Modèles de **régression** et de **séries chronologiques**

Exercices et solutions

Modèles de **régression** et de **séries chronologiques**

Exercices et solutions

Vincent Goulet

École d'actuariat, Université Laval

Seconde édition

© 2009 Vincent Goulet

Cette création est mise à disposition selon le contrat Paternité-Partage à l'identique 2.5 Canada disponible en ligne http://creativecommons.org/licenses/by-sa/2.5/ca/ ou par courrier postal à Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Historique de publication
Septembre 2009 : Seconde édition
Septembre 2006 : Première édition

Code source
Le code source LaTeX de ce document est disponible à l'adresse

http://vgoulet.act.ulaval.ca/methodes_statistiques/

ou en communiquant directement avec l'auteur.

ISBN 978-2-9811416-0-6
Dépôt légal – Bibliothèque et Archives nationales du Québec, 2009
Dépôt légal – Bibliothèque et Archives Canada, 2009

Introduction

Ce document est une collection des exercices distribués par l'auteur dans ses cours de Méthodes statistiques en actuariat entre 2003 et 2005, cours donnés à l'École d'actuariat de l'Université Laval. Certains exercices sont le fruit de l'imagination de l'auteur, alors que plusieurs autres sont des adaptations d'exercices tirés des ouvrages cités dans la bibliographie.

C'est d'ailleurs afin de ne pas usurper de droits d'auteur que ce document est publié selon les termes du contrat Paternité-ShareAlike 2.5 Canada de Creative Commons. Il s'agit donc d'un document «libre» que quiconque peut réutiliser et modifier à sa guise, à condition que le nouveau document soit publié avec le même contrat.

Cette seconde édition intègre les solutions des exercices qui faisaient l'objet d'un recueil séparé lors de la première édition. Les errata de la première édition sont corrigées et le document ne fait plus référence à S-Plus puisque le produit est aujourd'hui à toute fin pratique disparu au profit de R.

Le document est separé en deux parties correspondant aux deux sujets faisant l'objet d'exercices : d'abord la régression linéaire (simple et multiple), puis les séries chronologiques (lissage, modèles ARMA, ARIMA et SARIMA). Nous invitons le lecteur à consulter, entre autres, Abraham et Ledolter (1983), Draper et Smith (1998) et Brockwell et Davis (1996) pour d'excellents exposés sur la théorie des modèles de régression et des modèles de séries chronologiques.

L'estimation des paramètres, le calcul de prévisions et l'analyse des résultats — aussi bien en régression qu'en analyse de séries chronologiques — sont toutes des procédures à forte composante numérique. Il serait tout à fait artificiel de se restreindre, dans les exercices, à de petits ensembles de données se prêtant au calcul manuel. Dans cette optique, plusieurs des exercices de ce recueil requièrent l'utilisation du système statistique R. D'ailleurs, les annexes A et B présentent les principales fonctions de R pour la régression et l'analyse de séries chronologiques, dans l'ordre.

Le format de ces deux annexes est inspiré de Goulet (2007) : la présentation des fonctions compte peu d'exemples. Par contre, le lecteur est invité à lire et exécuter le code informatique des sections d'exemples A.7 et B.8. Le texte des sections d'exemples est disponible en format électronique dans le site Internet

`http://vgoulet.act.ulaval.ca/methodes_statistiques/`

L'annexe C contient quelques résultats d'algèbre matricielle utiles pour résoudre certains exercices.

Les réponses des exercices se trouvent à la fin de chacun des chapitres, alors que les solutions complètes sont regroupées à l'annexe D.

Tous les jeux de données mentionnés dans ce document sont disponibles en format électronique à l'adresse

`http://vgoulet.act.ulaval.ca/donnees/`

Ces jeux de données sont importés dans R avec l'une ou l'autre des commandes `scan` ou `read.table`. Certains jeux de données sont également fournis avec R ; la commande

```
> data()
```

en fournit une liste complète.

Nous remercions d'avance les lecteurs qui voudront bien nous faire part de toute erreur ou omission dans les exercices ou leurs réponses.

Enfin, nous tenons à remercier M. Michaël Garneau pour sa précieuse collaboration lors de la préparation de ce document, ainsi que tous les auxiliaires d'enseignement ayant, au cours des années, contribué à la rédaction d'exercices et de solutions.

Vincent Goulet <vincent.goulet@act.ulaval.ca>
Québec, septembre 2009

Table des matières

Première partie

Régression linéaire

2 Régression linéaire simple

2.1 Considérer les données suivantes et le modèle de régression linéaire $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$:

t	1	2	3	4	5	6	7	8	9	10
X_t	65	43	44	59	60	50	52	38	42	40
Y_t	12	32	36	18	17	20	21	40	30	24

a) Placer ces points ci-dessus sur un graphique.

b) Calculer les équations normales.

c) Calculer les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ en résolvant le système d'équations obtenu en b).

d) Calculer les prévisions $\hat{Y}_t$ correspondant à X_t pour $t = 1, \ldots, n$. Ajouter la droite de régression au graphique fait en a).

e) Vérifier empiriquement que $\sum_{t=1}^{10} e_t = 0$.

2.2 On vous donne les observations ci-dessous.

t	X_t	Y_t
1	2	6
2	3	4
3	5	6
4	7	3
5	4	6
6	4	4
7	1	7
8	6	4

$$\sum_{t=1}^{8} X_t = 32 \qquad \sum_{t=1}^{8} X_t^2 = 156$$

$$\sum_{t=1}^{8} Y_t = 40 \qquad \sum_{t=1}^{8} Y_t^2 = 214$$

$$\sum_{t=1}^{8} X_t Y_t = 146$$

a) Calculer les coefficients de la régression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, $\mathrm{Var}[\varepsilon_t] = \sigma^2$.

b) Construire le tableau d'analyse de variance de la régression en a) et calculer le coefficient de détermination R^2. Interpréter les résultats.

2.3 Le jeu de données `women.dat`, disponible à l'URL mentionnée dans l'introduction et inclus dans R, contient les tailles et les poids moyens de femmes américaines âgées de 30 à 39 ans. Importer les données dans dans R ou rendre le jeu de données disponible avec `data(women)`, puis répondre aux questions suivantes.

a) Établir graphiquement une relation entre la taille (*height*) et le poids (*weight*) des femmes.

b) À la lumière du graphique en a), proposer un modèle de régression approprié et en estimer les paramètres.

c) Ajouter la droite de régression calculée en b) au graphique. Juger visuellement de l'ajustement du modèle.

d) Obtenir, à l'aide de la fonction `summary` la valeur du coefficient de détermination R^2. La valeur est-elle conforme à la conclusion faite en c) ?

e) Calculer les statistiques SST, SSR et SSE, puis vérifier que SST $=$ SSR $+$ SSE. Calculer ensuite la valeur de R^2 et la comparer à celle obtenue en d).

2.4 Dans le contexte de la régression linéaire simple, démontrer que

$$\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})e_t = 0.$$

2.5 Considérer le modèle de régression linéaire par rapport au temps $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, $t = 1,\ldots,n$. Écrire les équations normales et obtenir les estimateurs des moindres carrés des paramètres β_0 et β_1. *Note* : $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$.

2.6 a) Trouver l'estimateur des moindres carrés du paramètre β dans le modèle de régression linéaire passant par l'origine $Y_t = \beta X_t + \varepsilon_t$, $t = 1,\ldots,n$, $E[\varepsilon_t] = 0$, $\mathrm{Cov}(\varepsilon_t,\varepsilon_s) = \delta_{ts}\sigma^2$.

b) Démontrer que l'estimateur en a) est sans biais.

c) Calculer la variance de l'estimateur en a).

2.7 Démontrer que l'estimateur des moindres carrés $\hat{\beta}$ trouvé à l'exercice 2.6 est l'estimateur sans biais à variance (uniformément) minimale du paramètre β. En termes mathématiques : soit

$$\beta^* = \sum_{t=1}^{n} c_t Y_t$$

un estimateur linéaire du paramètre β. Démontrer qu'en déterminant les coefficients $c_1,\ldots,c_n$ de façon à minimiser

$$\mathrm{Var}[\beta^*] = \mathrm{Var}\left[\sum_{t=1}^{n} c_t Y_t\right]$$

sous la contrainte que

$$E[\beta^*] = E\left[\sum_{t=1}^{n} c_t Y_t\right] = \beta,$$

on obtient $\beta^* = \hat{\beta}$.

2.8 Dans le contexte de la régression linéaire simple, démontrer que

a) $E[\text{MSE}] = \sigma^2$

b) $E[\text{MSR}] = \sigma^2 + \beta_1^2 \sum_{t=1}^{n} (X_t - \bar{X})^2$

2.9 Supposons que les observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont soumises à une transformation linéaire, c'est-à-dire que Y_t devient $Y_t' = a + bY_t$ et que X_t devient $X_t' = c + dX_t$, $t = 1, \ldots, n$.

a) Trouver quel sera l'impact sur les estimateurs des moindres carrés des paramètres β_0 et β_1 dans le modèle de régression linéaire $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$.

b) Démontrer que le coefficient de détermination R^2 n'est pas affecté par la transformation linéaire.

2.10 On sait depuis l'exercice 2.6 que pour le modèle de régression linéaire simple passant par l'origine $Y_t = \beta X_t + \varepsilon_t$, l'estimateur des moindres carrés de β est

$$\hat{\beta} = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}.$$

Démontrer que l'on peut obtenir ce résultat en utilisant la formule pour $\hat{\beta}_1$ dans la régression linéaire simple usuelle ($Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$) en ayant d'abord soin d'ajouter aux données un $(n+1)^{\text{e}}$ point $(m\bar{X}, m\bar{Y})$, où

$$m = \frac{n}{\sqrt{n+1} - 1} = \frac{n}{a}.$$

2.11 Soit le modèle de régression linéaire simple

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2).$$

Construire un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre β_1 si la variance σ^2 est connue.

2.12 Vous analysez la relation entre la consommation de gaz naturel *per capita* et le prix du gaz naturel. Vous avez colligé les données de 20 grandes villes et proposé le modèle

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

où Y représente la consommation de gaz *per capita*, X le prix et ε est le terme d'erreur aléatoire distribué selon une loi normale. Vous avez

obtenu les résultats suivants :

$$\hat{\beta}_0 = 138{,}581 \qquad \sum_{t=1}^{20}(X_t - \bar{X})^2 = 10\,668$$

$$\hat{\beta}_1 = -1{,}104 \qquad \sum_{t=1}^{20}(Y_t - \bar{Y})^2 = 20\,838$$

$$\sum_{t=1}^{20} X_t^2 = 90\,048 \qquad \sum_{t=1}^{20} e_t^2 = 7\,832.$$

$$\sum_{t=1}^{20} Y_t^2 = 116\,058$$

Trouver le plus petit intervalle de confiance à 95 % pour le paramètre β_1.

2.13 Le tableau ci-dessous présente les résultats de l'effet de la température sur le rendement d'un procédé chimique.

X	Y
-5	1
-4	5
-3	4
-2	7
-1	10
0	8
1	9
2	13
3	14
4	13
5	18

a) On suppose une relation linéaire simple entre la température et le rendement. Calculer les estimateurs des moindres carrés de l'ordonnée à l'origine et de la pente de cette relation.

b) Établir le tableau d'analyse de variance et tester si la pente est significativement différente de zéro avec un niveau de confiance de 0,95.

c) Quelles sont les limites de l'intervalle de confiance à 95 % pour la pente ?

d) Y a-t-il quelque indication qu'un meilleur modèle devrait être employé ?

2.14 Y a-t-il une relation entre l'espérance de vie et la longueur de la «ligne de vie» dans la main ? Dans un article de 1974 publié dans le *Journal of the American Medical Association*, Mather et Wilson dévoilent les 50 observations contenues dans le fichier `lifeline.dat`. À la lumière de ces

données, y a-t-il, selon vous, une relation entre la «ligne de vie» et l'espérance de vie ? Vous pouvez utiliser l'information partielle suivante :

$$\sum_{t=1}^{50} X_t = 3\,333 \qquad \sum_{t=1}^{50} X_t^2 = 231\,933 \qquad \sum_{t=1}^{50} X_t Y_t = 30\,549{,}75$$
$$\sum_{t=1}^{50} Y_t = 459{,}9 \qquad \sum_{t=1}^{50} Y_t^2 = 4\,308{,}57.$$

2.15 Considérer le modèle de régression linéaire passant par l'origine présenté à l'exercice 2.6. Soit X_0 une valeur de la variable indépendante, Y_0 la vraie valeur de la variable indépendante correspondant à X_0 et $\hat{Y}_0$ la prévision (ou estimation) de Y_0. En supposant que

i) $\varepsilon_t \sim N(0, \sigma^2)$;

ii) $\text{Cov}(\varepsilon_0, \varepsilon_t) = 0$ pour tout $t = 1, \ldots, n$;

iii) $\text{Var}[\varepsilon_t] = \sigma^2$ est estimé par s^2,

construire un intervalle de confiance de niveau $1 - \alpha$ pour Y_0. Faire tous les calculs intermédiaires.

2.16 La masse monétaire et le produit national brut (en millions de *snouks*) de la Fictinie (Asie postérieure) sont reproduits dans le tableau ci-dessous.

Année	Masse monétaire	PNB
1987	2,0	5,0
1988	2,5	5,5
1989	3,2	6,0
1990	3,6	7,0
1991	3,3	7,2
1992	4,0	7,7
1993	4,2	8,4
1994	4,6	9,0
1995	4,8	9,7
1996	5,0	10,0

a) Établir une relation linéaire dans laquelle la masse monétaire explique le produit national brut (PNB).

b) Construire des intervalles de confiance pour l'ordonnée à l'origine et la pente estimées en a). Peut-on rejeter l'hypothèse que la pente est nulle ? Égale à 1 ?

c) Si, en tant que ministre des Finances de la Fictinie, vous souhaitez que le PNB soit de 12,0 en 1997, à combien fixeriez-vous la masse monétaire ?

d) Pour une masse monétaire telle que fixée en c), déterminer les bornes inférieure et supérieure à l'intérieur desquelles devrait, avec une probabilité de 95 %, se trouver le PNB moyen. Répéter pour la valeur du PNB de l'année 1997.

2.17 Le fichier `house.dat` contient diverses données relatives à la valeur des maisons dans la région métropolitaine de Boston. La signification des différentes variables se trouve dans le fichier. Comme l'ensemble de données est plutôt grand (506 observations pour chacune des 13 variables), répondre aux questions suivantes à l'aide de R.

a) Déterminer à l'aide de graphiques à laquelle des variables suivantes le prix médian des maisons (`medv`) est le plus susceptible d'être lié par une relation linéaire : le nombre moyen de pièces par immeuble (`rm`), la proportion d'immeubles construits avant 1940 (`age`), le taux de taxe foncière par 10 000 $ d'évaluation (`tax`) ou le pourcentage de population sous le seuil de la pauvreté (`lstat`).

Astuce : en supposant que les données se trouvent dans le *data frame* `house`, essayer les commandes suivantes :

```
> plot(house)
> attach(house)
> plot(data.frame(rm, age, lstat, tax, medv))
> detach(house)
> plot(medv ~ rm + age + lstat + tax, data = house)
```

b) Faire l'analyse complète de la régression entre le prix médian des maisons et la variable choisie en a), c'est-à-dire : calcul de la droite de régression, tests d'hypothèses sur les paramètres afin de savoir si la régression est significative, mesure de la qualité de l'ajustement et calcul de l'intervalle de confiance de la régression.

c) Répéter l'exercice en b) en utilisant une variable ayant été rejetée en a). Observer les différences dans les résultats.

2.18 On veut prévoir la consommation de carburant d'une automobile à partir de ses différentes caractéristiques physiques, notamment le type du moteur. Le fichier `carburant.dat` contient des données tirées de *Consumer Reports* pour 38 automobiles des années modèle 1978 et 1979. Les caractéristiques fournies sont

- `mpg` : consommation de carburant en milles au gallon ;
- `nbcyl` : nombre de cylindres (remarquer la forte représentation des 8 cylindres !) ;
- `cylindree` : cylindrée du moteur, en pouces cubes ;
- `cv` : puissance en chevaux vapeurs ;
- `poids` : poids de la voiture en milliers de livres.

Utiliser R pour faire l'analyse ci-dessous.

a) Convertir les données du fichier en unités métriques, le cas échéant. Par exemple, la consommation de carburant s'exprime en $\ell/100$ km. Or, un gallon américain correspond à 3,785 litres et 1 mille à 1,6093 kilomètre. La consommation en litres aux 100 km s'obtient donc en divisant 235,1954 par la consommation en milles au gallon. De plus, 1 livre correspond à 0,45455 kilogramme.

b) Établir une relation entre la consommation de carburant d'une voiture et son poids. Vérifier la qualité de l'ajustement du modèle et si le modèle est significatif.

c) Trouver un intervalle de confiance à 95 % pour la consommation en carburant d'une voiture de 1 350 kg.

Réponses

2.1 c) $\hat{\beta}_0 = 66.44882$ et $\hat{\beta}_1 = -0.8407468$ d) $\hat{Y}_1 = 11{,}80, \hat{Y}_2 = 30{,}30, \hat{Y}_3 = 29{,}46, \hat{Y}_4 = 16{,}84, \hat{Y}_5 = 16{,}00, \hat{Y}_6 = 24{,}41, \hat{Y}_7 = 22{,}73, \hat{Y}_8 = 34{,}50, \hat{Y}_9 = 31{,}14, \hat{Y}_{10} = 32{,}82$

2.2 a) $\hat{\beta}_0 = 7$ et $\hat{\beta}_1 = -0{,}5$ b) SST = 14, SSR = 7, SSE = 7, MSR = 7, MSE = 7/6, F = 6, R^2 = 0,5

2.3 b) $\hat{\beta}_0 = -87{,}5167$ et $\hat{\beta}_1 = 3{,}45$ d) $R^2 = 0{,}991$ e) SSR $= 3332{,}7$ SSE $= 30{,}23$ et SST $= 3362{,}93$

2.5 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1(n+1)/2$, $\hat{\beta}_1 = (12\sum_{t=0}^{n} tY_t - 6n(n+1)\bar{Y})/(n(n^2-1)$

2.6 a) $\hat{\beta} = \sum_{t=1}^{n} X_t Y_t / \sum_{t=1}^{n} X_t^2$ c) $\text{Var}[\hat{\beta}] = \sigma^2 / \sum_{t=1}^{n} X_t^2$

2.9 a) $\hat{\beta}_1' = (b/d)\hat{\beta}_1$

2.11 $\beta_1 \in \hat{\beta}_1 \pm z_{\alpha/2}\sigma \left(\sum_{t=1}^{n}(X_t - \bar{X})^2\right)^{-1/2}$

2.12 $(-1{,}5, -0{,}7)$

2.13 a) $\hat{\beta}_0 = 9{,}273$, $\hat{\beta}_1 = 1{,}436$ b) $t = 9{,}809$ c) $(1{,}105, 1{,}768)$

2.14 $F = 0{,}73$, valeur p : 0,397

2.15 $\hat{Y}_0 \pm t_{\alpha/2}(n-1)\, s \sqrt{1 + X_0^2 / \sum_{t=1}^{n} X_t^2}$

2.16 a) PNB $= 1{,}168 + 1{,}716$ MM b) $\beta_0 \in (0{,}060, 2{,}276)$, $\beta_1 \in (1{,}427, 2{,}005)$ c) 6,31 d) $(11{,}20, 12{,}80)$ et $(10{,}83, 13{,}17)$

2.18 b) $R^2 = 0{,}858$ et $F = 217{,}5$ c) $10{,}57 \pm 2{,}13$

3 Régression linéaire multiple

3.1 Considérer le modèle de régression linéaire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\mathbf{X}$ est une matrice $n \times (p+1)$. Démontrer, en dérivant

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{t=1}^{n} (Y_t - \mathbf{x}_t'\boldsymbol{\beta})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

par rapport à $\boldsymbol{\beta}$, que les équations normales à résoudre pour obtenir l'estimateur des moindres carrés de $\boldsymbol{\beta}$ sont, sous forme matricielle,

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y},$$

Déduire l'estimateur des moindres carrés de ces équations. *Astuce* : utiliser le théorème C.5 de l'annexe C.

3.2 Pour chacun des modèles de régression ci-dessous, spécifier la matrice de schéma $\mathbf{X}$ dans la représentation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ du modèle, puis obtenir, si possible, les formules explicites des estimateurs des moindres carrés des paramètres.

a) $Y_t = \beta_0 + \varepsilon_t$

b) $Y_t = \beta_1 X_t + \varepsilon_t$

c) $Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t$

3.3 Vérifier, pour le modèle de régression linéaire simple, que les valeurs trouvées dans la matrice de variance-covariance $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ correspondent à celles calculées au chapitre 2.

3.4 Démontrer les relations ci-dessous dans le contexte de la régression linéaire multiple et trouver leur équivalent en régression linéaire simple. Utiliser $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

a) $\mathbf{X}'\mathbf{e} = \mathbf{0}$

b) $\hat{\mathbf{y}}'\mathbf{e} = 0$

c) $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$

3.5 Considérer le modèle de régression linéaire multiple présenté à l'exercice 3.1. Soit $\hat{Y}_0$ la prévision de la variable dépendante correspondant aux valeurs du vecteur ligne $\mathbf{x}_0 = (1, X_{01}, \ldots, X_{0p})$ des p variables indépendantes. On a donc

$$\hat{Y}_0 = \mathbf{x}_0 \hat{\boldsymbol{\beta}}.$$

a) Démontrer que $E[\hat{Y}_0] = E[Y_0]$.

b) Démontrer que l'erreur dans la prévision de la valeur moyenne de Y_0 est

$$E[(\hat{Y}_0 - E[Y_0])^2] = \sigma^2 \mathbf{x}_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0'.$$

Construire un intervalle de confiance de niveau $1 - \alpha$ pour $E[Y_0]$.

c) Démontrer que l'erreur dans la prévision de Y_0 est

$$E[(Y_0 - \hat{Y}_0)^2] = \sigma^2 (1 + \mathbf{x}_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0').$$

Construire un intervalle de confiance de niveau $1 - \alpha$ pour Y_0.

3.6 En ajustant le modèle

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t$$

à un ensemble de données, on a obtenu les statistiques suivantes :

$$R^2 = 0{,}521$$
$$F = 5{,}438.$$

Déterminer la valeur p approximative du test global de validité du modèle.

3.7 On vous donne les observations suivantes :

Y	X_1	X_2
17	4	9
12	3	10
14	3	11
13	3	11

De plus, si $\mathbf{X}$ est la matrice de schéma du modèle

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t, \quad t = 1,2,3,4,$$

où $\varepsilon_t \sim N(0, \sigma^2)$, alors

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{2} \begin{bmatrix} 765 & -87 & -47 \\ -87 & 11 & 5 \\ -47 & 5 & 3 \end{bmatrix}$$

et

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \frac{1}{2} \begin{bmatrix} -6 & 34 & -13 & -13 \\ 2 & -4 & 1 & 1 \\ 0 & -2 & 1 & 1 \end{bmatrix}$$

a) Trouver, par la méthode des moindres carrés, les estimateurs des paramètres du modèle mentionné ci-dessus.

b) Construire le tableau d'analyse de variance du modèle obtenu en a) et calculer le coefficient de détermination.

c) Vérifier si les variables X_1 et X_2 sont significatives dans le modèle.

d) Trouver un intervalle de confiance à 95 % pour la valeur de Y lorsque $X_1 = 3{,}5$ et $X_2 = 9$.

3.8 Répéter l'exercice 2.18 en ajoutant la cylindrée du véhicule en litres dans le modèle. La cylindrée est exprimée en pouces cubes dans les données. Or, 1 pouce correspond à 2,54 cm et un litre est définit comme étant 1 dm^3, soit 1 000 cm^3. Trouver un intervalle de confiance pour la consommation en carburant d'une voiture de 1 350 kg ayant un moteur de 1,8 litre.

3.9 Dans un exemple du chapitre 2 des notes de cours, nous avons tâché d'expliquer les sinistres annuels moyens par véhicule pour différents types de véhicules uniquement par la puissance du moteur (en chevaux-vapeur). Notre conclusion était à l'effet que la régression était significative — rejet de H_0 dans les tests t et F — mais l'ajustement mauvais — R^2 petit.

Examiner les autres variables fournies dans le fichier `auto-price.dat` et choisir deux autres caractéristiques susceptibles d'expliquer les niveaux de sinistres. Par exemple, peut-on distinguer une voiture sport d'une minifourgonnette ?

Une fois les variables additionnelles choisies, calculer les différentes statistiques propres à une régression en ajoutant d'abord une, puis deux variables au modèle de base. Quelles sont vos conclusions ?

3.10 En bon étudiant(e), vous vous intéressez à la relation liant la demande pour la bière, Y, aux variables indépendantes X_1 (le prix de celle-ci), X_2 (le revenu disponible) et X_3 (la demande de l'année précédente). Un total de 20 observations sont disponibles. Vous postulez le modèle

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t,$$

où $E[\varepsilon_t] = 0$ et $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = \delta_{ts}\sigma^2$. Les résultats de cette régression, tels que calculés dans R, sont fournis ci-dessous.

```
> fit <- lm(Y ~ X1 + X2 + X3, data = biere)
> summary(fit)

Call: lm(formula = Y ~ X1 + X2 + X3, data = biere)
Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.014e+04 -5.193e-03 -2.595e-03  4.367e-03  2.311e-02

Coefficients:
              Value Std. Error t value  Pr(>|t|)
```

```
(Intercept)  1.5943  1.0138     1.5726    0.1354
         X1 -0.0480  0.1479    -0.3243    0.7499
         X2  0.0549  0.0306     1.7950    0.0916
         X3  0.8130  0.1160     7.0121 2.933e-06

Residual standard error: 0.0098 on 16 degrees of freedom
Multiple R-Squared: 0.9810      Adjusted R-squared: 0.9774
F-statistic: 275.49 on 3 and 16 degrees of freedom,
the p-value is 7.160e-14
```

a) Indiquer les dimensions des matrices et vecteurs dans la représentation matricielle $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ du modèle.

b) La régression est-elle significative ? Expliquer.

c) On porte une attention plus particulière au paramètre β_2. Est-il significativement différent de zéro ? Quelle est l'interprétation du test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$?

d) Quelle est la valeur et l'interprétation de R^2, le coefficient de détermination ? De manière générale, est-il envisageable d'obtenir un R^2 élevé et, simultanément, toutes les statistiques t pour les tests $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ et $H_0 : \beta_3 = 0$ non significatives ? Expliquer brièvement.

3.11 Au cours d'une analyse de régression, on a colligé les valeurs de trois variables explicatives X_1, X_2 et X_3 ainsi que celles d'une variable dépendante Y. Les résultats suivants ont par la suite été obtenus avec R.

```
> anova(lm(Y ~ X2 + X3, data = foo))

Analysis of Variance Table

Response: Y

          Df Sum of Sq  Mean Sq  F Value       Pr(>F)
       X2  1  45.59085 45.59085 106.0095 0.0000000007 ***
       X3  1   8.76355  8.76355  20.3773 0.0001718416 ***
Residuals 22   9.46140  0.43006
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(Y ~ X1 + X2 + X3, data = foo))

Analysis of Variance Table

Response: Y

          Df Sum of Sq  Mean Sq  F Value    Pr(>F)
       X1  1  45.59240 45.59240 101.6681 0.0000000 ***
       X2  1   0.01842  0.01842   0.0411 0.8413279
       X3  1   8.78766  8.78766  19.5959 0.0002342 ***
Residuals 21   9.41731  0.44844
```

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

a) On considère le modèle complet $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$. À partir de l'information ci-dessus, calculer la statistique appropriée pour compléter chacun des tests suivants. Indiquer également le nombre de degrés de liberté de cette statistique. Dans tous les cas, l'hypothèse alternative H_1 est la négation de l'hypothèse H_0.
   i) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
   ii) $H_0 : \beta_1 = 0$
   iii) $H_0 : \beta_2 = \beta_3 = 0$

b) À la lumière des résultats en a), quelle(s) variable(s) devrait-on inclure dans la régression ? Justifier votre réponse.

3.12 Dans une régression multiple avec quatre variables explicatives et 506 données, on a obtenu :

$$\begin{aligned} \text{SSR}(X_1|X_4) &= 21\,348 \\ \text{SSR}(X_4) &= 2\,668 \\ R^2 &= 0{,}6903 \\ s^2 &= 26{,}41. \end{aligned}$$

Calculer la statistique appropriée pour le test

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

3.13 En régression linéaire multiple, on a $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ et $\text{SSE}/\sigma^2 \sim \chi^2(n-p-1)$.

a) Vérifier que

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t(n-p-1), \quad i = 0, 1, \ldots, p,$$

où c_{ii} est le $(i+1)^{\text{e}}$ élément de la diagonale de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$ et $s^2 = \text{MSE}$.

b) Que vaut c_{11} en régression linéaire simple ? Adapter le résultat ci-dessus à ce modèle.

3.14 Considérer le modèle de régression linéaire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\mathbf{X}$ est une matrice $n \times (p+1)$, $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{W}^{-1}$ et $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$. Démontrer, en dérivant

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{t=1}^{n} w_t (\mathbf{y}_t - \mathbf{x}_t'\boldsymbol{\beta})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

par rapport à $\boldsymbol{\beta}$, que les équations normales à résoudre pour obtenir l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$ sont, sous forme matricielle,

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}}^* = \mathbf{X}'\mathbf{W}\mathbf{y},$$

puis en déduire cet estimateur. *Astuce* : cette preuve est simple si l'on utilise le théorème C.5 de l'annexe C avec $\mathbf{A} = \mathbf{W}$ et $f(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

3.15 Considérer le modèle de régression linéaire simple passant par l'origine $Y_t = \beta X_t + \varepsilon_t$. Trouver l'estimateur linéaire sans biais à variance minimale du paramètre β, ainsi que sa variance, sous chacune des hypothèses suivantes.

a) $\text{Var}[\varepsilon_t] = \sigma^2$

b) $\text{Var}[\varepsilon_t] = \sigma^2 / w_t$

c) $\text{Var}[\varepsilon_t] = \sigma^2 X_t$

d) $\text{Var}[\varepsilon_t] = \sigma^2 X_t^2$

3.16 Proposer, à partir des données ci-dessous, un modèle de régression complet (incluant la distribution du terme d'erreur) pouvant expliquer le comportement de la variable Y en fonction de celui de X.

Y	X
32,83	25
9,70	3
29,25	24
15,35	11
13,25	10
24,19	20
8,59	6
25,79	21
24,78	19
10,23	9
8,34	4
22,10	18
10,00	7
18,64	16
18,82	15

3.17 On vous donne les 23 données dans le tableau ci-dessous.

t	Y_t	X_t	t	Y_t	X_t	t	Y_t	X_t
12	2,3	1,3	19	1,7	3,7	6	2,8	5,3
23	1,8	1,3	20	2,8	4,0	10	2,1	5,3
7	2,8	2,0	5	2,8	4,0	4	3,4	5,7
8	1,5	2,0	2	2,2	4,0	9	3,2	6,0
17	2,2	2,7	21	3,2	4,7	13	3,0	6,0
22	3,8	3,3	15	1,9	4,7	14	3,0	6,3
1	1,8	3,3	18	1,8	5,0	16	5,9	6,7
11	3,7	3,7	3	3,5	5,3			

a) Calculer l'estimateur des moindres carrés ordinaires $\hat{\boldsymbol{\beta}}$.

b) Supposons que la variance de Y_{16} est $4\sigma^2$ plutôt que σ^2. Recalculer la régression en a) en utilisant cette fois les moindres carrés pondérés.

c) Refaire la partie b) en supposant maintenant que la variance de l'observation Y_{16} est $16\sigma^2$. Quelles différences note-t-on ?

3.18 Une coopérative de taxi new-yorkaise s'intéresse à la consommation de carburant des douze véhicules de sa flotte en fonction de leur âge. Hormis leur âge, les véhicules sont identiques et utilisent tous le même type d'essence. La seule chose autre différence notable d'un véhicule à l'autre est le sexe du conducteur : la coopérative emploie en effet des hommes et des femmes. La coopérative a recueilli les données suivantes afin d'établir un modèle de régression pour la consommation de carburant :

Consommation (mpg)	Âge du véhicule	Sexe du conducteur
12,3	3	M
12,0	4	F
13,7	3	F
14,2	2	M
15,5	1	F
11,1	5	M
10,6	4	M
14,0	1	M
16,0	1	F
13,1	2	M
14,8	2	F
10,2	5	M

a) En plaçant les points sur un graphique de la consommation de carburant en fonction de l'âge du véhicule, identifier s'il existe ou non une différence entre la consommation de carburant des femmes et celle des hommes. *Astuce* : utiliser un symbole (`pch`) différent pour chaque groupe.

b) Établir un modèle de régression pour la consommation de carburant. Afin de pouvoir intégrer la variable qualitative «sexe du conducteur» dans le modèle, utiliser une variable indicatrice du type

$$X_{t2} = \begin{cases} 1, & \text{si le conducteur est un homme} \\ 0, & \text{si le conducteur est une femme.} \end{cases}$$

c) Quelle est, selon le modèle établi en b), la consommation moyenne d'une voiture taxi de quatre ans conduite par une femme ? Fournir un intervalle de confiance à 90 % pour cette prévision.

Réponses

3.2 a) $\hat{\beta}_0 = \bar{Y}$ b) $\hat{\beta}_1 = (\sum_{t=1}^{n} X_t Y_t) / (\sum_{t=1}^{n} X_t^2)$

3.6 $p \approx 0{,}01$

3.7 a) $\hat{\boldsymbol{\beta}} = (-22{,}5,\ 6{,}5,\ 1{,}5)$ b) $F = 13{,}5$, $R^2 = 0{,}9643$ c) $t_1 = 3{,}920$, $t_2 = 1{,}732$ d) $13{,}75 \pm 13{,}846$

3.8 b) $R^2 = 0{,}8927$ et $F = 145{,}6$ c) $12{,}04 \pm 2{,}08$

3.10 a) $\mathbf{y}_{20\times 1}$, $\mathbf{X}_{20\times 4}$, $\boldsymbol{\beta}_{4\times 1}$ et $\boldsymbol{\varepsilon}_{20\times 1}$

3.11 a) i) 40,44, 3 et 21 degrés de liberté ii) 0,098, 1 et 21 degrés de liberté iii) 9,82, 2 et 21 degrés de liberté b) X_1 et X_3, ou X_2 et X_3

3.12 103,67

3.15 a) $\hat{\beta}^* = \sum_{t=1}^{n} X_t Y_t / \sum_{t=1}^{n} X_t^2$, $\text{Var}[\hat{\beta}^*] = \sigma^2 / \sum_{t=1}^{n} X_t^2$

b) $\hat{\beta}^* = \sum_{t=1}^{n} w_t X_t Y_t / \sum_{t=1}^{n} w_t X_t^2$, $\text{Var}[\hat{\beta}^*] = \sigma^2 / \sum_{t=1}^{n} w_t X_t^2$

c) $\hat{\beta}^* = \bar{Y} / \bar{X}$, $\text{Var}[\hat{\beta}^*] = \sigma^2 / (n\bar{X})$

d) $\hat{\beta}^* = \sum_{t=1}^{n} Y_t / X_t$, $\text{Var}[\hat{\beta}^*] = \sigma^2 / n$

3.16 $Y_t = 18{,}12 + 29{,}68 X_t + 4{,}09 X_t^2 + \varepsilon_t$, $\varepsilon_t \sim N(0, 1{,}373)$

3.17 a) $\hat{\boldsymbol{\beta}} = (1{,}4256, 0{,}3158)$ b) $\hat{\boldsymbol{\beta}}^* = (1{,}7213, 0{,}2243)$ c) $\hat{\boldsymbol{\beta}}^* = (1{,}808, 0{,}1975)$

3.18 b) $\text{mpg} = 16{,}687 - 1{,}04\ \text{age} - 1{,}206\ \text{sexe}$ c) $12{,}53 \pm 0{,}58$ mpg

Deuxième partie

Séries chronologiques

4 Lissage de séries chronologiques

4.1 Éliminer la tendance des données du nombre de grèves aux États-Unis, 1951–1980 (`strikes.dat`) par une application appropriée de l'opérateur de différenciation ∇.

4.2 Comparer l'effet, tant sur l'estimateur de la tendance que sur les résidus, du lissage exponentiel des données du fichier `sales.dat` avec $\alpha = 0{,}2$, $\alpha = 0{,}5$ et $\alpha = 0{,}7$.

4.3 a) Décomposer les données du nombre mensuel de morts accidentelles, 1973–1978 (`deaths.dat`) en leurs composantes additives de tendance, de saisonnalité et de résidus en utilisant la fonction `stl` de R.

b) La fonction `stl` retourne une liste. Le premier élément, nommé `time.series`, est une matrice qui regroupe les composantes de la série. On extrait donc les résidus avec :

```
> stl(x)$time.series[, "remainder"]
```

Vérifier à l'aide du corrélogramme des résidus calculés en a) que ceux-ci forment une série stationnaire.

4.4 a) La production de bière australienne 1956–1990 (`beer.dat`) est une série comportant une tendance ainsi que de la saisonnalité. De plus, l'amplitude de la saisonnalité augmente avec le niveau du processus. Par conséquent, un modèle multiplicatif de la forme

$$Y_t = m_t \, s_t \, X_t$$

serait possiblement plus approprié pour cette série que le modèle additif usuel

$$Y_t = m_t + s_t + X_t.$$

Le modèle multiplicatif pour la série originale est toutefois équivalent à un modèle additif pour le logarithme des données. Faire un graphique de ces deux séries et vérifier lequel des deux modèles ci-dessus (multiplicatif ou additif) est le plus approprié.

b) Éliminer la tendance et la saisonnalité de la série choisie en a) à l'aide de différences.

4.5 On dit du filtre à moyenne mobile $\sum_{j=-\infty}^{\infty} a_j Y_{t-j}$ avec

$$a_j = \begin{cases} \dfrac{1}{2q+1}, & |j| \leq q \\ 0, & |j| > q \end{cases}$$

que c'est un *filtre linéaire* parce qu'une tendance linéaire le traverse sans distorsion. Démontrer que, en effet, une tendance $m_t = c_0 + c_1 t$ passe à travers le filtre ci-dessus sans être affectée, c'est-à-dire que $\sum_{j=-\infty}^{\infty} a_j m_{t-j} = m_t$.

4.6 Démontrer que le filtre $\sum_{j=-q}^{q} a_j Y_{t-j}$ avec coefficients $[a_{-2}, a_{-1}, a_0, a_1, a_2] = \frac{1}{9}[-1, 4, 3, 4, -1]$ laisse passer les polynômes du troisième degré (c'est-à-dire $m_t = c_0 + c_1 t + c_2 t^2 + c_3 t^3$) sans distorsion et élimine la saisonnalité de périodicité 3 (c'est-à-dire $s_t = s_{t+3}$).

4.7 Soit le filtre $a_j = 1/(2q+1)$, $-q \leq j \leq q$. Si Y_t, $t = 0, \pm 1, \pm 2, \ldots$, sont des variables aléatoires indépendantes de moyenne 0 et de variance σ^2, montrer que la moyenne mobile $A_t = \sum_{j=-q}^{q} a_j Y_{t-j}$ est «petite» pour de grandes valeurs de q dans le sens où $E[A_t] = 0$ et $\mathrm{Var}[A_t] = \sigma^2/(2q+1)$. Interpréter ce résultat.

4.8 Trouver un filtre de la forme $1 + \alpha B + \beta B^2 + \gamma B^3$ (c'est-à-dire trouver α, β et γ) qui laisse passer une tendance linéaire sans distorsion, mais élimine la saisonnalité de période 2. *Note* : B est l'opérateur de rétrodécalage.

4.9 On vous donne neuf observations d'une série chronologique $\{X_t\}$:

```
Time Series:
Start = 1
End = 9
Frequency = 1
[1]   4  -5  93  12  76 143 163 164 158
```

On vous donne également l'information suivante pour quelques séries apparentées à $\{X_t\}$, où T est la variable aléatoire du nombre de changements de direction dans une série chronologique.

	$\{X_t\}$	$\{\nabla X_t\}$	$\{\nabla_3 X_t\}$	$\{\nabla\nabla_3 X_t\}$	$\{\nabla^2 X_t\}$
$E[T]$	4,67	4	2,67	2	3,33
$\mathrm{Var}[T]$	1,28	1,1	0,74	0,57	0,92
$\sum_{h=1}^{4} \hat{\rho}(h)^2$	0,45	0,29	0,14	0,13	0,45

a) Éliminer la tendance et/ou la saisonnalité de cette série à l'aide de différences.

b) Vérifier si la série obtenue en a) forme un bruit blanc à l'aide des tests portmanteau et des changements de direction.

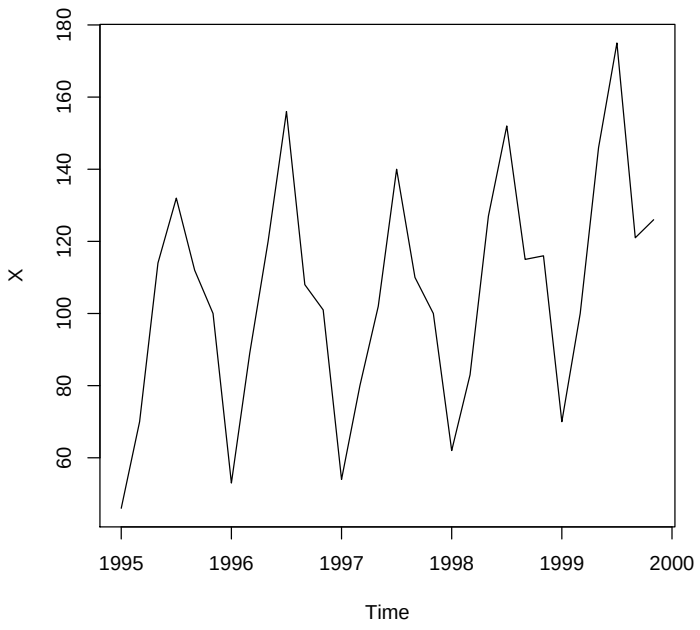


FIGURE 4.1: Ventes bimestrielles de Guinness, janvier 1995 – novembre 1999

4.10 On vous donne ci-dessous les valeurs des ventes bimestrielles (à tous les deux mois) de la bière Guinness, de janvier 1995 à novembre 1999. Un graphique de cette série vous est également fourni à la figure 4.1.

```
> X

Time Series:
Start = c(1995, 1)
End = c(1999, 6)
Frequency = 6
 [1]  46  70 114 132 112 100  53  89 120 156 108 101
[13]  54  80 102 140 110 100  62  83 127 152 115 116
[25]  70 100 146 175 121 126
```

Éliminer la tendance et/ou la saisonnalité de cette série. Si nécessaire, estimer la tendance à l'aide d'une moyenne mobile centrée à trois points et éliminer la saisonnalité à l'aide de différences.

4.11 Soit la série de 30 observations suivante :

$x_1,\ldots,x_6$:	486	474	434	441	435	401
$x_7,\ldots,x_{12}$:	414	414	386	405	411	389
$x_{13},\ldots,x_{18}$:	414	426	410	441	459	449
$x_{19},\ldots,x_{24}$:	486	510	506	549	579	581
$x_{25},\ldots,x_{30}$:	630	666	674	729	771	785

Cette série est en fait la somme d'une tendance quadratique et d'une composante de saisonnalité de période 3. Appliquer à cette série le filtre obtenu à l'exercice 4.8 et discuter des résultats.

4.12 On vous donne les 10 observations d'un processus $\{Y_t\}$ ainsi que les fonctions d'autocorrélations empiriques de $\{Y_t\}$, $\{\nabla Y_t\}$ et $\{\nabla^2 Y_t\}$.

```
> Y

Time Series:
Start = 1
End = 10
Frequency = 1
 [1]  4.6  6.1  7.5  7.6  9.2 10.3  9.3  8.9 12.6
[10] 12.5

> acf(Y, plot = FALSE)

Autocorrelations of series 'Y', by lag

     0      1      2      3      4      5      6
 1.000  0.539  0.153  0.152  0.008 -0.185 -0.198
     7      8      9
-0.263 -0.442 -0.264

> acf(diff(Y), plot = FALSE)

Autocorrelations of series 'diff(Y)', by lag

     0      1      2      3      4      5      6
 1.000 -0.306 -0.375  0.224  0.124 -0.298  0.090
     7      8
 0.078 -0.038

> acf(diff(Y, differences = 2), plot = FALSE)

Autocorrelations of series 'diff(Y, differences = 2)', by lag

     0      1      2      3      4      5      6
 1.000 -0.388 -0.351  0.225  0.183 -0.270  0.110
     7
-0.009
```

a) Éliminer la tendance de cette série à l'aide de différences.

b) Tester si la série obtenue en a) est un bruit blanc à l'aide des tests portmanteau et des changements de direction.

Réponses

4.8 $\alpha = \gamma = 1/4$, $\beta = -1/2$

4.9 b) $Q^* = 2{,}3267$, $T = 0{,}95$

4.10 $\nabla_6(X_t - \frac{1}{3}\sum_{j=-1}^{1} X_{t-j})$

4.12 $Q^* = 3{,}685$, $T = 0{,}5898$

5 Stationnarité et modèles stochastiques de séries chronologiques

5.1 Soit $\{Z_t\}$ une suite de variables aléatoires indépendantes distribuées selon une loi normale de moyenne 0 et de variance σ^2, et soit a, b et c des constantes. Déterminer lequel ou lesquels des processus ci-dessous sont stationnaires. Pour chaque processus stationnaire, calculer la moyenne et la fonction d'autocovariance.

a) $X_t = a + bZ_t + cZ_{t-2}$

b) $X_t = Z_1 \cos(ct) + Z_2 \sin(ct)$

c) $X_t = Z_t \cos(ct) + Z_{t-1} \sin(ct)$

d) $X_t = a + bZ_0$

e) $X_t = Z_t Z_{t-1}$

Astuce : $\cos(u+v) = \cos u \cos v - \sin u \sin v$ et $\sin(u+v) = \sin u \cos v + \cos u \sin v$.

5.2 Soit $\{X_t\}$ une série chronologique stationnaire de moyenne nulle et a, b, des constantes.

a) Si $Y_t = a + bt + s_t + X_t$, où s_t est une composante de saisonnalité de période 12, démontrer que $\nabla\nabla_{12} Y_t = (1-B)(1-B^{12})Y_t$ est stationnaire.

b) Si $Y_t = (a+bt)s_t + X_t$, où s_t est toujours une composante de saisonnalité de période 12, démontrer que $\nabla_{12}^2 Y_t = (1-B^{12})^2 Y_t$ est stationnaire.

5.3 Soit $\{X_t\}$ et $\{Y_t\}$ deux séries stationnaires et non corrélées, c'est-à-dire que $\mathrm{Cov}(X_r, Y_s) = 0$ pour tous r et s. Démontrer que $\{X_t + Y_t\}$ est stationnaire avec fonction d'autocovariance égale à la somme des fonctions d'autocovariance de $\{X_t\}$ et $\{Y_t\}$.

5.4 Les données `lake.dat` donnent le niveau du Lac Huron moins 570 pieds entre les années 1875 et 1972. L'ensemble contient donc 98 données. Faire une modélisation préliminaire (à être complétée plus tard) de cette série en suivant les étapes suivantes.

i) Tracer le graphique de la série et identifier visuellement une tendance et/ou de la saisonnalité.

ii) Si nécessaire, estimer la tendance par régression et éliminer la saisonnalité à l'aide de différences.

iii) Proposer un modèle pour les résidus obtenus en ii). Justifier votre réponse à l'aide du corrélogramme des résidus et des résultats des tests de détection du bruit blanc.

5.5 Considérer le processus à moyenne mobile $\{X_t\}$ suivant :

$$X_t = Z_t + \theta Z_{t-2},$$

où $\{Z_t\} \sim \text{WN}(0,1)$.

a) Calculer les fonctions d'autocovariance et d'autocorrélation de ce processus.

b) À l'aide de la fonction `arima.sim` de R, simuler 300 observations du processus ci-dessus avec $\theta = 0{,}8$. Calculer et tracer le corrélogramme du processus ainsi obtenu.

c) Répéter la partie b) avec $\theta = -0{,}8$.

d) Les corrélogrammes obtenus en b) et c) correspondent-ils à la fonction d'autocorrélation théorique calculée en a) ?

e) On remarquera que la série en b) fluctue moins rapidement que celle en c). Expliquer cet état de fait à l'aide de la fonction d'autocorrélation.

5.6 Soit $\{X_t\}$ un processus AR(1).

a) Calculer la variance de $(X_1 + X_2 + X_3 + X_4)/4$ quand $\phi = 0{,}9$ et $\sigma^2 = 1$.

b) Répéter la partie a) avec $\phi = -0{,}9$ et comparer le résultat avec celui obtenu en a). Interpréter.

5.7 Soit $\{Z_t\}$ un bruit IID avec $Z_t \sim N(0,1)$. On définit

$$X_t = \begin{cases} Z_t, & \text{si } t \text{ est pair,} \\ \dfrac{Z_{t-1}^2 - 1}{\sqrt{2}}, & \text{si } t \text{ est impair.} \end{cases}$$

Démontrer que $\{X_t\}$ est $\text{WN}(0,1)$ mais non $\text{IID}(0,1)$.

5.8 On vous donne les cinq valeurs suivantes d'un bruit blanc de moyenne 0 et de variance 1 :

$$0{,}18 \quad -1{,}61 \quad 3{,}00 \quad 1{,}33 \quad 0{,}37.$$

Calculer quatre valeurs des processus ci-dessous.

a) $\text{AR}(1)$ avec $\phi = 0{,}6$.

b) $\text{MA}(1)$ avec $\theta = -0{,}4$.

c) $\text{ARMA}(1,1)$ avec $\phi = 0{,}6$ et $\theta = -0{,}4$.

5.9 a) Vérifier que le processus stationnaire (causal)

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j} = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \ldots,$$

où $|\phi| < 1$, est bien une solution de l'équation $X_t - \phi X_{t-1} = Z_t$ définissant le processus AR(1).

b) Vérifier que le processus

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} = -\phi^{-1} Z_{t+1} - \phi^{-2} Z_{t+2} - \phi^{-3} Z_{t+3} - \ldots,$$

où $|\phi| > 1$ est aussi une solution de l'équation ci-dessus, mais que cette solution n'est pas causale.

5.10 Pour chaque processus ARMA ci-dessous, déterminer si le processus est stationnaire et s'il est réversible. (Dans chaque cas $\{Z_t\}$ est un bruit blanc.)

a) $X_t + 0{,}2X_{t-1} - 0{,}48X_{t-2} = Z_t$.

b) $X_t + 1{,}9X_{t-1} - 0{,}88X_{t-2} = Z_t + 0{,}2Z_{t-1} + 0{,}7Z_{t-2}$.

c) $X_t + 0{,}6X_{t-1} = Z_t + 1{,}2Z_{t-1}$.

d) $X_t + 1{,}8X_{t-1} - 0{,}81X_{t-2} = Z_t$.

e) $X_t + 1{,}6X_{t-1} = Z_t - 0{,}4Z_{t-1} + 0{,}04Z_{t-2}$.

5.11 Soit $\{Y_t\}$ la somme d'un processus AR(1) et d'un bruit blanc, c'est-à-dire

$$Y_t = X_t + W_t,$$

où $\{W_t\} \sim \mathrm{WN}(0, \sigma_W^2)$ et $\{X_t\}$ est le processus AR(1) avec $|\phi| < 1$

$$X_t - \phi X_{t-1} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma_Z^2).$$

On suppose de plus que $E[W_s Z_t] = 0$ pour tous s et t.

a) Démontrer que $\{Y_t\}$ est stationnaire et calculer sa fonction d'autocovariance.

b) Démontrer que la série chronologique $U_t \equiv Y_t - \phi Y_{t-1}$ est 1–corrélée (c'est-à-dire que $\gamma_U(h) = 0$ pour tout $|h| > 1$) et que, par conséquent, elle peut s'écrire comme un processus MA(1).

c) Conclure de b) que $\{Y_t\}$ est un processus ARMA$(1,1)$ et exprimer les trois paramètres de ce modèle en fonction de ϕ, σ_W^2 et σ_Z^2.

5.12 a) Les équations de Yule–Walker généralisées sont obtenues en multipliant

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

de part et d'autre par X_{t-h} et en prenant par la suite l'espérance. Démontrer que les équations ainsi obtenues sont les suivantes : pour $0 \leq h \leq q$

$$\gamma_X(h) - \phi_1\gamma_X(h-1) - \cdots - \phi_p\gamma_X(h-p) = \sigma^2 \sum_{j=0}^{\infty} \psi_j\theta_{h+j},$$

et, pour $h > q$,

$$\gamma_X(h) - \phi_1\gamma_X(h-1) - \cdots - \phi_p\gamma_X(h-p) = 0.$$

b) Utiliser les équations de Yule–Walker généralisées ci-dessus pour calculer la fonction d'autocovariance d'un modèle ARMA$(1,1)$.

5.13 Pour chaque modèle ci-dessous :

i) classer le modèle parmi la famille des processus ARIMA(p,d,q) ;

ii) calculer les quatre premiers coefficients ψ_0, ψ_1, ψ_2 et ψ_3 de la représentation MA(∞) de $\{X_t\}$;

iii) calculer les quatre premiers coefficients π_0, π_1, π_2 et π_3 de la représentation AR(∞) de $\{Z_t\}$.

Dans tous les cas, $\{Z_t\}$ est un bruit blanc.

a) $X_t - 0{,}5X_{t-1} = Z_t$

b) $X_t = Z_t - 1{,}3Z_{t-1} + 0{,}4Z_{t-2}$

c) $X_t - 0{,}5X_{t-1} = Z_t - 1{,}3Z_{t-1} + 0{,}4Z_{t-2}$

d) $X_t - 1{,}2X_{t-1} + 0{,}2X_{t-2} = Z_t - 0{,}5Z_{t-1}$

5.14 Démontrer que la valeur à $h = 2$ de la fonction d'autocorrélation partielle d'un modèle MA(1)

$$X_t = Z_1 + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0,\sigma^2),$$

est

$$\phi_{22} = -\frac{\theta^2}{1+\theta^2+\theta^4}.$$

5.15 On souhaite ajuster le modèle de régression suivant à un ensemble de n données :

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t,$$

où $\{\varepsilon_t\} \sim \text{AR}(1)$. Suite à de longues incantations proférées dans une langue gutturale proche de celle du Mordor, on a appris que les paramètres du processus AR(1) sont $\phi = 0{,}8$ et $\sigma^2 = 9$.

a) Expliquer brièvement pourquoi l'emploi des moindres carrés généralisés s'avère approprié pour l'estimation des paramètres de la régression β_0, β_1 et β_2.

b) Préciser la forme de la matrice $\mathbf{V} = \mathbf{V}[\boldsymbol{\varepsilon}]$ à utiliser dans les moindres carrés généralisés.

5.16 Le processus SARIMA$(p,d,q) \times (P,D,Q)_s$ est défini comme la solution stationnaire $\{X_t\}$ des équations

$$\phi(B)\Phi(B^s)W_t = \theta(B)\Theta(B^s)Z_t, \quad W_t = \nabla^d \nabla_s^D X_t,$$

où $\{Z_t\} \sim \text{WN}(0,\sigma^2)$ et $\phi(z)$, $\Phi(z)$, $\theta(z)$ et $\Theta(z)$ sont des polynômes de degré p, P, q et Q, respectivement. Ainsi, si le processus $\{X_t\}$ une tendance et de la saisonnalité de période s, alors $\{W_t\}$ est le processus stationnaire obtenu en éliminant la tendance puis la saisonnalité à l'aide de d différences (d'ordre 1) et D (normalement $D = 1$) différences d'ordre s. De plus,

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

est une composante AR saisonnière (entre les années) et

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$$

est une composante MA saisonnière. Remarquer que l'on peut obtenir un modèle saisonnier même si $D = 0$, en autant que $P > 0$ ou $Q > 0$.

Trouver l'ordre des modèles SARIMA ci-dessous.

a) $(1-B)(1-B^{12})X_t = (1+0{,}5B)(1-0{,}6B^{12})Z_t$

b) $(1-0{,}7B)(1-B^{12})X_t = (1+0{,}45B^{12})Z_t$

c) $(1-0{,}7B)(1+0{,}3B^4)(1-B)X_t = (1+0{,}37B^4)Z_t$

Réponses

5.1 Sont stationnaires : a), b), d), e).

5.4 $\hat{m}_t = 55{,}555 - 0{,}024t$, $Q = 109{,}24$, $T = 40$.

5.5 $\gamma_X(0) = 1+\theta^2$, $\gamma_X(\pm 2) = \theta$, $\gamma_X(h) = 0$ ailleurs.

5.6 $\text{Var}[\frac{1}{4}(X_1+X_2+X_3+X_4)] = \sigma^2(\frac{1}{4} + \frac{3}{8}\phi + \frac{1}{4}\phi^2 + \frac{1}{8}\phi^3)/(1-\phi^2)$.

5.8 a) $\{0{,}180, -1{,}502, 2{,}099, 2{,}589\}$

b) $\{0{,}180, -1{,}682, 3{,}644, 0{,}130, -0{,}162\}$

c) $\{0{,}180, -1{,}574, 2{,}700, 1{,}750\}$

5.10 a) Stationnaire et réversible b) Réversible seulement c) Stationnaire seulement d) Réversible seulement e) Réversible seulement

5.13 a) AR(1) b) MA(2) c) ARMA$(1,2)$ d) ARIMA$(1,1,1)$

5.16 a) SARIMA$(0,1,1) \times (0,1,1)_{12}$

b) SARIMA$(1,0,0) \times (0,1,1)_{12}$

c) SARIMA$(1,1,0) \times (1,0,1)_4$

6 Estimation

6.1 On vous donne les valeurs suivantes provenant d'un processus autorégressif d'ordre 1 :

$$-1{,}1 \quad 2{,}6 \quad 4{,}3 \quad -1{,}1 \quad 9{,}7 \quad 4{,}1 \quad -0{,}6 \quad 2{,}2.$$

a) Estimer la valeur de ϕ_{11} à partir des données ci-dessus.

b) Si on vous dit que $\phi = 0{,}85$, que vaut ϕ_{22} (la valeur théorique) ? Justifier votre réponse.

6.2 Trouver les estimateurs de Yule–Walker des paramètres θ et σ^2 du modèle MA(1) en supposant que $|\rho(1)| < \frac{1}{2}$.

6.3 a) Calculer la fonction d'autocovariance $\gamma(\cdot)$ de la série stationnaire

$$X_t = \mu + Z_t + \theta_1 Z_{t-1} + \theta_{12} Z_{t-12}, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

b) Calculer la moyenne et les autocovariances empiriques $\hat{\gamma}(h)$, $0 \leq h \leq 20$ de $\{\nabla\nabla_{12} X_t\}$, où $\{X_t, t = 1, \ldots, 72\}$ est la série du nombre de morts accidentelles `deaths.dat`.

c) En égalant les valeurs de $\hat{\gamma}(1)$, $\hat{\gamma}(11)$ et $\hat{\gamma}(12)$ trouvées en b) à $\gamma(1)$, $\gamma(11)$ et $\gamma(12)$, respectivement, de la partie a), trouver un modèle de la forme en a) pour la série $\{\nabla\nabla_{12} X_t\}$.

6.4 Soit le processus AR(2) défini comme la solution stationnaire de

$$X_t - \phi X_{t-1} - \phi^2 X_{t-2} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

a) Pour quelles valeurs de ϕ une solution stationnaire existe-t-elle ?

b) Les estimateurs des moments suivants ont été obtenus après l'observation de $X_1, \ldots, X_{200}$:

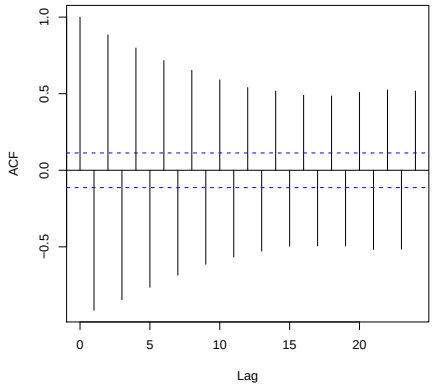
$$\hat{\gamma}(0) = 6{,}06, \quad \hat{\rho}(1) = 0{,}687, \quad \hat{\rho}(2) = 0{,}610.$$

Trouver des estimateurs de ϕ et σ^2 à l'aide des équations de Yule–Walker. (Si vous trouvez plus d'une solution, retenir celle qui est stationnaire.)

6.5 On vous donne ci-dessous les valeurs (arrondies) des fonctions d'autocovariance et d'autocorrélation partielle empiriques d'un processus stationnaire.

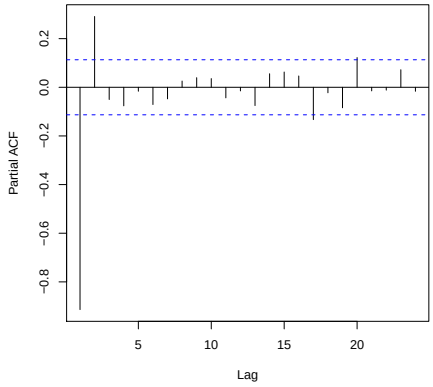
```
> acvf <- acf(X, type = "covariance", lag.max = 9)
> cbind(h = acvf$lag, `gamma(h)` = round(acvf$acf,
+     2))
```

```
      h gamma(h)
 [1,] 0     6.50
 [2,] 1    -5.94
 [3,] 2     5.74
 [4,] 3    -5.50
 [5,] 4     5.19
 [6,] 5    -4.96
 [7,] 6     4.65
 [8,] 7    -4.45
 [9,] 8     4.24
[10,] 9    -3.99
```




```
> pacf <- acf(X, type = "partial", lag.max = 10)
> cbind(h = pacf$lag, phi_hh = round(pacf$acf,
+     2))
```

```
       h phi_hh
 [1,]  1  -0.91
 [2,]  2   0.29
 [3,]  3  -0.05
 [4,]  4  -0.08
 [5,]  5  -0.02
 [6,]  6  -0.07
 [7,]  7  -0.05
 [8,]  8   0.03
 [9,]  9   0.04
[10,] 10   0.04
```



a) Trouver un modèle adéquat pour ce processus étant donné les informations ci-dessus. Justifier votre réponse.

b) Trouver des estimateurs des moments pour tous les paramètres du modèle proposé en a).

6.6 Trouver l'autocorrélation partielle de pas 3, ϕ_{33}, à l'aide des informations suivantes :

t	1	2	3	4	5
X_t	2,2	1,2	2,4	4,1	3,0

```
> ar(X, aic = FALSE, order.max = 1)

Call:
ar(x = X, aic = FALSE, order.max = 1)

Coefficients:
    1
0.249

Order selected 1  sigma^2 estimated as  1.428

> ar(X, aic = FALSE, order.max = 2)

Call:
ar(x = X, aic = FALSE, order.max = 2)

Coefficients:
      1        2
 0.3878  -0.5574

Order selected 2  sigma^2 estimated as  1.477

> ar(X, aic = FALSE, order.max = 3)

Call:
ar(x = X, aic = FALSE, order.max = 3)

Coefficients:
      1        2        3
 0.4432  -0.5959   0.0993

Order selected 3  sigma^2 estimated as  2.925
```

6.7 Les estimateurs de Yule–Walker sont sans biais. Vérifier si cette affirmation est vraie pour un modèle AR(2) par la petite expérience suivante.

i) Choisir des valeurs pour les paramètres du modèle

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

ii) Simuler 200 observations d'un processus AR(2) avec les paramètres choisis en i).

iii) Calculer et sauvegarder les estimateurs de Yule–Walker des paramètres (utiliser la fonction `ar` avec `order.max = 2`).

iv) Répéter les étapes ii) et iii) un grand nombre de fois (au moins 1000).

v) Calculer la moyenne des estimateurs de chaque paramètre.

Comparer les valeurs obtenues en v) aux vraies valeurs choisies en i). Quelle est votre conclusion ?

Réponses

6.1 a) $-0{,}2695$ b) 0

6.2 $\hat{\theta} = (1 - \sqrt{1 - 4\hat{\rho}(1)^2})/(2\hat{\rho}(1))$

6.3 c) $\nabla\nabla_{12}X_t = 28{,}83 + Z_t - 0{,}5859Z_{t-1} - 0{,}5486Z_{t-12}$, $\{Z_t\} \sim \mathrm{WN}(0, 92730)$

6.4 a) $(1 - \sqrt{5})/2 < \phi < (-1 + \sqrt{5})/2$ b) $\hat{\phi} = 0{,}509$, $\hat{\sigma}^2 = 2{,}983$

6.5 a) AR(2) b) $\hat{\phi}_1 = -0{,}6479$, $\hat{\phi}_2 = 0{,}2911$, $\hat{\sigma}^2 = 0{,}9888$

6.6 0,0993

6.7 Biais négatif

7 Prévision de séries chronologiques

7.1 Soit $\hat{X}_t(1)$ la prévision pour la période $t+1$ faite depuis le temps t à l'aide du lissage exponentiel. Calculer $\hat{X}_5(1)$ à partir des informations ci-dessous.

t	1	2	3	4	5
X_t	56	55	42	48	39
$\hat{X}_{t-1}(1)$	48,99	49,83	50,45	49,44	49,27

7.2 Soit $\{X_t\}$ un processus ARIMA$(2,1,0)$ solution de

$$(1-\phi_1 B-\phi_2 B^2)\nabla X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0,\sigma^2).$$

Trouver une expression pour $\hat{X}_n(h)$, $h < n$.

7.3 On vous donne les valeurs suivantes d'une série $\{X_t\}$:

```
Time Series:
Start = 1
End = 10
Frequency = 1
 [1] 0 0 3 4 5 7 6 3 8 9
```

Après analyse, le modèle suivant est jugé adéquat pour cette série :

$$(1+0{,}6B)(X_t - 2X_{t-1} + X_{t-2}) = Z_t \quad \{Z_t\} \sim \text{WN}(0,9),$$

où B est l'opérateur de rétro-décalage, comme d'habitude.

a) Identifier le modèle ci-dessus.

b) Soit $W_t = \nabla^2 X_t$. Identifier le processus $\{W_t\}$ et montrer qu'il est stationnaire.

c) Calculer la meilleure prévision de X_{12} et l'erreur quadratique moyenne de $\hat{X}_{10}(2)$.

7.4 Soit les vingt dernières valeurs d'une série chronologique :

t	X_t	t	X_t
$n-19$	2890	$n-9$	3602
$n-18$	2955	$n-8$	3678
$n-17$	3023	$n-7$	3750
$n-16$	3098	$n-6$	3828
$n-15$	3163	$n-5$	3912
$n-14$	3234	$n-4$	3999
$n-13$	3304	$n-3$	4080
$n-12$	3375	$n-2$	4166
$n-11$	3451	$n-1$	4249
$n-10$	3521	n	4339

On a ajusté aux n données un modèle dont l'équation caractéristique est

$$(1+0{,}6B)(1-B)^2(1-B^{12})X_t = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0,5).$$

a) Expliquer si la série $\{X_t\}$ contient de la tendance et/ou de la saisonnalité. Le cas échéant, décrire brièvement ces composantes (type, périodicité).

b) Calculer un intervalle de confiance à 95 % pour la prévision de la période $n+2$.

Réponses

7.2 $\hat{X}_n(1) = (\phi_1+1)X_n + (\phi_2-\phi_1)X_{n-1} - \phi_2 X_{n-2}$, $\hat{X}_n(2) = (\phi_1+1)\hat{X}_{n+1} + (\phi_2-\phi_1)X_n - \phi_2 X_{n-1}$, $\hat{X}_n(3) = (\phi_1+1)\hat{X}_{n+2} + (\phi_2-\phi_1)\hat{X}_{n+1} - \phi_2 X_n$, $\hat{X}_n(h) = (\phi_1+1)\hat{X}_n(h-1) + (\phi_2-\phi_1)\hat{X}_n(h-2) - \phi_2\hat{X}_n(h-3)$, $h>3$

7.3 a) ARIMA$(1,2,0)$ b) AR(1) c) 14,36 et 26,64

7.4 b) $4517{,}96 \pm 1{,}96\sqrt{14{,}8}$

A R et la régression linéaire

Comme tous les grands logiciels statistiques — et même plusieurs calculatrices scientifiques — R comporte des fonctions permettant de calculer les coefficients d'une régression simple ou multiple. Les outils disponibles vont toutefois bien au-delà de ce calcul relativement simple. Ainsi, par l'entremise de quelques fonctions génériques simples à utiliser, il est possible de générer différents graphiques relatifs à la régression, d'en calculer le tableau ANOVA et d'en extraire les informations principales, de calculer des prévisions ainsi que des intervalles de confiance. Bref, l'analyse complète d'un ensemble de données tient en quelques lignes de code ; il suffit de connaître les fonctions à utiliser.

Cette annexe présente les principales fonctions — dont la liste se trouve au tableau A.1 — utiles lors de l'analyse de données et la modélisation par régression. Il n'a cependant aucune prétention d'exhaustivité. Consulter l'aide en ligne de R, ainsi que Venables et Ripley (2002) pour de plus amples détails.

A.1 Importation de données

La modélisation statistique en R — par exemple, l'analyse de régression — repose souvent sur l'utilisation de *data frames* pour le stockage des données. On se reportera à la section 2.7 de Goulet (2007) pour une présentation de ce type d'objet.

La principale fonction utilisée pour importer des données dans R en vue d'une analyse de régression est `read.table`. Celle-ci retourne un *data frame*. Les arguments de `read.table` les plus souvent utilisés sont :

`file` le nom ou l'URL du fichier de données à importer ;

`header` `TRUE` si la première ligne du fichier à être lue contient les étiquettes des colonnes ;

`comment.char` le caractère (`#` par défaut) représentant le début d'un commentaire dans le fichier ;

`skip` le nombre de lignes à sauter au début du fichier.

Phase de l'analyse	Fonctions

Création et manipulation de *data frames*	`data.frame`
	`as.data.frame`
	`read.table`
	`cbind`
	`rbind`
	`names`, `colnames`
	`row.names`, `rownames`
	`attach`
	`detach`

Modélisation	`lm` `add1`, `addterm`[1] `drop1`, `dropterm`[1] `step`, `stepAIC`[1]

Analyse des résultats et diagnostics	`summary`
	`anova`
	`coef`, `coefficients`
	`confint`
	`residuals`
	`fitted`
	`deviance`
	`df.residual`

Mise à jour et prévisions	`update` `predict`

Graphiques	`plot`
	`abline`
	`matplot`
	`matlines`

[1] Dans le package `MASS`.

TABLE A.1: Principales fonctions R pour la régression linéaire

A.2 Formules

Lorsque l'on fait une régression, il faut informer R des variables que l'on entend inclure dans celle-ci et leurs relations entre elles. La convention utilisée dans le langage S est celle dite des «formules». Le tableau A.2 présente quelques exemples de formulation de modèles linéaires simples en S.

Pour une utilisation de base des fonctions de régression, la connaissance des règles suivantes suffit.

1. Les opérateurs + et – prennent une nouvelle signification dans les formules : + signifie «inclusion» et –, «exclusion».

Modèle mathématique	Formule S
$y_t = \alpha + \beta x_t + \varepsilon_t$	`y ~ x` `y ~ 1 + x`
$y_t = \beta x_t + \varepsilon_t$	`y ~ -1 + x` `y ~ x - 1`
$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t$	`y ~ x1 + x2` `y ~ x` où `x <- cbind(x1, x2)`

TABLE A.2: Modèles linéaires simples et leur formulation en S

2. Le terme constant d'une régression est inclus implicitement. Pour l'exclure explicitement (pour la régression passant par l'origine), il faut donc ajouter un terme `-1` du côté droit de la formule.
3. Dans une régression multiple, on peut soit lister toutes les variables à inclure du côté droit de la formule, soit ne spécifier qu'une matrice contenant ces variables (dans les colonnes).

Consulter les sections 6.2 de Venables et Ripley (2002) et 11.1 de Venables et collab. (2005) pour plus de détails.

A.3 Modélisation des données

Supposons que l'on souhaite étudier la relation entre la variable indépendante `x1` et la variable dépendante (ou réponse) `y1` du jeu de données `anscombe`. La première étape de la modélisation des données en régression linéaire simple consiste habituellement à représenter celles-ci graphiquement.

La fonction `plot` est une fonction générique comportant des méthodes pour un grand nombre de classes d'objets différentes. Puisqu'il existe une méthode pour les objets de classe `formula`, on peut tracer un graphique de `y1` en fonction de `x1` avec

```
> plot(y1 ~ x1, data = anscombe)
```

ou, si les colonnes du *data frame* `anscombe` sont visibles, simplement avec

```
> plot(y1 ~ x1)
```

Le résultat de ces commandes se trouve à la figure A.1.

Le graphique nous montre qu'il est raisonnable de postuler une relation linéaire entre les éléments de `y1` et `x1`. On pose donc le modèle

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

où y_t et x_t, $t = 1, \ldots, 11$ sont les éléments des vecteurs `y1` et `x1`, respectivement, et ε_t est le terme d'erreur.

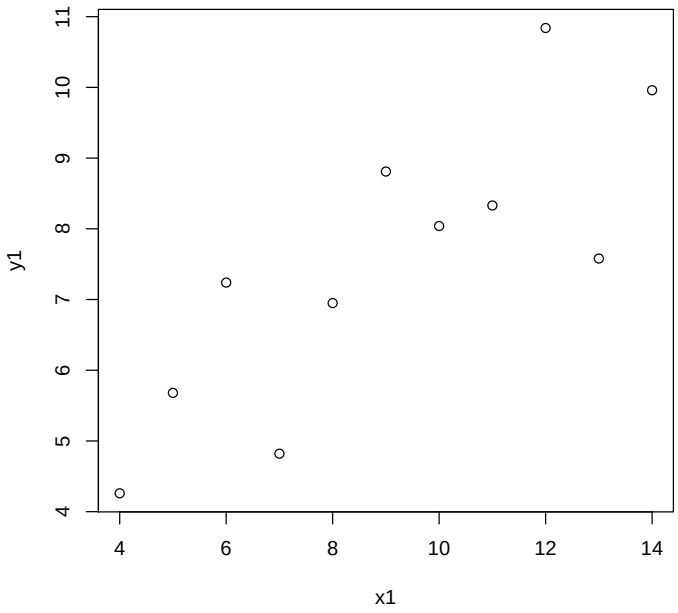


FIGURE A.1: Relation entre `y1` et `x1` des données `anscombe`

C'est avec la fonction `lm` (pour *linear model*) que l'on calcule les estimateurs des coefficients de la régression β_0 et β_1. De façon simplifiée, cette fonction prend en arguments une formule et un *data frame* comprenant les données relatives aux termes de la formule. La fonction `lm` retourne un objet de classe `lm`, classe pour laquelle il existe de nombreuses méthodes.

```
> (fit <- lm(y1 ~ x1, data = anscombe))

Call:
lm(formula = y1 ~ x1, data = anscombe)

Coefficients:
(Intercept)           x1
     3.0001       0.5001

> class(fit)

[1] "lm"
```

Lorsque plusieurs variables explicatives sont disponibles, l'analyste doit souvent choisir les variables les plus significatives pour la régression. Les

techniques d'élimination successive, de sélection successive et de sélection pas à pas, qui reposent toutes sur les tests F partiels, sont alors populaires pour parvenir au modèle le plus utile. Ces techniques sont mises en œuvre, respectivement, dans les fonctions `dropterm`, `addterm` et `stepAIC` du package `MASS` (Venables et Ripley, 2002).

A.4 Analyse des résultats

Le résultat de la fonction `lm` est une liste dont on peut extraire manuellement les différents éléments (consulter la rubrique d'aide). Grâce à quelques fonctions génériques disposant d'une méthode pour les objets de classe `lm`, il est toutefois facile et intuitif d'extraire les principaux résultats d'une régression :

1. `coef` ou `coefficients` extraient les coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$ de la régression ;
2. `fitted` extrait les valeurs ajustées $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$;
3. `residuals` extrait les résidus $y_t - \hat{y}_t$;
4. `deviance` retourne la somme des carrés des résidus $\text{SSR} = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$;
5. `df.residual` extrait le nombre de degrés de liberté de la somme des carrés des résidus.

La fonction générique `summary` présente les informations ci-dessus de manière facile à consulter. Plus précisément, le sommaire de la régression contient, outre le modèle utilisé et les estimateurs des coefficients de la régression : les résultats des tests t, la valeur du coefficient de détermination

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

et celle du coefficient de détermination ajusté

$$R_a^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1},$$

ainsi que le résultat du test F global.

La fonction `confint` calcule les intervalles de confiance des paramètres de la régression.

D'autre part, le tableau d'analyse de variance (séquentiel, en régression multiple) est calculé avec la fonction générique `anova`.

Pour ajouter la droite de régression au graphique créé au début de l'analyse, utiliser la fonction `abline`, qui dispose elle aussi d'une méthode pour les objets de classe `lm`.

A.5 Diagnostics

Les statistiques servant à mesurer la qualité d'un modèle de régression (R^2, R^2 ajusté, statistiques t et F) sont calculées par les fonctions `summary` et `anova`.

La méthode de la fonction `plot` pour les objets de classe `lm` produit une série de six graphiques (quatre dans R avant la version 2.2.0) permettant de juger de la qualité d'une régression. Consulter la rubrique d'aide de la fonction `plot.lm` pour plus de détails.

A.6 Mise à jour des résultats et prévisions

Il peut arriver que, une fois la modélisation d'un ensemble de données effectuée, l'on doive ajouter ou modifier une ou plusieurs données ou variables. Plutôt que de reprendre toute la modélisation avec la fonction `lm`, il peut alors s'avérer plus simple et élégant d'utiliser la fonction `update` :

```
> update(fit, ~. + x4)


Call:
lm(formula = y1 ~ x1 + x4, data = anscombe)

Coefficients:
(Intercept)           x1           x4
    4.33291      0.45073     -0.09873
```

Le calcul de prévisions et d'intervalles de confiance pour la régression et pour les prévisions se fait avec la fonction générique `predict` et sa méthode pour les objets de classe `lm`. Par défaut, `predict` calculera les prévisions pour les valeurs x_t, $t = 1, \ldots, n$. Par conséquent, le résultat de `predict` sera le même que celui de `fitted` :

```
> all.equal(predict(fit), fitted(fit))


[1] TRUE
```

Comme on souhaite généralement prévoir la réponse pour d'autres valeurs de la variable indépendante, on spécifiera celles-ci par le biais d'un *data frame* passé à `predict` avec l'option `newdata`.

La fonction `predict` peut également servir au calcul des bornes d'intervalles de confiance et de prévision. Pour calculer les bornes d'un intervalle de confiance, on ajoutera l'argument `interval = "confidence"`, alors que pour les bornes d'un intervalle de prévision on utilise `interval = "prediction"`. Le niveau de confiance est déterminé avec l'argument `level` (0,95 par défaut). Le résultat est une matrice de trois colonnes dont la première contient les prévisions et les deux autres les bornes inférieures (`lwr`) et supérieures (`upr`) des intervalles de confiance.

On ajoute les limites des intervalles de confiance au graphique des données avec les fonctions `matlines` ou `matplot`. Consulter les rubriques d'aide et les exemples pour de plus amples détails.

A.7 Exemples

```
###
### IMPORTATION DE DONNÉES
###

## On importe les données du fichier anscombe.dat. On peut
## lire le fichier directement depuis Internet. De plus, les
## lignes débutant par # sont automatiquement reconnues comme
## des lignes de commentaires.
anscombe <- read.table(
  "http://vgoulet.act.ulaval.ca/pub/donnees/anscombe.dat")

## Ce jeu de données se trouve en fait déjà dans R et il est
## chargé en mémoire avec 'data'.
data(anscombe)

## Le résultat est un data frame, soit
mode(anscombe)             # ... une liste...
class(anscombe)            # ... de classe "data.frame"

## Extraction des étiquettes des colonnes et des lignes.
names(anscombe)            # étiquettes des colonnes
row.names(anscombe)        # étiquettes des lignes

###
### MODÉLISATION DES DONNÉES
###

## Relation graphique entre les variables y1 et x1 des données
## anscombe.
plot(y1 ~ x1, data = anscombe)

## On peut aussi rendre les colonnes du data frame visibles
## dans l'espace de travail et référer ensuite à celles-ci
## directement.
attach(anscombe)
plot(y1 ~ x1)

## Estimation des coefficients de la régression. Il est
## recommandé de sauvegarder les résultats dans un objet (de
## classe "lm") puisqu'il existe de multiples méthodes pour de
## tels objets.
( fit <- lm(y1 ~ x1, data = anscombe) )
class(fit)

###
### ANALYSE DES RÉSULTATS
###
```

```
## Le sommaire de la régression contient, outre le modèle
## utilisé, les résultats des tests t, la valeur des
## coefficients de détermination et de détermination ajusté,
## ainsi que le résultat du test F global.
summary(fit)

## Calcul du coefficient de détermination à la main.
attach(anscombe)
1 - sum(residuals(fit)^2)/sum((y1 - mean(y1))^2)
1 - deviance(fit)/sum((y1 - mean(y1))^2)
detach(anscombe)

## Intervalles de confiance pour les paramètres de la
## régression.
confint(fit)

## Le tableau d'analyse de variance (séquentiel, en régression
## multiple) est calculé avec la fonction générique 'anova'.
anova(fit)

## Pour ajouter la droite de régression au graphique créé
## précédemment, utiliser la fonction générique
## 'abline'. L'ordonnée à l'origine et la pente sont extraites
## de l'objet 'fit'.
abline(fit)

###
### MISE À JOUR DES RÉSULTATS ET PRÉVISION
###

## La fonction 'update' est utilisé pour modifier une ou
## plusieurs données dans le modèle ou pour enlever ou ajouter
## une ou plusieurs variables dans le modèle.
anscombe$x1[11] <- 6       # modification d'une donnée
update(fit)                # modèle mis à jour
update(fit, . ~ . + x4)    # ajout de la variable "x4"

## Retour au modèle d'origine
fit <- lm(y1 ~ x1, data = anscombe)

## Prévisions du modèle pour des valeurs de la variables "x1"
## de 3 et 15:
predict(fit, newdata = data.frame(x1 = c(3, 15)))

## Calcul des intervalles de confiance et de prévision pour
## les prévisions ci-dessus avec un niveau de confiance de
## 90%.
predict(fit, newdata = data.frame(x1 = c(3, 15)),
        interval = "confidence", level = 0.90)
```

```
predict(fit, newdata = data.frame(x1 = c(3, 15)),
        interval = "prediction", level = 0.90)

## Ajout des limites supérieures et inférieures des
## intervalles de confiance au graphique des données. On
## utilise la fonction 'matplot' qui prend en argument deux
## matrices 'x' et 'y' et produit un graphique des coordonnées
## de la première colonne de 'x' avec la première colonne de
## 'y', la seconde de 'x' avec la seconde de 'y', etc.
##
## Afin d'obtenir un beau graphique, il faut s'assurer de
## mettre les valeurs de 'x' en ordre croissant et de classer
## celles de 'y' en conséquence.
##
## En fait, on utilise la fonction 'matlines' qui ajoute à un
## graphique existant. La fonction 'matplot' créerait un
## nouveau graphique. (Note: il est possible de combiner les
## deux commandes matlines() ci-dessous en une seule.)
##
## Rendre les colonnes visibles.
attach(anscombe)

## Calcul des prévisions et des intervalles pour toutes les
## valeurs de "x1".
pred.ci <- predict(fit, interval = "confidence")
pred.pi <- predict(fit, interval = "prediction")
matlines(sort(x1), pred.ci[order(x1), -1],
         lty = 2, col = "red")
matlines(sort(x1), pred.pi[order(x1), -1],
         lty = 2, col = "green")

## Pour éviter que des lignes ne dépassent à extérieur du
## graphique, il faut trouver, avant de faire le graphique,
## les limites inférieure et supérieure des ordonnées. La
## fonction 'matplot' peut combiner des lignes et des points,
## ce qui permet de faire tout le graphique avec une seule
## commande.
y <- cbind(y1, pred.ci, pred.pi[, -1])
matplot(sort(x1), y[order(x1),],
     pch = 19, type = c("p", rep("l", 5)),
     lty = c(0, 1, rep(2, 4)),
     col = c("black", "blue", "red", "red", "green", "green"))
```

A.8 Exercices

1.1 Importer dans S-Plus ou R le jeu de données `steam.dat` à l'aide de la fonction `read.table`. Les trois première lignes du fichier sont des lignes de commentaires débutant par le caractère `#`. La quatrième ligne contient les étiquettes des colonnes.

1.2 Rendre les colonnes individuelles de l'ensemble de données `steam` visibles dans l'espace de travail.

1.3 Faire (même à l'aveuglette) l'analyse de régression de la variable `Y` en fonction de la variable `X1` des données `steam`.

a) Évaluer visuellement le type de relation pouvant exister entre `Y` et `X1`.

b) Évaluer les coefficients d'une régression linéaire entre `Y` et `X1` et ajouter la droite de régression ainsi obtenue au graphique créé en a).

c) Répéter la partie b) en forçant la droite de régression à passer par l'origine $(0,0)$. Quel modèle semble le plus approprié ?

d) Le coefficient de détermination R^2 mesure la qualité de l'ajustement d'une droite de régression aux données. Calculer le R^2 pour les modèles en b) et c). Obtient-on les mêmes résultats que ceux donnés par `summary` ? Semble-t-il y avoir une anomalie ?

e) Calculer les bornes d'intervalles de confiance pour la droite de régression des deux modèles.

f) Calculer les prévisions de chaque modèle pour toutes les valeurs de `X1` ainsi que les bornes d'intervalles de confiance pour ces prévisions.

g) Ajouter au graphique créé précédemment les bornes inférieures et supérieures des intervalles de confiance calculées en e) et f). Utiliser des types de lignes (option `lty`) et des couleurs (option `col`) différents pour chaque ensemble de limites.

1.4 Répéter l'exercice précédent en ajoutant la variable `X5` à l'analyse, transformant ainsi le modèle de régression linéaire simple en un modèle de régression multiple.

B R et les séries chronologiques

R offre toutes les fonctions nécessaires pour faire l'analyse complète de séries chronologiques : création et manipulation d'objets de classe «série chronologique», identification et définition d'un modèle, estimation des paramètres, calcul de prévisions et simulation de séries. La liste des principales fonctions utilisées pour l'analyse de séries chronologiques se trouve au tableau B.1. Quelques autres fonctions sont disponibles, notamment pour le traitement des séries multivariées ; voir Venables et Ripley (2002, chapitre 14).

B.1 Importation des données

Les séries chronologiques sont typiquement créées à partir de vecteurs simples. Or, la fonction `scan` lit justement l'intégralité des données du fichier dont le nom est donné en premier argument, puis retourne un vecteur. Elle constitue donc le meilleur choix pour importer des séries chronologiques dans R.

Contrairement à `read.table`, la fonction `scan` ne reconnaît pas les commentaires par défaut. Toutefois, il suffit de spécifier le caractère représentant le début d'un commentaire avec l'argument `comment.char`.

B.2 Création et manipulation de séries

La façon la plus simple de créer des séries chronologiques est d'utiliser la fonction `ts`. Les fonctions `rts` (séries régulières), `cts` (séries avec dates) et `its` (séries irrégulières) sont plus récentes et parfois nécessaires.

La fonction `window` permet d'extraire un sous-ensemble d'une série chronologique en spécifiant des dates de début et de fin plutôt que des positions dans le vecteur des observations.

B.3 Identification

La première chose à faire dans l'analyse d'une série chronologique consiste à tracer le graphique de la série et son corrélogramme. Le premier graphique

Phase de l'analyse	Fonctions

Création et manipulation de séries	`ts`, `rts`, `cts`, `its`
	`time`
	`start`
	`end`
	`frequency`
	`cycle`
	`window`
	`diff`
	`filter`
	`stl`

Identification	`plot`, `ts.plot`
	`acf`
	`pacf`

Estimation	`ar`
	`arima`
	`ARMAacf`
	`ARMAtoMA`

Diagnostics	`tsdiag`
Calcul de prévisions	`predict`
Simulation	`arima.sim`

TABLE B.1: Principales fonctions R pour l'analyse de séries chronologiques

est obtenu avec la fonction spécialisée `ts.plot` ou, plus simplement, avec `plot`.

La fonction `acf` peut calculer et tracer les fonctions (échantillonnales) d'autocovariance $\hat{\gamma}_X(h)$, d'autocorrélation $\hat{\rho}_X(h)$ ou d'autocorrélation partielle $\hat{\phi}_{hh}$ selon la valeur de son argument `type` (spécifier `covariance`, `correlation` et `partial`, respectivement). Par défaut, `acf` trace le corrélogramme de la série. Si l'on souhaite obtenir les valeurs de la fonction d'autocorrélation sans graphique, ajouter l'option `plot = FALSE` dans l'appel de la fonction.

La fonction d'autocorrélation partielle s'obtient aussi plus directement avec la fonction `pacf`.

B.4 Estimation

Un processus ARMA d'ordre (p,q) est défini comme la solution $\{X_t\}$ des équations

$$\phi(B)X_t = \theta(B)Z_t, \quad t = 0, \pm 1, \pm 2, \ldots$$

où

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$
$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

$BX_t = X_{t-1}$ et $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. C'est là précisément la paramétrisation retenue dans R.

Un processus ARIMA est un processus non stationnaire qui, une fois la d^{e} différence appliquée sur la série, est un processus ARMA. Autrement dit, $\{X_t\} \sim \text{ARIMA}(p,d,q)$ si $\{\nabla^d X_t\} \sim \text{ARMA}(p,q)$ et donc $\{X_t\}$ est la solution stationnaire de

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t.$$

L'étape de la modélisation consiste donc à ajuster un modèle ARIMA aux observations d'une série chronologique en estimant les paramètres $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ et σ^2. C'est le rôle des fonctions `ar` et `arima`.

La fonction `ar` est très pratique pour une première estimation : elle ajuste un modèle $\text{AR}(p)$ aux données pour plusieurs valeurs de p à l'aide des équations de Yule–Walker (par défaut) et retourne le modèle avec la plus faible statistique AIC. Cette statistique est égale à moins deux fois la fonction de log-vraisemblance pénalisée par le nombre de paramètres dans le modèle.

D'autre part, la fonction `arima` estime les paramètres d'un modèle ARIMA d'ordre (p,d,q) par la technique du maximum de vraisemblance (par défaut). Contrairement à `ar`, la fonction `arima` ne fait pas un choix parmi plusieurs modèles — il y en aurait beaucoup trop. Il faut donc spécifier les valeurs de p, d et q à l'aide de l'argument `order` (un vecteur de trois éléments). À noter que la fonction `arima` inclut une moyenne μ dans le modèle lorsque $d = 0$.

Finalement, les séries comportant de la saisonnalité sont modélisées à l'aide des très généraux processus SARIMA. Le processus SARIMA d'ordre $(p,d,q) \times (P,D,Q)_s$ est défini comme la solution stationnaire $\{X_t\}$ des équations

$$\phi(B)\Phi(B^s) W_t = \theta(B)\Theta(B^s) Z_t, \quad W_t = \nabla^d \nabla_s^D X_t,$$

où

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$
$$\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$$
$$\theta(z) = 1 + \theta_1 z + \cdots + \theta z^q$$
$$\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$$

et $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

Les paramètres d'un modèle SARIMA sont toujours estimés à l'aide de la fonction `arima` en spécifiant les valeurs de P, D, Q et s par l'argument `seasonal`.

La fonction `ARMAacf` permet de calculer la fonction d'autocorrélation ou d'autocorrélation partielle théorique d'un processus ARMA quelconque. La

fonction `ARMAtoMA`, comme son nom l'indique, permet quant à elle d'inverser un processus ARMA quelconque. Toutes deux peuvent s'avérer utiles pour vérifier ses calculs.

B.5 Diagnostics

La fonction `tsdiag` permet de juger rapidement de la qualité d'ajustement d'un modèle. La fonction crée trois graphiques : la série des résidus $\{Z_t\}$, le corrélogramme de cette même série et un graphique de la valeur p de la statistique de Ljung–Box pour des valeurs de $H = 1,2,\ldots$ La statistique de Ljung–Box est simplement une version améliorée de la statistique du test portmanteau :

$$Q_{LB} = n(n+2)\sum_{h=1}^{H}\frac{\hat{\rho}^2(h)}{n-h}.$$

Si l'ajustement du modèle est bon, les résidus forment un bruit blanc. Le corrélogramme généré par `tsdiag` devrait donc ressembler à celui d'un bruit blanc et les valeurs p devraient être grandes (on ne rejette pas l'hypothèse de bruit blanc).

B.6 Calcul de prévisions

La prévision de modèles ARIMA repose sur une méthode de la fonction générique `predict` pour les objets de classe `Arima` (créés par la fonction `arima`). Les prévisions sont donc calculées exactement comme en régression, outre que l'argument principal de `predict` devient le nombre de périodes pour lesquelles l'on veut une prévision, et non les valeurs d'une ou plusieurs variables indépendantes. L'écart type de chaque prévision est également calculé par `predict`, ce qui permet de calculer des bornes d'intervalles de prévision.

B.7 Simulation

La simulation de séries chronologiques ARIMA est très simple avec la fonction `arima.sim`. Il suffit de savoir comment spécifier le modèle à simuler. L'argument `model` de la fonction `arima.sim` est une liste comportant un ou plusieurs des éléments `ar`, `ma` et `order`. Le premier de ces éléments est le vecteur des paramètres $\phi_1,\ldots,\phi_p$; le second, le vecteur des paramètres $\theta_1,\ldots,\theta_q$; le troisième, le vecteur (p,d,q) — utilisé seulement si $d > 0$.

Par défaut, le bruit blanc est généré avec une loi normale centrée réduite. On peut changer la distribution à utiliser avec l'argument `rand.gen` ou passer des arguments différents à la fonction de simulation du bruit blanc directement dans l'appel de `arima.sim`. Voir les exemples à la section B.8.

B.8 Exemples

```
###
### IMPORTATION DE DONNÉES
###

## On utilise la fonction 'scan' pour importer des données
## sous forme de vecteur. Les fichiers 'deaths.dat' et
## 'strikes.dat' comptent chacun trois lignes de commentaires
## en début de fichier. On spécifie le caractère délimitant
## les commentaires avec l'argument 'comment.char'. De plus,
## on peut lire les fichiers directement depuis Internet.
deaths <- scan(
  "http://vgoulet.act.ulaval.ca/pub/donnees/deaths.dat",
  comment.char = "#")
strikes <- scan(
  "http://vgoulet.act.ulaval.ca/pub/donnees/strikes.dat",
  comment.char = "#")

###
### CRÉATION ET MANIPULATION DE SÉRIES
###

## Le fichier deaths.dat contient le nombre mensuel de morts
## accidentelles, 1973-1978. On transforme l'objet 'deaths'
## en une série chronologique aux propriétés correspondantes
## avec la fonction 'ts'.
( deaths <- ts(deaths, start = 1973, frequency = 12) )

## Le résultat est une série chronologique.
mode(deaths)               # un vecteur...
class(deaths)              # ... de classe "ts"

## Même chose avec l'objet 'strikes', qui contient le nombre
## de grèves aux États-Unis entre 1951-1980. L'argument
## 'frequency' n'est pas nécessaire: les séries sont
## annuelles par défaut.
( strikes <- ts(strikes, start = 1951) )

## La fonction 'window' est la façon élégante d'extraire des
## observations d'une série. Ici, on extrait les données
## 'deaths' du mois de février 1974 au mois d'octobre 1974,
## inclusivement.
window(deaths, start = c(1974, 2), end = c(1974, 10))

###
### IDENTIFICATION
###
```

```
## Graphiques des séries 'deaths' et 'strikes'.
plot(deaths)
plot(strikes)

## Corrélogramme de la série 'deaths'. Par défaut, 'acf'
## trace le corrélogramme.
acf(deaths)

## Pour obtenir les valeurs numériques de la fonction
## d'autocorrélation empirique, utiliser l'argument
## 'plot = FALSE'.
acf(deaths, plot = FALSE)

###
### MODÉLISATION
###

## On ajuste d'abord un modèle autorégressif pur aux données
## 'strikes' avec la fonction 'ar'.
( modele <- ar(strikes) )  # modèle AR(2) choisi

## On peut comparer les statistiques AIC des divers modèles.
## La statistique AIC du modèle AR(2) ne vaut pas vraiment 0;
## les statistiques sont simplement mise à l'échelle avec
## cette valeur comme référence.
modele$aic

## Ajustement d'un modèle ARIMA(1, 2, 1) aux données
## 'strikes'.
( fit.strikes <- arima(strikes, order = c(1, 2, 1)) )

## Ajustement d'un modèle SARIMA(0, 1, 1) x (0, 1, 1)_{12}
## aux données 'deaths'. Par défaut, la fréquence de la série
## (s = 12) est supposée identique à celle spécifiée dans
## l'objet. Il n'est donc pas nécessaire de préciser la
## valeur de s dans l'appel de 'arima', ici, puisque la série
## a été correctement définie dès le départ.
( fit.deaths <- arima(deaths, order = c(0, 1, 1),
                      seasonal = c(0, 1, 1)) )

## Cinq premières valeurs de la fonction d'autocorrélation
## théorique d'un processus ARMA(1, 1) avec phi = 0,6 et
## theta = -0,4.
ARMAacf(ar = 0.6, ma = -0.4, lag.max = 5)

## Cinq premiers coefficients de la représentation MA(infini)
## d'un processus AR(1) avec phi = 0,8.
ARMAtoMA(ar = 0.8, lag.max = 3)

###
```

```
### DIAGNOSTICS
###

## Vérification graphique de la qualité de l'ajustement du
## modèle ARIMA(1, 2, 1) aux données 'strikes' à l'aide de la
## fonction 'tsdiag'.
tsdiag(fit.strikes)

## Idem pour le modèle des données 'deaths'.
tsdiag(fit.deaths)

###
### PRÉVISIONS
###

## Prévision des six prochaines valeurs de la série 'deaths'
## à partir du modèle SARIMA.
( pred <- predict(fit.deaths, n.ahead = 6) )

## Graphique présentant la série originale, les prévisions
## des six prochaines années et les intervalles de prévision.
ts.plot(deaths,
        pred$pred,
        pred$pred + 1.96 * pred$se,
        pred$pred - 1.96 * pred$se,
        col = c(1, 2, 4, 4), lty = c(1, 3, 2, 2))

###
### SIMULATION
###

## Simulation de 10 observations d'un modèle ARMA(1, 1) avec
## phi = 0,8, theta = 0,5 et sigma^2 = 1.
arima.sim(10, model = list(ar = 0.8, ma = -0.5))

## Simulation de 10 observations d'un modèle ARIMA(2, 1, 1)
## avec phi_1 = 0,6, phi_2 = 0,3, theta = -0,2 et
## sigma^2 = 25.
arima.sim(10, model = list(ar = c(0.6, 0.3), ma = 0.2,
               order = c(2, 1, 1), sd = 5))
```

B.9 Exercices

Avant de faire les exercices ci-dessous, importer dans R les ensembles de données `deaths`, `strikes`, `uspop` et `wine`. Utiliser pour ce faire les commandes suivantes :

```
> deaths <- ts(scan("deaths.dat", comment.char = "#"),
+     start = 1973, frequency = 12)
> strikes <- ts(scan("strikes.dat", comment.char = "#"),
```

```
+     start = 1951)
> uspop <- ts(scan("uspop.dat", comment.char = "#"),
+     start = 1790, deltat = 10)
> wine <- ts(scan("wine.dat", comment.char = "#"),
+     start = 1980, frequency = 12)
```

Il est possible d'afficher plus d'un graphique à la fois sur un périphérique graphique en le subdivisant à l'aide des options `mfrow` (remplissage par ligne) et `mfcol` (remplissage par colonne) de la fonction `par`. Par exemple,

```
> par(mfrow = c(2, 1))
```

divisera la «page» en deux lignes et une colonne. Les deux prochains graphiques se retrouveront donc l'un au-dessus de l'autre.

2.1 Exécuter les commandes `par` ci-dessous. Après chacune, exécuter les commandes suivantes pour constater l'effet de `par` sur le périphérique graphique :

```
> plot(deaths)
> plot(strikes)
> plot(uspop)
> acf(wine)
```

a) `par(mfrow = c(2, 1))`

b) `par(mfrow = c(1, 2))`

c) `par(mfrow = c(2, 2))`

d) `par(mfcol = c(2, 2))`

2.2 Simuler 100 observations des processus suivants. Pour chacun, tracer sur un seul périphérique graphique le graphique de la série simulée ainsi que son corrélogramme (l'un au-dessus de l'autre). Comparer le corrélogramme à la fonction d'autocorrélation théorique.

a) $\{Z_t\} \sim \text{WN}(0,2)$ où chaque Z_t est une variable aléatoire normale de moyenne 0 et variance 2.

b) $\{X_t\} \sim \text{MA}(1)$ avec $\theta = 0{,}8$ et $\sigma^2 = 1$.

c) $\{X_t\} \sim \text{MA}(1)$ avec $\theta = -0{,}6$ et $\sigma^2 = 100$.

d) $\{X_t\} \sim \text{MA}(2)$ avec $\theta_1 = 0{,}5$, $\theta_2 = 0{,}4$ et $\sigma^2 = 1$.

e) $\{X_t\} \sim \text{AR}(1)$ avec $\phi = 0{,}8$ et $\sigma^2 = 1$.

f) $\{X_t\} \sim \text{AR}(1)$ avec $\phi = -0{,}9$ et $\sigma^2 = 100$.

g) $\{X_t\} \sim \text{AR}(2)$ avec $\phi = 0{,}7$, $\phi_2 = -0{,}1$ et $\sigma^2 = 1$.

2.3 Ajuster un modèle autorégressif pur aux données `lh` du package `MASS` à l'aide de la fonction `ar`.

2.4 L'exercice suivant, bien qu'un peu artificiel, illustre la procédure d'analyse d'une série chronologique.

a) Simuler 100 valeurs d'un processus ARMA$(1,1)$ avec $\phi = 0{,}7$, $\theta = 0{,}5$ et $\sigma^2 = 1$.

b) Tracer les graphiques suivants sur un même périphérique : la série, le corrélogramme et la fonction d'autocorrélation partielle empirique.

c) Ajuster un modèle ARMA$(1,1)$ aux données simulées au point a) en estimant les paramètres à l'aide de la fonction `arima`. Les estimateurs devraient être près des valeurs utilisées lors de la simulation.

d) Vérifier la qualité de l'ajustement du modèle obtenu en c) à l'aide de la fonction `tsdiag`.

e) Prévoir les 12 prochaines valeurs du processus. Tracer un graphique de la série originale et des prévisions en fournissant les deux séries en argument à la fonction `ts.plot`.

C Éléments d'algèbre matricielle

Cette annexe présente quelques résultats d'algèbre matricielle utiles en régression linéaire.

C.1 Trace

La *trace* d'une matrice est la somme des éléments de la diagonale.

Théorème C.1. *Soient* $\mathbf{A} = [a_{ij}]$ *et* $\mathbf{B} = [b_{ij}]$ *des matrices carrées* $k \times k$. *Alors*

a) $\text{tr}(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$

b) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.

Théorème C.2. *Soient les matrices* $\mathbf{A}_{p\times q}$ *et* $\mathbf{B}_{q\times p}$. *Alors* $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

Démonstration. Posons $\mathbf{C} = \mathbf{AB}$ et $\mathbf{D} = \mathbf{BA}$. Par définition du produit matriciel, l'élément c_{ij} de la matrice $\mathbf{C}$ est égal au produit scalaire entre la ligne i de $\mathbf{A}$ et de la colonne j de $\mathbf{B}$, soit

$$c_{ij} = \sum_{k=1}^{q} a_{ik} b_{kj}.$$

Les éléments de la diagonale de $\mathbf{C}$ sont donc $c_{ii} = \sum_{j=1}^{q} a_{ij} b_{ji}$ et, par symétrie, ceux de la diagonale de $\mathbf{D}$ sont $d_{jj} = \sum_{i=1}^{p} b_{ji} a_{ij}$. Or,

$$\begin{aligned}
\text{tr}(\mathbf{C}) &= \sum_{i=1}^{p} c_{ii} \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} a_{ij} b_{ji} \\
&= \sum_{j=1}^{q} \sum_{i=1}^{p} b_{ji} a_{ij} \\
&= \sum_{j=1}^{p} d_{jj} \\
&= \text{tr}(\mathbf{D}).
\end{aligned}$$

□

C.2 Formes quadratiques et dérivées

Soit $\mathbf{A} = [a_{ij}]$ une matrice $k \times k$ symétrique et $\mathbf{x} = (x_1, \ldots, x_k)'$ un vecteur. Alors

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}x_i x_j$$

est une forme quadratique.

Par exemple, si

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{et} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix},$$

alors

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}x_i x_j \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2. \end{aligned}$$

Remarque. Si $\mathbf{A}$ est diagonale, $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{k} a_{ii}x_i^2$.

Théorème C.3. *Soient* $\mathbf{x} = (x_1, \ldots, x_k)'$ *et* $\mathbf{a} = (a_1, \ldots, a_k)'$, *d'où* $\mathbf{x}'\mathbf{a} = a_1x_1 + \cdots + a_kx_k = \sum_{i=1}^{k} a_ix_i$. *Alors*

$$\begin{aligned} \frac{d}{d\mathbf{x}}\mathbf{x}'\mathbf{a} &= \frac{d}{d\mathbf{x}}\sum_{i=1}^{k} a_ix_i \\ &= \begin{bmatrix} \frac{d}{dx_1}\sum_{i=1}^{k} a_ix_i \\ \vdots \\ \frac{d}{dx_k}\sum_{i=1}^{k} a_ix_i \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} \\ &= \mathbf{a}. \end{aligned}$$

Théorème C.4. *Soit* $\mathbf{A}_{k\times k}$ *une matrice symétrique. Alors*

$$\frac{d}{d\mathbf{x}}\mathbf{x}'\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}.$$

Démonstration. On a

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}x_i x_j \\ &= \sum_{i=1}^{k} a_{ii}x_i^2 + \sum_{i=1}^{k}\sum_{\substack{j=1 \\ j\neq i}}^{k} a_{ij}x_i x_j. \end{aligned}$$

Par conséquent, pour $t = 1, \ldots, k$ et puisque $a_{ij} = a_{ji}$,

$$\begin{aligned}\frac{\partial}{\partial x_t}\mathbf{x}'\mathbf{A}\mathbf{x} &= 2a_{tt}x_t + \sum_{\substack{i=1\\ i\neq t}}^{k} a_{it}x_i + \sum_{\substack{j=1\\ j\neq t}}^{k} a_{tj}x_j \\ &= 2\sum_{i=1}^{k} a_{it}x_t,\end{aligned}$$

d'où

$$\frac{d}{d\mathbf{x}}\mathbf{x}'\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}.$$

□

Théorème C.5. *Si $f(\mathbf{x})$ est une fonction quelconque du vecteur $\mathbf{x}$, alors*

$$\frac{d}{d\mathbf{x}} f(\mathbf{x})'\mathbf{A}f(\mathbf{x}) = 2\left(\frac{d}{d\mathbf{x}}f(\mathbf{x})\right)'\mathbf{A}f(\mathbf{x}).$$

Vérifier en exercice les résultats ci-dessus pour une matrice $\mathbf{A}$ 3×3.

C.3 Vecteurs et matrices aléatoires

Soit $X_1, \ldots, X_n$ des variables aléatoires. Alors

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

est un *vecteur aléatoire*. On définit le vecteur espérance

$$E[\mathbf{x}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$$

et la matrice de variance-covariance

$$\begin{aligned}\mathbf{V}[\mathbf{x}] &= E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])'] \\ &= \begin{bmatrix} \mathrm{Var}[X_1] & \ldots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \ldots & \mathrm{Var}[X_n] \end{bmatrix}\end{aligned}$$

Théorème C.6. *Soit $\mathbf{x}$ un vecteur aléatoire et $\mathbf{A}$ une matrice de constantes. Alors*
a) $E[\mathbf{A}\mathbf{x}] = \mathbf{A}E[\mathbf{x}]$

b) $\mathbf{V}[\mathbf{Ax}] = \mathbf{AV}[\mathbf{x}]\mathbf{A}'$.

Démonstration de b).

$$\begin{aligned}\mathbf{V}[\mathbf{Ax}] &= E[(\mathbf{Ax} - E[\mathbf{Ax}])(\mathbf{Ax} - E[\mathbf{Ax}])'] \\ &= E[\mathbf{A}(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])'\mathbf{A}'] \\ &= \mathbf{AV}[\mathbf{x}]\mathbf{A}'.\end{aligned}$$

□

Exemple C.1. Soit $\mathbf{A} = [1\ 1]$, $\mathbf{x}' = [X_1\ X_2]$ et $Y = \mathbf{Ax}$, donc $Y = X_1 + X_2$. Alors

$$\begin{aligned}E[Y] &= \mathbf{A}E[\mathbf{x}] \\ &= [1\ \ 1] \begin{bmatrix} E[X_1] \\ E[X_2] \end{bmatrix} \\ &= E[X_1] + E[X_2]\end{aligned}$$

et

$$\begin{aligned}\mathbf{V}[Y] &= \mathbf{AV}[\mathbf{x}]\mathbf{A}' \\ &= [1\ \ 1] \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\mathrm{Cov}(X_1, X_2).\end{aligned}$$

D Solutions

Chapitre 2

2.1 a) Voir la figure D.1. Remarquer que l'on peut, dans la fonction `plot`, utiliser une formule pour exprimer la relation entre les variables.

b) Les équations normales sont les équations à résoudre pour trouver les estimateurs de β_0 et β_1 minimisant la somme des carrés

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

Or,

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

d'où les équations normales sont

$$\begin{aligned} \sum_{t=1}^{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) &= 0 \\ \sum_{t=1}^{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) X_t &= 0. \end{aligned}$$

c) Par la première des deux équations normales, on trouve

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

soit, en isolant $\hat{\beta}_0$,

$$\hat{\beta}_0 = \frac{\sum_{t=1}^{n} Y_t - \hat{\beta}_1 \sum_{t=1}^{n} X_t}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

```
> x <- c(65, 43, 44, 59, 60, 50, 52, 38, 42,
+     40)
> y <- c(12, 32, 36, 18, 17, 20, 21, 40, 30,
+     24)
> plot(y ~ x, pch = 16)
```

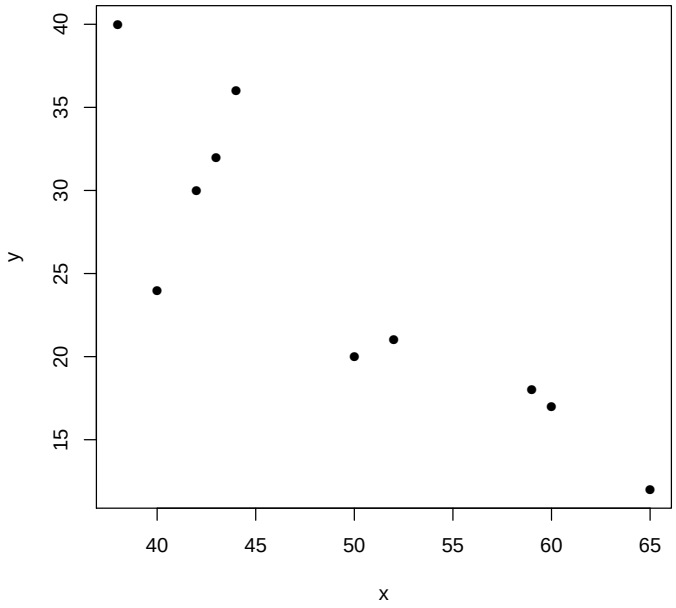


FIGURE D.1: Relation entre les données de l'exercice 2.1

De la seconde équation normale, on obtient

$$\sum_{t=1}^{n} X_t Y_t - \hat{\beta}_0 \sum_{t=1}^{n} X_t - \hat{\beta}_1 \sum_{t=1}^{n} X_t^2 = 0$$

puis, en remplaçant $\hat{\beta}_0$ par la valeur obtenue ci-dessus,

$$\hat{\beta}_1 \left(\sum_{t=1}^{n} X_t^2 - n\bar{X}^2 \right) = \sum_{t=1}^{n} X_t Y_t - n\bar{X}\bar{Y}.$$

Par conséquent,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{n} X_t Y_t - n\bar{X}\bar{Y}}{\sum_{t=1}^{n} X_t^2 - n\bar{X}^2} \\ &= \frac{11\,654 - (10)(49{,}3)(25)}{25\,103 - (10)(49{,}3)^2} \\ &= -0{,}8407\end{aligned}$$

et

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ &= 25 - (-0{,}8407)(49{,}3) \\ &= 66{,}4488.\end{aligned}$$

d) On peut calculer les prévisions correspondant à $X_1, \ldots, X_{10}$ — ou valeurs ajustées — à partir de la relation $\hat{Y}_t = 66{,}4488 - 0{,}8407 X_t$, $t = 1, 2, \ldots, 10$. Avec R, on crée un objet de type modèle de régression avec `lm` et on en extrait les valeurs ajustées avec `fitted` :

```
> fit <- lm(y ~ x)
> fitted(fit)
```

```
       1        2        3        4        5
11.80028 30.29670 29.45596 16.84476 16.00401
       6        7        8        9       10
24.41148 22.72998 34.50044 31.13745 32.81894
```

Pour ajouter la droite de régression au graphique de la figure D.1, il suffit d'utiliser la fonction `abline` avec en argument l'objet créé avec `lm`. L'ordonnée à l'origine et la pente de la droite seront extraites automatiquement. Voir la figure D.2.

e) Les résidus de la régression sont $e_t = Y_t - \hat{Y}_t$, $t = 1, \ldots, 10$. Dans R, la fonction `residuals` extrait les résidus du modèle :

```
> residuals(fit)
```

```
         1          2          3          4
 0.1997243  1.7032953  6.5440421  1.1552437
         5          6          7          8
 0.9959905 -4.4114773 -1.7299837  5.4995615
         9         10
-1.1374514 -8.8189450
```

On vérifie ensuite que la somme des résidus est (essentiellement) nulle :

```
> sum(residuals(fit))
```

```
[1] -4.440892e-16
```

```
> abline(fit)
```

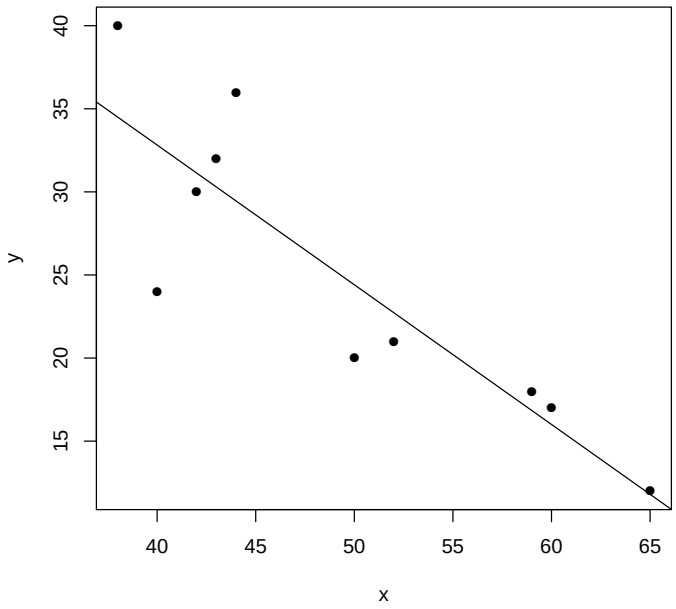


FIGURE D.2: Relation entre les données de l'exercice 2.1 et la droite de régression

2.2 a) Nous avons le modèle de régression usuel. Les coefficients de la régression sont

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{8} X_t Y_t - n\bar{X}\bar{Y}}{\sum_{t=1}^{8} X_t^2 - n\bar{X}^2} \\ &= \frac{146 - (8)(32/8)(40/8)}{156 - (8)(32/8)^2} \\ &= -0{,}5\end{aligned}$$

et

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ &= (40/8) - (-0{,}5)(32/8) \\ &= 7.\end{aligned}$$

b) Les sommes de carrés sont

$$\begin{aligned}
\text{SST} &= \sum_{t=1}^{8}(Y_t - \bar{Y})^2 \\
&= \sum_{t=1}^{8} Y_t^2 - n\bar{Y}^2 \\
&= 214 - (8)(40/8)^2 \\
&= 14, \\
\text{SSR} &= \sum_{t=1}^{8}(\hat{Y}_t - \bar{Y})^2 \\
&= \sum_{t=1}^{8} \hat{\beta}_1^2 (X_t - \bar{X})^2 \\
&= \hat{\beta}_1^2 (\sum_{t=1}^{8} X_t^2 - n\bar{X}^2) \\
&= (-1/2)^2 (156 - (8)(32/8)^2) \\
&= 7.
\end{aligned}$$

et $\text{SSE} = \text{SST} - \text{SSR} = 14 - 7 = 7$. Par conséquent, $R^2 = SSR/SST = 7/14 = 0{,}5$, donc la régression explique 50 % de la variation des Y_t par rapport à leur moyenne $\bar{Y}$. Le tableau ANOVA est le suivant :

Source	SS	d.l.	MS	Ratio F
Régression	7	1	7	6
Erreur	7	6	7/6	
Total	14	7		

2.3 a) Voir la figure D.3.

b) Le graphique montre qu'un modèle linéaire serait excellent. On estime les paramètres de ce modèle avec `lm` :

```
> (fit <- lm(weight ~ height, data = women))

Call:
lm(formula = weight ~ height, data = women)

Coefficients:
(Intercept)       height
     -87.52         3.45
```

c) Voir la figure D.4. On constate que l'ajustement est excellent.

d) Le résultat de la fonction `summary` appliquée au modèle `fit` est le suivant :

```
> summary(fit)
```

```
> data(women)
> plot(weight ~ height, data = women, pch = 16)
```

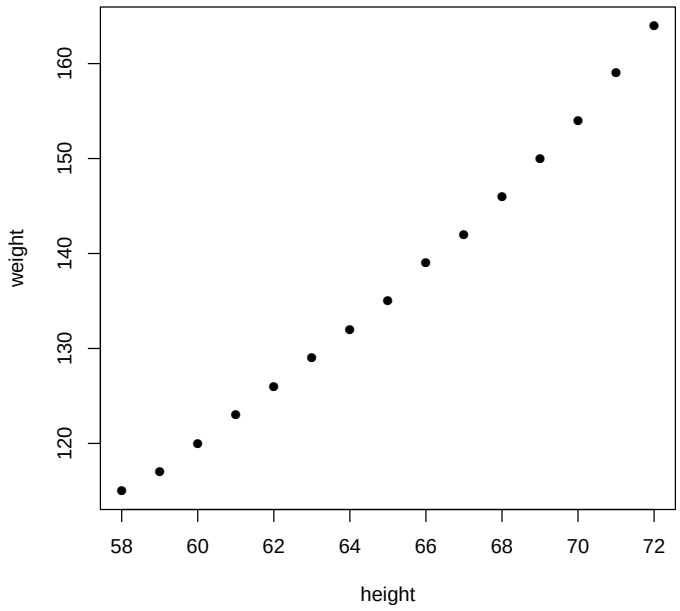


FIGURE D.3: Relation entre la taille et le poids moyen de femmes américaines âgées de 30 à 39 ans (données `women`)

```
Call:
lm(formula = weight ~ height, data = women)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09
height        3.45000    0.09114   37.85 1.09e-14

(Intercept) ***
height      ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> abline(fit)
```

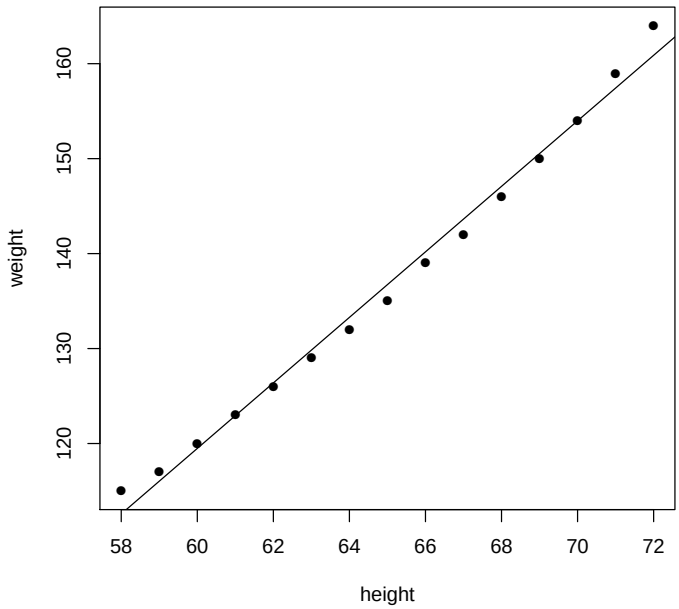


FIGURE D.4: Relation entre les données `women` et droite de régression linéaire simple

```
Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared: 0.991,        Adjusted R-squared: 0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Le coefficient de détermination est donc $R^2 = 0{,}991$, ce qui est près de 1 et confirme donc l'excellent ajustement du modèle évoqué en c).

e) On a

```
> attach(women)
> SST <- sum((weight - mean(weight))^2)
> SSR <- sum((fitted(fit) - mean(weight))^2)
> SSE <- sum((weight - fitted(fit))^2)
> all.equal(SST, SSR + SSE)

[1] TRUE

> all.equal(summary(fit)$r.squared, SSR/SST)

[1] TRUE
```

2.4 Puisque $\hat{Y}_t = (\bar{Y} - \hat{\beta}_1\bar{X}) + \hat{\beta}_1 X_t = \bar{Y} + \hat{\beta}_1(X_t - \bar{X})$ et que $e_t = Y_t - \hat{Y}_t = (Y_t - \bar{Y}) - \hat{\beta}_1(X_t - \bar{X})$, alors

$$\begin{aligned}\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})e_t &= \hat{\beta}_1\left(\sum_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y}) - \hat{\beta}_1\sum_{t=1}^{n}(X_t - \bar{X})^2\right)\\ &= \hat{\beta}_1\left(S_{XY} - \frac{S_{XY}}{S_{XX}}S_{XX}\right)\\ &= 0.\end{aligned}$$

2.5 On a un modèle de régression linéaire simple usuel avec $X_t = t$. Les estimateurs des moindres carrés des paramètres β_0 et β_1 sont donc

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\frac{\sum_{t=1}^{n} t}{n}$$

et

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} tY_t - \bar{Y}\sum_{t=1}^{n} t}{\sum_{t=1}^{n} t^2 - n^{-1}(\sum_{t=1}^{n} t)^2}.$$

Or, puisque $\sum_{t=1}^{n} t = n(n+1)/2$ et $\sum_{t=1}^{n} t^2 = n(n+1)(2n+1)/6$, les expressions ci-dessus se simplifient en

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\frac{n+1}{2}$$

et

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{n} tY_t - n(n+1)\bar{Y}/2}{n(n+1)(2n+1)/6 - n(n+1)^2/4}\\ &= \frac{12\sum_{t=1}^{n} tY_t - 6n(n+1)\bar{Y}}{n(n^2-1)}.\end{aligned}$$

2.6 a) L'estimateur des moindres carrés du paramètre β est la valeur $\hat{\beta}$ minimisant la somme de carrés

$$\begin{aligned}S(\beta) &= \sum_{t=1}^{n}\varepsilon_t^2\\ &= \sum_{t=1}^{n}(Y_t - \beta X_t)^2.\end{aligned}$$

Or,

$$\frac{d}{d\beta}S(\beta) = -2\sum_{t=1}^{n}(Y_t - \hat{\beta}X_t)X_t,$$

d'où l'unique équation normale de ce modèle est

$$\sum_{t=1}^{n} X_t Y_t - \hat{\beta} \sum_{t=1}^{n} X_t^2 = 0.$$

L'estimateur des moindres carrés de β est donc

$$\hat{\beta} = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}.$$

b) On doit démontrer que $E[\hat{\beta}] = \beta$. On a

$$\begin{aligned} E[\hat{\beta}] &= E\left[\frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}\right] \\ &= \frac{1}{\sum_{t=1}^{n} X_t^2} \sum_{t=1}^{n} X_t E[Y_t] \\ &= \frac{1}{\sum_{t=1}^{n} X_t^2} \sum_{t=1}^{n} X_t \beta X_t \\ &= \beta \frac{\sum_{t=1}^{n} X_t^2}{\sum_{t=1}^{n} X_t^2} \\ &= \beta. \end{aligned}$$

c) Des hypothèses du modèle, on a

$$\begin{aligned} \mathrm{Var}[\hat{\beta}] &= \mathrm{Var}\left[\frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}\right] \\ &= \frac{1}{(\sum_{t=1}^{n} X_t^2)^2} \sum_{t=1}^{n} X_t^2 \mathrm{Var}[Y_t] \\ &= \frac{\sigma^2}{(\sum_{t=1}^{n} X_t^2)^2} \sum_{t=1}^{n} X_t^2 \\ &= \frac{\sigma^2}{\sum_{t=1}^{n} X_t^2}. \end{aligned}$$

2.7 On veut trouver les coefficients $c_1, \ldots, c_n$ tels que $E[\beta^*] = \beta$ et $\mathrm{Var}[\beta^*]$ est minimale. On cherche donc à minimiser la fonction

$$\begin{aligned} f(c_1, \ldots, c_n) &= \mathrm{Var}[\beta^*] \\ &= \sum_{t=1}^{n} c_t^2 \mathrm{Var}[Y_t] \\ &= \sigma^2 \sum_{t=1}^{n} c_t^2 \end{aligned}$$

sous la contrainte $E[\beta^*] = \sum_{t=1}^n c_t E[Y_t] = \sum_{t=1}^n c_t \beta X_t = \beta \sum_{t=1}^n c_t X_t = \beta$, soit $\sum_{t=1}^n c_t X_t = 1$ ou $g(c_1, \ldots, c_n) = 0$ avec

$$g(c_1, \ldots, c_n) = \sum_{t=1}^n c_t X_t - 1.$$

Pour utiliser la méthode des multiplicateurs de Lagrange, on pose

$$\begin{aligned}\mathcal{L}(c_1, \ldots, c_n, \lambda) &= f(c_1, \ldots, c_n) - \lambda g(c_1, \ldots, c_n), \\ &= \sigma^2 \sum_{t=1}^n c_t^2 - \lambda \left(\sum_{t=1}^n c_t X_t - 1 \right),\end{aligned}$$

puis on dérive la fonction $\mathcal{L}$ par rapport à chacune des variables $c_1, \ldots, c_n$ et λ. On trouve alors

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial c_u} &= 2\sigma^2 c_u - \lambda X_u, \quad u = 1, \ldots, n \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= -\sum_{t=1}^n c_t X_t + 1.\end{aligned}$$

En posant les n premières dérivées égales à zéro, on obtient

$$c_t = \frac{\lambda X_t}{2\sigma^2}.$$

Or, de la contrainte,

$$\sum_{t=1}^n c_t X_t = \frac{\lambda}{2\sigma^2} \sum_{t=1}^n X_t^2 = 1,$$

d'où

$$\frac{\lambda}{2\sigma^2} = \frac{1}{\sum_{t=1}^n X_t^2}$$

et, donc,

$$c_t = \frac{X_t}{\sum_{t=1}^n X_t^2}.$$

Finalement,

$$\begin{aligned}\beta^* &= \sum_{t=1}^n c_t Y_t \\ &= \frac{\sum_{t=1}^n X_t Y_t}{\sum_{t=1}^n X_t^2} \\ &= \hat{\beta}.\end{aligned}$$

2.8 a) Tout d'abord, puisque $\text{MSE} = \text{SSE}/(n-2) = \sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2/(n-2)$ et que $E[Y_t] = E[\hat{Y}_t]$, alors

$$\begin{aligned}
E[MSE] &= \frac{1}{n-2} E\left[\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2\right] \\
&= \frac{1}{n-2}\sum_{t=1}^{n} E[(Y_t - \hat{Y}_t)^2] \\
&= \frac{1}{n-2}\sum_{t=1}^{n} E[((Y_t - E[Y_t]) - (\hat{Y}_t - E[\hat{Y}_t]))^2] \\
&= \frac{1}{n-2}\sum_{t=1}^{n} \left(\text{Var}[Y_t] + \text{Var}[\hat{Y}_t] - 2\text{Cov}(Y_t, \hat{Y}_t)\right).
\end{aligned}$$

Or, on a par hypothèse du modèle que $\text{Cov}(Y_t, Y_s) = \text{Cov}(\varepsilon_t, \varepsilon_s) = \delta_{ts}\sigma^2$, d'où $\text{Var}[Y_t] = \sigma^2$ et $\text{Var}[\bar{Y}] = \sigma^2/n$. D'autre part,

$$\begin{aligned}
\text{Var}[\hat{Y}_t] &= \text{Var}[\bar{Y} + \hat{\beta}_1(X_t - \bar{X})] \\
&= \text{Var}[\bar{Y}] + (X_t - \bar{X})^2\text{Var}[\hat{\beta}_1] + 2(X_t - \bar{X})\text{Cov}(\bar{Y}, \hat{\beta}_1)
\end{aligned}$$

et l'on sait que

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

et que

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

puisque $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Ainsi,

$$\text{Var}[\hat{Y}_t] = \frac{\sigma^2}{n} + \frac{(X_t - \bar{X})^2\sigma^2}{\sum_{t=1}^{n}(X_t - \bar{X})^2}.$$

De manière similaire, on détermine que

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

Par conséquent,

$$E[(Y_t - \hat{Y}_t)^2] = \frac{n-1}{n}\sigma^2 - \frac{(X_t - \bar{X})^2\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

et

$$\sum_{t=1}^n E[(Y_t - \hat{Y}_t)^2] = (n-2)\sigma^2,$$

d'où $E[\mathrm{MSE}] = \sigma^2$.

b) On a

$$\begin{aligned}E[\mathrm{MSR}] &= E[\mathrm{SSR}] \\ &= E\left[\sum_{t=1}^n (\hat{Y}_t - \bar{Y})^2\right] \\ &= \sum_{t=1}^n E[\hat{\beta}_1^2 (X_t - \bar{X})^2] \\ &= \sum_{t=1}^n (X_t - \bar{X})^2 E[\hat{\beta}_1^2] \\ &= \sum_{t=1}^n (X_t - \bar{X})^2 (\mathrm{Var}[\hat{\beta}_1] + E[\hat{\beta}_1]^2) \\ &= \sum_{t=1}^n (X_t - \bar{X})^2 \left(\frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X})^2} + \beta_1^2\right) \\ &= \sigma^2 + \beta_1^2 \sum_{t=1}^n (X_t - \bar{X})^2.\end{aligned}$$

2.9 a) Il faut exprimer $\hat{\beta}_0'$ et $\hat{\beta}_1'$ en fonction de $\hat{\beta}_0$ et $\hat{\beta}_1$. Pour ce faire, on trouve d'abord une expression pour chacun des éléments qui entrent dans la définition de $\hat{\beta}_1'$. Tout d'abord,

$$\begin{aligned}\bar{X}' &= \frac{1}{n}\sum_{t=1}^n X_t' \\ &= \frac{1}{n}\sum_{t=1}^n (c + dX_t) \\ &= c + d\bar{X},\end{aligned}$$

et, de manière similaire, $\bar{Y}' = a + b\bar{Y}$. Ensuite,

$$\begin{aligned} S'_{XX} &= \sum_{t=1}^{n} (X'_t - \bar{X}')^2 \\ &= \sum_{t=1}^{n} (c + dX_t - c + d\bar{X})^2 \\ &= d^2 S_{XX} \end{aligned}$$

et $S'_{YY} = b^2 S_{YY}$, $S'_{XY} = bdS_{XY}$. Par conséquent,

$$\begin{aligned} \hat{\beta}'_1 &= \frac{S'_{XY}}{S'_{XX}} \\ &= \frac{bdS_{XY}}{d^2 S_{XX}} \\ &= \frac{b}{d}\hat{\beta}_1 \end{aligned}$$

et

$$\begin{aligned} \hat{\beta}'_0 &= \bar{Y}' - \hat{\beta}'_1 \bar{X}' \\ &= a + b\bar{Y} - \frac{b}{d}\hat{\beta}_1 (c + d\bar{X}) \\ &= a - \frac{bc}{d}\hat{\beta}_1 + b(\bar{Y} - \hat{\beta}_1 \bar{X}) \\ &= a - \frac{bc}{d}\hat{\beta}_1 + b\hat{\beta}_0. \end{aligned}$$

b) Tout d'abord, on établit que

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{\sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \\ &= \hat{\beta}_1^2 \frac{\sum_{t=1}^{n} (X_t - \bar{X})^2}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \\ &= \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}}. \end{aligned}$$

Maintenant, avec les résultats obtenus en a), on démontre directement

que

$$\begin{aligned}(R^2)' &= (\hat{\beta}_1')^2 \frac{S'_{XX}}{S'_{YY}} \\ &= \left(\frac{b}{d}\right)^2 \hat{\beta}_1^2 \frac{d^2 S_{XX}}{b^2 S_{YY}} \\ &= \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} \\ &= R^2.\end{aligned}$$

2.10 Considérons un modèle de régression usuel avec l'ensemble de données $(X_1, Y_1), \ldots, (X_n, Y_n), (m\bar{X}, m\bar{Y})$, où $\bar{X} = n^{-1}\sum_{t=1}^{n} X_t$, $\bar{Y} = n^{-1}\sum_{t=1}^{n} Y_t$, $m = n/a$ et $a = \sqrt{n+1} - 1$. On définit

$$\begin{aligned}\bar{X}' &= \frac{1}{n+1}\sum_{t=1}^{n+1} X_t \\ &= \frac{1}{n+1}\sum_{t=1}^{n} X_t + \frac{m}{n+1}\bar{X} \\ &= k\bar{X}\end{aligned}$$

et, de manière similaire,

$$\bar{Y}' = k\bar{Y},$$

où

$$\begin{aligned}k &= \frac{n+m}{n+1} \\ &= \frac{n(a+1)}{a(n+1)}.\end{aligned}$$

L'expression pour l'estimateur des moindres carrés de la pente de la droite de régression pour cet ensemble de données est

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{n+1} X_t Y_t - (n+1)\bar{X}'\bar{Y}'}{\sum_{t=1}^{n+1} X_t^2 - (n+1)(\bar{X}')^2} \\ &= \frac{\sum_{t=1}^{n} X_t Y_t + m^2\bar{X}\bar{Y} - (n+1)k^2\bar{X}\bar{Y}}{\sum_{t=1}^{n} X_t^2 + m^2\bar{X}^2 - (n+1)k^2\bar{X}^2}.\end{aligned}$$

Or,

$$\begin{aligned}m^2 - k^2(n+1) &= \frac{n^2}{a^2} - \frac{n^2(a+1)^2}{a^2(n+1)} \\ &= \frac{n^2(n+1) - n^2(n+1)}{a^2(n+1)} \\ &= 0.\end{aligned}$$

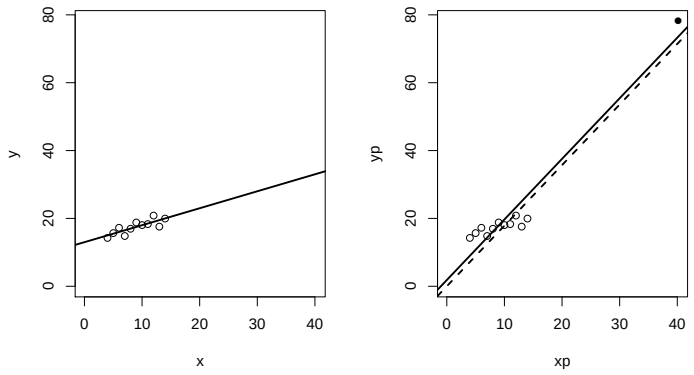


FIGURE D.5: Illustration de l'effet de l'ajout d'un point spécial à un ensemble de données. À gauche, la droite de régression usuelle. À droite, le même ensemble de points avec le point spécial ajouté (cercle plein), la droite de régression avec ce nouveau point (ligne pleine) et la droite de régression passant par l'origine (ligne pointillée). Les deux droites sont parallèles.

Par conséquent,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^n X_t Y_t}{\sum_{t=1}^n X_t^2} \\ &= \hat{\beta}.\end{aligned}$$

Interprétation : en ajoutant un point bien spécifique à n'importe quel ensemble de données, on peut s'assurer que la pente de la droite de régression sera la même que celle d'un modèle passant par l'origine. Voir la figure D.5 pour une illustration du phénomène.

2.11 Puisque, selon le modèle, $\varepsilon_t \sim N(0, \sigma^2)$ et que $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, alors $Y_t \sim N(\beta_0 + \beta_1 X_t, \sigma^2)$. De plus, on sait que

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

donc l'estimateur $\hat{\beta}_1$ est une combinaison linéaire des variables aléatoires $Y_1, \ldots, Y_n$. Par conséquent, $\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \text{Var}[\hat{\beta}_1])$, où $E[\hat{\beta}_1] = \beta_1$ et $\text{Var}[\hat{\beta}_1] = \sigma^2 / S_{XX}$ et, donc,

$$\Pr\left[-z_{\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{XX}}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Un intervalle de confiance de niveau $1-\alpha$ pour le paramètre β_1 lorsque la variance σ^2 est connue est donc

$$\beta_1 \in \hat{\beta}_1 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{\sum_{t=1}^{n}(X_t-\bar{X})^2}}.$$

2.12 L'intervalle de confiance pour β_1 est

$$\begin{aligned}\beta_1 &\in \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}\\ &\in \hat{\beta}_1 \pm t_{0,025}(20-2)\sqrt{\frac{MSE}{S_{XX}}}.\end{aligned}$$

On nous donne $\text{SST} = S_{YY} = 20\,838$ et $S_{XX} = 10\,668$. Par conséquent,

$$\begin{aligned}\text{SSR} &= \hat{\beta}_1^2\sum_{t=1}^{20}(X_t-\bar{X})^2\\ &= (-1{,}104)^2(10\,668)\\ &= 13\,002{,}33\\ \text{SSE} &= \text{SST}-\text{SSR}\\ &= 7\,835{,}67\end{aligned}$$

et

$$\begin{aligned}\text{MSE} &= \frac{\text{SSE}}{18}\\ &= 435{,}315.\end{aligned}$$

De plus, on trouve dans une table de quantiles de la loi de Student (ou à l'aide de la fonction `qt` dans R) que $t_{0,025}(18) = 2{,}101$. L'intervalle de confiance recherché est donc

$$\begin{aligned}\beta_1 &\in -1{,}104 \pm 2{,}101\sqrt{\frac{435{,}315}{10\,668}}\\ &\in (-1{,}528, -0{,}680).\end{aligned}$$

2.13 a) On trouve aisément les estimateurs de la pente et de l'ordonnée à l'origine de la droite de régression :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{n}X_tY_t - n\bar{X}\bar{Y}}{\sum_{t=1}^{n}X_t^2 - n\bar{X}^2}\\ &= 1{,}436\\ \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X}\\ &= 9{,}273.\end{aligned}$$

b) Les sommes de carrés sont

$$\begin{aligned}
\text{SST} &= \sum_{t=1}^{n} Y_t^2 - n\bar{Y}^2 \\
&= 1194 - 11(9{,}273)^2 \\
&= 248{,}18 \\
\text{SSR} &= \hat{\beta}_1^2 \left(\sum_{t=1}^{n} X_t^2 - n\bar{X}^2 \right) \\
&= (1{,}436)^2(110 - 11(0)) \\
&= 226{,}95
\end{aligned}$$

et $\text{SSE} = \text{SST} - \text{SSR} = 21{,}23$. Le tableau d'analyse de variance est donc le suivant :

Source	SS	d.l.	MS	Ratio F
Régression	226,95	1	226,95	96,21
Erreur	21,23	9	2,36	
Total	248,18	10		

Or, puisque $t = \sqrt{F} = 9{,}809 > t_{\alpha/2}(n-2) = t_{0{,}025}(9) = 2{,}26$, on rejette l'hypothèse $H_0 : \beta_1 = 0$ soit, autrement dit, la pente est significativement différente de zéro.

c) Puisque la variance σ^2 est inconnue, on l'estime par $s^2 = \text{MSE} = 2{,}36$. On a alors

$$\begin{aligned}
\beta_1 &\in \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\sqrt{\widehat{\text{Var}}[\hat{\beta}_1]} \\
&\in 1{,}436 \pm 2{,}26\sqrt{\frac{2{,}36}{110}} \\
&\in (1{,}105, 1{,}768).
\end{aligned}$$

d) Le coefficient de détermination de la régression est $R^2 = \text{SSR}/\text{SST} = 226{,}95/248{,}18 = 0{,}914$, ce qui indique que l'ajustement du modèle aux données est très bon. En outre, suite au test effectué à la partie b), on conclut que la régression est globalement significative. Toutes ces informations portent à conclure qu'il n'y a pas lieu d'utiliser un autre modèle.

2.14 On doit déterminer si la régression est significative, ce qui peut se faire à l'aide de la statistique F. Or, à partir de l'information donnée dans

l'énoncé, on peut calculer

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{t=1}^{50} X_t Y_t - 50\bar{X}\bar{Y}}{\sum_{t=1}^{50} X_t^2 - 50\bar{X})^2} \\
&= -0{,}0110 \\
\text{SST} &= \sum_{t=1}^{50} Y_t^2 - 50\bar{Y}^2 \\
&= 78{,}4098 \\
\text{SSR} &= \hat{\beta}_1^2 \sum_{t=1}^{50} (X_t - \bar{X})^2 \\
&= 1{,}1804 \\
\text{SSE} &= \text{SST} - \text{SSR} \\
&= 77{,}2294
\end{aligned}$$

d'où

$$\begin{aligned}
\text{MSR} &= 1{,}1804 \\
\text{MSE} &= \frac{\text{SSE}}{50 - 2} \\
&= 1{,}6089
\end{aligned}$$

et, enfin,

$$\begin{aligned}
F &= \frac{\text{MSR}}{\text{MSE}} \\
&= 0{,}7337.
\end{aligned}$$

Soit F une variable aléatoire ayant une distribution de Fisher avec 1 et 48 degrés de liberté, soit la même distribution que la statistique F sous l'hypothèse $H_0 : \beta_1 = 0$. On a que $\Pr[F > 0{,}7337] = 0{,}3959$, donc la valeur p du test $H_0 : \beta_1 = 0$ est 0,3959. Une telle valeur p est généralement considérée trop élevée pour rejeter l'hypothèse H_0. On ne peut donc considérer la relation entre la ligne de vie et l'espérance de vie comme significative. (Ou on ne la considère significative qu'avec un niveau de confiance de $1 - p = 60{,}41$ %.)

2.15 Premièrement, selon le modèle de régression passant par l'origine, $Y_0 = \beta X_0 + \varepsilon_0$ et $\hat{Y}_0 = \hat{\beta} X_0$. Considérons, pour la suite, la variable aléatoire $Y_0 - \hat{Y}_0$. On voit facilement que $E[\hat{\beta}] = \beta$, d'où $E[Y_0 - \hat{Y}_0] = E[\beta X_0 + \varepsilon_0 - \hat{\beta} X_0] = \beta X_0 - \beta X_0 = 0$ et

$$\text{Var}[Y_0 - \hat{Y}_0] = \text{Var}[Y_0] + \text{Var}[\hat{Y}_0] - 2\text{Cov}(Y_0, \hat{Y}_0).$$

Or, $\text{Cov}(Y_0, \hat{Y}_0) = 0$ par l'hypothèse ii) de l'énoncé, $\text{Var}[Y_0] = \sigma^2$ et $\text{Var}[\hat{Y}_0] = X_0^2 \text{Var}[\hat{\beta}]$. De plus,

$$\begin{aligned}\text{Var}[\hat{\beta}] &= \frac{1}{(\sum_{t=1}^n X_t^2)^2} \sum_{t=1}^n X_t^2 \text{Var}[Y_t] \\ &= \frac{\sigma^2}{\sum_{t=1}^n X_t^2}\end{aligned}$$

d'où, finalement,

$$\text{Var}[Y_0 - \hat{Y}_0] = \sigma^2 \left(1 + \frac{X_0^2}{\sum_{t=1}^n X_t^2}\right).$$

Par l'hypothèse de normalité et puisque $\hat{\beta}$ est une combinaison linéaire de variables aléatoires normales,

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2 \left(1 + \frac{X_0^2}{\sum_{t=1}^n X_t^2}\right)\right)$$

ou, de manière équivalente,

$$\frac{Y_0 - \hat{Y}_0}{\sigma\sqrt{1 + X_0^2 / \sum_{t=1}^n X_t^2}} \sim N(0,1).$$

Lorsque la variance σ^2 est estimée par s^2, alors

$$\frac{Y_0 - \hat{Y}_0}{s\sqrt{1 + X_0^2 / \sum_{t=1}^n X_t^2}} \sim t(n-1).$$

La loi de Student a $n-1$ degrés de liberté puisque le modèle passant par l'origine ne compte qu'un seul paramètre. Les bornes de l'intervalle de confiance pour la vraie valeur de Y_0 sont donc

$$\hat{Y}_0 \pm t_{\alpha/2}(n-1)\, s \sqrt{1 + \frac{X_0^2}{\sum_{t=1}^n X_t^2}}.$$

2.16 a) Soit $X_1, \ldots, X_{10}$ les valeurs de la masse monétaire et $Y_1, \ldots, Y_{10}$ celles du PNB. On a $\bar{X} = 3{,}72$, $\bar{Y} = 7{,}55$, $\sum_{t=1}^{10} X_t^2 = 147{,}18$, $\sum_{t=1}^{10} Y_t^2 = 597{,}03$ et $\sum_{t=1}^{10} X_t Y_t = 295{,}95$. Par conséquent,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{t=1}^{10} X_t Y_t - 10\bar{X}\bar{Y}}{\sum_{t=1}^{10} X_t^2 - 10\bar{X}^2} \\ &= 1{,}716\end{aligned}$$

et

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ &= 1{,}168.\end{aligned}$$

On a donc la relation linéaire PNB $= 1{,}168 + 1{,}716$ MM.

b) Tout d'abord, on doit calculer l'estimateur s^2 de la variance car cette quantité entre dans le calcul des intervalles de confiance demandés. Pour les calculs à la main, on peut éviter de calculer les valeurs de $\hat{Y}_1, \ldots, \hat{Y}_{10}$ en procédant ainsi :

$$\begin{aligned}\text{SST} &= \sum_{t=1}^{10} Y_t^2 - 10\bar{Y}^2 \\ &= 27{,}005 \\ \text{SSR} &= \hat{\beta}_1^2 \left(\sum_{t=1}^{10} X_t^2 - 10\bar{X}^2 \right) \\ &= 25{,}901,\end{aligned}$$

puis SSE $=$ SST $-$ SSR $= 1{,}104$ et $s^2 = \text{MSE} = \text{SSE}/(10-2) = 0{,}1380$. On peut maintenant construire les intervalles de confiance :

$$\begin{aligned}\beta_0 &\in \hat{\beta}_0 \pm t_{\alpha/2}(n-2)\, s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}} \\ &\in 1{,}168 \pm (2{,}306)(0{,}3715)\sqrt{\frac{1}{10} + \frac{3{,}72^2}{8{,}796}} \\ &\in (0{,}060, 2{,}276) \\ \beta_1 &\in \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\, s \sqrt{\frac{1}{S_{XX}}} \\ &\in 1{,}716 \pm (2{,}306)(0{,}3715)\sqrt{\frac{1}{8{,}796}} \\ &\in (1{,}427, 2{,}005).\end{aligned}$$

Puisque l'intervalle de confiance pour la pente β_1 ne contient ni la valeur 0, ni la valeur 1, on peut rejeter, avec un niveau de confiance de 95 %, les hypothèses $H_0 : \beta_1 = 0$ et $H_0 : \beta_1 = 1$.

c) Par l'équation obtenue en a) liant le PNB à la masse monétaire (MM), un PNB de 12,0 correspond à une masse monétaire de

$$\begin{aligned}\text{MM} &= \frac{12{,}0 - 1{,}168}{1{,}716} \\ &= 6{,}31.\end{aligned}$$

d) On cherche un intervalle de confiance pour la droite de régression en $\text{MM}_{1997} = 6{,}31$ ainsi qu'un intervalle de confiance pour la prévision $\text{PNB} = 12{,}0$ associée à cette même valeur de la masse monétaire. Avec une probabilité de $\alpha = 95\ \%$, le PNB moyen se trouve dans l'intervalle

$$12{,}0 \pm t_{\alpha/2}(n-2)\, s \sqrt{\frac{1}{n} + \frac{(6{,}31 - \bar{X})^2}{S_{XX}}} = (11{,}20, 12{,}80),$$

alors que la vraie valeur du PNB se trouve dans l'intervalle

$$12{,}0 \pm t_{\alpha/2}(n-2)\, s \sqrt{1 + \frac{1}{n} + \frac{(6{,}31 - \bar{X})^2}{S_{XX}}} = (10{,}83, 13{,}17).$$

2.17 a) Les données du fichier `house.dat` sont importées dans R avec la commande

```
> house <- read.table("house.dat", header = TRUE)
```

La figure D.6 contient les graphiques de `medv` en fonction de chacune des variables `rm`, `age`, `lstat` et `tax`. Le meilleur choix de variable explicative pour le prix médian semble être le nombre moyen de pièces par immeuble, `rm`.

b) Les résultats ci-dessous ont été obtenus avec R.

```
> fit1 <- lm(medv ~ rm, data = house)
> summary(fit1)

Call:
lm(formula = medv ~ rm, data = house)

Residuals:
      Min        1Q    Median        3Q       Max
-23.34590  -2.54748   0.08976   2.98553  39.43314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16
rm             9.102      0.419   21.72   <2e-16

(Intercept) ***
rm          ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared: 0.4835,        Adjusted R-squared: 0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

On peut voir que tant l'ordonnée à l'origine que la pente sont très significativement différentes de zéro. La régression est donc elle-même

```
> par(mfrow = c(2, 2))
> plot(medv ~ rm + age + lstat + tax, data = house,
+     ask = FALSE)
```

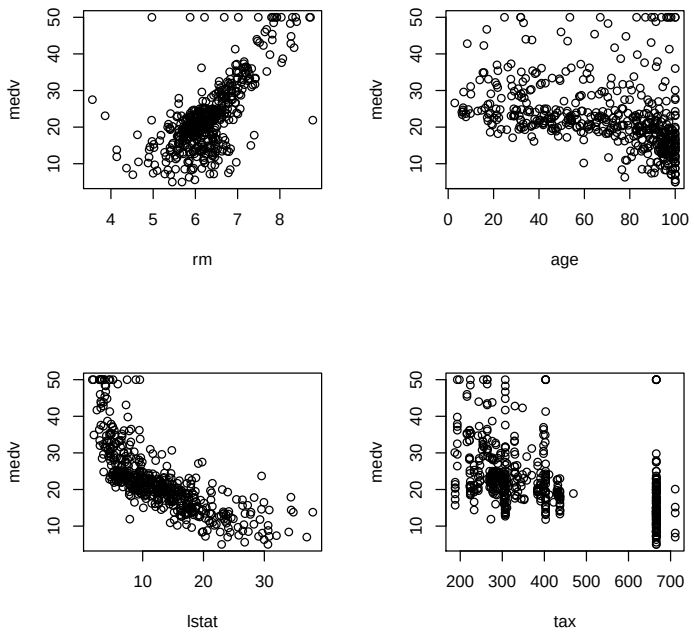


FIGURE D.6: Relation entre la variable `medv` et les variables `rm`, `age`, `lstat` et `tax` des données `house.dat`

significative. Cependant, le coefficient de détermination n'est que de $R^2 = 0{,}4835$, ce qui indique que d'autres facteurs pourraient expliquer la variation dans `medv`.

On calcule les bornes de l'intervalle de confiance de la régression avec la fonction `predict` :

```
> pred.ci <- predict(fit1, interval = "confidence",
+     level = 0.95)
```

La droite de régression et ses bornes d'intervalle de confiance inférieure et supérieure sont illustrée à la figure D.7.

c) On reprend la même démarche, mais cette fois avec la variable `age` :

```
> fit2 <- lm(medv ~ age, data = house)
> summary(fit2)

Call:
lm(formula = medv ~ age, data = house)
```

```
> ord <- order(house$rm)
> plot(medv ~ rm, data = house, ylim = range(pred.ci))
> matplot(house$rm[ord], pred.ci[ord, ], type = "l",
+     lty = c(1, 2, 2), lwd = 2, col = "black",
+     add = TRUE)
```

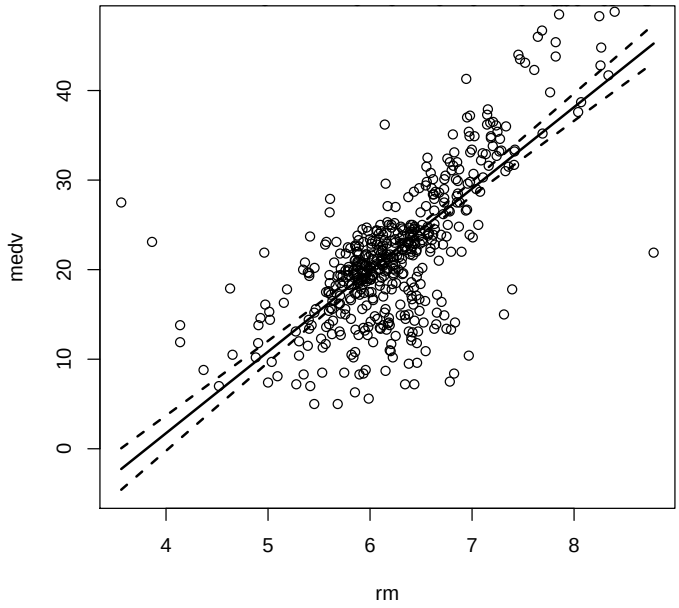


FIGURE D.7: Résultat de la régression de la variable `rm` sur la variable `medv` des données `house.dat`

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.097  -5.138  -1.957   2.398  31.338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.97868    0.99911  31.006   <2e-16
age         -0.12316    0.01348  -9.137   <2e-16

(Intercept) ***
age         ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.527 on 504 degrees of freedom
```

```
> ord <- order(house$age)
> plot(medv ~ age, data = house, ylim = range(pred.ci))
> matplot(house$age[ord], pred.ci[ord, ],
+     type = "l", lty = c(1, 2, 2), lwd = 2,
+     col = "black", add = TRUE)
```

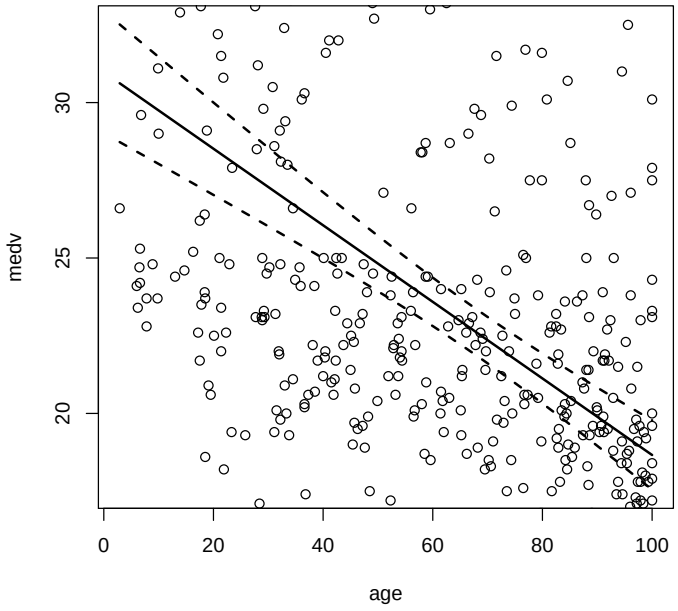


FIGURE D.8: Résultat de la régression de la variable `age` sur la variable `medv` des données `house.dat`

```
Multiple R-squared: 0.1421,        Adjusted R-squared: 0.1404
F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
> pred.ci <- predict(fit2, interval = "confidence",
+     level = 0.95)
```

La régression est encore une fois très significative. Cependant, le R^2 est encore plus faible qu'avec la variable `rm`. Les variables `rm` et `age` contribuent donc chacune à expliquer les variations de la variable `medv` (et `rm` mieux que `age`), mais aucune ne sait le faire seule de manière satisfaisante. La droite de régression et l'intervalle de confiance de celle-ci sont reproduits à la figure D.8. On constate que l'intervalle de confiance est plus large qu'en b).

2.18 a) On importe les données dans R, puis on effectue les conversions demandées. La variable `consommation` contient la consommation des

voitures en $\ell/100$ km et la variable `poids` le poids en kilogrammes.

```
> carburant <- read.table("carburant.dat",
+     header = TRUE)
> consommation <- 235.1954/carburant$mpg
> poids <- carburant$poids * 0.45455 * 1000
```

b) La fonction `summary` fournit l'information essentielle pour juger de la validité et de la qualité du modèle :

```
> fit <- lm(consommation ~ poids)
> summary(fit)

Call:
lm(formula = consommation ~ poids)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07123 -0.68380  0.01488  0.44802  2.66234

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0146530  0.7118445  -0.021    0.984
poids        0.0078382  0.0005315  14.748   <2e-16

(Intercept)
poids       ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.039 on 36 degrees of freedom
Multiple R-squared: 0.858,        Adjusted R-squared: 0.854
F-statistic: 217.5 on 1 and 36 DF,  p-value: < 2.2e-16
```

Le modèle est donc le suivant : $Y_t = -0{,}01465 + 0{,}007838 X_t + \varepsilon_t$, $\varepsilon_t \sim N(0, 1{,}039^2)$, où Y_t est la consommation en litres aux 100 kilomètres et X_t le poids en kilogrammes. La faible valeur p du test F indique une régression très significative. De plus, le R^2 de 0,858 confirme que l'ajustement du modèle est assez bon.

c) On veut calculer un intervalle de confiance pour la consommation en carburant prévue d'une voiture de 1 350 kg. On obtient, avec la fonction `predict` :

```
> predict(fit, newdata = data.frame(poids = 1350),
+     interval = "prediction")

       fit      lwr      upr
1 10.56690 8.432089 12.70170
```

Chapitre 3

3.1 Tout d'abord, selon le théorème C.5 de l'annexe C,

$$\frac{d}{d\mathbf{x}} f(\mathbf{x})'\mathbf{A}f(\mathbf{x}) = 2\left(\frac{d}{d\mathbf{x}} f(\mathbf{x})\right)'\mathbf{A}f(\mathbf{x}).$$

Il suffit, pour faire la démonstration, d'appliquer directement ce résultat à la forme quadratique

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

avec $f(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ et $\mathbf{A} = \mathbf{I}$, la matrice identité. On a alors

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2\left(\frac{d}{d\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)'\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ &= 2(-\mathbf{X})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

En posant ces dérivées exprimées sous forme matricielle simultanément égales à zéro, on obtient les équations normales à résoudre pour calculer l'estimateur des moindres carrés du vecteur $\boldsymbol{\beta}$, soit

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

En isolant $\hat{\boldsymbol{\beta}}$ dans l'équation ci-dessus, on obtient, finalement, l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

3.2 a) On a un modèle sans variable explicative. Intuitivement, la meilleure prévision de Y_t sera alors $\bar{Y}$. En effet, pour ce modèle,

$$\mathbf{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}$$

et

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= n^{-1}\sum_{t=1}^{n} Y_t \\ &= \bar{Y}.\end{aligned}$$

b) Il s'agit du modèle de régression linéaire simple passant par l'origine, pour lequel la matrice de schéma est

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

Par conséquent,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \left(\sum_{t=1}^{n} X_t^2 \right)^{-1} \sum_{t=1}^{n} X_t Y_t \\ &= \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2},\end{aligned}$$

tel qu'obtenu à l'exercice 2.6.

c) On est ici en présence d'un modèle de régression multiple ne passant pas par l'origine et ayant deux variables explicatives. La matrice de schéma est alors

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}_{n\times 3}.$$

Par conséquent,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ X_{12} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ X_{12} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \begin{bmatrix} n & n\bar{X}_1 & n\bar{X}_2 \\ n\bar{X}_1 & \sum_{t=1}^{n} X_{t1}^2 & \sum_{t=1}^{n} X_{t1}X_{t2} \\ n\bar{X}_2 & \sum_{t=1}^{n} X_{t1}X_{t2} & \sum_{t=1}^{n} X_{t2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{n} Y_t \\ \sum_{t=1}^{n} X_{t1}Y_t \\ \sum_{t=1}^{n} X_{t2}Y_t \end{bmatrix}.\end{aligned}$$

L'inversion de la première matrice et le produit par la seconde sont laissés aux bons soins du lecteur plus patient que les rédacteurs de ces solutions.

3.3 Dans le modèle de régression linéaire simple, la matrice schéma est

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

Par conséquent,

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}] &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 \left(\begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \right)^{-1} \\
&= \sigma^2 \begin{bmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{t=1}^n X_t^2 \end{bmatrix}^{-1} \\
&= \frac{\sigma^2}{n\sum_{t=1}^n X_t^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{t=1}^n X_t^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix} \\
&= \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X}^2)} \begin{bmatrix} n^{-1}\sum_{t=1}^n X_t^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix},
\end{aligned}$$

d'où

$$\begin{aligned}
\text{Var}[\hat{\beta}_0] &= \sigma^2 \frac{\sum_{t=1}^n X_t^2}{n\sum_{t=1}^n (X_t - \bar{X}^2)} \\
&= \sigma^2 \frac{\sum_{t=1}^n (X_t - \bar{X}^2) + n\bar{X}^2}{n\sum_{t=1}^n (X_t - \bar{X}^2)}
\end{aligned}$$

et

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

Ceci correspond aux résultats antérieurs.

3.4 Dans les démonstrations qui suivent, trois relations de base seront utilisées : $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

a) On a

$$\begin{aligned}
\mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) \\
&= \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} \\
&= \mathbf{0}.
\end{aligned}$$

En régression linéaire simple, cela donne

$$\begin{aligned}\mathbf{X}'\mathbf{e} &= \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_{t=1}^{n} e_t \\ \sum_{t=1}^{n} X_t e_t \end{bmatrix}.\end{aligned}$$

Par conséquent, $\mathbf{X}'\mathbf{e} = \mathbf{0}$ se simplifie en $\sum_{t=1}^{n} e_t = 0$ et $\sum_{t=1}^{n} X_t e_t = 0$ soit, respectivement, la condition pour que l'estimateur des moindres carrés soit sans biais et la seconde équation normale obtenue à la partie b) de l'exercice 2.1.

b) On a

$$\begin{aligned}\hat{\mathbf{y}}'\mathbf{e} &= (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= 0.\end{aligned}$$

Pour tout modèle de régression cette équation peut aussi s'écrire sous la forme plus conventionnelle $\sum_{t=1}^{n} \hat{Y}_t e_t = 0$. Cela signifie que le produit scalaire entre le vecteur des prévisions et celui des erreurs doit être nul ou, autrement dit, que les vecteurs doivent être orthogonaux. C'est là une condition essentielle pour que l'erreur quadratique moyenne entre les vecteurs $\mathbf{y}$ et $\hat{\mathbf{y}}$ soit minimale. (Pour de plus amples détails sur l'interprétation géométrique du modèle de régression, consulter Draper et Smith (1998, chapitres 20 et 21).) D'ailleurs, on constate que $\hat{\mathbf{y}}'\mathbf{e} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e}$ et donc, en supposant sans perte de généralité que $\hat{\boldsymbol{\beta}} \neq \mathbf{0}$, que $\hat{\mathbf{y}}'\mathbf{e} = 0$ et $\mathbf{X}'\mathbf{e} = \mathbf{0}$ sont des conditions en tous points équivalentes.

c) On a

$$\begin{aligned}\hat{\mathbf{y}}'\hat{\mathbf{y}} &= (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.\end{aligned}$$

Cette équation est l'équivalent matriciel de l'identité

$$\begin{aligned}\text{SSR} &= \hat{\beta}_1^2 \sum_{t=1}^{n} (X_t - \bar{X})^2 \\ &= \frac{S_{XY}^2}{S_{XX}}\end{aligned}$$

utilisée à plusieurs reprises dans les solutions du chapitre 2. En effet, en régression linéaire simple, $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \sum_{t=1}^{n} \hat{Y}_t^2 = \sum_{t=1}^{n} (\hat{Y} - \bar{Y})^2 + n\bar{Y}^2 = \text{SSR} + n\bar{Y}^2$ et

$$\begin{aligned}\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} &= \hat{\beta}_0 n\bar{Y} + \hat{\beta}_1 \sum_{t=1}^{n} X_t Y_t \\ &= (\bar{Y} - \hat{\beta}_1 \bar{X}) n\bar{Y} + \hat{\beta}_1 \sum_{t=1}^{n} X_t Y_t \\ &= \hat{\beta}_1 \sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) + n\bar{Y}^2 \\ &= \frac{S_{XY}^2}{S_{XX}} + n\bar{Y}^2,\end{aligned}$$

d'où $\text{SSR} = S_{XY}^2 / S_{XX}$.

3.5 a) Premièrement, $Y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$ avec $E[\varepsilon_0] = 0$. Par conséquent, $E[Y_0] = E[x_0\boldsymbol{\beta} + \varepsilon_0] = \mathbf{x}_0\boldsymbol{\beta}$. Deuxièmement, $E[\hat{Y}_0] = E[\mathbf{x}_0\hat{\boldsymbol{\beta}}] = \mathbf{x}_0 E[\hat{\boldsymbol{\beta}}] = x_0\boldsymbol{\beta}$ puisque l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est sans biais. Ceci complète la preuve.

b) Tout d'abord, $E[(\hat{Y}_0 - E[Y_0])^2] = \mathbf{V}[\hat{Y}_0] = \text{Var}[\hat{Y}_0]$ puisque la matrice de variance-covariance du vecteur aléatoire $\hat{Y}_0$ ne contient, ici, qu'une seule valeur. Or, par le théorème C.6,

$$\begin{aligned}\text{Var}[\hat{Y}_0] &= \mathbf{V}[\mathbf{x}_0\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0 \mathbf{V}[\hat{\boldsymbol{\beta}}]\mathbf{x}_0' \\ &= \sigma^2 \mathbf{x}_0 (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'.\end{aligned}$$

Afin de construire un intervalle de confiance pour $E[Y_0]$, on ajoute au modèle l'hypothèse $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Par linéarité de l'estimateur des moindres carrés, on a alors $\hat{Y}_0 \sim N(E[Y_0], \text{Var}[\hat{Y}_0])$. Par conséquent,

$$\Pr\left[-z_{\alpha/2} \leq \frac{\hat{Y} - E[\hat{Y}_0]}{\sqrt{\text{Var}[\hat{Y}_0]}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

d'où un intervalle de confiance de niveau $1 - \alpha$ pour $E[Y_0]$ est

$$E[Y_0] \in \hat{Y}_0 \pm z_{\alpha/2}\sigma\sqrt{\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

Si la variance σ^2 est inconnue et estimée par s^2, alors la distribution normale est remplacée par une distribution de Student avec $n-p-1$ degrés de liberté. L'intervalle de confiance devient alors

$$E[Y_0] \in \hat{Y}_0 \pm t_{\alpha/2}(n-p-1)\, s \sqrt{\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

c) Par le résultat obtenu en a) et en supposant que $\text{Cov}(\varepsilon_0, \varepsilon_t) = 0$ pour tout $t = 1, \ldots, n$, on a

$$\begin{aligned} E[(Y_0 - \hat{Y}_0)^2] &= \text{Var}[Y_0 - \hat{Y}_0] \\ &= \text{Var}[Y_0] + \text{Var}[\hat{Y}_0] \\ &= \sigma^2(1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'). \end{aligned}$$

Ainsi, avec l'hypothèse sur le terme d'erreur énoncée en b), $Y_0 - \hat{Y}_0 \sim N(0, \text{Var}[Y_0 - \hat{Y}_0])$. En suivant le même cheminement qu'en b), on détermine qu'un intervalle de confiance de niveau $1-\alpha$ pour Y_0 est

$$Y_0 \in \hat{Y}_0 \pm z_{\alpha/2}\sigma \sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

ou, si la variance σ^2 est inconnue et estimée par s^2,

$$Y_0 \in \hat{Y}_0 \pm t_{\alpha/2}(n-p-1)\, s \sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

3.6 On a la relation suivante liant la statistique F et le coefficient de détermination R^2 :

$$F = \frac{R^2}{1-R^2}\,\frac{n-p-1}{p}$$

La principale inconnue dans le problème est n, le nombre de données. Or,

$$\begin{aligned} n &= pF\left(\frac{1-R^2}{R^2}\right) + p + 1 \\ &= 3(5{,}438)\left(\frac{1-0{,}521}{0{,}521}\right) + 3 + 1 \\ &= 19. \end{aligned}$$

Soit F une variable aléatoire dont la distribution est une loi de Fisher avec 3 et $19-3-1=15$ degrés de liberté, soit la même distribution que la statistique F du modèle. On obtient la valeur p du test global de validité du modèle dans un tableau de quantiles de la distribution F ou avec la fonction `pf` dans R :

$$\Pr[F > 5{,}438] = 0{,}0099$$

3.7 a) On a

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \frac{1}{2}\begin{bmatrix} -6 & 34 & -13 & -13 \\ 2 & -4 & 1 & 1 \\ 0 & -2 & 1 & 1 \end{bmatrix}\begin{bmatrix} 17 \\ 12 \\ 14 \\ 13 \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} -45 \\ 13 \\ 3 \end{bmatrix} = \begin{bmatrix} -22{,}5 \\ 6{,}5 \\ 1{,}5 \end{bmatrix}\end{aligned}$$

b) Avec les résultats de la partie a), on a

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 17 \\ 12 \\ 13{,}5 \\ 13{,}5 \end{bmatrix},$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 0 \\ 0 \\ 0{,}5 \\ -0{,}5 \end{bmatrix}$$

et $\bar{Y} = 14$. Par conséquent,

$$\begin{aligned}\text{SST} &= \mathbf{y}'\mathbf{y} - n\bar{Y}^2 = 14 \\ \text{SSE} &= \mathbf{e}'\mathbf{e} = 0{,}5 \\ \text{SSR} &= \text{SST} - \text{SSR} = 13{,}5,\end{aligned}$$

d'où le tableau d'analyse de variance est le suivant :

Source	SS	d.l.	MS	F
Régression	13,5	2	6,75	13,5
Erreur	0,5	1	0,5	
Total	14			

Le coefficient de détermination est

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 0{,}9643.$$

c) On sait que $\text{Var}[\hat{\beta}_i] = \sigma^2 c_{ii}$, où c_{ii} est l'élément en position $(i+1, i+1)$ de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Or, $\hat{\sigma}^2 = s^2 = \text{MSE} = 0{,}5$, tel que calculé en b). Par conséquent, la statistique t du test $H_0 : \beta_1 = 0$ est

$$t = \frac{\hat{\beta}_1}{s\sqrt{c_{11}}} = \frac{6{,}5}{\sqrt{0{,}5(\frac{11}{2})}} = 3{,}920,$$

alors que celle du test $H_0 : \beta_2 = 0$ est

$$t = \frac{\hat{\beta}_2}{s\sqrt{c_{22}}} = \frac{1{,}5}{\sqrt{0{,}5(\frac{3}{2})}} = 1{,}732.$$

À un niveau de signification de 5 %, la valeur critique de ces tests est $t_{0{,}025}(1) = 12{,}706$. Dans les deux cas, on ne rejette donc pas H_0, les variables X_1 et X_2 ne sont pas significatives dans le modèle.

d) Soit $\mathbf{x}_0 = \begin{bmatrix} 1 & 3{,}5 & 9 \end{bmatrix}$ et Y_0 la valeur de la variable dépendante correspondant à $\mathbf{x}_0$. La prévision de Y_0 donnée par le modèle trouvé en a) est

$$\begin{aligned} \hat{Y}_0 &= \mathbf{x}_0\hat{\boldsymbol{\beta}} \\ &= -22{,}5 + 6{,}5(3{,}5) + 1{,}5(9) \\ &= 13{,}75. \end{aligned}$$

D'autre part,

$$\begin{aligned} \widehat{\mathrm{Var}}[Y_0 - \hat{Y}_0] &= s^2(1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0') \\ &= 1{,}1875. \end{aligned}$$

Par conséquent, un intervalle de confiance à 95 % pour Y_0 est

$$\begin{aligned} E[Y_0] &\in \hat{Y}_0 \pm t_{0{,}025}(1)s\sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'} \\ &\in 13{,}75 \pm 12{,}706\sqrt{1{,}1875} \\ &\in (-0{,}096, 27{,}596). \end{aligned}$$

3.8 a) On importe les données dans R, puis on effectue les conversions nécessaires. Comme précédemment, la variable `consommation` contient la consommation des voitures en $\ell/100$ km et la variable `poids` le poids en kilogrammes. On ajoute la variable `cylindree`, qui contient la cylindrée des voitures en litres.

```
> carburant <- read.table("carburant.dat",
+     header = TRUE)
> consommation <- 235.1954/carburant$mpg
> poids <- carburant$poids * 0.45455 * 1000
> cylindree <- carburant$cylindree * 2.54^3/1000
```

b) La fonction `summary` fournit l'information essentielle pour juger de la validité et de la qualité du modèle :

```
> fit <- lm(consommation ~ poids + cylindree)
> summary(fit)

Call:
lm(formula = consommation ~ poids + cylindree)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.8799 -0.5595  0.1577  0.6051  1.7900

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.049304   1.098281  -2.776  0.00877
poids        0.012677   0.001512   8.386 6.85e-10
cylindree   -1.122696   0.333479  -3.367  0.00186

(Intercept) **
poids       ***
cylindree   **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9156 on 35 degrees of freedom
Multiple R-squared: 0.8927,        Adjusted R-squared: 0.8866
F-statistic: 145.6 on 2 and 35 DF,  p-value: < 2.2e-16
```

Le modèle est donc le suivant :

$$Y_t = -3{,}049 + 0{,}01268X_{t1} + -1{,}123X_{t2} + \varepsilon_t, \quad \boldsymbol{\varepsilon}_t \sim N(0, 0{,}9156^2\mathbf{I})$$

où Y_t est la consommation en litres aux 100 kilomètres, X_{t1} le poids en kilogrammes et X_{t2} la cylindrée en litres. La faible valeur p du test F indique une régression globalement très significative. Les tests t des paramètres individuels indiquent également que les deux variables du modèle sont significatives. Enfin, le R^2 de 0,8927 confirme que l'ajustement du modèle est toujours bon.

c) On veut calculer un intervalle de confiance pour la consommation prévue d'une voiture de 1 350 kg ayant un moteur d'une cylindrée de 1,8 litres. On obtient, avec la fonction `predict` :

```
> predict(fit, newdata = data.frame(poids = 1350,
+     cylindree = 1.8), interval = "prediction")
       fit      lwr      upr
1 12.04325 9.959855 14.12665
```

3.9 Il y a plusieurs réponses possibles pour cet exercice. Si l'on cherche, tel que suggéré dans l'énoncé, à distinguer les voitures sport des minifourgonnettes (en supposant que ces dernières ont moins d'accidents que les premières), alors on pourrait s'intéresser, en premier lieu, à la variable `peak.rpm`. Il s'agit du régime moteur maximal, qui est en général beaucoup plus élevé sur les voitures sport. Puisque l'on souhaite expliquer le montant total des sinistres de différents types de voitures, il devient assez naturel de sélectionner également la variable `price`, soit le prix du véhicule. Un véhicule plus luxueux coûte en général plus cher à faire réparer à dommages égaux. Voyons l'effet de l'ajout, pas à pas, de ces deux variables au modèle précédent ne comportant que la variable `horsepower` :

```
> autoprice <- read.table("auto-price.dat",
+     header = TRUE)
> fit1 <- lm(losses ~ horsepower + peak.rpm,
+     data = autoprice)
> summary(fit1)

Call:
lm(formula = losses ~ horsepower + peak.rpm, data = autoprice)

Residuals:
    Min      1Q  Median      3Q     Max
-67.973 -24.074  -6.373  18.049 130.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.521414  29.967570   0.184 0.854060
horsepower   0.318477   0.086840   3.667 0.000336
peak.rpm     0.016639   0.005727   2.905 0.004205

(Intercept)
horsepower  ***
peak.rpm    **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.44 on 156 degrees of freedom
Multiple R-squared: 0.1314,        Adjusted R-squared: 0.1203
F-statistic:  11.8 on 2 and 156 DF,  p-value: 1.692e-05

> anova(fit1)

Analysis of Variance Table

Response: losses
            Df Sum Sq Mean Sq F value    Pr(>F)
horsepower   1  16949   16949 15.1573 0.0001463 ***
peak.rpm     1   9437    9437  8.4397 0.0042049 **
Residuals  156 174435    1118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La variable `peak.rpm` est significative, mais le R^2 demeure faible. Ajoutons maintenant la variable `price` au modèle :

```
> fit2 <- lm(losses ~ horsepower + peak.rpm +
+     price, data = autoprice)
> summary(fit2)

Call:
lm(formula = losses ~ horsepower + peak.rpm + price, data = autoprice)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-66.745 -25.214  -5.867  18.407 130.032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6972172 31.3221462  -0.022  0.98227
horsepower   0.2414922  0.1408272   1.715  0.08838
peak.rpm     0.0181386  0.0061292   2.959  0.00357
price        0.0005179  0.0007451   0.695  0.48803

(Intercept)
horsepower  .
peak.rpm    **
price
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.49 on 155 degrees of freedom
Multiple R-squared: 0.1341,        Adjusted R-squared: 0.1173
F-statistic: 8.001 on 3 and 155 DF,  p-value: 5.42e-05

> anova(fit2)

Analysis of Variance Table

Response: losses
            Df Sum Sq Mean Sq F value    Pr(>F)
horsepower   1  16949   16949 15.1071 0.0001502 ***
peak.rpm     1   9437    9437  8.4118 0.0042702 **
price        1    542     542  0.4832 0.4880298
Residuals  155 173893    1122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Du moins avec les variables `horsepower` et `peak.rpm`, la variable `price` n'est pas significative. D'ailleurs, l'augmentation du R^2 suite à l'ajout de cette variable est minime. À ce stade de l'analyse, il vaudrait sans doute mieux reprendre tout depuis le début avec d'autres variables. Des méthodes de sélection des variables seront étudiées plus avant dans le chapitre.

3.10 a) On a $p = 3$ variables explicatives et, du nombre de degrés de liberté de la statistique F, on apprend que $n - p - 1 = 16$. Par conséquent, $n = 16 + 3 + 1 = 20$. Les dimensions des vecteurs et de la matrice de schéma dans la représentation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ sont donc : $n \times 1 = 20 \times 1$ pour les vecteurs $\mathbf{y}$ et $\boldsymbol{\varepsilon}$, $n \times (p+1) = 20 \times 4$ pour la matrice $\mathbf{X}$, $(p+1) \times 1$ pour le vecteur $\boldsymbol{\beta}$.

b) La valeur p associée à la statistique F est, à toute fin pratique, nulle. Cela permet de rejeter facilement l'hypothèse nulle selon laquelle la régression n'est pas significative.

c) On doit se fier ici au résultat du test t associé à la variable X_2. Dans les résultats obtenus avec R, on voit que la valeur p de la statistique t du paramètre β_2 est 0,0916. Cela signifie que jusqu'à un seuil de signification de 9,16 % (ou un niveau de confiance supérieur à 90,84 %), on ne peut rejeter l'hypothèse $H_0 : \beta_2 = 0$ en faveur de $H_1 : \beta_2 \neq 0$. Il s'agit néanmoins d'un cas limite et il est alors du ressort de l'analyste de décider d'inclure ou non le revenu disponible dans le modèle.

d) Le coefficient de détermination est de $R^2 = 0{,}981$. Cela signifie que le prix de la bière, le revenu disponible et la demande de l'année précédente expliquent plus de 98 % de la variation de la demande en bière. L'ajustement du modèle aux données est donc particulièrement bon. Il est tout à fait possible d'obtenir un R^2 élevé et, simultanément, toutes les statistiques t non significatives : comme chaque test t mesure l'impact d'une variable sur la régression étant donné la présence des autres variables, il suffit d'avoir une bonne variable dans un modèle pour obtenir un R^2 élevé et une ou plusieurs autres variables redondantes avec la première pour rendre les tests t non significatifs.

3.11 a) L'information demandée doit évidemment être extraite des deux tableaux d'analyse de variance fournis dans l'énoncé. Il importe, ici, de savoir que le résultat de la fonction `anova` de R est un tableau d'analyse de variance séquentiel, où chaque ligne identifiée par le nom d'une variable correspond au test F partiel résultant de l'ajout de cette variable au modèle. Ainsi, du premier tableau on obtient les sommes de carrés

$$\begin{aligned} \text{SSR}(X_2) &= 45{,}59085 \\ \text{SSR}(X_3|X_2) &= 8{,}76355 \end{aligned}$$

alors que du second tableau on a

$$\begin{aligned} \text{SSR}(X_1) &= 45{,}59240 \\ \text{SSR}(X_2|X_1) &= 0{,}01842 \\ \text{SSR}(X_3|X_1, X_2) &= 8{,}78766, \end{aligned}$$

ainsi que

$$\begin{aligned} \text{MSE} &= \frac{SSE(X_1, X_2, X_3)}{n - p - 1} \\ &= 0{,}44844. \end{aligned}$$

i) Le test d'hypothèse $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ est le test global de

validité du modèle. La statistique F pour ce test est

$$\begin{aligned} F &= \frac{\text{SSR}(X_1, X_2, X_3)/3}{\text{MSE}} \\ &= \frac{(\text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1,X_2))/3}{\text{MSE}} \\ &= \frac{(45{,}5924 + 0{,}01842 + 8{,}78766)/3}{0{,}44844} \\ &= 40{,}44. \end{aligned}$$

Puisque la statistique MSE a 21 degrés de liberté, la statistique F en a 3 et 21.

ii) Pour tester cette hypothèse, il faut utiliser un test F partiel. On teste si la variable X_1 est significative dans la régression globale. La statistique du test est alors

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_1|X_2,X_3)/1}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_2, X_3)}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_2) - \text{SSR}(X_3|X_2)}{\text{MSE}} \\ &= \frac{54{,}39848 - 45{,}59085 - 8{,}76355}{0{,}44844} \\ &= 0{,}098, \end{aligned}$$

avec 1 et 21 degrés de liberté.

iii) Cette fois, on teste si les variables X_2 et X_3 (les deux ensemble) sont significatives dans la régression globale. On effectue donc encore un test F partiel avec la statistique

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_2, X_3|X_1)/2}{\text{MSE}} \\ &= \frac{(\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1))/2}{\text{MSE}} \\ &= \frac{(54{,}39848 - 45{,}5924)/2}{0{,}44844} \\ &= 9{,}819, \end{aligned}$$

avec 2 et 21 degrés de liberté.

b) À la lecture du premier tableau d'analyse de variance que tant les variables X_2 que X_3 sont significatives dans le modèle. Par contre, comme on le voit dans le second tableau, la variable X_2 devient non significative dès lors que la variable X_1 est ajoutée au modèle. (L'impact de la variable X_3 demeure, lui, inchangé.) Cela signifie que les

variables X_1 et X_2 sont redondantes et qu'il faut choisir l'une ou l'autre, mais pas les deux. Par conséquent, les choix de modèle possibles sont X_1 et X_3, ou X_2 et X_3.

3.12 La statistique à utiliser pour faire ce test F partiel est

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_2, X_3 | X_1, X_4)/2}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3, X_4) - \text{SSR}(X_1, X_4)}{2\text{MSE}} \\ &= \frac{\text{SSR} - \text{SSR}(X_4) - \text{SSR}(X_1 | X_4)}{2s^2} \end{aligned}$$

où $\text{SSR} = \text{SSR}(X_1, X_2, X_3, X_4)$. Or,

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{\text{SSR}}{\text{SSR} + \text{SSE}}, \end{aligned}$$

d'où

$$\begin{aligned} \text{SSR} &= \frac{R^2}{1 - R^2} \text{SSE} \\ &= \frac{R^2}{1 - R^2} \text{MSE}(n - p - 1) \\ &= \frac{0{,}6903}{1 - 0{,}6903} (26{,}41)(506 - 4 - 1) \\ &= 29\,492. \end{aligned}$$

Par conséquent,

$$\begin{aligned} F^* &= \frac{29\,492 - 2\,668 - 21\,348}{(2)(26{,}41)} \\ &= 103{,}67. \end{aligned}$$

3.13 a) Tout d'abord, si $Z \sim N(0,1)$ et $V \sim \chi^2(r)$ alors, par définition,

$$\frac{Z}{\sqrt{V/r}} \sim t(r).$$

Tel que mentionné dans l'énoncé, $\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$ ou, de manière équivalente,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1).$$

Par conséquent,

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}}{\sqrt{\frac{\text{SSE}}{\sigma^2(n-p-1)}}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t(n - p - 1).$$

b) En régression linéaire simple, $c_{11} = 1/\sum_{t=1}^{n}(X_t - \bar{X})^2 = 1/S_{XX}$ et $\sigma^2 c_{11} = \text{Var}[\hat{\beta}_1]$. Le résultat général en a) se réduit donc, en régression linéaire simple, au résultat bien connu du test t sur le paramètre β_1

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{1/S_{XX}}} \sim t(n-1-1).$$

3.14 En suivant les indications donnée dans l'énoncé, on obtient aisément

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2\left(\frac{d}{d\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)' \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2(\mathbf{X}'\mathbf{W}\mathbf{y} - \mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

Par conséquent, les équations normales à résoudre pour trouver l'estimateur $\hat{\boldsymbol{\beta}}^*$ minimisant la somme de carrés pondérés $S(\boldsymbol{\beta})$ sont $(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}}^* = \mathbf{X}'\mathbf{W}\mathbf{y}$ et l'estimateur des moindres carrés pondérés est

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

3.15 De manière tout à fait générale, l'estimateur linéaire sans biais à variance minimale dans le modèle de régression linéaire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{W}^{-1}$ est

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

et sa variance est, par le théorème C.6,

$$\begin{aligned}\mathbf{V}[\hat{\boldsymbol{\beta}}^*] &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{V}[\mathbf{y}]\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{W}^{-1}\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\end{aligned}$$

puisque les matrices $\mathbf{W}$ et $\mathbf{X}'\mathbf{W}\mathbf{X}$ sont symétriques. Dans le cas de la régression linéaire simple passant par l'origine et en supposant que $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$, ces formules se réduisent en

$$\hat{\beta}^* = \frac{\sum_{t=1}^{n} w_t X_t Y_t}{\sum_{t=1}^{n} w_t X_t^2}$$

et

$$\text{Var}[\hat{\beta}^*] = \frac{\sigma^2}{\sum_{t=1}^{n} w_t X_t^2}.$$

a) Cas déjà traité à l'exercice 2.6 où $\mathbf{W} = \mathbf{I}$ et, donc,

$$\hat{\beta}^* = \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}$$

et

$$\text{Var}[\hat{\beta}^*] = \frac{\sigma^2}{\sum_{t=1}^n w_t X_t^2}.$$

b) Cas général traité ci-dessus.

c) Si $\text{Var}[\varepsilon_t] = \sigma^2 X_t$, alors $w_t = X_t^{-1}$. Le cas général se simplifie donc en

$$\begin{aligned}\hat{\beta}^* &= \frac{\sum_{t=1}^n Y_t}{\sum_{t=1}^n X_t} \\ &= \frac{\bar{Y}}{\bar{X}}, \\ \text{Var}[\hat{\beta}^*] &= \frac{\sigma^2}{\sum_{t=1}^n X_t} \\ &= \frac{\sigma^2}{n\bar{X}}.\end{aligned}$$

d) Si $\text{Var}[\varepsilon_t] = \sigma^2 X_t^2$, alors $w_t = X_t^{-2}$. On a donc

$$\begin{aligned}\hat{\beta}^* &= \frac{1}{n}\sum_{t=1}^n \frac{Y_t}{X_t} \\ \text{Var}[\hat{\beta}^*] &= \frac{\sigma^2}{n}.\end{aligned}$$

3.16 Le graphique des valeurs de Y en fonction de celles de X, à la figure D.9, montre clairement une relation quadratique. On postule donc le modèle

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2).$$

Par la suite, on peut estimer les paramètres de ce modèle avec la fonction `lm` de R :

```
> fit <- lm(Y ~ poly(X, 2), data = donnees)
> summary(fit)

Call:
lm(formula = Y ~ poly(X, 2), data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9123 -0.6150 -0.1905  0.6367  1.6921

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.1240     0.3025   59.91 3.10e-16
poly(X, 2)1  29.6754     1.1717   25.33 8.72e-12
poly(X, 2)2   4.0899     1.1717    3.49  0.00446
```

```
> plot(Y ~ X, data = donnees)
```

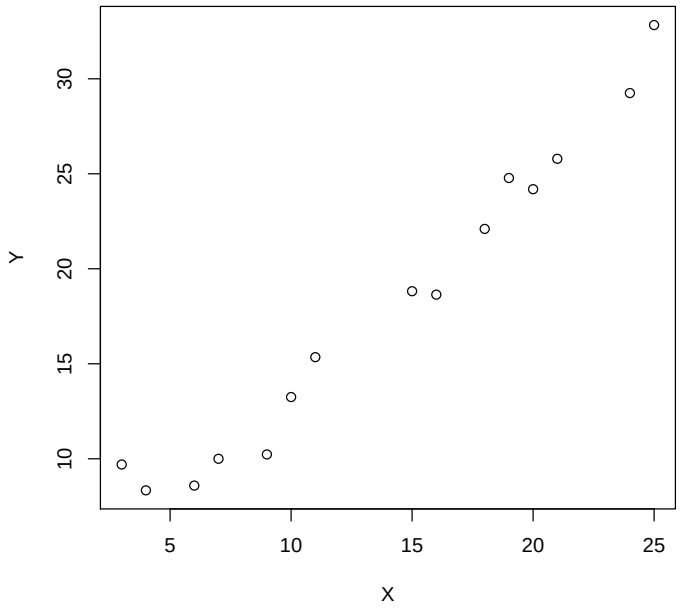


FIGURE D.9: Graphique des données de l'exercice 3.16

```
(Intercept) ***
poly(X, 2)1 ***
poly(X, 2)2 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.172 on 12 degrees of freedom
Multiple R-squared: 0.982,        Adjusted R-squared: 0.979
F-statistic: 326.8 on 2 and 12 DF,  p-value: 3.434e-11

> anova(fit)

Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
poly(X, 2)  2 897.36  448.68  326.79 3.434e-11 ***
Residuals  12  16.48    1.37
```

```
> plot(Y ~ X, data = donnees)
> x <- seq(min(donnees$X), max(donnees$X),
+     length = 200)
> lines(x, predict(fit, data.frame(X = x),
+     lwd = 2))
```

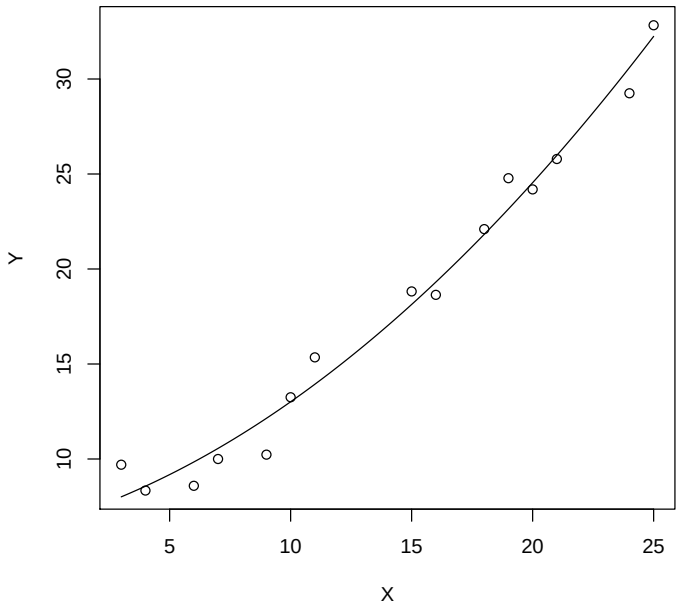


FIGURE D.10: Graphique des données de l'exercice 3.16 et courbe obtenue par régression

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

Tant le test F global que les tests t individuels sont concluants, le coefficient de détermination est élevé et l'on peut constater à la figure D.10 que l'ajustement du modèle est bon. On conclut donc qu'un modèle adéquat pour cet ensemble de données est

$$Y_t = 18{,}12 + 29{,}68X_t + 4{,}09X_t^2 + \varepsilon_t, \quad \varepsilon_t \sim N(0, 1{,}373).$$

3.17 Comme on peut le constater à la figure D.11, le point (X_{16}, Y_{16}) est plus éloigné des autres. En b) et c), on diminue son poids dans la régression.

```
> plot(Y ~ X, data = donnees)
> points(donnees$X[16], donnees$Y[16], pch = 16)
```

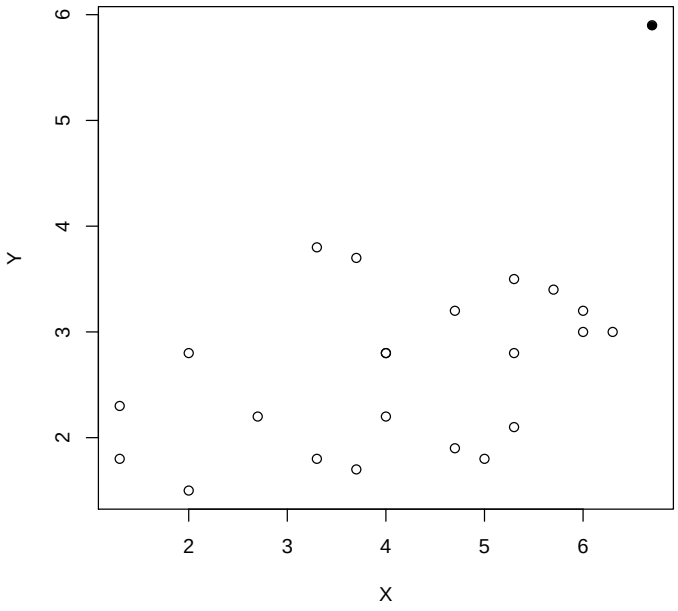


FIGURE D.11: Graphique des données de l'exercice 3.17. Le cercle plein représente la donnée (X_{16}, Y_{16}).

a) On calcule d'abord l'estimateur des moindres carrés ordinaires :

```
> (fit1 <- lm(Y ~ X, data = donnees))

Call:
lm(formula = Y ~ X, data = donnees)

Coefficients:
(Intercept)            X
     1.4256       0.3158
```

b) Si l'on suppose que la variance de la données (X_{16}, Y_{16}) est quatre fois plus élevée que la variance des autres données, alors il convient d'accorder un point quatre fois moins grand à cette donnée dans la régression. Cela requiert les moindres carrés pondérés. Pour calculer les estimateurs avec `lm` dans R, on utilise l'argument `weights` :

```
> w <- rep(1, nrow(donnees))
> w[16] <- 0.25
> (fit2 <- update(fit1, weights = w))
```

```
Call:
lm(formula = Y ~ X, data = donnees, weights = w)

Coefficients:
(Intercept)            X
     1.7213       0.2243
```

c) On répète la procédure en b) avec un poids de encore plus petit pour la donnée (X_{16}, Y_{16}) :

```
> w[16] <- 0.0625
> (fit3 <- update(fit1, weights = w))

Call:
lm(formula = Y ~ X, data = donnees, weights = w)

Coefficients:
(Intercept)            X
     1.8080       0.1975
```

Plus le poids accordé à la donnée (X_{16}, Y_{16}) est faible, moins la droite de régression est attirée vers ce point (voir la figure D.12).

3.18 a) Voir la figure D.13 pour le graphique. Il y a effectivement une différence entre la consommation de carburant des hommes et des femmes : ces dernières font plus de milles avec un gallon d'essence.

b) Remarquer que la variable `sexe` est un facteur et peut être utilisée telle quelle dans `lm` :

```
> (fit <- lm(mpg ~ age + sexe, data = donnees))

Call:
lm(formula = mpg ~ age + sexe, data = donnees)

Coefficients:
(Intercept)          age        sexeM
     16.687       -1.040       -1.206
```

c) Calcul d'une prévision pour la valeur moyenne de la variable `mpg` :

```
> predict(fit, newdata = data.frame(age = 4,
+     sexe = "F"), interval = "confidence",
+     level = 0.9)

       fit      lwr      upr
1 12.52876 11.94584 13.11168
```

Chapitre 4

4.1 La série `strikes` comporte une composante de tendance relativement compliquée (voir la figure D.14(a)). Cette tendance n'est certes pas linéaire, ni même quadratique, mais un polynôme du troisième degré constituerait une bonne approximation. Les figures D.14(b)–D.14(d) montrent

```
> plot(Y ~ X, data = donnees)
> points(donnees$X[16], donnees$Y[16], pch = 16)
> abline(fit1, lwd = 2, lty = 1)
> abline(fit2, lwd = 2, lty = 2)
> abline(fit3, lwd = 2, lty = 3)
> legend(1.2, 6, legend = c("Modèle a)",
+     "Modèle b)", "Modèle c)"), lwd = 2,
+     lty = 1:3)
```

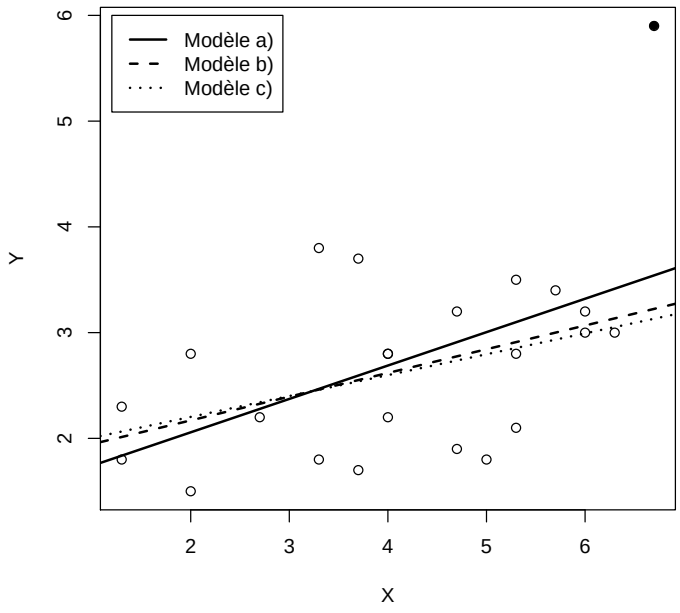


FIGURE D.12: Graphique des données de l'exercice 3.17 avec les droites de régression obtenues à l'aide des moindres carrés pondérés.

l'effet des différences de premier, second et troisième ordre, respectivement, sur la série. On constate que seule la dernière série ne semble contenir aucune tendance. Il faut donc différencier trois fois :

```
> diff(strikes, differences = 3)

Time Series:
Start = 1954
End = 1980
Frequency = 1
 [1] -1191  4072 -3822  1690  -170  -180  -382   798
 [9]  -196  -712  1044  -530   119  -386   512   -55
```

```
> hommes <- subset(donnees, sexe == "M")
> femmes <- subset(donnees, sexe == "F")
> plot(mpg ~ age, data = hommes, xlim = range(donnees$age),
+     ylim = range(donnees$mpg))
> points(mpg ~ age, data = femmes, pch = 16)
> legend(4, 16, legend = c("Hommes", "Femmes"),
+     pch = c(1, 16))
```

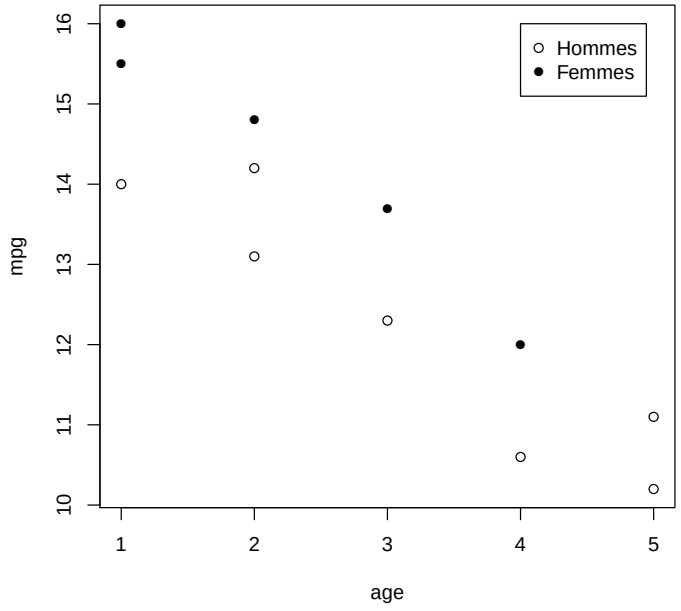
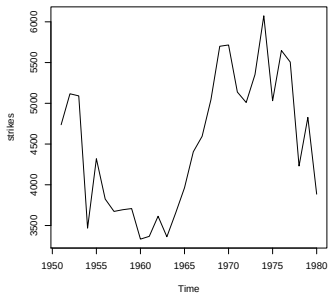


FIGURE D.13: Graphique des données de l'exercice 3.18

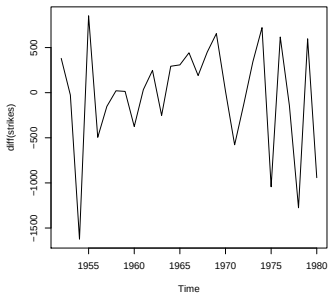
```
[17]  -844    45  1044    21   -93 -2142  3424 -2419
[25]  -375  3007 -3412
```

4.2 La série `sales` ainsi que les estimations de sa tendance obtenues par lissage exponentiel avec $\alpha = 0{,}2$, $\alpha = 0{,}5$ et $\alpha = 0{,}7$ sont présentées graphiquement à la figure D.15. On constate que plus la valeur de α augmente, plus l'estimation de la tendance est proche de la série originale. Cela n'est pas nécessairement souhaitable, puisque la soustraction de la tendance de la série originale résultera en une série de résidus contenant peu d'informations.

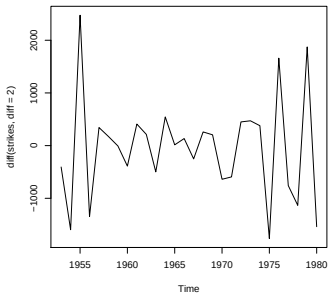
4.3 a) On peut considérer la série `deaths` comme formée de composantes de tendance, de saisonnalité et de bruit aléatoire (ou résidus). La fonction `stl` décompose la série en ces trois composantes :



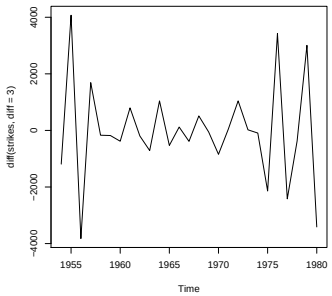
(a) Série originale



(b) Série différenciée une fois



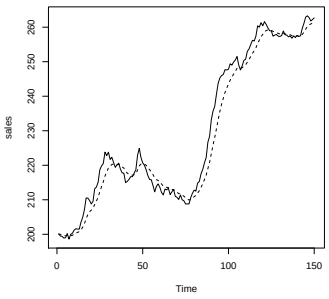
(c) Série différenciée deux fois



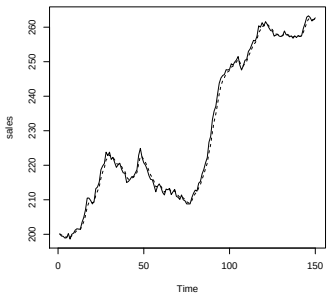
(d) Série différenciée trois fois

FIGURE D.14: Série `strikes`

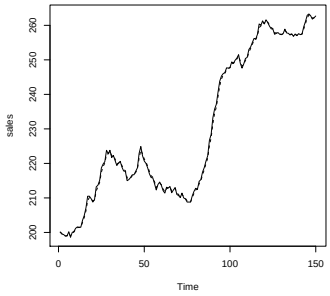
```
> deaths.stl <- stl(deaths, "periodic")
> summary(deaths.stl)

 Call:
 stl(x = deaths, s.window = "periodic")

 Time.series components:
    seasonal                trend
 Min.   :-1.557915e+03   Min.   :8353.534
 1st Qu.:-5.941842e+02   1st Qu.:8520.346
 Median :-4.378816e+01   Median :8712.571
 Mean   :-1.212946e-05   Mean   :8796.119
 3rd Qu.: 4.570118e+02   3rd Qu.:8894.177
 Max.   : 1.682646e+03   Max.   :9934.492
   remainder
 Min.   :-473.061194
 1st Qu.:-162.853032
 Median : -30.432160
 Mean   :  -8.382972
 3rd Qu.: 120.005132
```



(a) $\alpha = 0{,}2$



(b) $\alpha = 0{,}5$



(c) $\alpha = 0{,}7$

FIGURE D.15: Série `sales` (ligne pleine) et lissage exponentiel pour différentes valeurs de α (ligne pointillée)

```
Max.   : 602.016731
IQR:
    STL.seasonal STL.trend STL.remainder data
    1051.2        373.8     282.9        1234.3
  %  85.2         30.3      22.9         100.0

Weights: all == 1

Other components: List of 5
$ win  : Named num [1:3] 721 19 13
$ deg  : Named int [1:3] 0 1 1
$ jump : Named num [1:3] 73 2 2
$ inner: int 2
$ outer: int 0
```

Se reporter à la figure D.16 pour la représentation graphique de cette décomposition.

b) Le corrélogramme des résidus se trouve à la figure D.17. Il s'agit essen-

```
> plot(deaths.stl)
```

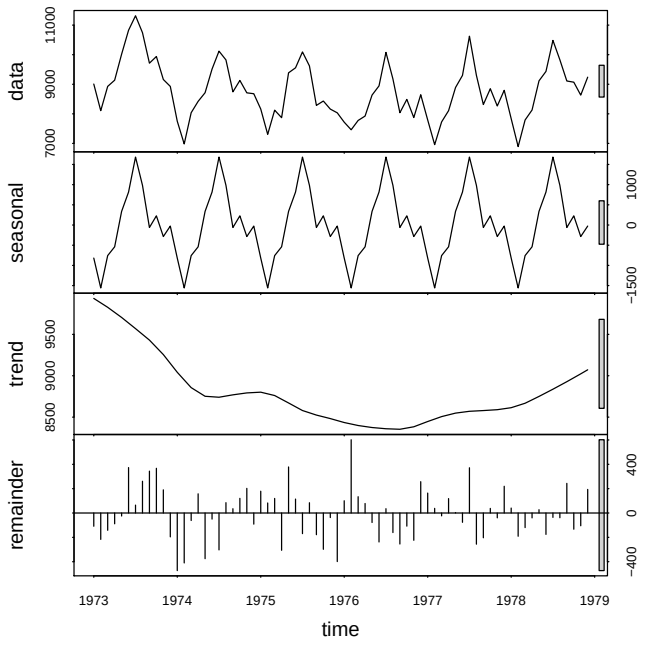


FIGURE D.16: Décomposition de la série `deaths` avec la fonction `stl`

tiellement du corrélogramme d'un bruit blanc, un processus stationnaire. Ceci indique que la tendance et la saisonnalité ont correctement été éliminées de la série.

4.4 a) La série originale se trouve à la figure D.18(a), alors que le logarithme de la série est représenté à la figure D.18(b) (page 114). Dans cette dernière série, l'amplitude de la composante saisonnière est davantage constante dans le temps. On préférera donc utiliser le modèle $Y_t = m_t + s_t + X_t$ pour le logarithme des données de vente de bière.

b) La période de la série `log(beer)` est d'environ 12 mois. On élimine donc la saisonnalité avec une différence de pas 12 :

```
> diff(log(beer), lag = 12)
```

La série résultante est présentée à la figure D.18(c). La moyenne de cette série n'est pas stationnaire. On élimine donc une composante de tendance à l'aide de la première différence :

```
> diff(diff(log(beer), lag = 12))
```

Voir la figure D.18(d) pour la série résultante. Celle-ci est maintenant stationnaire.

```
> acf(deaths.stl$time.series[, "remainder"])
```

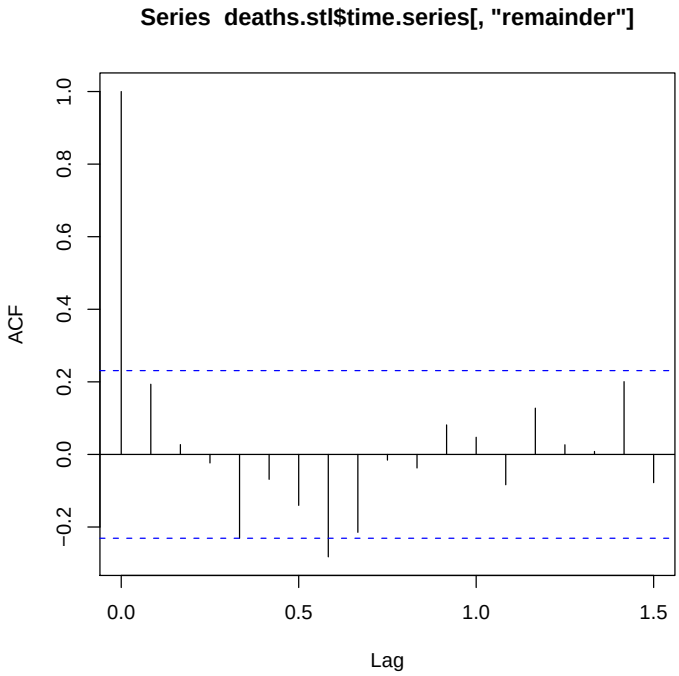
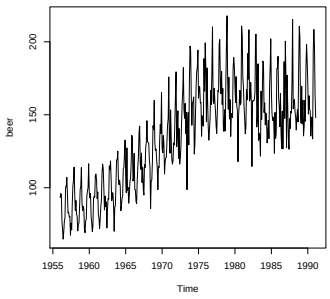


FIGURE D.17: Corrélogramme des résidus de la décomposition de la série `deaths` avec la fonction `stl`

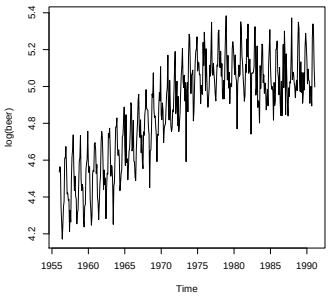
4.5 Soit $m_t = c_0 + c_1 t$. On a

$$\begin{aligned}\sum_{j=-\infty}^{\infty} a_j m_{t-j} &= \frac{1}{2q+1} \sum_{j=-q}^{q} (c_0 + c_1(t-j)) \\ &= \frac{1}{2q+1} \left[(2q+1)(c_0 + c_1 t) - \sum_{j=-q}^{q} j \right] \\ &= c_0 + c_1 t \\ &= m_t.\end{aligned}$$

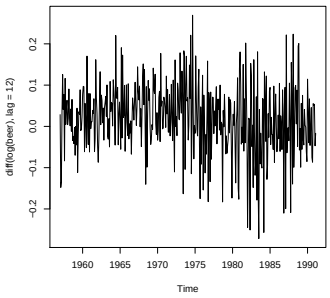
4.6 On a $[a_{-2}, a_{-1}, a_0, a_1, a_2] = \frac{1}{9}[-1, 4, 3, 4, -1]$, $m_t = c_0 + c_1 t + c_2 t^2 + c_3 t^3$, $s_{t+3} = s_t$ et $s_t + s_{t+1} + s_{t+2} = 0$ pour tout t. Premièrement, on démontre



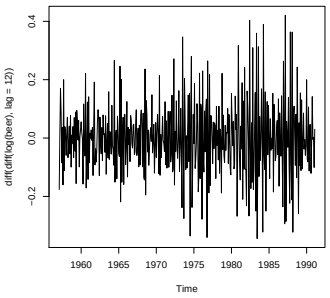
(a) Série originale



(b) Logarithme de la série



(c) `diff(log(beer), lag = 12)`



(d) `diff(diff(log(beer), lag = 12))`

FIGURE D.18: Graphiques de la série `beer` (exercice 4.4)

que le filtre laisse passer la tendance sans distorsion :

$$\begin{aligned}
\sum_{j=-2}^{2} a_j m_{t-j} &= \frac{1}{9}\left[-m_{t+2} + 4m_{t+1} + 3m_t + 4m_{t-1} - m_{t-2}\right] \\
&= \frac{1}{9}\left[c_0(-1+4+3+4-1)\right. \\
&\quad + c_1(-(t+2) + 4(t+1) + 3t + 4(t-1) - (t-2)) \\
&\quad + c_1(-(t+2)^2 + 4(t+1)^2 + 3t^2 + 4(t-1)^2 - (t-2)^2) \\
&\quad \left. + c_1(-(t+2)^3 + 4(t+1)^3 + 3t^3 + 4(t-1)^3 - (t-2)^3)\right] \\
&= \frac{1}{9}\left[9c_0 + 9c_1 t + 9c_2 t^2 + 9c_3 t^3\right] = m_t.
\end{aligned}$$

Deuxièmement, on démontre que le filtre élimine une composante de

saisonnalité de période 3 :

$$\begin{aligned}\sum_{j=-2}^{2} a_j s_{t-j} &= \frac{1}{9}\left[-s_{t+2} + 4s_{t+1} + 3s_t + 4s_{t-1} - s_{t-2}\right] \\ &= \frac{1}{9}\left[3s_{t+2} + 3s_{t+1} + 3s_t\right] \\ &= \frac{1}{3}\left[s_{t+2} + s_{t+1} + s_t\right] = 0.\end{aligned}$$

4.7 Que $E[A_t] = 0$ et $\mathrm{Var}[A_t] = \sigma^2/(2q+1)$ est évident. Or, $\lim_{q\to\infty} \mathrm{Var}[A_t] = 0$, d'où l'on dit que $\{A_t\}$ est «petite» pour de grandes valeurs de q.

4.8 On a $B^k X_t = X_{t-k}$, $m_t = c_0 + c_1 t$, $s_t = {s_{t+2}}^{(*)}$, $s_t + s_{t+1} = 0^{(**)}$. On souhaite trouver les valeurs α, β et γ tel que $(1 + \alpha B + \beta B^2 + \gamma B^3)m_t = m_t \Leftrightarrow (\alpha B + \beta B^2 + \gamma B^3)m_t = 0$ et $(1 + \alpha B + \beta B^2 + \gamma B^3)s_t = 0$. Or,

$$\begin{aligned}(\alpha B + \beta B^2 + \gamma B^3)m_t &= \alpha m_{t-1} + \beta m_{t-2} + \gamma m_{t-3} \\ &= (\alpha + \beta + \gamma)(c_0 + c_1 t) - (\alpha + 2\beta + 3\gamma)c_1 \\ &= (\alpha + \beta + \gamma)m_t - (\alpha + 2\beta + 3\gamma)c_1\end{aligned}$$

et, en supposant que $m_t \neq 0$ et que $c_1 \neq 0$ (sans intérêt sinon), cette expression est égale à 0 si

$$\begin{aligned}\alpha + \beta + \gamma &= 0 \\ \alpha + 2\beta + 3\gamma &= 0.\end{aligned}$$

De plus,

$$\begin{aligned}(1 + \alpha B + \beta B^2 + \gamma B^3)s_t &= s_t + \alpha s_{t-1} + \beta s_{t-2} + \gamma s_{t-3} \\ &\overset{(*)}{=} (1 + \beta)s_t + (\alpha + \gamma)s_{t-1} \\ &\overset{(**)}{=} (1 - \alpha + \beta - \gamma)s_t\end{aligned}$$

et, en supposant que $s_t \neq 0$ (ce qui est raisonnable), cette expression est égale à 0 si

$$\alpha - \beta + \gamma = 1.$$

La résolution du système de trois équations donne $\alpha = \gamma = 1/4$ et $\beta = -1/2$.

4.9 a) Il n'y a qu'une tendance approximativement linéaire à éliminer.

```
> diff(x)

Time Series:
Start = 2
End = 9
Frequency = 1
[1]  -9  98 -81  64  67  20   1  -6
```

b) Test portmanteau :

```
> Box.test(diff(x), lag = 4)

        Box-Pierce test

data:  diff(x)
X-squared = 2.3267, df = 4, p-value = 0.6759
```

Pour le test des changements de direction, $T = (3-4)/\sqrt{1{,}1} = 0{,}95 < 1{,}96$. On ne rejette donc pas l'hypothèse de bruit blanc.

4.10 En effectuant une inspection visuelle, on remarque que la série des ventes de la Guinness de janvier 1995 à novembre 1999 (un exemple fictif obtenu à partir de la série `wines.dat`, par ailleurs) montre une composante de saisonnalité de période $d = 6$ (c'est-à-dire une année) et une tendance linéaire croissante.

```
> (m.t <- filter(X, filter = rep(1/3, 3),
+     sides = 2))

Time Series:
Start = c(1995, 1)
End = c(1999, 6)
Frequency = 6
 [1]        NA  76.66667 105.33333 119.33333
 [5] 114.66667  88.33333  80.66667  87.33333
 [9] 121.66667 128.00000 121.66667  87.66667
[13]  78.33333  78.66667 107.33333 117.33333
[17] 116.66667  90.66667  81.66667  90.66667
[21] 120.66667 131.33333 127.66667 100.33333
[25]  95.33333 105.33333 140.33333 147.33333
[29] 140.66667        NA

> diff(X - m.t, lag = 6)

Time Series:
Start = c(1996, 1)
End = c(1999, 6)
Frequency = 6
 [1]          NA   8.3333333 -10.3333333  15.3333333
 [5] -11.0000000   1.6666667   3.3333333  -0.3333333
 [9]  -3.6666667  -5.3333333   7.0000000  -4.0000000
[13]   4.6666667  -9.0000000  11.6666667  -2.0000000
[17]  -6.0000000   6.3333333  -5.6666667   2.3333333
[21]  -0.6666667   7.0000000  -7.0000000          NA
```

L'ordre d'application du filtre ou des différences n'a pas d'importance.

```
> diff(X, lag = 6) - filter(diff(X, lag = 6),
+     filter = rep(1/3, 3), sides = 2)

Time Series:
Start = c(1996, 1)
```

```
End = c(1999, 6)
Frequency = 6
 [1]          NA   8.3333333 -10.3333333  15.3333333
 [5] -11.0000000   1.6666667   3.3333333  -0.3333333
 [9]  -3.6666667  -5.3333333   7.0000000  -4.0000000
[13]   4.6666667  -9.0000000  11.6666667  -2.0000000
[17]  -6.0000000   6.3333333  -5.6666667   2.3333333
[21]  -0.6666667   7.0000000  -7.0000000          NA
```

4.11 Le filtre est essentiellement sans aucun effet puisqu'il a été conçu pour laisser passer une tendance linéaire et pour éliminer une saisonnalité de période 2.

4.12 a) La série $\{Y_t\}$ montre une tendance linéaire croissante. Cette tendance peut être éliminée en différenciant la série une fois pour obtenir une nouvelle série $X_t = \nabla Y_t$:

```
> diff(Y)

Time Series:
Start = 2
End = 10
Frequency = 1
[1]  1.5  1.4  0.1  1.6  1.1 -1.0 -0.4  3.7 -0.1
```

b) On teste l'hypothèse selon laquelle la série $X_t = \nabla Y_t$ est un bruit blanc. La statistique du test portmanteau pour les 9 observations de $\{X_t\}$ et les 8 autocorrélations empiriques $\hat{\rho}(h)$ fournies dans l'énoncé est

$$\begin{aligned} Q^* &= 9 \sum_{h=1}^{8} \hat{\rho}(h)^2 \\ &= 3{,}685 < 15{,}51 = \chi^2_{0{,}05,8}. \end{aligned}$$

Selon le test portmanteau, on ne rejette pas l'hypothèse de bruit blanc.

Il y a 4 changements de direction dans la série $\{X_t\}$ (observations x_3, x_4, x_6 et x_8). La statistique du test du nombre de changements de direction est

$$\begin{aligned} T &= \left| \frac{4 - \frac{2}{3}(7)}{\sqrt{\frac{1}{90}(16(9) - 29)}} \right| \\ &= 0{,}5898 < 1{,}96. \end{aligned}$$

On ne rejette pas plus l'hypothèse de bruit blanc avec ce test qu'avec le test effectué précédemment. On peut dès lors conclure qu'un modèle approprié pour la série originale $\{Y_t\}$ est

$$Y_t = c_0 + c_1 t + Z_t,$$

où $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

Chapitre 5

5.1 a) On a

$$\begin{aligned}
E[X_t] &= a \quad \text{est indépendant de } t \\
\gamma_X(t,t+h) &= b^2\gamma_Z(h) + bc\gamma_Z(h-2) \\
&\quad + bc\gamma_Z(h+2) + c^2\gamma_Z(h) \\
&= \begin{cases} (b^2+c^2)\sigma^2, & h=0 \\ bc\sigma^2, & h=\pm 2 \\ 0, & \text{ailleurs} \end{cases} \\
&= \gamma_X(h).
\end{aligned}$$

Le processus $X_t = a + bZ_t + cZ_{t-2}$ est donc stationnaire.

b) On a

$$\begin{aligned}
E[X_t] &= 0 \\
\gamma_X(t,t+h) &= \cos(ct)\cos(ct+ch)\sigma^2 + \sin(ct)\sin(ct+ch)\sigma^2 \\
&= \sigma^2[\cos^2(ct)\cos(ch) - \cos(ct)\sin(ct)\sin(ch) \\
&\quad + \sin^2(ct)\cos(ch) + \cos(ct)\sin(ch)\sin(ct)] \\
&= \cos(ch)\sigma^2 \\
&= \gamma_X(h).
\end{aligned}$$

Le processus $X_t = Z_1\cos(ct) + Z_2\sin(ct)$ est donc stationnaire.

c) On a

$$\begin{aligned}
E[X_t] &= 0 \\
\gamma_X(t,t+h) &= \cos(ct)\cos(ct+ch)\gamma_Z(h) \\
&\quad + \cos(ct)\sin(ct+ch)\gamma_Z(h-1) \\
&\quad + \sin(ct)\cos(ct+ch)\gamma_Z(h+1) \\
&\quad + \sin(ct)\sin(ct+ch)\gamma_Z(h) \\
&= \begin{cases} \sigma^2, & h=0 \\ \cos(ct)\sin(ct+ch)\sigma^2, & h=1 \\ \sin(ct)\cos(ct+ch)\sigma^2, & h=-1 \\ 0, & \text{ailleurs.} \end{cases}
\end{aligned}$$

Le processus $X_t = Z_t\cos(ct) + Z_{t-1}\sin(ct)$ n'est donc pas stationnaire.

d) On a

$$\begin{aligned}
E[X_t] &= a \\
\gamma_X(t,t+h) &= b^2\text{Var}[Z_0] \\
&= b^2\sigma^2 \\
&= \gamma_X(h).
\end{aligned}$$

Le processus $X_t = a + bZ_0$ est donc stationnaire.

e) On a

$$\begin{aligned} E[X_t] &= \gamma_Z(1) = 0 \\ \gamma_X(t, t+h) &= E[X_t X_{t+h}] - E[X_t]E[X_{t+h}] \\ &= E[Z_t Z_{t-1} Z_{t+h} Z_{t+h-1}] \\ &= \begin{cases} \sigma^2\sigma^2 = \sigma^4, & h = 0 \\ 0, & h = \pm 1 \\ 0, & h \neq 0 \end{cases} \\ &= \gamma_X(h). \end{aligned}$$

Le processus $X_t = Z_t Z_{t-1}$ est donc stationnaire.

5.2 On sait que $\nabla_k X_t = X_t - X_{t-k}$, $k = 1, 2, \ldots$ et $\gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h})$.

a) On a $s_t = s_{t-12}$. Ainsi, avec $Y_t = a + bt + s_t + X_t$,

$$\begin{aligned} \nabla_{12} Y_t &= Y_t - Y_{t-12} \\ &= 12b + X_t - X_{t-12} \end{aligned}$$

et

$$\nabla \nabla_{12} Y_t = X_t - X_{t-1} - X_{t-12} + X_{t-13} = W_t.$$

Maintenant, il est clair que l'espérance du processus $\{W_t\}$ est indépendante de t, car celle du processus $\{X_t\}$ l'est. En outre,

$$\begin{aligned} \gamma_W(h) = \mathrm{Cov}(&X_t - X_{t-1} - X_{t-12} + X_{t-13}, \\ &X_{t+h} - X_{t+h-1} - X_{t+h-12} + X_{t+h-13}) \\ = 4\gamma_X(h) &- 2\gamma_X(h-1) - 2\gamma_X(h-12) \\ &+ \gamma_X(h-13) - 2\gamma_X(h+1) + \gamma_X(h-11) \\ &- 2\gamma_X(h+12) + \gamma_X(h+11) + \gamma_X(h+13), \end{aligned}$$

ce qui est clairement indépendant de t. Le processus $\{W_t\}$ est donc stationnaire.

b) Avec $Y_t = (a + bt)s_t + X_t$ et, encore une fois, $s_t = s_{t-12}$,

$$\nabla_{12} Y_t = 12bs_t + X_t - X_{t-12}$$

et

$$\nabla_{12}^2 Y_t = X_t - 2X_{t-12} + X_{t-24} = W_t.$$

Avec une démarche semblable à celle effectuée en a), on démontre alors que le processus $\{W_t\}$ est stationnaire.

5.3 Soit μ_X la moyenne et $\gamma_X(h)$ la FACV du processus stationnaire $\{X_t\}$, puis μ_Y et $\gamma_Y(h)$ les fonctions correspondantes pour le processus $\{Y_t\}$. Les deux processus sont indépendants. Par conséquent,

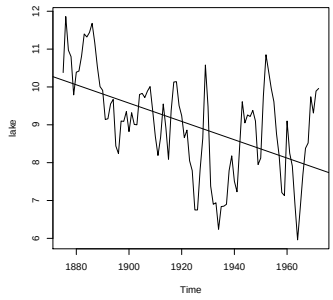
$$E[X_t + Y_t] = \mu_X + \mu_Y \quad \text{est indépendant de } t,$$

et

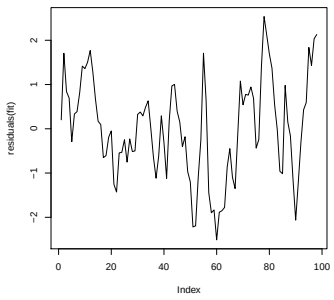
$$\begin{aligned}\gamma_{X+Y}(t, t+h) &= \text{Cov}(X_t + Y_t, X_{t+h} + Y_{t+h}) \\ &= \text{Cov}(X_t, X_{t+h}) + \text{Cov}(Y_t, Y_{t+h}) \\ &= \gamma_X(h) + \gamma_Y(h),\end{aligned}$$

ce qui complète la preuve.

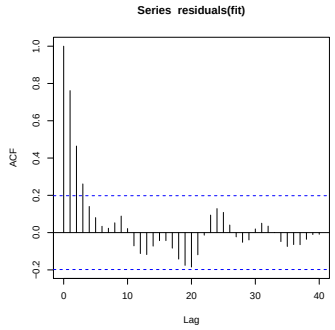
5.4 La série `lake.dat` est représentée à la figure D.19(a). Il n'y a pas de composante de saisonnalité apparente dans cette série, mais il y a une tendance linéaire décroissante. On peut alors postuler le modèle suivant



(a) Série originale avec la droite de régression



(b) Résidus de régression



(c) FAC des résidus de régression

FIGURE D.19: Graphiques reliés à la série `lake` (exercice 5.4)

pour la série `lake` $\{Y_t\}$:

$$Y_t = m_t + X_t, \quad t = 1875, \ldots, 1972,$$

où $m_t = \beta_0 + \beta_1 t$ et le modèle des résidus $\{X_t\}$ est à déterminer. La valeur de β_0 et β_1 est estimée par les moindres carrés en utilisant la fonction `lm` dans R :

```
> (fit <- lm(lake ~ time(lake)))

Call:
lm(formula = lake ~ time(lake))

Coefficients:
(Intercept)   time(lake)
    55.5549      -0.0242
```

Cette droite de régression est incorporée au graphique de la série à la figure D.19(a). Le graphique de la série des résidus de la régression

$$\hat{X}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 t$$

se trouve à la figure D.19(b). La fonction d'autocorrélation empirique correspondante $\hat{\rho}(h)$ se trouve, quant à elle, à la figure D.19(c) pour $h = 0, \ldots, 40$.

Puisque plus de $0{,}95(40) = 2$ valeurs excèdent les bornes de l'intervalle de confiance à 95 %, il est clair que les résidus $\{\hat{X}_t\}$ ne proviennent pas d'un bruit blanc. Cette assertion est confirmée par les tests portmanteau et des changements de direction, qui tous les deux rejettent l'hypothèse d'un bruit blanc :

```
> res <- residuals(fit)
> n <- length(res)
> Box.test(res, lag = 40)

        Box-Pierce test

data:  res
X-squared = 109.2377, df = 40, p-value =
2.418e-08

> TP <- sum(max.col(embed(res, 3)) == 2, na.rm = TRUE) +
+     sum(max.col(embed(-res, 3)) == 2, na.rm = TRUE)
> abs((TP - 2 * (n - 2)/3)/sqrt((16 * n -
+     29)/90)) > 1.96

[1] TRUE
```

La forme générale de la FAC empirique suggère que des modèles potentiels pour les résidus $\{\hat{X}_t\}$ seraient un AR(1) ou un AR(2).

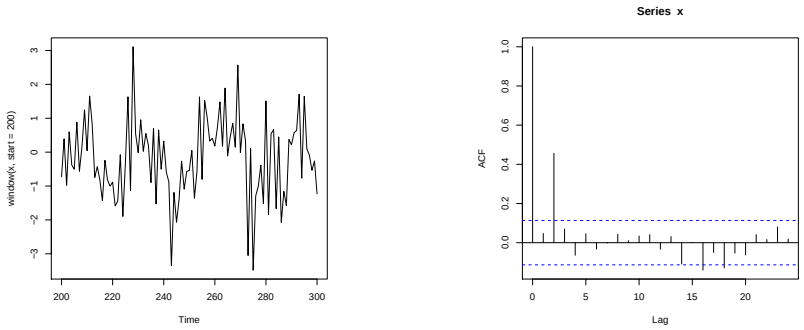


FIGURE D.20: Simulation (100 dernières valeurs) et FAC empirique de la série $\{X_t\}$ de l'exercice 5.5 b)

5.5 a) La FACV est donnée par

$$\begin{aligned}\gamma_X(h) &= \text{Cov}(Z_t + \theta Z_{t-2}, Z_{t+h} + \theta Z_{t+h-2}) \\ &= \gamma_Z(h) + \theta\gamma_Z(h-2) + \theta\gamma_Z(h+2) + \theta^2\gamma_Z(h) \\ &= \begin{cases} 1+\theta^2, & h = 0 \\ \theta, & h = \pm 2 \\ 0, & \text{ailleurs} \end{cases}\end{aligned}$$

puisque $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. La FAC est alors

$$\rho_X(h) = \begin{cases} 1, & h = 0 \\ \frac{\theta}{1+\theta^2}, & h = \pm 2 \\ 0, & \text{ailleurs.} \end{cases}$$

b) Le modèle de la série $\{X_t\}$ est simplement un processus MA(2) avec $\theta_1 = 0$, un modèle simple à simuler avec la fonction `arima.sim` :

```
> x <- arima.sim(list(ma = c(0, 0.8)), n = 300)
```

La série simulée ainsi que sa FAC empirique se trouvent à la figure D.20. Afin d'améliorer la lisibilité du graphique, seules les 100 dernières valeurs de la série sont affichées.

c) La fonction `arima.sim` est maintenant appelée avec $\theta_2 = -0.8$:

```
> x <- arima.sim(list(ma = c(0, -0.8)), n = 300)
```

La série simulée ainsi que sa FAC empirique se retrouvent à la figure D.21.

d) Oui, les corrélogrammes obtenus en b) et c) correspondent à la fonction d'autocorrélation théorique calculée en a).

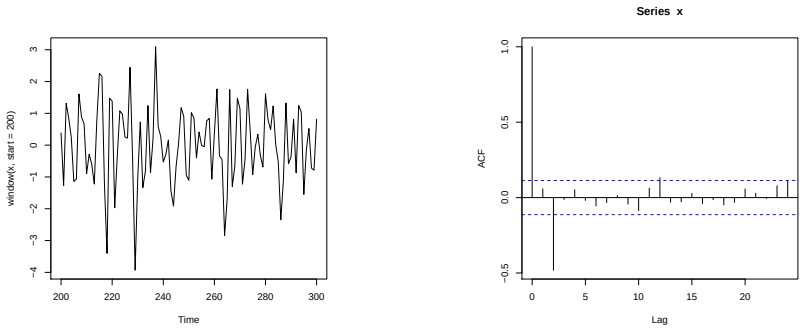


FIGURE D.21: Simulation (100 dernières valeurs) et FAC empirique de la série $\{X_t\}$ de l'exercice 5.5 c)

e) La série avec $\theta = 0.8$ fluctue moins rapidement puisque les observations distantes d'un pas de 2 sont corrélées positivement. Elles ont donc tendance à aller dans la même direction.

5.6 On a $X_t = \phi X_{t-1} + Z_t$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ et $\gamma_X(h) = \phi^h \gamma_X(0)$, $\gamma_X(0) = \sigma^2/(1-\phi^2)$.

a) Des hypothèses ci-dessus,

$$\begin{aligned}\text{Var}\left[\frac{X_1 + X_2 + X_3 + X_4}{4}\right] &= \frac{1}{16}(4\text{Var}[X_1] + 6\gamma_X(1) + \\ &\quad + 4\gamma_X(2) + 2\gamma_X(3)) \\ &= \frac{\sigma^2}{1-\phi^2}\left[\frac{1}{4} + \frac{3}{8}\phi + \frac{1}{4}\phi^2 + \frac{1}{8}\phi^3\right].\end{aligned}$$

Si $\phi = 0.9$ et $\sigma^2 = 1$, alors la variance est égale à 4,638.

b) Maintenant, si $\phi = -0.9$, alors la variance est égale à 0,126. L'explication de cette plus faible variance lorsque $\phi < 0$ est similaire à celle donnée en 5.5 e).

5.7 Tout d'abord, puisque $Z_t \sim N(0,1)$, alors $Z_t^2 \sim \chi^2(1)$, d'où $E[Z_t^2] = 1$ et $\text{Var}[Z_t^2] = 2$. On a donc

$$\begin{aligned}E[X_t] &= \begin{cases} E[Z_t], & t \text{ pair} \\ E\left[\frac{Z_{t-1}^2 - 1}{\sqrt{2}}\right], & t \text{ impair} \end{cases} \\ &= 0\end{aligned}$$

et

$$\begin{aligned}\text{Var}[X_t] &= \begin{cases}\text{Var}[Z_t], & t \text{ pair}\\ \text{Var}\left[\frac{Z_{t-1}^2-1}{\sqrt{2}}\right], & t \text{ impair}\end{cases}\\ &= 1.\end{aligned}$$

De plus,

$$\begin{aligned}\gamma_X(1) &= \text{Cov}\left(Z_t, \frac{Z_t^2-1}{\sqrt{2}}\right)\\ &= \frac{1}{\sqrt{2}}\text{Cov}(Z_t, Z_t^2)\\ &= \frac{1}{\sqrt{2}}(E[Z_t^3] - E[Z_t]E[Z_t^2])\\ &= 0\end{aligned}$$

car tous les moments impairs d'une variable aléatoire normale sont nuls. Il est clair que $\gamma_X(h) = 0$ pour $|h| > 1$. Par conséquent,

$$\mu_X = 0 \quad \text{et} \quad \gamma_X(h) = \begin{cases}1, & h = 0\\ 0, & h \neq 0.\end{cases}$$

Ainsi, $\{X_t\}$ est WN$(0,1)$. Cependant, $X_t = Z_t$ et $X_{t+1} = (Z_t^2 - 1)/\sqrt{2}$ ne sont pas des variables aléatoires indépendantes, d'où $\{X_t\}$ n'est pas un bruit IID.

5.8 Cet exercice vise simplement à illustrer comment diverses combinaisons de valeurs d'un processus de bruit blanc peuvent générer des observations de processus ARMA.

a) On peut calculer les valeurs de $\{X_t\}$ à partir de la définition $X_t = 0{,}6X_{t-1} + Z_t$ avec $X_1 = Z_1$. On a donc

$$\begin{aligned}X_1 &= Z_1 = 0{,}180\\ X_2 &= 0{,}6X_1 + Z_2 = -1{,}502\\ X_3 &= 0{,}6X_2 + Z_3 = 2{,}099\\ X_4 &= 0{,}6X_3 + Z_4 = 2{,}589.\end{aligned}$$

De manière équivalente, la solution de l'équation caractéristique d'un processus AR(1) est $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. On a donc

$$\begin{aligned}X_1 &= Z_1 = 0{,}180\\ X_2 &= Z_2 + 0{,}6Z_1 = -1{,}502\\ X_3 &= Z_3 + 0{,}6Z_2 + 0{,}36Z_1 = 2{,}099\\ X_4 &= Z_4 + 0{,}6Z_3 + 0{,}36Z_2 + 0{,}216Z_3 = 2{,}099.\end{aligned}$$

b) On a $X_t = Z_t - 0{,}4Z_{t-1}$, donc

$$\begin{aligned} X_1 &= Z_1 = 0{,}180 \\ X_2 &= Z_2 - 0{,}4Z_1 = -1{,}682 \\ X_3 &= Z_3 - 0{,}4Z_2 = 3{,}644 \\ X_4 &= Z_4 - 0{,}4Z_3 = 0{,}130. \end{aligned}$$

c) Ici, $X_t = 0{,}6X_{t-1} + Z_t - 0{,}4Z_{t-1}$. Par conséquent,

$$\begin{aligned} X_1 &= Z_1 = 0{,}180 \\ X_2 &= 0{,}6X_1 + Z_2 - 0{,}4Z_1 = -1{,}574 \\ X_3 &= 0{,}6X_2 + Z_3 - 0{,}4Z_2 = 2{,}700 \\ X_4 &= 0{,}6X_3 + Z_4 - 0{,}4Z_3 = 1{,}750. \end{aligned}$$

On peut également démontrer que la réprésentation MA(∞) d'un processus ARMA$(1,1)$ est $X_t = Z_t + (\phi + \theta)\sum_{j=1}^{\infty} \phi^{j-1} Z_{t-j}$. On a donc, de manière équivalente,

$$\begin{aligned} X_1 &= Z_1 = 0{,}180 \\ X_2 &= Z_2 + 0{,}2Z_1 = -1{,}574 \\ X_3 &= Z_3 + 0{,}2(Z_2 + 0{,}6Z_1) = 2{,}700 \\ X_4 &= Z_4 + 0{,}2(Z_3 + 0{,}6Z_2 + 0{,}36Z_3) = 1{,}750. \end{aligned}$$

5.9 a) On a $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$, d'où

$$\begin{aligned} X_t - \phi X_{t-1} &= \sum_{j=0}^{\infty} \phi^j Z_{t-j} - \phi \sum_{j=0}^{\infty} \phi^j Z_{t-1-j} \\ &= \sum_{j=0}^{\infty} \phi^j Z_{t-j} - \sum_{j=0}^{\infty} \phi^{j+1} Z_{t-1-j} \\ &= \sum_{j=0}^{\infty} \phi^j Z_{t-j} - \sum_{j=1}^{\infty} \phi^j Z_{t-j} \\ &= Z_t. \end{aligned}$$

b) On a cette fois $X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$, d'où

$$\begin{aligned} X_t - \phi X_{t-1} &= -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} + \phi \sum_{j=1}^{\infty} \phi^{-j} Z_{t-1+j} \\ &= -\sum_{j=1}^{\infty} \phi^{-j} Z_{t-j} + \sum_{j=1}^{\infty} \phi^{-j+1} Z_{t-1+j} \\ &= -\sum_{j=1}^{\infty} \phi^{-j} Z_{t-j} + \sum_{j=0}^{\infty} \phi^{-j} Z_{t-j} \\ &= Z_t. \end{aligned}$$

La solution n'est évidemment pas causale puisque les valeurs du processus $\{X_t\}$ sont déterminées par des valeurs futures du processus $\{Z_t\}$.

5.10 Dans ce qui suit, z_k, $k = 1,2$ représentent les racines du polynôme en z.

a) Processus AR(2) : $X_t + 0.2X_{t-1} - 0.48X_{t-2} = Z_t \Rightarrow \phi(z) = 1 + 0.2z - 0.48z^2 \Rightarrow |z_1| = |5/3| > 1$ et $|z_2| = |-1.25| > 1$. De plus, $\theta(z) \equiv 1$. Le processus $\{X_t\}$ est donc stationnaire et réversible.

b) Processus ARMA(2,2) : $X_t + 1.9X_{t-1} - 0.88X_{t-2} = Z_t + 0.2Z_{t-1} + 0.7Z_{t-2} \Rightarrow \phi(z) = 1 + 1.9z - 0.88z^2 \Rightarrow |z_1| = |2.5967| > 1$ et $|z_2| = |-0.4376| < 1$. De plus, $\theta(z) = 1 + 0.2z + 0.7z^2 \Rightarrow z_{1,2} = -0.1429 \pm 1.1867i \Rightarrow |z_{1,2}| = \sqrt{(-0.1429)^2 + (1.1867)^2} > 1$. Le processus $\{X_t\}$ n'est donc pas stationnaire, mais il est réversible.

c) Processus ARMA(1,1) : $X_t + 0.6X_{t-1} = Z_t + 1.2Z_{t-1} \Rightarrow \phi(z) = 1 + 0.6z \Rightarrow |\phi| = |-0.6| < 1$. De plus, $\theta(z) = 1 + 1.2z \Rightarrow |\theta| = |1.2| > 1$. Le processus $\{X_t\}$ est alors stationnaire, mais non réversible.

d) Processus AR(2) : $X_t + 1.8X_{t-1} - 0.81X_{t-2} = Z_t \Rightarrow \phi(z) = 1 + 1.8z - 0.81z^2 \Rightarrow |z_1| = |-0.4602| < 1$ et $|z_2| = |2.6825| > 1$. De plus, $\theta(z) \equiv 1$. Le processus $\{X_t\}$ n'est donc pas stationnaire, mais il est réversible.

e) Processus ARMA(1,2) : $X_t + 1.6X_{t-1} = Z_t - 0.4Z_{t-1} + 0.04Z_{t-2} \Rightarrow \phi(z) = 1 + 1.6z \Rightarrow |\phi| = |-1.6| > 1$. De plus, $\theta(z) = 1 - 0.4z + 0.04z^2 \Rightarrow |z_1| = |z_2| = |5| > 1$. Le processus $\{X_t\}$ n'est donc pas stationnaire, mais il est réversible.

5.11 a) Pour démontrer que la série $\{Y_t\}$ est stationnaire, on doit connaître la covariance entre X_t et W_s pour tout t et s. Comme $\{X_t\}$ est un processus AR(1) stationnaire, on peut l'écrire sous la forme $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. On constate donc que

$$\mathrm{Cov}(X_t, W_s) = \sum_{j=0}^{\infty} \phi^j \mathrm{Cov}(Z_{t-j}, W_s) = 0,$$

car $\mathrm{Cov}(Z_t, W_s) = E[Z_t W_s] = 0$. Par conséquent, $E[X_t] = E[Y_t + W_t] = 0$ et

$$\begin{aligned}
\gamma_Y(h) &= \mathrm{Cov}(X_{t+h} + W_{t+h}, X_t + W_t) \\
&= \gamma_X(h) + \gamma_W(h) \\
&= \begin{cases} \dfrac{\sigma_Z^2}{1-\phi^2} + \sigma_W^2, & h = 0 \\ \dfrac{\phi^{|h|}}{1-\phi^2}\sigma_Z^2, & h \neq 0, \end{cases}
\end{aligned}$$

car $\{X_t\} \sim \mathrm{AR}(1)$ et $\{W_t\} \sim \mathrm{WN}(0, \sigma_W^2)$. Puisque les fonctions $E[Y_t]$ et $\gamma_Y(h)$ ne dépendent pas du temps, on conclut que le processus $\{Y_t\}$ est stationnaire.

b) On a $U_t = Y_t - \phi Y_{t-1}$, d'où

$$\begin{aligned}\gamma_U(h) &= \mathrm{Cov}(Y_t - \phi Y_{t-1}, Y_{t+h} - \phi Y_{t+h-1}) \\ &= (1+\phi^2)\gamma_Y(h) - \phi\gamma_Y(h-1) - \phi\gamma_Y(h+1) \\ &= \begin{cases} (1+\phi^2)\sigma_W^2 + \sigma_Z^2, & h = 0 \\ -\phi\sigma_W^2, & h = \pm 1 \\ 0, & |h| > 1. \end{cases}\end{aligned}$$

Ainsi, $\{U_t\}$ est 1–corrélée et, en raison de la correspondance biunivoque entre les modèles ARMA et leur FACV, il s'agit d'un processus MA(1) avec paramètres θ et σ^2.

c) Si $\{Y_t - \phi Y_{t-1}\} \sim \mathrm{MA}(1)$, alors $\{Y_t\} \sim \mathrm{ARMA}(1,1)$, car la série $\{Y_t\}$ peut être exprimée comme étant la solution de $Y_t - \phi Y_{t-1} = V_t + \theta V_{t-1}$, où $\{V_t\} \sim \mathrm{WN}(0, \sigma^2)$. Les paramètres de ce modèle ARMA(1,1) sont ϕ, θ et σ^2, où θ et σ^2 sont les solutions du système d'équations non linéaires

$$\begin{aligned}(1+\phi^2)\sigma_W^2 + \sigma_Z^2, &= (1+\theta^2)\sigma^2 \\ -\phi\sigma_W^2 &= \theta\sigma^2.\end{aligned}$$

5.12 a) En multipliant l'équation caractéristique d'un processus ARMA(p,q) de part et d'autre par X_{t-h}, puis en prenant l'espérance, on obtient

$$\sum_{k=0}^{p} \phi_k E[X_{t-k}X_{t-h}] = \sum_{k=0}^{p} \theta_k E[Z_{t-k}X_{t-h}],$$

avec $\phi_0 = \theta_0 = 1$. Or, $E[X_{t-k}X_{t-h}] = \gamma_X(h-k)$ et, en utilisant l'identité $X_{t-h} = \sum_{j=0}^{\infty} \psi_j Z_{t-h-j}$,

$$\begin{aligned}E[Z_{t-k}X_{t-h}] &= \sum_{j=0}^{\infty} \psi_j E[Z_{t-k}Z_{t-h-j}] \\ &= \sum_{j=0}^{\infty} \psi_j \delta_{k,h+j}\sigma^2.\end{aligned}$$

On a donc

$$\begin{aligned}\sum_{k=0}^{p} \phi_k \gamma_X(h-k) &= \sum_{k=0}^{q}\sum_{j=0}^{\infty} \theta_k\psi_j\delta_{k,h+j}\sigma^2 \\ &= \sigma^2 \sum_{j=0}^{\infty} \psi_j\theta_{h+j}\end{aligned}$$

pour $h = 0, \ldots, q$. Lorsque $h > q$, le côté droit de l'équation est égal à 0.

b) Pour le processus ARMA$(1,1)$, on a, pour $h = 0$ et $h = 1$,

$$\begin{aligned}\gamma_X(0) - \phi\gamma_X(1) &= \sigma^2(1 + \psi_1\theta)\\ \gamma_X(1) - \phi\gamma_X(0) &= \sigma^2\theta.\end{aligned}$$

Or, on sait que, pour le processus ARMA$(1,1)$, $\psi_1 = \phi + \theta$. La solution de ce système d'équations est, par conséquent,

$$\begin{aligned}\gamma_X(0) &= \sigma^2\left(\frac{1 + \psi_1\theta + \phi\theta}{1 - \phi^2}\right)\\ &= \sigma^2\left(1 + \frac{(\phi + \theta)^2}{1 - \phi^2}\right)\end{aligned}$$

et

$$\begin{aligned}\gamma_X(1) &= \sigma^2\theta + \phi\gamma_X(0)\\ &= \sigma^2\left(\phi + \theta + \frac{(\phi + \theta)^2\phi}{1 - \phi^2}\right).\end{aligned}$$

Pour $h > 1$, on a

$$\begin{aligned}\gamma_X(2) &= \phi\gamma_X(1)\\ \gamma_X(3) &= \phi\gamma_X(2)\\ &= \phi^2\gamma_X(1)\end{aligned}$$

soit, de manière générale,

$$\gamma_X(h) = \phi^{h-1}\gamma_X(1), \quad h > 1.$$

5.13 a) Il s'agit d'un processus ARMA$(1,0)$, ou plus simplement AR(1). Les coefficients ψ_1, ψ_2 et ψ_3 satisfont l'égalité

$$(1 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \ldots)(1 - 0{,}5z) = 1,$$

soit

$$\begin{aligned}\psi_1 &= 0{,}5\\ \psi_2 &= 0{,}5\psi_1 = 0{,}25\\ \psi_3 &= 0{,}5\psi_2 = 0{,}125.\end{aligned}$$

On confirme cette réponse avec la fonction `ARMAtoMA` de R :

```
> ARMAtoMA(ar = 0.5, lag.max = 3)
[1] 0.500 0.250 0.125
```

Puisque le processus est déjà inversé, on sait que les coefficients de la représentation AR(∞) sont simplement $\pi_1 = -0{,}5$ et $\pi_j = 0$, $j > 1$.

b) On a un processus MA(2) avec $\psi_1 = -1{,}3$, $\psi_2 = 0{,}4$ et $\psi_j = 0$, $j > 2$. Les trois premiers coefficients de la représentation AR(∞) satisfont l'équation

$$(1 + \pi_1 z + \pi_2 z^2 + \pi_3 z^3 + \ldots)(1 - 1{,}3z + 0{,}4z^2) = 1,$$

soit

$$\begin{aligned}
\pi_1 &= 1{,}3 \\
\pi_2 &= -0{,}4 + 1{,}3\pi_1 = 1{,}29 \\
\pi_3 &= -0{,}4\pi_1 + 1{,}3\pi_2 = 1{,}157.
\end{aligned}$$

On peut calculer les coefficients π_j avec la fonction `ARMAtoMA` dans R en inversant simplement le rôle des coefficients ϕ_j et θ_j (ainsi que leur signe). En d'autres mots, trouver les coefficients π_j du processus $X_t = Z_t - 1{,}3Z_{t-1} + 0{,}4Z_{t-2}$ est en tous points équivalent à trouver les coefficients ψ_j du processus $X_t - 1{,}3X_{t-1} + 0{,}4X_{t-2} = Z_t$. On a donc

```
> ARMAtoMA(ar = c(1.3, -0.4), lag.max = 3)
[1] 1.300 1.290 1.157
```

c) On a un processus ARMA(1,2). Les trois premiers coefficients ψ_j sont obtenus en égalant les coefficients des puissances de z dans

$$(1 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \ldots)(1 - 0{,}5z) = 1 - 1{,}3z + 0{,}4z^2.$$

On obtient

$$\begin{aligned}
\psi_1 &= 0{,}5 - 1{,}3 = -0{,}8 \\
\psi_2 &= 0{,}5\psi_1 + 0{,}4 = 0 \\
\psi_3 &= 0{,}5\psi_2 = 0.
\end{aligned}$$

Confirmation avec R :

```
> ARMAtoMA(ar = 0.5, ma = c(-1.3, 0.4), lag.max = 3)
[1] -0.8  0.0  0.0
```

Les trois premiers coefficients π_j sont obtenus à partir de l'équation

$$(1 + \pi_1 z + \pi_2 z^2 + \pi_3 z^3 + \ldots)(1 - 1{,}3z + 0{,}4z^2) = 1 - 0{,}5z,$$

soit

$$\begin{aligned}
\pi_1 &= 1{,}3 - 0{,}5 = 0{,}8 \\
\pi_2 &= -0{,}4 + 1{,}3\pi_1 = 0{,}64 \\
\pi_3 &= -0{,}4\pi_1 + 1{,}3\pi_2 = 0{,}512.
\end{aligned}$$

En effet,

```
> ARMAtoMA(ar = c(1.3, -0.4), ma = -0.5, lag.max = 3)
[1] 0.800 0.640 0.512
```

d) On a en fait $(1-0{,}2B)(1-B)X_t = Z_t - 0{,}5Z_{t-1}$, soit un processus ARIMA$(1,1,1)$. Les trois premiers coefficients ψ_j sont obtenus à partir de l'équation

$$(1+\psi_1 z+\psi_2 z^2+\psi_3 z^3+\ldots)(1-1{,}2z+0{,}2z^2) = 1-0{,}5z,$$

d'où

$$\begin{aligned}\psi_1 &= 1{,}2-0{,}5 = 0{,}7\\ \psi_2 &= 1{,}2\psi_1 - 0{,}2 = 0{,}640\\ \psi_3 &= 1{,}2\psi_2 - 0{,}2\psi_1 = 0{,}628.\end{aligned}$$

On obtient le même résultat avec `ARMAtoMA` :

```
> ARMAtoMA(ar = c(1.2, -0.2), ma = -0.5, lag.max = 3)
[1] 0.700 0.640 0.628
```

Pour les trois premiers coefficients π_j, on résout

$$(1+\pi_1 z+\pi_2 z^2+\pi_3 z^3+\ldots)(1-0{,}5z) = 1-1{,}2z+0{,}2z^2,$$

ce qui donne

$$\begin{aligned}\pi_1 &= 0{,}5-1{,}2 = -0{,}7\\ \pi_2 &= 0{,}5\pi_1 + 0{,}2 = -0{,}15\\ \pi_3 &= 0{,}5\pi_2 = -0{,}075.\end{aligned}$$

En effet :

```
> ARMAtoMA(ar = 0.5, ma = c(-1.2, 0.2), lag.max = 3)
[1] -0.700 -0.150 -0.075
```

5.14 L'autocorrélation partielle de pas 2, ϕ_{22}, est donnée par

$$\begin{aligned}\phi_{21}+\phi_{22}\rho(1) &= \rho(1)\\ \phi_{21}\rho(1)+\phi_{22} &= \rho(2).\end{aligned}$$

Or, pour un processus MA(1), $\rho(1) = \theta/(1+\theta^2)$ et $\rho(2) = 0$. On a donc

$$\begin{aligned}\phi_{22} &= -\frac{\rho(1)^2}{1-\rho(1)^2}\\ &= -\frac{\theta^2}{1+\theta^2+\theta^4}.\end{aligned}$$

5.15 a) Puisque $\mathbf{V}[\varepsilon] \neq \sigma^2\mathbf{I}$, l'estimateur des moindres carrés généralisés a une plus faible variance que l'estimateur des moindres carrés ordinaires.

b) On sait que $\{\varepsilon_t\} \sim \text{AR}(1)$ avec $\phi = 0{,}8$ et $\sigma^2 = 9$. Par conséquent,

$$\begin{aligned}\gamma_\varepsilon(h) &= \frac{\sigma^2}{1-\phi^2}\phi^h \\ &= 25(0{,}8)^h, \quad h = 0,1,2,\ldots\end{aligned}$$

d'où la matrice $\mathbf{V} = \mathbf{V}[\boldsymbol{\varepsilon}]$ à utiliser dans les moindres carrés généralisés est

$$\mathbf{V} = 25\begin{bmatrix} 1 & 0{,}8 & 0{,}8^2 & \ldots & 0{,}8^{n-1} \\ 0{,}8 & 1 & 0{,}8 & \ldots & 0{,}8^{n-2} \\ \vdots & & \ddots & & \vdots \\ 0{,}8^{n-1} & 0{,}8^{n-2} & 0{,}8^{n-3} & \ldots & 1 \end{bmatrix}.$$

Chapitre 6

6.1 a) La première autocorrélation partielle est toujours égale à la première autocorrélation. Par conséquent,

$$\begin{aligned}\hat{\phi}_{11} &= \hat{\rho}(1) \\ &= \frac{\sum_{t=1}^{n-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\ &= -0{,}2695.\end{aligned}$$

b) Pour un processus AR(1), on a toujours $\phi_{22} = 0$.

6.2 On sait des notes de cours que les estimateurs de Yule–Walker des paramètres θ et σ^2 d'un modèle $\text{MA}(1)$ sont obtenus avec

$$\hat{\sigma}^2 = \frac{\hat{\gamma}(1)}{\hat{\theta}}$$
$$\hat{\rho}(1) = \frac{\hat{\theta}}{1+\hat{\theta}^2}.$$

En isolant $\hat{\theta}$ dans la seconde équation et en supposant que $|\hat{\rho}(1)| < 1$ et $|\hat{\theta}| < 1$, on obtient

$$\hat{\theta} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}.$$

6.3 a) Par définition de la fonction d'autocovariance,

$$\begin{aligned}\gamma_X(h) &= \text{Cov}(X_t, X_{t+h}) \\ &= \begin{cases} (1+\theta_1^2+\theta_{12}^2)\sigma^2 & h = 0 \\ \theta_1\sigma^2 & h = \pm 1 \\ \theta_1\theta_{12}\sigma^2 & h = \pm 11 \\ \theta_{12}\sigma^2 & h = \pm 12. \end{cases}\end{aligned}$$

b) On a

```
> y <- diff(diff(deaths, lag = 12))
> mean(y)
[1] 28.83051
> acf(y, lag.max = 12, type = "covariance",
+     plot = FALSE)$acf
, , 1

            [,1]
 [1,] 152669.632
 [2,] -54326.528
 [3,] -15071.682
 [4,]  14584.568
 [5,] -17177.694
 [6,]   6340.251
 [7,]  17420.908
 [8,] -31164.460
 [9,]  -1087.513
[10,]  15277.175
[11,] -12434.670
[12,]  29801.969
[13,] -50866.898
```

c) En suivant la procédure mentionnée dans l'énoncé, on obtient les estimateurs des paramètres θ_1, θ_{12} et σ^2 suivants :

$$\hat{\theta}_1 = \frac{\hat{\gamma}(11)}{\hat{\gamma}(12)} = -0{,}5859$$
$$\hat{\theta}_{12} = \frac{\hat{\gamma}(11)}{\hat{\gamma}(1)} = -0{,}5486$$
$$\hat{\sigma}^2 = \frac{\hat{\gamma}(1)\hat{\gamma}(12)}{\hat{\gamma}(11)} = 92730.$$

Un modèle pour la série $\{\nabla\nabla_{12}X_t\}$ est donc

$$\nabla\nabla_{12}X_t = 28{,}83 + Z_t - 0{,}5859Z_{t-1} - 0{,}5486Z_{t-12},$$

où $\{Z_t\} \sim \text{WN}(0, 92730)$.

6.4 a) On sait que le processus AR(2) est stationnaire si

$$\phi_2 + \phi_1 < 1$$
$$\phi_2 - \phi_1 < 1$$
$$-1 < \phi_2 < 1.$$

Dans le présent cas où $\phi_1 = \phi$ et $\phi_2 = \phi^2$, sera stationnaire si

$$\phi^2 + \phi < 1$$
$$\phi^2 - \phi < 1$$
$$-1 < \phi^2 < 1.$$

On vérifie alors aisément que les trois inégalités sont satisfaites dès lors que

$$\frac{1-\sqrt{5}}{2} < \phi < \frac{-1+\sqrt{5}}{2}.$$

b) Il y a seulement deux paramètres à estimer dans ce modèle AR(2) spécial. Les estimateurs de Yule–Walker des paramètres ϕ et σ^2 sont les solutions de ce système d'équations :

$$\begin{aligned}\hat{\rho}(1) &= \hat{\phi} + \hat{\phi}^2\hat{\rho}(1)\\ \hat{\sigma}^2 &= \hat{\gamma}(0)(1 - \hat{\phi}\hat{\rho}(1) - \hat{\phi}^2\hat{\rho}(2)).\end{aligned}$$

Or, en utilisant les valeurs de $\hat{\gamma}(0)$, $\hat{\rho}(1)$ et $\hat{\rho}(2)$ fournies dans l'énoncé et en choisissant la solution stationnaire, on obtient

$$\begin{aligned}\hat{\phi} &= 0{,}509\\ \hat{\sigma}^2 &= 2{,}983.\end{aligned}$$

6.5 a) La FACP tombe (statistiquement) à zéro après un pas de 2, donc le processus est un AR(2).

b) Par les équations de Yule–Walker,

$$\begin{aligned}\gamma(1) &= \phi_1\gamma(0) + \phi_2\gamma(1)\\ \gamma(2) &= \phi_1\gamma(1) + \phi_2\gamma(0)\end{aligned}$$

et

$$\sigma^2 = \gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2).$$

En remplaçant les valeurs de la FACV par $\hat{\gamma}(0) = 6{,}5$, $\hat{\gamma}(1) = -5{,}94$ et $\hat{\gamma}(2) = 5{,}74$, puis en résolvant le système d'équations linéaires, on obtient : $\hat{\phi}_1 = -0{,}6479$, $\hat{\phi}_2 = 0{,}2911$ et $\hat{\sigma}^2 = 0{,}9888$.

6.6 L'autocorrélation partielle de pas 3 correspond à la valeur de ϕ_3 lorsque l'on ajuste un modèle AR(3) aux données. Cette valeur est donnée dans le troisième appel de la fonction `ar`. On a donc directement $\phi_{33} = 0{,}0993$.

6.7 On peut faire l'expérience avec les quelques lignes de codes suivantes :

```
> f <- function(ar) {
+     ar(arima.sim(n = 200, model = list(ar = ar)),
+         aic = FALSE, order.max = 2)$ar
+ }
> rowMeans(replicate(1000, f(c(0.3, 0.6))))

[1] 0.2944099 0.5640856
```

Selon ces résultats, les estimateurs de Yule–Walker sont légèrement biaisés négativement.

Chapitre 7

7.1 La prévision pour la période $n+1$ calculée avec le lissage exponentiel est $\hat{X}_n(1) = \alpha X_n + (1-\alpha)\hat{X}_{(n-1)}(1)$, pour α fixé. Ici, on doit d'abord trouver la valeur de α utilisée pour calculer les prévisions aux temps $t = 1,\ldots,5$. On a

$$49{,}27 = 48\alpha + 49{,}44(1-\alpha),$$

d'où $\alpha = 0{,}1181$. Par conséquent,

$$\begin{aligned}\hat{X}_5(1) &= 0{,}1181(39) + (1-0{,}1181)(49{,}27)\\ &= 48{,}0571.\end{aligned}$$

7.2 On réécrit tout d'abord l'équation caractéristique sous la forme

$$X_t = (\phi_1+1)X_{t-1} + (\phi_2-\phi_1)X_{t-2} - \phi_2 X_{t-3} + Z_t.$$

Par les propriétés de l'opérateur de prévision, on a alors

$$\begin{aligned}\hat{X}_n(1) &= (\phi_1+1)X_n + (\phi_2-\phi_1)X_{n-1} - \phi_2 X_{n-2}\\ \hat{X}_n(2) &= (\phi_1+1)\hat{X}_n(1) + (\phi_2-\phi_1)X_n - \phi_2 X_{n-1}\\ \hat{X}_n(3) &= (\phi_1+1)\hat{X}_n(2) + (\phi_2-\phi_1)\hat{X}_n(1) - \phi_2 X_n\end{aligned}$$

et

$$\hat{X}_n(h) = (\phi_1+1)\hat{X}_n(h-1) + (\phi_2-\phi_1)\hat{X}_n(h-2) - \phi_2\hat{X}_n(h-3)$$

pour $h > 3$.

7.3 a) On peut réécrire l'équation caractéristique comme

$$(1+0{,}6B)(1-B)^2 X_t = Z_t,$$

et donc le modèle est un ARIMA$(1,2,0)$ avec $\phi = -0{,}6$ et $\sigma^2 = 9$.

b) Le processus $W_t = \nabla^2 X_t$ satisfait les équations

$$(1+0{,}6B)W_t = Z_t,$$

d'où $\{W_t\} \sim$ AR(1). En écrivant ce processus sous la forme MA(∞), on trouve

$$W_t = \sum_{j=0}^{\infty}(-0{,}6)^j Z_{t-j}$$

et $\sum_{j=0}^{\infty}|(-0{,}6)^j| = 2{,}5 < \infty$. Par conséquent, $\{W_t\}$ est un processus linéaire, qui est toujours stationnaire par la Proposition 5.1 des notes de cours.

c) En isolant X_t dans l'équation caractéristique du processus $\{X_t\}$, on obtient

$$X_t = 1{,}4X_{t-1} + 0{,}2X_{t-2} - 0{,}6X_{t-3} + Z_t.$$

Par conséquent,

$$\hat{X}_n(1) = 1{,}4X_n + 0{,}2X_{n-1} - 0{,}6X_{n-2}$$

et

$$\hat{X}_n(2) = 1{,}4\hat{X}_n(1) + 0{,}2X_n - 0{,}6X_{n-1},$$

d'où $\hat{X}_{10}(1) = 12{,}4$ et $\hat{X}_{10}(2) = 14{,}36$. De plus, on a que $\text{Var}[X_{12} - \hat{X}_{10}(2)] = \sigma^2(1+\psi_1^2)$. Or, en posant

$$(1 - 1{,}4z - 0{,}2z^2 + 0{,}6z^3)(1 + \psi_1 z + \psi_2 z^2 + \dots) = 1,$$

on trouve $\psi_1 = 1{,}4$, d'où $\text{Var}[X_{12} - \hat{X}_{10}(2)] = 9(1 + 1{,}4^2) = 26{,}64$.

7.4 a) Le terme $(1-B)^2 \equiv \nabla^2$ indique une tendance quadratique, alors que le terme $(1 - B^{12}) \equiv \nabla_{12}$ montre que la série contient de la saisonnalité de période 12.

b) On a

$$\begin{aligned}\nu(z) &= (1 + 0{,}6z)(1-z)^2(1-z^{12}) \\ &= 1 - 1{,}4z - 0{,}2z^2 + 0{,}6z^3 - z^{12} + 1{,}4z^{13} + 0{,}2z^{14} - 0{,}6z^{15}.\end{aligned}$$

Par conséquent,

$$\begin{aligned}X_t = {} & 1{,}4X_{t-1} + 0{,}2X_{t-2} - 0{,}6X_{t-3} + X_{t-12} \\ & - 1{,}4X_{t-13} - 0{,}2X_{t-14} + 0{,}6X_{t-15} + Z_t,\end{aligned}$$

d'où

$$\begin{aligned}\hat{X}_n(1) &= 1{,}4X_n + 0{,}2X_{n-1} - 0{,}6X_{n-2} + X_{n-11} \\ &\quad - 1{,}4X_{n-12} - 0{,}2X_{n-13} + 0{,}6X_{n-14} \\ &= 4430{,}4\end{aligned}$$

et

$$\begin{aligned}\hat{X}_n(2) &= 1{,}4\hat{X}_n(1) + 0{,}2X_n - 0{,}6X_{n-1} + X_{n-10} \\ &\quad - 1{,}4X_{n-11} - 0{,}2X_{n-12} + 0{,}6X_{n-13} \\ &= 4517{,}96.\end{aligned}$$

D'autre part, on trouve à partir de $\nu(z)(1 + \psi_1 z + \psi_2 z^2 + \dots) = 1$ que le premier coefficient dans la représentation MA(∞) de $\{X_t\}$ est $\psi_1 = 1{,}4$, d'où $\text{Var}[X_{n+2} - \hat{X}_n(2)] = \sigma^2(1+\psi_1^2) = 14{,}8$. Un intervalle de prévision à 95 % pour la période $n+2$ est donc

$$4517.96 \pm 1{,}96\sqrt{14{,}8} \Leftrightarrow (4510{,}42,\ 4525{,}5).$$

Bibliographie

Abraham, B. et J. Ledolter. 1983, *Statistical Methods for Forecasting*, Wiley, New York, ISBN 0-4718676-4-0.

Brockwell, P. J. et R. A. Davis. 1996, *Introduction to Time Series and Forecasting*, Springer, New York, ISBN 0-3879471-9-1.

Draper, N. R. et H. Smith. 1998, *Applied Regression Analysis*, 3e éd., Wiley, New York, ISBN 0-4711708-2-8.

Goulet, V. 2007, *Introduction à la programmation en S*, 2e éd., Document libre publié sous contrat GNU FDL, ISBN 978-2-9809136-7-9. URL http://vgoulet.act.ulaval.ca/intro_S.

Miller, R. B. et D. W. Wichern. 1977, *Intermediate Business Statistics*, HRW, Orlando, FL, ISBN 0-0308910-1-9.

Venables, W. N. et B. D. Ripley. 2002, *Modern Applied Statistics with S*, 4e éd., Springer, New York, ISBN 0-3879545-7-0.

Venables, W. N., D. M. Smith et the R Development Core Team. 2005, *An Introduction to R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

Index

Cet index contient des entrées pour les annexes A et B seulement. Les numéros de page en caractères gras indiquent les pages où les concepts sont introduits, définis ou expliqués.

ISBN 978-2-9811416-0-6