# Modèles linéaires en actuariat

## Exercices et solutions

# Modèles linéaires en actuariat

## Exercices et solutions

**Marie-Pier Côté**

**Vincent Mercier**

**École d'actuariat, Université Laval**

**Seconde édition**

**Historique de publication**
Septembre 2019 : Première édition

**Code source**
Le code source LaTeX de la première édition de ce document est disponible en communiquant directement avec les auteurs.

# Introduction

Ce document contient les exercices proposés par Marie-Pier Côté pour le cours ACT-2003 Modèles linéaires en actuariat, donné à l'École d'actuariat de l'Université Laval. Certains exercices sont le fruit de l'imagination des auteurs ou de ceux des versions précédentes, 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-Partage des conditions initiales à l'identique 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.

Le document est séparé en deux parties correspondant aux deux sujets faisant l'objet d'exercices : d'abord la régression linéaire (simple, multiple et régularisée), puis les modèles linéaires généralisés.

L'estimation des paramètres, le calcul de prévisions et l'analyse des résultats 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 logiciel statistique R. D'ailleurs, l'annexe **??** présente les principales fonctions de R pour la régression.

Le format de cet annexe est inspiré de **?** ] : 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 **??**.

L'annexe **??** 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 A.

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

???? à régler

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.

Marie-Pier Côté <marie-pier.cote@act.ulaval.ca>
Vincent Mercier <vincent.mercier.7@ulaval.ca>
Québec, septembre 2019

# 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 $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, $\text{var}[\varepsilon_t] = \sigma^2$.

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 $\beta_0$ et $\beta_1$. *Note* : $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$.

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

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 $\beta$. En termes mathématiques : soit

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

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

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

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 $\beta_0$ et $\beta_1$ 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 $\beta$ 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 $\beta_1$ si la variance $\sigma^2$ 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 $\varepsilon$ 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 $\beta_1$.

**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 :

$$\sum_{t=1}^{50} X_t = 3333 \qquad \sum_{t=1}^{50} X_t^2 = 231933 \qquad \sum_{t=1}^{50} X_t Y_t = 30549{,}75$$
$$\sum_{t=1}^{50} Y_t = 459{,}9 \qquad \sum_{t=1}^{50} Y_t^2 = 4308{,}57.$$

**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) $10{,}57 \pm 2{,}13$

# A Solutions

## Chapitre 2

**2.1** a) Voir la figure A.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 $\beta_0$ et $\beta_1$ 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 $\hat{\beta}_0$,

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

De la seconde équation normale, on obtient

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

```r
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)
```

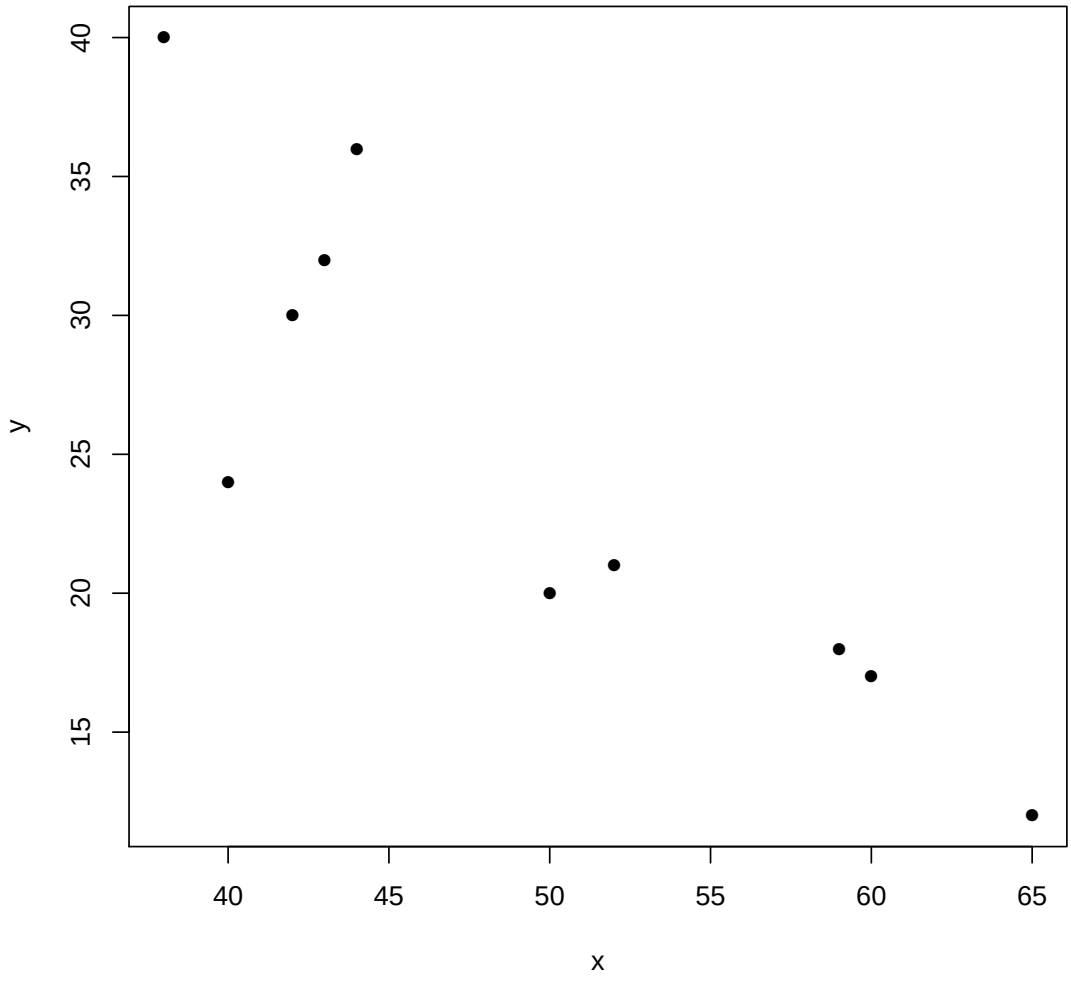


FIG. A.1 – Relation entre les données de l'exercice 2.1

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        6
## 11.80028 30.29670 29.45596 16.84476 16.00401 24.41148
##        7        8        9       10
## 22.72998 34.50044 31.13745 32.81894
```

Pour ajouter la droite de régression au graphique de la figure A.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 A.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          5
##  0.1997243  1.7032953  6.5440421  1.1552437  0.9959905
##          6          7          8          9         10
## -4.4114773 -1.7299837  5.4995615 -1.1374514 -8.8189450
```

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

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

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

```
abline(fit)
```

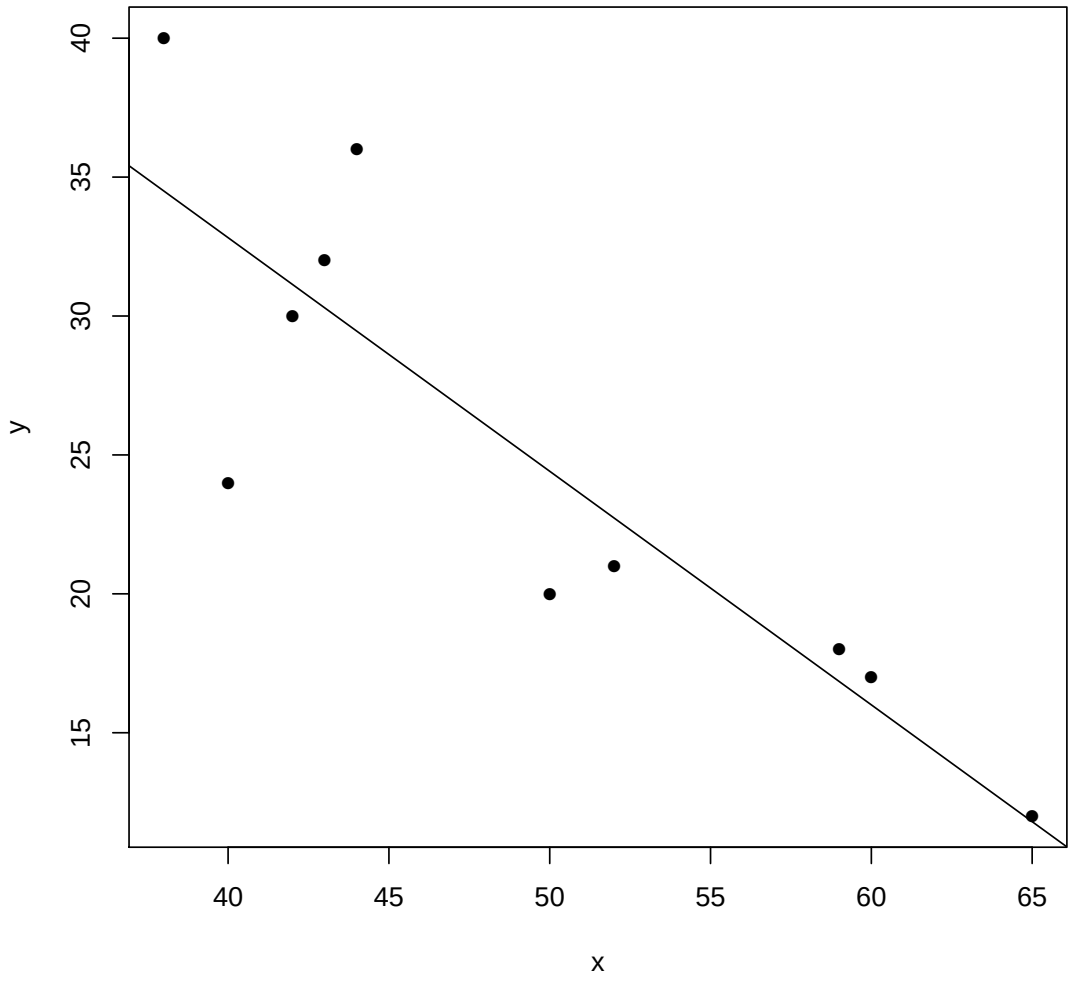


FIG. A.2 – Relation entre les données de l'exercice 2.1 et la droite de régression

**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 A.3.

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

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

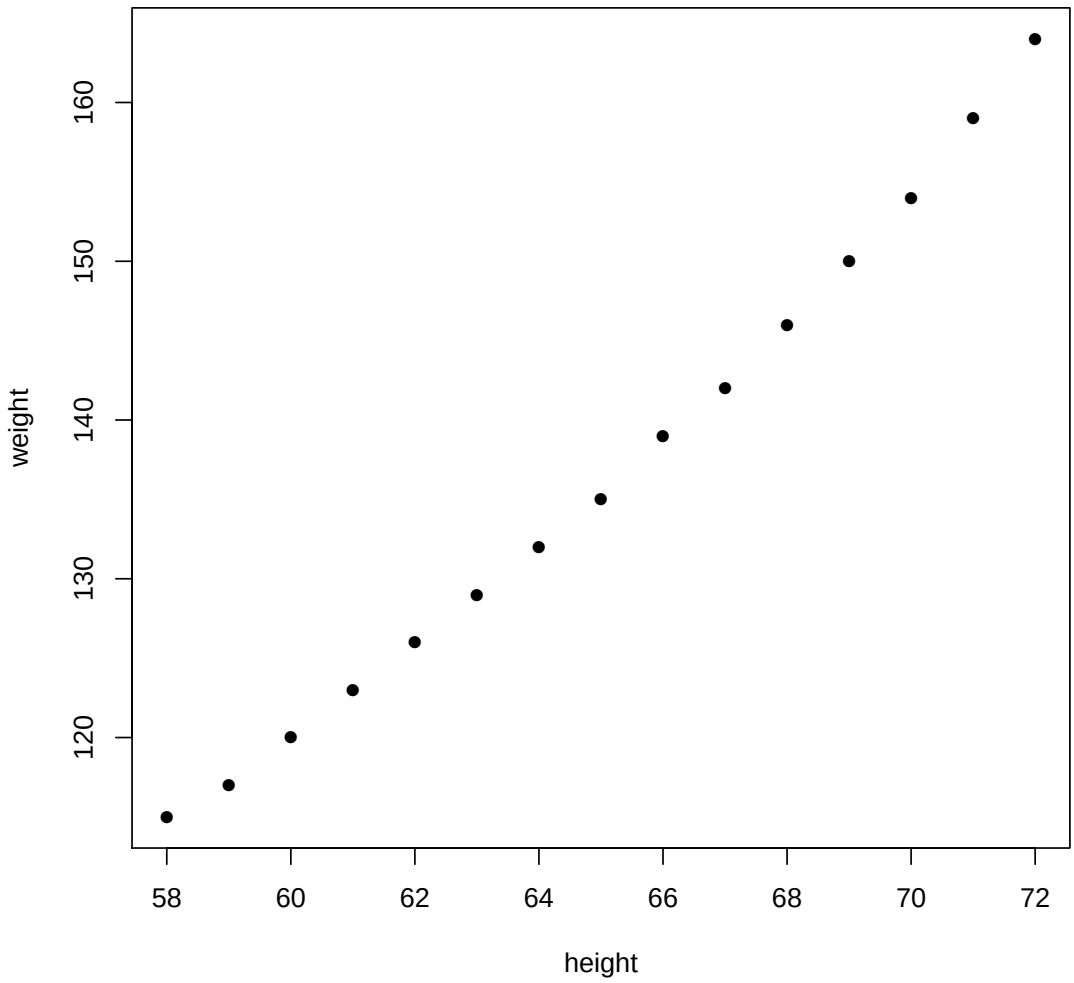


FIG. A.3 – Relation entre la taille et le poids moyen de femmes américaines âgées de 30 à 39 ans (données `women`)

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

##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Coefficients:
## (Intercept)       height
##      -87.52         3.45
```

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

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

```
summary(fit)

##
## 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 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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 $e_t = Y_t - \hat{Y}_t = (Y_t - \bar{Y}) - \hat{\beta}_1 (X_t - \bar{X})$,

```
abline(fit)
```

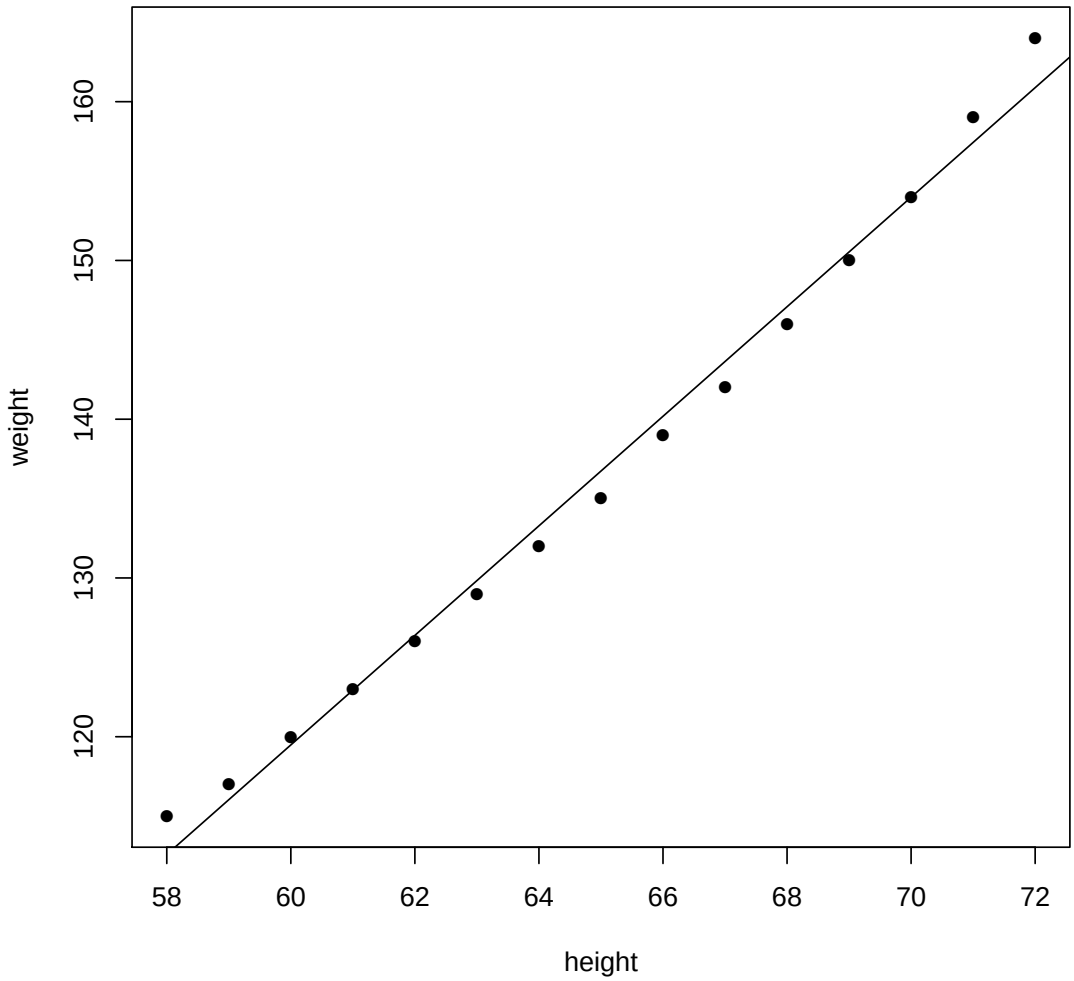


FIG. A.4 – Relation entre les données `women` et droite de régression linéaire simple

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 $\beta_0$ et $\beta_1$ 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 $\beta$ 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

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

L'estimateur des moindres carrés de $\beta$ est donc

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

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

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

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

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

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 $\lambda$. On trouve alors

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

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

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

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

b) On a

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

**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,

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

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 A.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 $\beta_1$ lorsque la variance $\sigma^2$ 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 $\beta_1$ est

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

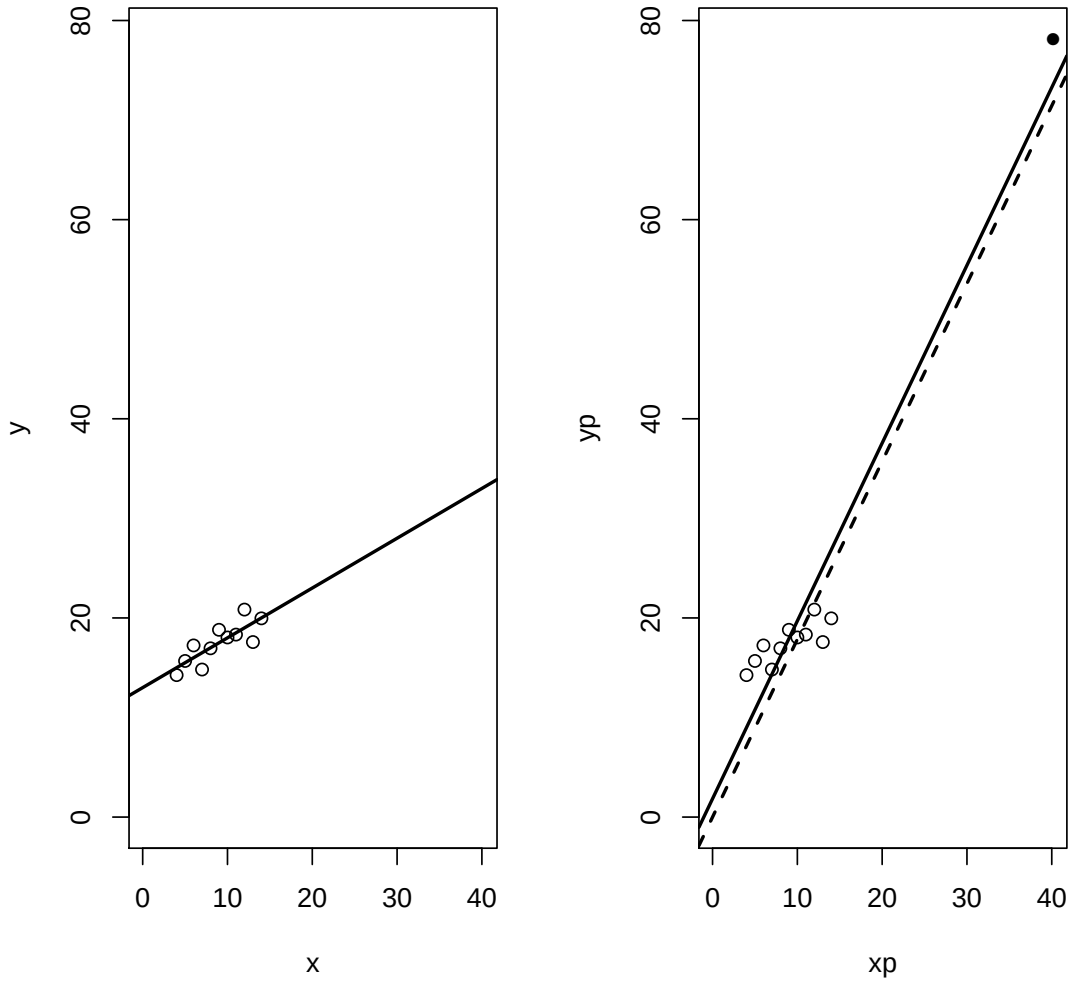


FIG. A.5 – Illustration de l'effet de l'ajout d'un point spécial à un ensemble de données. À gauche, la droite de régression usuelle. À droite, le même ensemble de points avec le point spécial ajouté (cercle plein), la droite de régression avec ce nouveau point (ligne pleine) et la droite de régression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

On nous donne 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 :

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

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 SSE $=$ SST $-$ 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 $\sigma^2$ 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 $\sigma^2$ 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 $\text{SSE} = \text{SST} - \text{SSR} = 1{,}104$ 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 $\beta_1$ 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

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

```
par(mfrow = c(2, 2))
plot(medv ~ rm + age + lstat + tax, data = house, ask = FALSE)
```

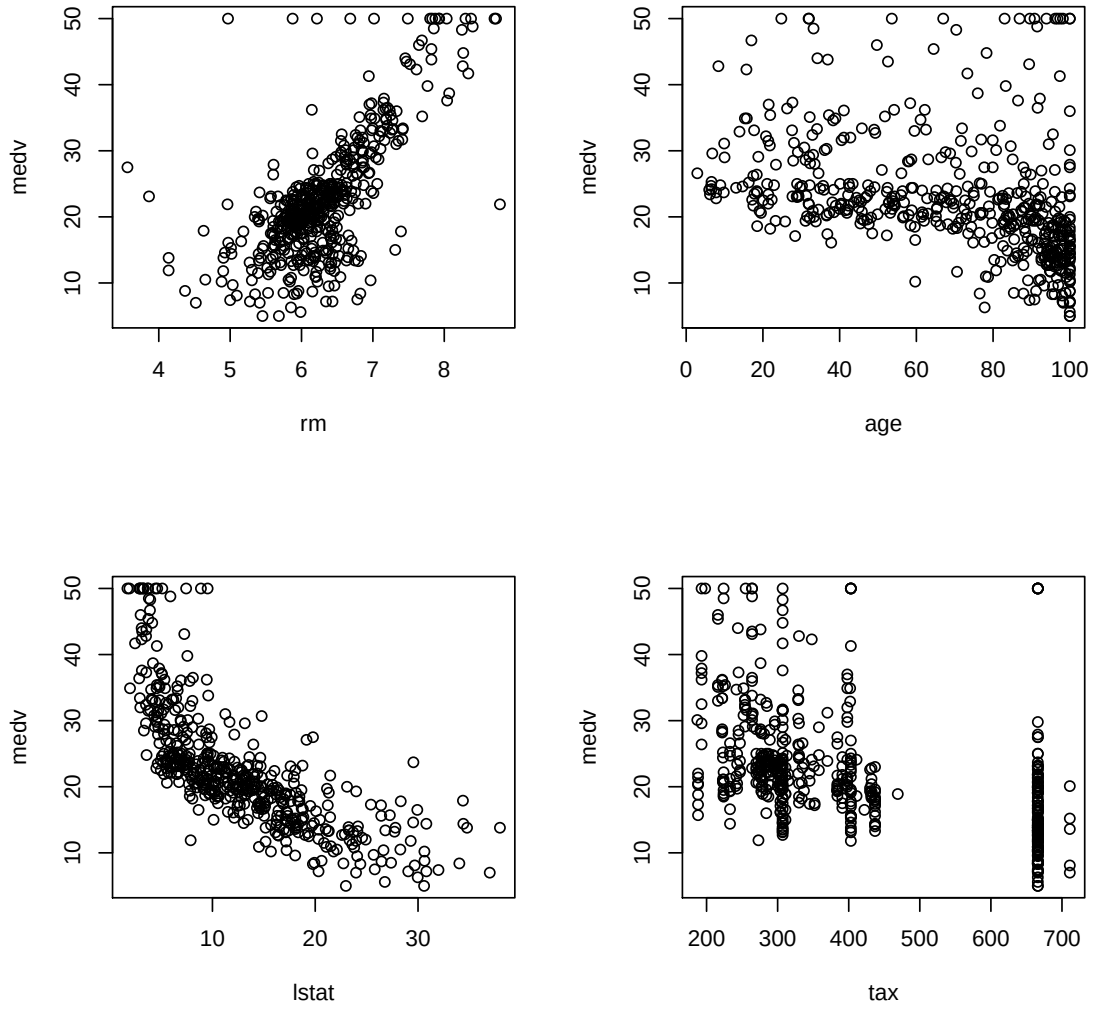


FIG. A.6 – Relation entre la variable `medv` et les variables `rm`, `age`, `lstat` et `tax` des données `house.dat`

**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 A.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
## -23.346  -2.547   0.090   2.986  39.433
##
## 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 ***
## ---
## 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 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 A.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)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  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 ***
## ---
```

```
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)
```

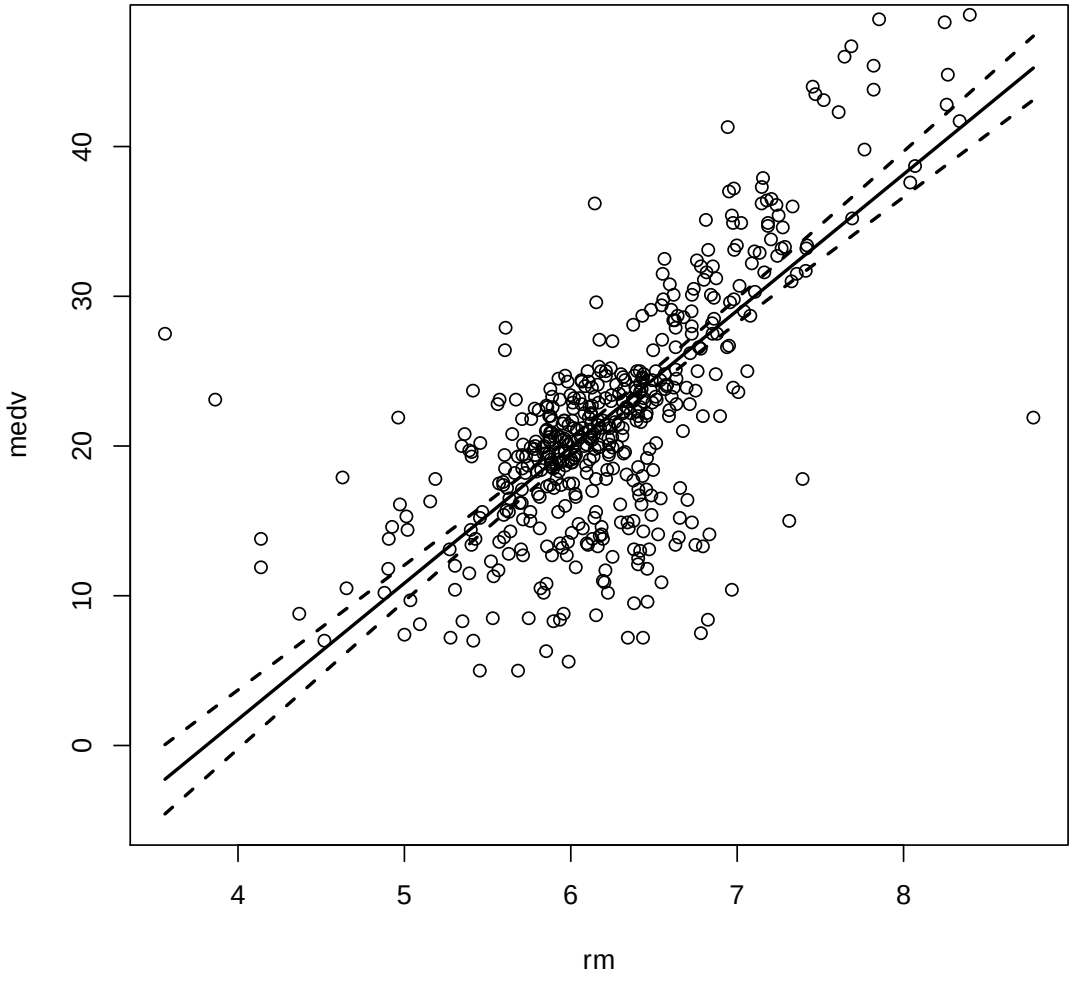


FIG. A.7 – Résultat de la régression de la variable `rm` sur la variable `medv` des données `house.dat`

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## 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 A.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 ***
## ---
## 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.

```
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)
```

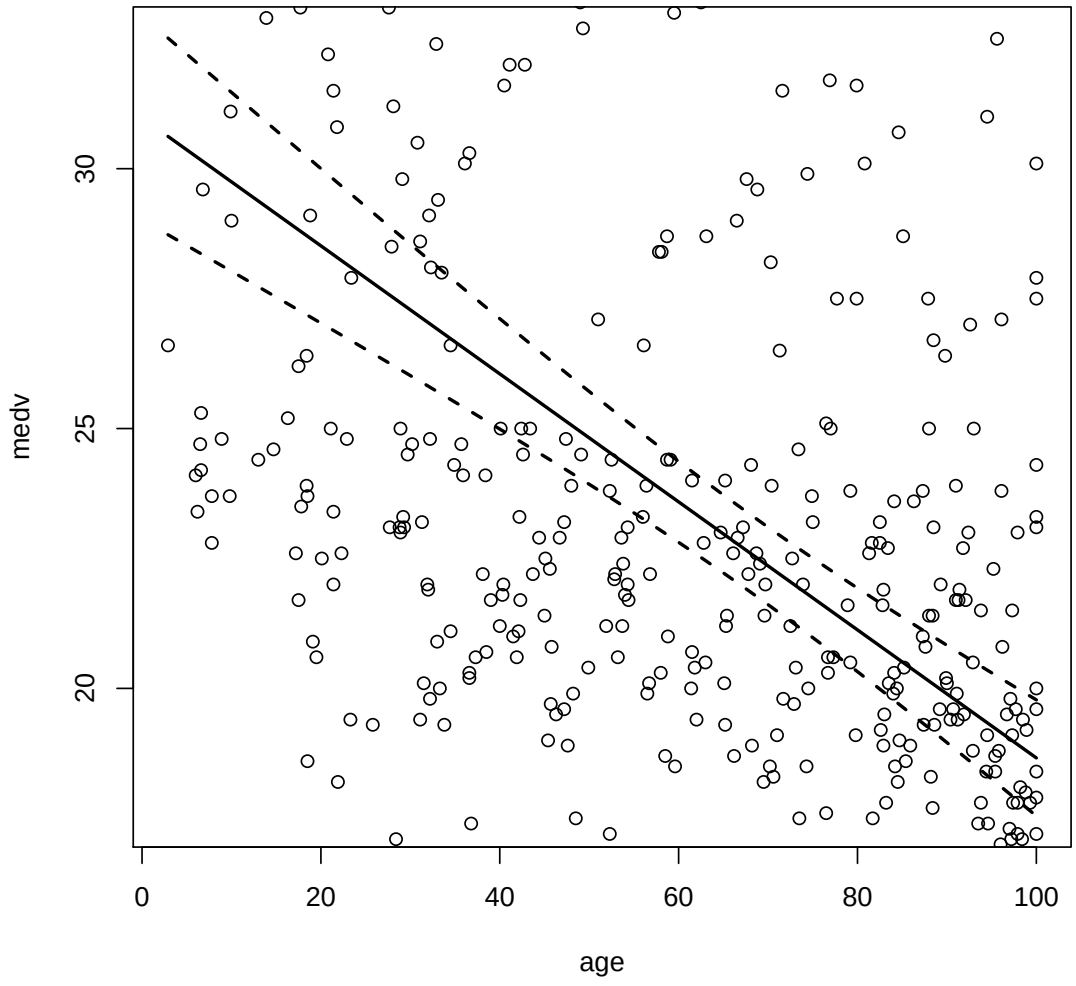


FIG. A.8 – Résultat de la régression de la variable `age` sur la variable `medv` des données `house.dat`

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 10.5669 8.432089 12.7017
```