$

- **Variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. In general, more flexible statistical methods have higher variance.
- **Bias** refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. Generally, more flexible methods result in less bias.

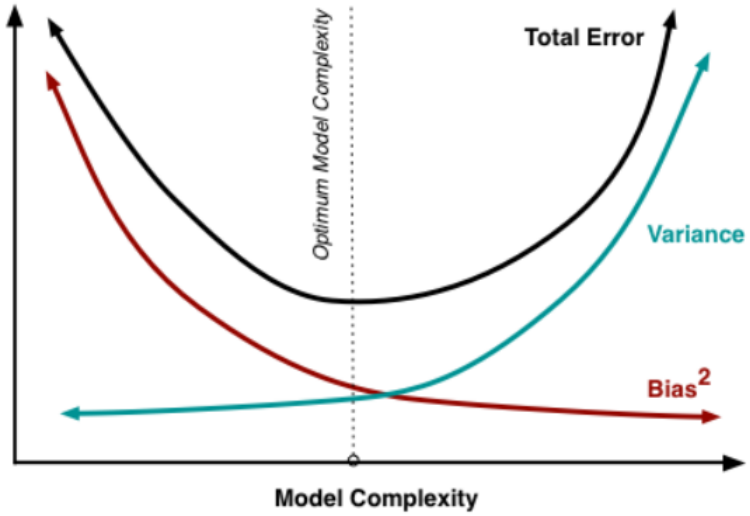

FIGURE 9.2 – Bias-Variance trade-Off

9.2 Linear Regression

9.2.1 Simple Linear Regression

Definition

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where the $\hat{\beta}$ are estimate using the **least squares** criterion.

Residual sum od squares We define the residual sum of square as

$$\text{RSS} = \text{SSE} = \sum_{i=0}^{n} \varepsilon_i = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

R^2 **Statistic** For simple linear regression, $R^2 = \text{Cov}(X, Y)$.
Backward selection cannot be used if p > n, while forward selection can always be used. Forward selection is a greedy approach, and might include variables early that later become redundant. Mixed selection can remedy this.

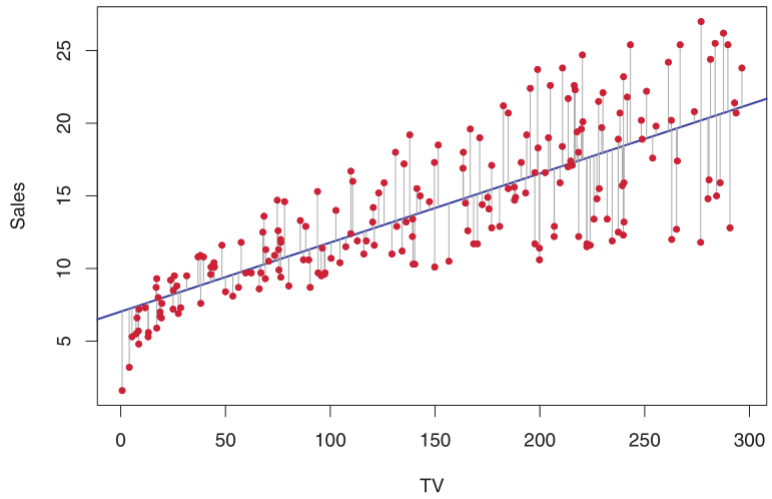

FIGURE 9.3 – Exemple of simple linear regression

9.3 Multiple Linear Regression

9.3.1 Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following :

1. Non-linearity of the Data The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.
Residual plots are a useful graphical tool for identifying non-linearity. If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\ln(X)$, $\sqrt{X}$, and X^2 , in the regression model.

2. Correlation of Error Terms An important assumption of the linear regression model is that the error terms, $\varepsilon_1, ..., \varepsilon_n$ are uncorrelated. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.

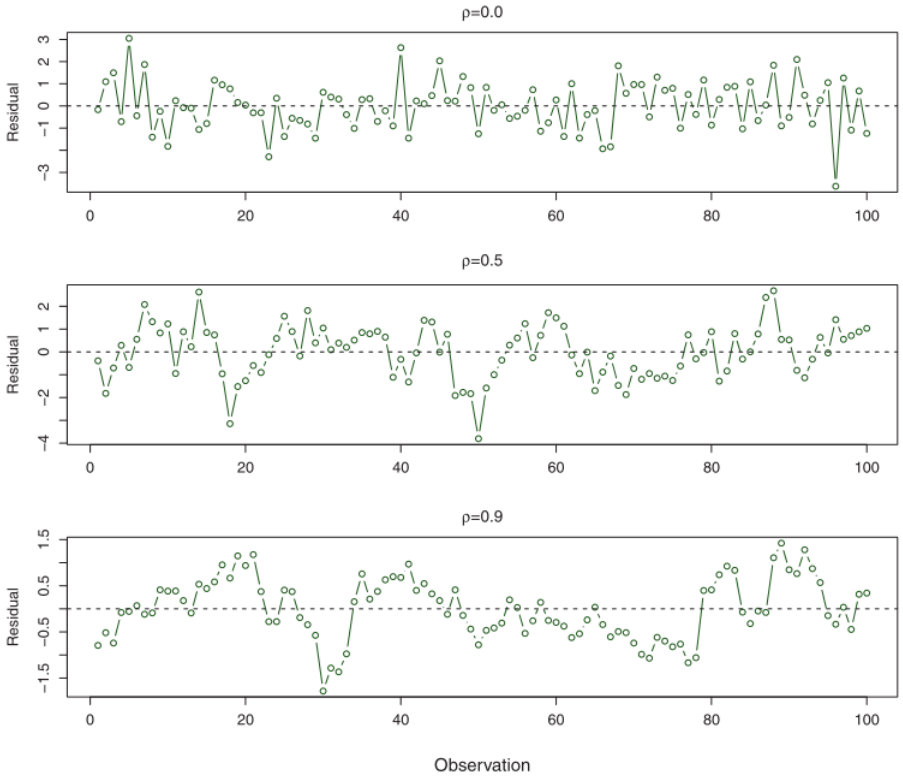

FIGURE 9.4 – Plots of residuals time series data sets with differing levels of correlation ρ between error terms for adjacent time points. The graphs on top is good for linear regression

3. Non-constant Variance of Error Terms Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{Var}(\varepsilon) = \sigma^2$. If not, we can recognize **heteroscedasticity** with a funnel shape in the residual plot.
When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\ln(Y)$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. Sometimes, we can also use **weighted least squares**.

4. Outliers An outlier is a point for which y_i is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect

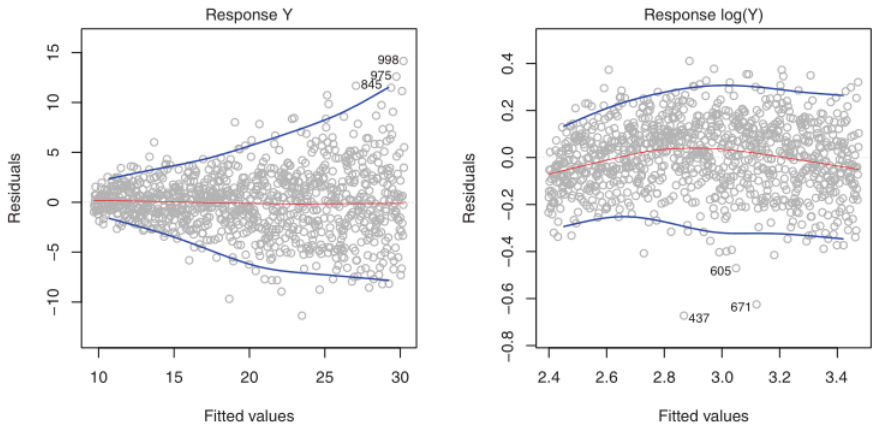

FIGURE 9.5 – Residual plots. Left : The funnel shape indicates heteroscedasticity. Right : The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.

recording of an observation during data collection.

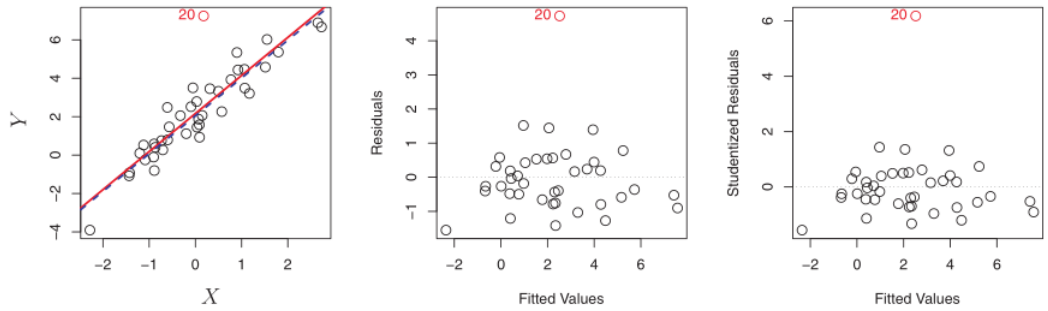

FIGURE 9.6 – Center : The residual plot clearly identifies the outlier. Right : The outlier has a studentized residual of 6 ; typically we expect values between -3 and 3.

Outlier can affect our MSE estimation, resolving in inadequate R^2 or confidence interval. We can remove the outlier to resolve this issue.

5. High Leverage Points In constrast of *outlier* that are unusual y_i, leverage are unusual value of x_i.

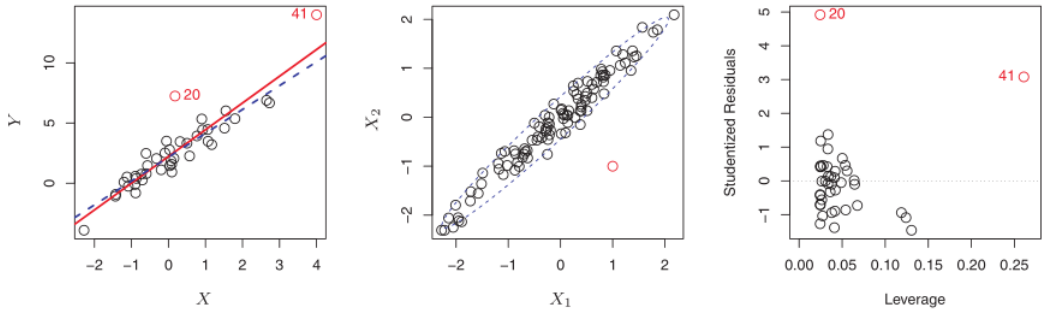

FIGURE 9.7 – Left : Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center : The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right : Observation 41 has a high leverage and a high residual.

For simple linear regression, we can compute the leverage statistic define as

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

where $\frac{1}{n} < h_i < 1$ and $\sum h_i = \frac{(p+1)}{n}$. So if a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, hen we may suspect that the corresponding point has high leverage.

6. Collinearity Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.
The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.

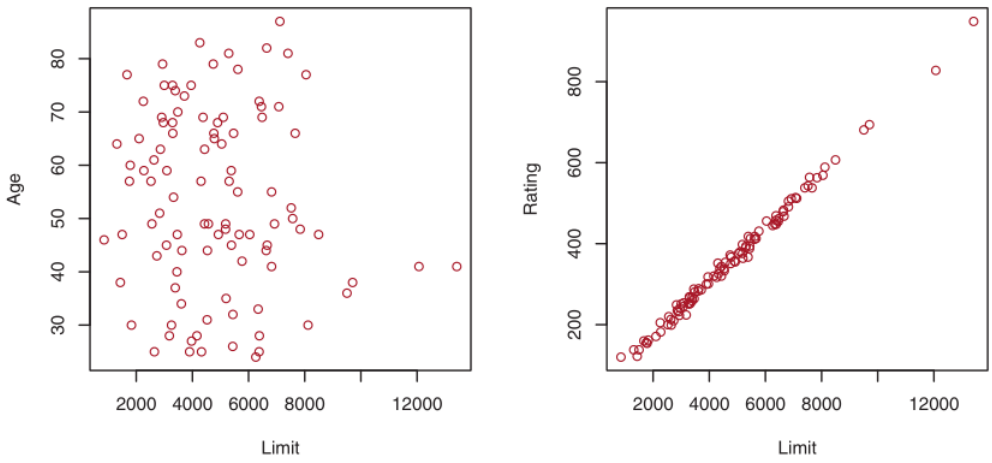

FIGURE 9.8 – Left : A plot of age versus limit. These two variables are not collinear. Right : A plot of rating versus limit. There is high collinearity.

9.4 Classification

9.4.1 Logistic Regression

In logistic regression, we use the logistic function to be sure that the probability of success given you are in group i, π, is between 0 and 1.

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

where $\eta = \hat{\beta}_0 + ... + \hat{\beta}_p X_p$.

Estimating the Regression Coefficients To fit the model, we use the **maximum likelihood** method. In the case of simple linear models, we can compute the estimation with

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} \pi_i \prod_{i:y_i=0} (1 - \pi_i)$$

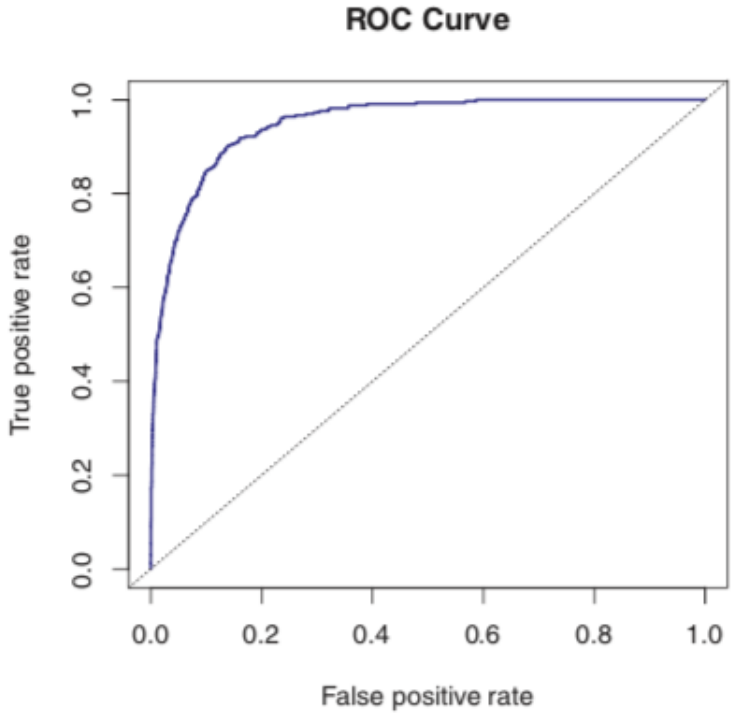

FIGURE 9.9 – ROC curve used to see the fit of a binomial model. We want to maximize the true rate and minimize the false rate. We want the graph to be as much as possible in the top-left corner of the graph.

Cross-Validation on Classification Problems Cross-validation works just as described earlier in this chapter, except that rather than using MSE to quantify test error, we instead use the number of misclassified observations.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=0}^{n} \mathrm{Err}_i$$

where $\mathrm{Err}_i = \mathbb{1}_{\{y_i \neq \hat{y}_i\}}$.

9.5 Resampling Methods

The process of evaluating a model's performance is known as model assessment, whereas the process of selecting the proper level of flexibility for a model is known as model selection.

9.5.1 Linear Model Selection and Regularization

We discuss in this chapter some ways in which the simple linear model can be improved with better prediction accuracy and model interpretability. We achieve this by replacing plain least squares fitting with some alternative fitting procedures

- **Prediction Accuracy** If n is not much larger than p, then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training. And if p > n, then there is no longer a unique least squares coefficient estimate : the variance is infinite so the method cannot be used at all. By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias.
- **Interpretability** It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables we can obtain a model that is more easily interpreted.

There are many alternatives, both classical and modern, to using least squares to fit :

- Subset Selection.
- Shrinkage.
- Dimension Reduction.

9.5.2 Subset Selection

Best Subset Selection In general, there are 2p models that involve subsets of p predictors.
We need to be careful because SSE of these $p+1$ models decreases monotonically, and the R2 increases monotonically, as the number of features included in the models increases. Consequently, best sub- set selection becomes computationally infeasible for values of p greater than around 40

Forward Stepwise Selection Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward stepwise selection involves fitting

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$$

Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$, but in this case, it only possible to fit $M_0, ..., M_{n-1}$ models.

Backward Stepwise Selection Backward stepwise selection begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

9.5.3 Choosing the Optimal Model

RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

C_p For a fitted least squares model containing d predictors, the C_p estimate of **test MSE** is computed using the equation

$$C_p = \frac{1}{n}(SSE + 2d\hat{sigma}^2)$$

AIC The AIC criterion is defined for a large class of models fit by maximum likelihood.

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(SSE + 2d\hat{\sigma}^2)$$

BIC BIC is derived from a Bayesian point of view.

$$\frac{1}{n}(SSE + \ln(n)d\hat{\sigma}^2)$$

Since $\ln(n) > 2$ for any $n > 7$, the BIC give a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Adjusted R^2 For a least squares model with d variables, the adjusted R^2 statistic is calculated as

$$R_a^2 = 1 - \frac{SSE/(n-d-1)}{SST/(n-1)}$$

9.5.4 Shrinkage Methods

As an alternative of using the least squares to fit a linear model, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates. shrinking the coefficient estimates can significantly **reduce their variance**.

Ridge Regression

The ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda \geq 0$ is a tunning parameter to be determined separately.
The notation $||\beta||_2$ denotes de ℓ_2 norm of a vector and is defined as

$$\sqrt{\sum_{j=1}^{n}\beta_j^2}$$

It measures the distance of β from zero. As λ decrease, the ℓ_2 norm of $\hat{\beta}_\lambda^R$ will **always** decrease, and so will $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$. We can see this amount as the amount that the ridge regression coefficient estimates have been shrunken towards zero.

Note

The standard least squares coefficient estimate are scale invariant : multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant. Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

Why Does Ridge Regression Improve Over Least Squares ? As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

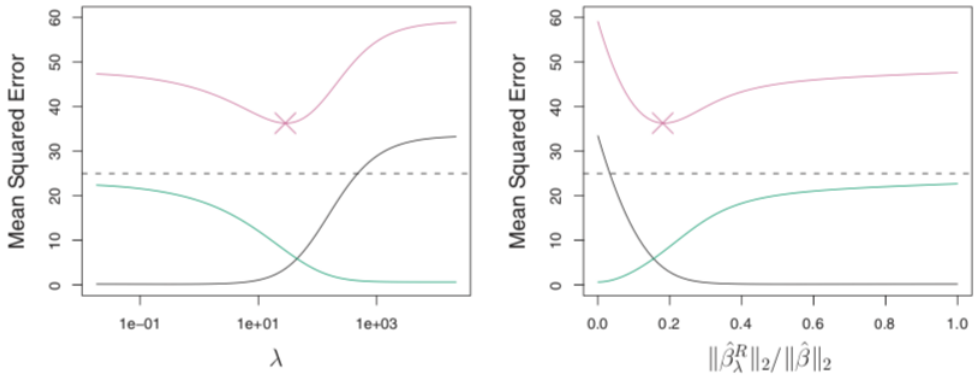

FIGURE 9.10 – Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

Lasso Regression

The advantage of Lasso is that the $\hat{\beta}$ can equal zero, while ridge will include all p parameters. The Lasso regression minimize

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\text{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

where $\lambda \geq 0$ is a tunning parameter to be determined separately. The notation $||\beta||_1$ denotes de ℓ_1 norm of a vector and is defined as

$$||\beta||_1=\sum|\beta_j|$$

Another Formulation for Ridge Regression and the Lasso

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}\quad\text{subject to}\quad\sum_{j=1}^{p}|\beta_j|\leq s$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

If s is small, the estimate of *hatbeta* will goes tzero. In contrast, if s is large enough that the least squares solution falls within the budget, then the least squares estimate will be generated.

Ridge versus Lasso

Neither ridge regression nor the lasso will universally dominate the other. Ridge regression more or less shrinks every dimension of the data by the same proportion, whereas the lasso more or less shrinks all coefficients toward zero by a similar amount, and sufficiently small co- efficients are shrunken all the way to zero. soft- thresholding

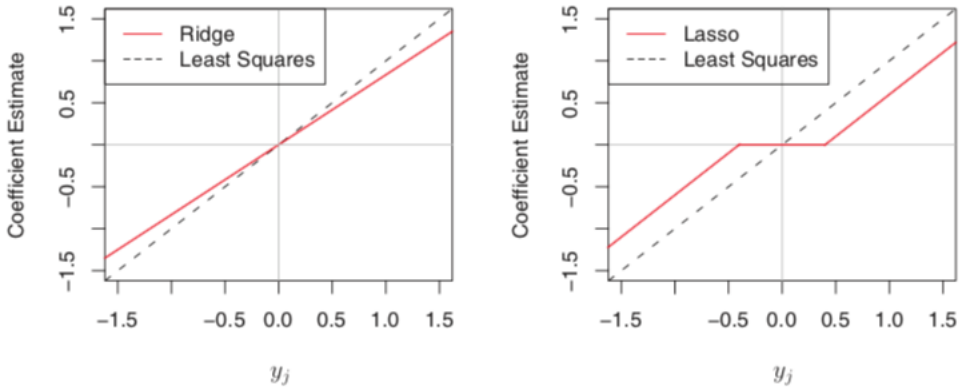

FIGURE 9.11 – Left : The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates. Right : The lasso coefficient estimates are soft-thresholded towards zero.

9.5.5 Dimension Reduction Methods

We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables.

Definition

Let $Z_1, ..., Z_M$ represent $M < p$ linear combinations of our original p predictor.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for some constants $\phi_{1m}, ..., \phi_{pm}$. We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1} \theta_m z_{im} + \varepsilon \quad i = i, ..., n$$

using least squares.

Note

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

This constraint on the form of the coefficients has the potential to bias the coefficient estimates. However, in situations where p is large relative to n, selecting a value of $M << p$ can significantly reduce the variance of the fitted coefficients.

Principal Components Analysis

We want to combine p variables into only one Z_1, that will capture the maximum of the **variability**. The first **component score** for the observation i is given by

$$z_{1,i} = \phi_{11}(x_1 - \bar{x}_2) + \phi_{12}(x_2 - \bar{x}_2)$$

under the constraint that the norm of the **component loading** equal unity

$$\phi_{i,1}^2 + \phi_{i,2}^2 = 1$$

The second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_2, and has largest variance subject to this constraint. Since Z_1 and Z_2 are orthogonal, the following equation must be true

$$\phi_{11}\phi_{12} + \phi_{21}\phi_{22} = 0$$

Note

Even though PCR provides a simple way to perform regression using $M < p$ predictors, it is not a feature selection method. In fact, one can show that PCR and ridge regression are very closely related.

Choosing the number of M parameters In PCR, the number of principal components, M, is typically chosen by cross-validation.

Standardized Predictor When performing PCR, we generally recommend standardizing each predictor. In the absence of standardization, the high-variance variables will tend to play a larger role in the principal components obtained, and the scale on which the variables are measured will ultimately have an effect on the final PCR model.

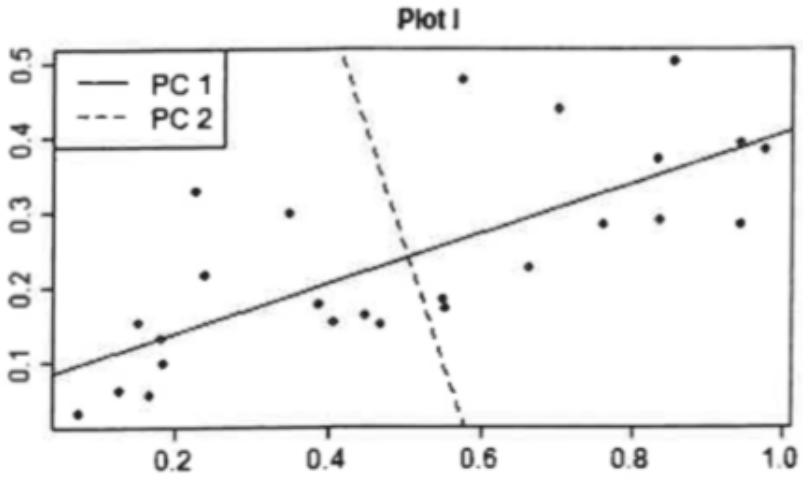

FIGURE 9.12 – Illustration of the first two principal components

Partial Least Squares

PCA was a unsupervised method of reducing dimension. Partial least square is a supervised method because it the response Y is involve. Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

Finding the First Direction After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{j1} equal to the coefficient from simple linear regression of Y in function of X_j. Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{j,1} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.

Finding the Second Direction To identify the second PLS direction we first adjust each of the variables for Z_1, by regressing each variable on Z_1 and taking residuals. These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction. We then compute Z_2 using this orthogonalized data in exactly the same fashion as Z_1 was computed based on the original data.

9.6 Moving Beyond Linearity

Linear model have advantage over other approaches in terms of interpretation and inference. However, their have significant limitations in terms of predictive power.

9.6.1 Polynomial Regression

A polynomial regression is define as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \varepsilon_i$$

where β can be estimated using least squares.

Note

We usually don't use d greater than 3 or 4 because otherwise the model can become over flexible and take strange shapes near boundary of the X variable.

9.6.2 Step Functions

To create a step function model we create cutpoints $c_1, ..., c_k$ in the range of X, and then construct $K + 1$ new dummy variable.

$$\begin{aligned} c_0(X) &= \mathbb{1}_{\{X<c_1\}} \\ c_1(X) &= \mathbb{1}_{\{c_1\leq X<c_2\}} \\ c_2(X) &= \mathbb{1}_{\{c_2\leq X<c_3\}} \\ &\vdots \\ c_{K-1}(X) &= \mathbb{1}_{\{c_{K-1}\leq X<c_K\}} \\ c_K(X) &= \mathbb{1}_{\{c_K\leq X\}} \end{aligned}$$

We can then define a step functions model is define as

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \varepsilon_i$$

where $C_i(X)$ are dummy variable.

9.6.3 Piecewise Polynomials

Instead of fitting a high-degree polynomial over the entire range of X, piecewise polynomial regression involves fitting separate low-degree polynomials over different regions of X.

Example A piecewise cubic polynomial with a single knot at a point c takes the form

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases}$$

In general, if we place K different knots throughout the range of X, then we will end up fitting K + 1 different cubic polynomials. In fact, our piecewise constant functions (step function) are piecewise polynomials of degree 0 !

9.6.4 Spline Model

We define a d spline with K knots as

$$y_i = \beta_0 + \beta_1 X + ... + \beta_d X^d + \beta_{d+1} h(X, \zeta_1) + ... + \beta_{d+K} h(X, \zeta_K) + \varepsilon$$

where

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

Then, a spline model have $1 + d + K$ degrees of freedom. Unfortunately, splines can have high variance at the outer range of the predictors

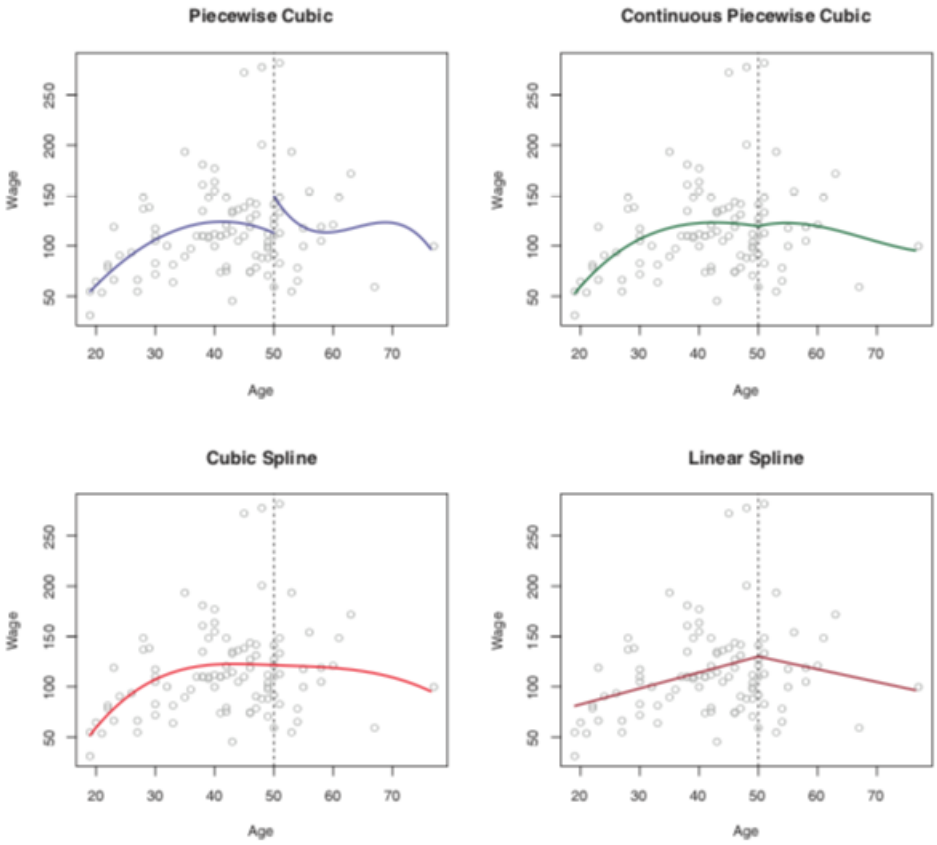

FIGURE 9.13 – Four different model fit in the same data

Natural Spline

A natural spline is a regression spline with additional boundary constraints : the function is required to be linear at the boundary (in the region where X is smaller than the smallest knot, or larger than the largest knot). The number of df is then reduce by 4 in the case of natural spline $(1+d+K-4)$.

Choosing the Location of the Knots The regression spline is most flexible in regions that contain a lot of knots, because in those regions the polynomial coefficients can change rapidly. Hence, one option is to place more knots in places where we feel the function might vary most rapidly, and to place fewer knots where it seems more stable. While this option can work well, in practice it is common to place knots in a uniform fashion.

Choosing the Number of Knots We also use cross-validation to select the best degree of freedoms.

9.6.5 Smoothing Splines

The function g that minimizes the equation below is known as a smoothing spline.

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where λ is a nonnegative tuning parameter. The penalty $\lambda \int g''(t)^2 dt$ encourages $g(x)$ to be smooth. The larger the value of λ, the smoother $g(x)$ will be. When $\lambda = 0$, then the penalty term will have no effect, and so the function $g(x)$ will be very jumpy and will exactly interpolate the training observation. When $\lambda \to \infty$, $g(x)$ will be perfectly smooth (i.e. a linear least squares line).

Note

The function $g(x)$ that result from the equation above is a **natural cubic spline with knots at** $x_1, ..., x_n$. However, it is not the same natural cubic spline that one would get if one applied the basis function approach described in Spline Model section with knots at $x_1, ..., x_n$, rather, it is a shrunken version of such a natural cubic spline, where the value of the tuning parameter λ controls the level of shrinkage.

Choosing the Smoothing Parameter λ It is possible to show that as λ increases from 0 to ∞, the effective degrees of freedom, which we write df_λ, decrease from n to 2.

9.6.6 Local Regression

Local regression is a different approach for fitting flexible non-linear functions, which involves computing the fit at a target point x_0 using only the nearby training observations.

Algorithm 7.1 *Local Regression At $X = x_0$*

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.
3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

The smaller the value of s, the more local and wiggly will be our fit ; alternatively, a very large value of s will lead to a global fit to the data using all of the training observations.

9.6.7 Generalized Additive Models

In previous section, we present a number of approaches for flexibly predicting a response Y on the basis of a single predictor X. These approaches can be seen as extensions of simple linear regression. Here we explore the problem of flexibly predicting Y on the basis of several predictors, $X_1, ..., X_P$.

$$y_i = \beta_0 + f_1(x_{i,1}) + ... + f_p(x_{i,p}) + \varepsilon_i$$

where $f_j(x_{i,j})$ can be any model described in previous section. It is called **additive** because we calculate a separate f_j for each X_j and then add together

all of their contributions. We can also fit interaction with $f_{i,j}(X_j, X_k)$ with two-dimension model such as local regression.

Advantages

- GAMs allow us to fit a non-linear f_i to each X_j, so that we can automatically model non-linear relationships.
- The non-linear fits can potentially make more accurate predictions for the response Y .
- Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed. Hence if we are interested in inference, GAMs provide a useful representation.
- The smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom.

Disadvantage

- The main limitation of GAMs is that the model is restricted to be additive.

Backfitting This method fits a model involving multiple predictors by repeatedly updating the fit for each predictor in turn, holding the others fixed. The beauty of this approach is that each time we update a function, we simply apply the fitting method for that variable to a partial residual.

Note

A partial residual for X_3, for exemple, has the form $r_i = y_i - f_1(x_{i,1}) - f_2(x_{i,2})$. If we know f_1 and f_2, then we can fit f_3 by treating this residual as a response in a non-linear regression on X_3.

Annexe A

Principales distribution de probabilité utilisées

introduction

Annexe B

Résultats (et démonstrations) utiles

B.1 Stop-Loss ($\pi_X(d)$)

Dans un contexte continu,

$$\pi_X(d) = \int_d^\infty \overline{F}(d)du$$

Démonstration.

$$\begin{aligned}
\pi_X(d) &= E[\max(X-d,0)] \\
&= \int_0^\infty \max(x-d,0)F_X(x)dx \\
&= \int_0^\infty (x-d)1_{\{X>d\}}f_X(x)dx \\
&= \int_d^\infty (x-d)f_X(x)dx \\
&= \int_d^\infty xf_X(x)dx - \int_d^\infty df_X(x)dx
\end{aligned}$$

On doit alors faire une intégration par partie, en posant

$$\begin{aligned}
u &= x & du &= dx \\
dv &= dF_X(x) & v &= -S(x)
\end{aligned}$$

Note : si on fait tendre $S(x)$ vers l'infini, ça va tendre plus rapidement vers 0 que x seul.

$$\begin{aligned}\pi_X(d) &= -xS(x)\Big|_d^\infty - \int_d^\infty -S(x)dx - d\Big(F(\infty) - F(d)\Big) \\ &= 0 + \cancel{dS(d)} + \int_d^\infty S(x)dx - \cancel{dS(d)} \\ &= \int_d^\infty S(x)dx\end{aligned}$$

□

Il existe aussi le contexte discret :

$$\pi_X(d) = \sum_{k=d}^{\infty} S(k)$$

Démonstration.

$$\begin{aligned}\pi_X(k) &= E[\max(N-k), 0)] \\ &= \sum_{j=k}^{\infty}(j-k)P(N=j) \\ &= (k-k)P(N=k) + ((k+1)-k)P(N=k+1) + P((k+2)-k)P(N=k+2) + ... \\ &= P(N=k+1) + 2P(N=k+2) + 3P(N=k+3) + ... \\ &= \underbrace{(P(N=k+1) + P(N=k+2) + P(N=k+3) + ...)}_{S(k)} \\ &+ \underbrace{(P(N=k+2) + P(N=k+3) + P(N=k+4) + ...)}_{S(k+1)} \\ &+ \underbrace{(P(N=k+3) + (PN=k+4) + P(N=k+5) + ...)}_{S(k+2)} \\ &+ ... \\ &= \sum_{i=k}^{\infty} S(i)\end{aligned}$$

□

B.2 TVaR

B.2.1 Les 3 formes explicites de la $TVaR$

Pour la $TVaR$, il y a 3 preuves à bien connaître :

$$TVaR_\kappa(X) = \frac{1}{1-\kappa}\pi_X(VaR_\kappa(X)) + VaR_\kappa(X)$$

Démonstration.

$$\begin{aligned}
TvaR_\kappa(X) &= \frac{1}{1-\kappa}\int_\kappa^1 VaR_u(X)du \\
&= \frac{1}{1-\kappa}\int_\kappa^1 (VaR_u(X) - VaR_\kappa(X) + VaR_\kappa(X))du \\
&= \frac{1}{1-\kappa}\int_\kappa^1 (\underbrace{VaR_u(X)}_{\substack{\text{fonction}\\ \text{quantile}}} - VaR_\kappa(X))du + \underbrace{\int_\kappa^1 VaR_\kappa(X)du}_{\text{intégration d'une constante}} \\
&= \frac{1}{1-\kappa}\int_\kappa^1 (F_X^{-1}(u) - VaR_\kappa(X)) \underbrace{f_U(u)}_{U\backsim Unif(0,1)} du + \frac{1}{\cancel{1-\kappa}} VaR_\kappa(X)(\cancel{1-\kappa}) \\
&= \frac{1}{1-\kappa}E[\max(\underbrace{F_X^{-1}(U)}_{F_X^{-1}\backsim X} - VaR_\kappa(X); 0)] + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}E[\max(X - VaR_\kappa(X); 0)] + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}\pi_X(VaR_\kappa(X)) + VaR_\kappa(X)
\end{aligned}$$

□

à partir de la preuve ci-dessus, on peut démontrer celle-ci :

$$TVaR_\kappa(X) = \frac{E[X \times 1_{\{X > VaR_\kappa(X)\}}] + VaR_\kappa(X)(F_X(VaR_\kappa(X)) - \kappa)}{1-\kappa}$$

Démonstration.

$$
\begin{aligned}
TVaR_\kappa(X) &= \frac{1}{1-\kappa}\pi_X(VaR_\kappa(X)) + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}E[\max(X - VaR_\kappa(X); 0)] + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}E[(X - VaR_\kappa(X)) \times 1_{\{X>VaR_\kappa(X)\}}] + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}E[X \times 1_{\{X>VaR_\kappa(X)\}}] - \frac{1}{1-\kappa}E[VaR_\kappa(X) \times \underbrace{1_{\{X>VaR_\kappa(X)\}}}_{=S_X(VaR_\kappa(X))}] + VaR_\kappa(X) \\
&= \frac{1}{1-\kappa}E[X \times 1_{\{X>VaR_\kappa(X)\}}] - \frac{1}{1-\kappa}VaR_\kappa(X)(1 - F_X(VaR_\kappa(X))) + \frac{1-\kappa}{1-\kappa}VaR_\kappa(X) \\
&= \frac{E[X \times 1_{\{X>VaR_\kappa(X)\}}] + VaR_\kappa(X)(-1 + F_X(VaR_\kappa(X)) + 1 - \kappa)}{1-\kappa} \\
&= \frac{E[X \times 1_{\{X>VaR_\kappa(X)\}}] + VaR_\kappa(X)(F_X(VaR_\kappa(X)) - \kappa)}{1-\kappa}
\end{aligned}
$$

□

Une dernière preuve fortement utilisée pour la $TVaR$, qui découle directement de la dernière : :

$$TVaR_\kappa(X) = \frac{E[X \times 1_{\{X>VaR_\kappa(X)\}}]}{1-\kappa}$$

Démonstration. Étant donné que cette formule ne fonctionne seulement que pour une v.a. continue, elle est très facile à prouver :

$$\text{si } X \text{ est continue}, \quad \forall x, F_X(VaR_\kappa(X))) = \kappa$$

Alors, on peut enlever la partie de droite de l'équation. □

B.3 Sous-additivité de la $TVaR$

Il y a plusieurs façons de prouver la sous-additivité de la $TVaR$.

B.3.1 À l'aide de la fonction convexe $\varphi(x)$

On sait que la fonction $\varphi(x)$ est convexe :

$$\varphi(x) = x + \frac{1}{1-\kappa}\pi_X(x)$$

Et on sait aussi que

$$TVaR_\kappa(X) = \inf\left\{\varphi(x)\right\}$$

Il faut prouver que $TVaR_\kappa(X+Y) \leq TVaR_\kappa(X) + TVaR_\kappa(Y)$

Démonstration. Puisque $\varphi(x)$ est une fonction convexe, on peut dire que

$$\begin{aligned} TVaR_\kappa(X) &\leq \varphi(x) \\ &\leq x + \frac{1}{1-\kappa}\pi_X(x) \end{aligned}$$

On pose le changement de variable $\boxed{X^* = \alpha X + (1-\alpha)Y}$

On peut donc remplacer x dans $\varphi(x)$ par

$$\begin{aligned} x_0 &= VaR_\kappa(X^*) \\ &= VaR_\kappa(\alpha X + (1-\alpha)Y) \\ &= \alpha VaR_\kappa(X) + (1-\alpha)VaR_\kappa(Y) \end{aligned}$$

Alors,

$$
\begin{aligned}
TVaR_\kappa(\alpha X + (1-\alpha)Y) &\leq \alpha VaR_\kappa(X) + (1-\alpha)VaR_\kappa(Y) \\
&+ \frac{1}{1-\kappa} E[\max(\alpha X + (1-\alpha)Y - \alpha VaR_\kappa(X) - (1-\alpha)VaR_\kappa(Y); 0)] \\
&= \alpha VaR_\kappa(X) + (1-\alpha)VaR_\kappa(Y) \\
&+ \frac{1}{1-\kappa} E[\max(\alpha(X - VaR_\kappa(X)) + (1-\alpha)(Y - VaR_\kappa(Y)); 0)] \\
&\leq \alpha VaR_\kappa(X) + (1-\alpha)VaR_\kappa(Y) \\
&+ \alpha \left(\frac{1}{1-\kappa} E[\max(X - VaR_\kappa(X); 0)] \right) \\
&+ (1-\alpha) \left(\frac{1}{1-\kappa} E[\max(X - VaR_\kappa(X); 0)] \right) \\
&\text{Si on met en commun, on retrouve les expressions de la } TVaR \\
&= \alpha \left(\frac{1}{1-\kappa} \pi_X(VaR_\kappa(X)) + VaR_\kappa(X) \right) \\
&+ (1-\alpha) \left(\frac{1}{1-\kappa} \pi_Y(VaR_\kappa(Y)) + VaR_\kappa(Y) \right) \\
TVaR_\kappa(\alpha X + (1-\alpha)Y) &\leq \alpha TVaR_\kappa(X) + (1-\alpha)TVaR_\kappa(Y)
\end{aligned}
$$

La relation se vérifie très bien avec le cas où $\alpha = 0,5$:

$$
\begin{aligned}
TVaR_\kappa(0,5X + (1-0,5)Y) &\leq 0,5TVaR_\kappa(X) + (1-0,5)TVaR_\kappa(Y) \\
0,5TVaR_\kappa(X+Y) &\leq 0,5TVaR_\kappa(X) + 0,5TVaR_\kappa(Y) \\
&\text{on multiplie par 2 pour enlever les 0,5} \\
\mathbf{TVaR_\kappa(X+Y)} &\leq \mathbf{TVaR_\kappa(X) + TVaR_\kappa(Y)}
\end{aligned}
$$

□

B.3.2 Avec les fonctions indicatrices

Si on a les v.a. continues X et Y (les espérances existent) avec les fonctions de répartition respectives F_X et F_Y, alors

$$TVaR_\kappa(X) = \frac{E[X \times 1_{\{X>VaR_\kappa(X)\}}]}{1-\kappa}$$
$$(1-\kappa)TVaR_\kappa(X) = E[X \times 1_{\{X>VaR_\kappa(X)\}}]$$

est valide pour toute v.a. continue X.

On veut alors démontrer que

$$TVaR_\kappa(X) + TVaR_\kappa(Y) - TVaR_\kappa(X+Y) \geq 0 \tag{B.1}$$

Démonstration. .

(1) On peut écrire le membre de gauche de l'inégalité (B.1) comme

$$\underbrace{(1-\kappa)TVaR_\kappa(X)}_{E[X\times 1_{\{X>VaR_\kappa(X)\}}]} + \underbrace{(1-\kappa)TVaR_\kappa(Y)}_{E[Y\times 1_{\{Y>VaR_\kappa(Y)\}}]} - \underbrace{(1-\kappa)TVaR_\kappa(X+Y)}_{E[(X+Y)\times 1_{\{X+Y>VaR_\kappa(X+Y)\}}]}$$
$$= E[X \times 1_{\{X>VaR_\kappa(X)\}}] + E[Y \times 1_{\{Y>VaR_\kappa(Y)\}}] - \underbrace{E[(X+Y) \times 1_{\{X+Y>VaR_\kappa(X+Y)\}}]}_{\text{On } \textit{split} \text{ cette espérance}}$$
$$= \underbrace{E[X \times 1_{\{X>VaR_\kappa(X)\}}] - E[X \times 1_{\{X+Y>VaR_\kappa(X+Y)\}}]}_{\text{On peut rassembler les indicatrices}}$$
$$+ \underbrace{E[Y \times 1_{\{Y>VaR_\kappa(Y)\}}] - E[Y \times 1_{\{X+Y>VaR_\kappa(X+Y)\}}]}_{\text{ici aussi}}$$

$$= E[X\times(1_{\{X>VaR_\kappa(X)\}}-1_{\{X+Y>VaR_\kappa(X+Y)\}})]+E[Y\times(1_{\{Y>VaR_\kappa(Y)\}}-1_{\{X+Y>VaR_\kappa(X+Y)\}})] \tag{B.2}$$

(2) Rendu ici, on veut prouver que chacun de ces espérance ≥ 0, pour que la somme des 2 soit ≥ 0 aussi. Étant donné que les 2 parties du membre de gauche sont identiques, on va le prouver seulement pour un côté.

(2.1) Pour nous aider, on va créer un terme *auxiliaire*, i.e un terme qui est égal à zéro, mais qui va nous aider à faire la preuve, soit le terme suivant :

$$\begin{aligned}
&E[VaR_\kappa(X) \times (1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}})] \\
&= VaR_\kappa(X) E[(1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}})] \\
&= VaR_\kappa(X)\Big((1-\kappa) - (1-\kappa)\Big) \\
&= 0
\end{aligned}$$

(2.2) Alors, l'équation (B.2) devient

$$E[(X - VaR_\kappa(X)) \times (1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}})]$$

(2.3) On va prouver que la quantité à l'intérieur de l'espérance ci-haut sera toujours ≥ 0, de sorte que l'espérance sera toujours positive aussi :

$$\begin{aligned}
&(X - VaR_\kappa(X))(1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}}) \geq 0 \text{ si} \quad X < VaR_\kappa(X) \\
&(X - VaR_\kappa(X))(1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}}) = 0 \text{ si} \quad X = VaR_\kappa(X) \\
&(X - VaR_\kappa(X))(1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}}) \geq 0 \text{ si} \quad X > VaR_\kappa(X)
\end{aligned}$$

(2.4) Alors, on déduit que

$$\begin{aligned}
&E[(X - VaR_\kappa(X)) \times (1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}})] \\
&= E[X \times (1_{\{X>VaR_\kappa(X)\}} - 1_{\{X+Y>VaR_\kappa(X+Y)\}})] \geq 0
\end{aligned}$$

(3) Par conséquent,

$$\begin{aligned}
(1-\kappa)TVaR_\kappa(X) + (1-\kappa)TVaR_\kappa(Y) - (1-\kappa)TVaR_\kappa(X+Y) &\geq 0 \\
\mathbf{TVaR}_\kappa(\mathbf{X}) + \mathbf{TVaR}_\kappa(\mathbf{Y}) - \mathbf{TVaR}_\kappa(\mathbf{X}+\mathbf{Y}) &\geq \mathbf{0}
\end{aligned}$$

□

B.3.3 À l'aide des statistiques d'ordre

Pour pouvoir prouver la sous-additivité de la $TVaR$, on peut aussi utiliser la relation des statistiques d'ordre [1] :

$$TVaR_\kappa(X) = \lim_{n\to\infty} \frac{\sum_{j=[n\kappa]+1}^{n} X_{j:n}}{[n(1-\kappa)]}$$

à compléter plus tard

B.4 Loi des grands nombres

Cette preuve était demandée à l'examen Intra traditionnel H2017 du cours ACT-2001.

Théorème

Soit les v.a. *iid* $X_1, ..., X_n$ avec $E[X^m] < \infty \quad , m = 1, 2, ...$ et $Var(X) < \infty$. Alors,

$$\lim_{n\to\infty} F_{W_n}(x) \longrightarrow F_Z(x) \tag{B.3}$$

où Z est une v.a. tel que $P(Z = E[X]) = 1$.

Démonstration. (1) Première étape, on va démontrer que $\lim_{n\to\infty} \mathcal{L}_{w_n}(t) \to \mathcal{L}_Z(t)$

(1.1) On sait que $\mathcal{L}_{w_n}(t) = \mathcal{L}_X\left(\frac{t}{n}\right)^n \quad n = 1, 2, ...$

(1.2) Soit une v.a. Y positive. On fixe t tout petit

(1.3) Alors

$$\begin{aligned}\mathcal{L}_Y(t) &= E[e^{-tY}] \\ &\approx E[1 - tY] \quad \text{par dév. de Taylor} \quad = E[1] - tE[Y]\end{aligned}$$

1. Relation qui d'ailleurs est utilisée dans le contexte de simulation Monte Carlo pour estimer la TVaR d'une variable aléatoire.

(1.4)

$$\begin{aligned}\mathcal{L}_{w_n}(t) &= \mathcal{L}_X\left(\frac{t}{n}\right)^n \\ &\simeq \left(1-\frac{t}{n}E[X]\right)^n\end{aligned}$$

(1.5) On prends la limite de part et d'autre de l'égalité en (1.3)

$$\begin{aligned}\lim_{n\to\infty}\mathcal{L}_{w_n}(t) &= \lim_{n\to\infty}\left(\mathcal{L}_X\left(\frac{t}{n}\right)\right)^n \\ &\simeq \lim_{n\to\infty}\left(1-\frac{t}{n}E[X]\right)^n \\ &= e^{-tE[X]} \\ &= \mathcal{L}_Z(t)\end{aligned}$$

Ce qui correspond à la Transformée de la v.a. Z où $P(Z = E[X]) = 1$

(2) On applique le résultat de (1.4)

$$\lim_{n\to\infty}F_{w_n}(x) = F_Z(x), \quad \forall x$$

□

B.5 Somme de v.a. indépendantes d'une loi Poisson Composée

Démonstration. Soit les v.a. indépendantes $X_1, ..., X_n$ où
$X_i \sim PoisComp(\lambda_i; F_{B_i}) \quad , i = 1, ..., n$
Ainsi,

$$\begin{aligned}\mathcal{L}_{X_1}(t) &= \mathcal{P}_{M_i}\left(\mathcal{L}_{B_i}(t)\right) \\ &= e^{\lambda\left(\mathcal{L}_{B_i}(t)-1\right)} \quad , i = 1, 2, ..., n\end{aligned}$$

On peut trouver la transformée de S,

$$\mathcal{L}_S(t) = \prod_{i=1}^{n} \mathcal{L}_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda\left(\mathcal{L}_{B_i}(t)-1\right)} \tag{B.4}$$

Le passage de l'équation (B.4) aux étapes suivantes résulte d'une propriété de la loi de Poisson, i.e.

$$\begin{aligned}
\mathcal{L}_S(t) &= e^{\sum_{i=1}^{n} \lambda_i\left(\mathcal{L}_{B_i}(t)-1\right)} \\
&= e^{\sum_{i=1}^{n} \lambda_i \mathcal{L}_{B_i}(t)-\lambda_i} \\
&= e^{\sum_{i=1}^{n} \lambda_i \mathcal{L}_{B_i}(t)-\sum_{i=1}^{n} \lambda_i} \\
&= e^{\sum_{i=1}^{n} \lambda_i \mathcal{L}_{B_i}(t)-\lambda_S}
\end{aligned}$$

Si on met en évidence le λ_S ...

$$\mathcal{L}_S(t) = e^{\lambda_S\left(\sum_{i=1}^{n} \frac{\lambda_i}{\lambda_S} \mathcal{L}_{B_i}(t)-1\right)} \tag{B.5}$$

Si on pose $c_i = \frac{\lambda_i}{\lambda_S}$, on observe que $0 < c_i < 1$ et que $\sum_{i=1}^{n} c_i = 1$.

On se définit une nouvelle v.a., D, où

$$\mathcal{L}_D(t) = \sum_{i=1}^{n} c_i \mathcal{L}_{B_i}(t) \tag{B.6}$$

Ce qui implique que D obéit à une loi mélange :

$$F_D(x) = \sum_{i=1}^{n} c_1 F_{B_i}(x) \quad , x \geq 0$$

en combinant (B.5) et (B.6), on obtient

$$\mathcal{L}_S(t) = e^{\lambda_S(\mathcal{L}_D(t)-1)} \tag{B.7}$$

On introduit une nouvelle v.a., $N_S \sim Pois(\lambda_S)$ et $P_N(s) = e^{\lambda_S(s-1)}$. Alors, (B.7) devient

$$\mathcal{L}_S(t) = \mathcal{P}_{N_S}(\mathcal{L}_D(t))$$

On peut donc représenter S comme

$$S = \begin{cases} \sum_{k=1}^{N_s} D_k & N_s > 0 \\ 0 & N_s = 0 \end{cases}$$

où $D_k, \quad k = 1, 2, ...$ forme une suite de v.a *iid*, et D et N_s sont indépendants. □

B.6 Théorème d'Euler

Définition Soit une fonction $\phi : \mathbb{R}^n \to \mathbb{R}$ homogène d'ordre n. Alors, pour toute fonction ϕ dérivable partout, on a

$$n\phi(x_1, ..., x_n) = \sum_{i=1}^{n} x_i \frac{\partial \phi(x_1, .., x_n)}{\partial x_i}$$

Démonstration. .

(1) Puisque ϕ est homogène d'ordre n, on a

$$\phi(x_1, ..., x_n) = \lambda^n \phi(x_1, ..., x_n) \tag{B.8}$$

(2) On dérive le terme de gauche de l'équation dans l'équation (B.8) par rapport à λ et on pose $\lambda = 1$.

$$\begin{aligned} \frac{\partial}{\partial \lambda} \lambda^n \phi(x_1, ..., x_n)\Big|_{\lambda=1} &= n\lambda^{n-1} \phi(x_1, ..., x_n)\Big|_{\lambda=1} \\ &= n\phi(x_1, ..., x_n) \end{aligned}$$

(3) On dérive le terme de droite de l'équation dans l'équation (B.8) par rapport à λ et on pose $\lambda = 1$.

$$\begin{aligned} \frac{\partial}{\partial \lambda} \lambda^n \phi(x_1, ..., x_n)\Big|_{\lambda=1} &= \sum_{i=1}^{n} \frac{\partial \phi(x_1, ..., x_n)}{\partial (\lambda x_i)} \frac{\partial (\lambda x_i)}{\partial \lambda}\Big|_{\lambda=1} \\ &= \sum_{i=1}^{n} \frac{\partial (\lambda x_i, ..., \lambda x_n)}{\partial \lambda x_i} x_i\Big|_{\lambda=1} \\ &= \sum_{i=1}^{n} x_i \frac{\partial (x_1, ..., x_n)}{\partial x_i} \end{aligned}$$

(4) On pose (4) = (3), et on obtient le résultat souhaité.

□

B.7 Dérivée de l'écart-type (générale)

Lorsqu'on prouve la contribution $C(X_i)$ pour $\rho(X) = \sqrt{Var(\sum_{i=1}^{n} X_i)}$, on doit dériver l'écart-type... voici le développement complet, avec un exemple où $n = 3$. Ce qui est important de suivre, c'est qu'on cherche ici la contribution de la v.a. X_i : alors, lorsqu'on dérive par rapport à λ_i, ça peut être n'importe quoi le $i : 1, 2, ..., n$.

Rappel d'ACT-1002 Pour les propriétés de la covariance, voir la sous-section **??**.

$$\frac{\partial}{\partial \lambda_i} \sqrt{Var\left(\sum_{i=1}^{n} \lambda_i X_i\right)} = \frac{1}{2} \left(\frac{1}{\sqrt{Var\left(\sum_{i=1}^{n} \lambda_i X_i\right)}} \right) \times$$

$$\frac{\partial}{\partial \lambda_i} \left(\sum_{i=1}^{n} \lambda_i^2 Var(X_i) + \sum_{i=1}^{n} \sum_{k=1, k \neq i}^{n} \lambda_i \lambda_k Cov(X_i, X_k) \right)$$

Explication de la forme générale de la variance

Avec un exemple $n = 3$, il est très facile de comprendre d'où vient la formule générale de la variance (qui est universelle si les X_i sont indépendants ou non).

$$\begin{aligned} Var(X_1 + X_2 + X_3) &= Cov(X_1 + X_2 + X_3, X_1 + X_2 + X_3) \\ &= Cov(X_1, X_1) + Cov(X_1, X_2) + Cov(X_1, X_3) \\ &+ Cov(X_2, X_1) + Cov(X_2, X_2) + Cov(X_2, X_3) \\ &+ Cov(X_3, X_1) + Cov(X_3, X_2) + Cov(X_3, X_3) \end{aligned}$$

On remarque en bleu les variances séparées pour chacun de nos X_i de notre exemple, qu'on va pouvoir rassembler ensemble dans une même somme. On remarque aussi que les covariances sont similaires. On remarque en orange les covariances reliées à X_1, en vert les covariances reliées à X_2 et finalement en violet les covariances qui sont reliées à X_3.

En étant attentif, on remarque qu'on peut sommer ensemble chaque *paquet* de covariance sur tout le support ($n = 3$), sauf la combinaison $Cov(X_i, X_i)$, car celle-ci a été prise pour rassembler les variances ensemble ($Var(X_i) = Cov(X_i, X_i)$).

Alors, on obtient (pour le cas $n = 3$) :

$$Var\left(\sum_{i=1}^{3} X_i\right) = \sum_{i=1}^{3} Var(X_i) + \sum_{i=1}^{3} \sum_{k=1, k\neq i}^{3} Cov(X_i, X_k)$$

Si on développe le la dérivée en rouge seule du reste (en prenant l'exemple

du cas $n-3$ et qu'on dérive par rapport à λ_1), on obtient :

$$\begin{aligned}
\frac{\partial}{\partial\lambda_1}\rho(X_1+X_2+X_3) &= \frac{\partial}{\partial\lambda_1}\Big[\lambda_1^2 Var(X_1)+\lambda_2^2 Var(X_2)+\lambda_3^2 Var(X_3)\\
&\quad+\lambda_1\lambda_2 Cov(X_1,X_2)+\lambda_1\lambda_3 Cov(X_1,X_3)\\
&\quad+\lambda_2\lambda_1 Cov(X_2,X_1)+\lambda_2\lambda_3 Cov(X_2,X_3)\\
&\quad+\lambda_3\lambda_1 Cov(X_3,X_1)+\lambda_3\lambda_2 Cov(X_3,X_2)\Big]\\
&= 2\lambda_1 Var(X_1)\\
&\quad+\lambda_2 Cov(X_1,X_2)+\lambda_3 Cov(X_1,X_3)\\
&\quad+\lambda_2 Cov(X_2,X_1)+\lambda_3 Cov(X_3,X_1)\\
&= 2\lambda_1 Var(X_1)+\sum_{k=1,k\neq 1}^{3}\lambda_k Cov(X_1,X_k)+\sum_{k=1,k\neq 1}^{3}\lambda_k Cov(X_k,X_1)\\
&= 2\lambda_1 Var(X_1)+2\sum_{k=1,k\neq 1}^{3}\lambda_k Cov(X_1,X_k)
\end{aligned}$$

Il ne reste plus qu'à remettre toute l'équation ensemble :

$$\frac{\partial}{\partial\lambda_i} = \frac{1}{\cancel{2}}\frac{\cancel{2}\lambda_i Var(X_i)+\cancel{2}\sum_{k=1,k\neq i}^{3}\lambda_k Cov(X_i,X_k)}{\sqrt{Var\left(\sum_{i=1}^{n}\lambda_i X_i\right)}}$$

Si on pose $\lambda_1 = ... = \lambda_i = ... = \lambda_n = 1$ et qu'on utilise les définitions des covariances pour rentrer les sommes dans la covariance, tel que

$$\begin{aligned}
Var(X_i)+\sum_{k=1,k\neq i}^{n} Cov(X_i,X_k) &= \sum_{k=1}^{n} Cov(X_i,X_k)\\
&= Cov\left(X_i,\sum_{k=1}^{n}X_k\right)
\end{aligned}$$

Alors,

$$C(X_i) = \frac{Cov\left(X_i,\sum_{k=1}^{n}X_k\right)}{\sqrt{Var\left(\sum_{k=1}^{n}X_k\right)}} = \frac{Cov(X_i,S)}{\sqrt{Var(S)}}$$

B.8 Distribution limite de W_n

Soit la v.a. Z où $\Pr(Z = E[X]) = 1$. On veut démontrer (à l'aide des transformées de Laplace) que

$$F_{W_n}(x) \longleftarrow F_Z(x) \quad x > 0$$

où $\Pr(Z = \gamma_j) = \Pr(\Theta = \theta_j)$ et $\gamma_j = E[X|\Theta = \theta_j]$.

Démonstration. Pour faire la preuve, il faut savoir les 2 résultats suivants :

$$e^{-x} \approx 1 - x \tag{B.9}$$

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x \tag{B.10}$$

Si on développe la transformée de Laplace :

$$\begin{aligned}
\mathcal{L}_{W_n}(t) &= E[e^{-tW_n}] \\
&= E_\Theta[E[e^{-tW_n}|\Theta=\theta]] \\
&= \int_0^\infty E[e^{-tW_n}|\Theta=\theta] f_\Theta(\theta)d\theta \\
&= \int_0^\infty E[e^{\frac{t}{n}(X_1+\ldots+X_n)}|\Theta=\theta] f_\Theta(\theta)d\theta \\
&= \int_0^\infty \prod_{i=1}^n E[e^{-\frac{t}{n}X_i}|\Theta=\theta] f_\Theta(\theta)d\theta \quad \text{(Car les risques sont cond. indép.)} \\
&= \int_0^\infty E[e^{-\frac{t}{n}X}|\Theta=\theta]^n f_\Theta(\theta)d\theta \quad \text{(car les v.a. sont } id\text{)} \\
&\approx \int_0^\infty E\left[\left(1-\frac{t}{n}X\right)|\Theta=\theta\right]^n f_\Theta(\theta)d\theta \quad \text{(par l'équation (B.9))} \\
&= \int_0^\infty \left(E[1|\Theta]-\frac{t}{n}E[X|\Theta]\right)^n f_\Theta(\theta)d\theta \\
&= \int_0^\infty \left(1-\frac{t}{n}E[X|\Theta]\right)^n f_\Theta(\theta)d\theta
\end{aligned}$$

Si on pose la limite $n \to \infty$,

$$\begin{aligned}
\lim_{n\to\infty} \mathcal{L}_{W_n}(t) &= \int_0^\infty \lim_{n\to\infty}\left(1-\frac{t}{n}E[X|\Theta]\right)^n f_\Theta(\theta)d\theta \\
&= \int_0^\infty e^{-tE[X|\Theta]} f_\Theta(\theta)d\theta \quad \text{(par l'équation (B.10))} \\
&= \int_0^\infty e^{-t\gamma} f_\Theta(\theta)d\theta \quad \text{, où } \gamma = E[X|\Theta] \\
&= \mathcal{L}_Z(t)
\end{aligned}$$

□

Annexe C

Travail collaboratif avec git

```
$ wget http://tex.stackexchange.com
```