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. Le recueil a été mis en forme par Vincent Mercier, auxiliaire d'enseignement, à l'aide du soutien financier de la Chaire de leadership en enseignement en analyse de données massives pour l'actuariat — Intact.

Plusieurs des exercices de la première partie proviennent d'une ancienne version du recueil, rédigée par Vincent Goulet, professeur 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 du cours : 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 requièrent l'utilisation du logiciel statistique R.

L'annexe A rappelle quelques concepts de statistique de base et contient les tables de la loi khi-carrée et Student, nécessaires pour quelques exercices. L'annexe B détaille la notation et les propriétés de la loi normale multivariée. 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 sur le site de cours ainsi qu'à l'adresse `https://github.com/mpcot24/ACT2003-exercices/tree/master/exercices_modeles_lineaires/data`

Ces jeux de données sont importés dans R avec une commande `scan`, `read.table` ou `read.csv`. 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 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$:

i	1	2	3	4	5	6	7	8	9	10
x_i	65	43	44	59	60	50	52	38	42	40
Y_i	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 $\hat{Y}_i$ correspondant à x_i pour $i = 1, \ldots, n$. Ajouter la droite de régression au graphique fait en a).

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

2.2 On vous donne les observations ci-dessous.

t	x_i	Y_i
1	2	6
2	3	4
3	5	6
4	7	3
5	4	6
6	4	4
7	1	7
8	6	4

$$\sum_{i=1}^{8} x_i = 32 \qquad \sum_{i=1}^{8} x_i^2 = 156$$

$$\sum_{i=1}^{8} Y_i = 40 \qquad \sum_{i=1}^{8} Y_i^2 = 214$$

$$\sum_{i=1}^{8} x_i Y_i = 146$$

a) Calculer les coefficients de la régression $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\text{var}(\varepsilon_i) = \sigma^2$.

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

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\varepsilon_i = 0.$$

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 $Y_i = \beta x_i + \varepsilon_i$, $i = 1,\ldots,n$, $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = \delta_{ij}\sigma^2$.

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

$$\beta^* = \sum_{i=1}^{n} c_i Y_i$$

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

$$\mathrm{var}(\beta^*) = \mathrm{var}\left(\sum_{i=1}^{n} c_i Y_i\right)$$

sous la contrainte que

$$\mathrm{E}[\beta^*] = \mathrm{E}\left[\sum_{i=1}^{n} c_i Y_i\right] = \beta,$$

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

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

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

b) $\mathrm{E}[\mathrm{MSR}] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$

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_i devient $Y_i' = a + bY_i$ et que x_i devient $x_i' = c + dx_i$, $i = 1,\ldots,n$.

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 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

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 $Y_i = \beta x_i + \varepsilon_i$, l'estimateur des moindres carrés de β est

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

Démontrer que l'on peut obtenir ce résultat en utilisant la formule pour $\hat{\beta}_1$ dans la régression linéaire simple usuelle ($Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) 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

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

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 :

$$\hat{\beta}_0 = 138{,}581 \qquad \sum_{i=1}^{20}(x_i - \bar{x})^2 = 10\,668$$

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

$$\sum_{i=1}^{20} x_i^2 = 90\,048 \qquad \sum_{i=1}^{20} \varepsilon_i^2 = 7\,832.$$

$$\sum_{i=1}^{20} Y_i^2 = 116\,058$$

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 :

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

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

ii) $\text{cov}(\varepsilon_0, \varepsilon_i) = 0$ pour tout $i = 1, \ldots, n$;

iii) $\text{var}(\varepsilon_i) = \sigma^2$ 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.

2.19 On s'intéresse à l'impact du sexe sur l'espérance de vie. On connaît les durées de vie de $n_F = 300$ femmes et $n_H = 200$ hommes. On choisit d'utiliser la variable indicatrice

$$x_i = \begin{cases} 0 & \text{, si } \texttt{SEXE}_i = \texttt{H} \\ 1 & \text{, si } \texttt{SEXE}_i = \texttt{F} \end{cases}.$$

On note $\bar{Y}_F$ la moyenne des durées de vie des femmes et $\bar{Y}_H$ la moyenne des durées de vie des hommes.

a) Montrer que l'estimateur des moindres carrés $\hat{\beta}_1$ (lié à la variable explicative x) est égal à $\bar{Y}_F - \bar{Y}_H$. *Indice : On peut exprimer $\bar{Y}$ en termes de $\bar{Y}_F$ et $\bar{Y}_H$.*

b) Ce résultat permet-il d'interpréter le coefficient relié à une variable catégorique binaire ? Expliquer.

c) Que représente $\hat{\beta}_0$ dans ce cas ?

2.20 On s'intéresse à la covariance entre deux résidus.

a) D'abord, trouver $\text{cov}(Y_i, \hat{Y}_j)$.

b) Puis, calculer $\text{cov}(\hat{Y}_i, \hat{Y}_j)$.

c) Déduire de a) et b) que

$$\text{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right).$$

2.21 Dans un graphiques des résidus en fonction des valeurs prédites, on observe de l'hétéroscédasticité. Après une analyse plus poussée, on note que la variance de $\hat{\varepsilon}_i$ est approximativement proportionnelle à $E[Y_i]^4$. Proposer une transformation g de la variable réponse qui permettra de stabiliser la variance.

2.22 Les données suivantes présentent le nombre moyen de bactéries vivantes dans une boîte de conserve de nourriture et le temps (en minutes) d'exposition à une chaleur de 300^oF. [1]

Nombre de bactéries	Temps d'exposition (min)
175	1
108	2
95	3
82	4
71	5
50	6
49	7
31	8
28	9
17	10
16	11
11	12

a) Tracer un nuage de points des données. Est-ce qu'un modèle de régression linéaire semble adéquat ?

b) Ajuster aux données un modèle de régression linéaire. Calculer les statistiques sommaires et produire les graphiques de résidus. Interpréter les résultats. Quelles sont vos conclusions par rapport à la validité du modèle de régression ?

c) Identifier une transformation pour ces données afin d'utiliser adéquatement les méthodes de régression. Ajuster ce nouveau modèle et tester la validité de la régression.

1. Source : D. Montgomery, E.A. Peck et G.G. Vining (2012). Introduction to Linear Regression Analysis. Fifth Edition. Wiley.

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) $\hat{\beta} = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2$ c) $\text{var}(\hat{\beta}) = \sigma^2 / \sum_{i=1}^{n} x_i^2$

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

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

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 $\hat{Y}_0 \pm t_{\alpha/2}(n-1)\, s \sqrt{1 + x_0^2 / \sum_{i=1}^{n} x_i^2}$

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

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\boldsymbol{\beta})^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

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,

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$

Déduire l'estimateur des moindres carrés de ces équations. *Astuce* : utiliser le théorème **Dérivée d'une fonction** de la section C.3.

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) $Y_i = \beta_0 + \varepsilon_i$

b) $Y_i = \beta_1 x_i + \varepsilon_i$

c) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

3.3 Vérifier, pour le modèle de régression linéaire simple, que les valeurs trouvées dans la matrice de variance-covariance $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ 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 $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$.

a) $\mathbf{X}^\top \boldsymbol{\varepsilon} = 0$

b) $\hat{\mathbf{Y}}^\top \boldsymbol{\varepsilon} = 0$

c) $\hat{\mathbf{Y}}^\top \hat{\mathbf{Y}} = \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{Y}$

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 colonne $\mathbf{x}_0^\top = (1, x_{01}, \ldots, x_{0p})$ des p variables indépendantes. On a donc

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

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

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

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

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

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

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

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

3.6 En ajustant le modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

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

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1,2,3,4,$$

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

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

et

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

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 $\mathrm{E}[\varepsilon_t] = 0$ 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 $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ 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 $(\mathbf{X}^\top\mathbf{X})^{-1}$ 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

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{t=1}^{n} w_t(\mathbf{Y}_t - \mathbf{X}_t^\top\boldsymbol{\beta})^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

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,

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

puis en déduire cet estimateur. *Astuce* : cette preuve est simple si l'on utilise le théorème **Dérivée d'une fonction** de la section C.3 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 $Y_i = \beta x_i + \varepsilon_i$. Trouver l'estimateur linéaire sans biais à variance minimale du paramètre β, ainsi que sa variance, sous chacune des hypothèses suivantes.

a) $\text{var}(\varepsilon_i) = \sigma^2$

b) $\text{var}(\varepsilon_i) = \sigma^2/w_i$

c) $\text{var}(\varepsilon_i) = \sigma^2 x_i$

d) $\text{var}(\varepsilon_i) = \sigma^2 x_i^2$

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.

3.19 Le modèle de régression linéaire multiple

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \text{ pour } i = 1,...,n$$

a été ajusté à des données avec la méthode des moindres carrés.

a) La figure 3.1 montre le QQ-plot des résidus studentisés. À la lumière de ce graphique, y a-t-il un postulat du modèle qui n'est pas vérifié ? Si oui, lequel et pourquoi ? S'il y a lieu, expliquer l'impact de la violation de ce postulat.

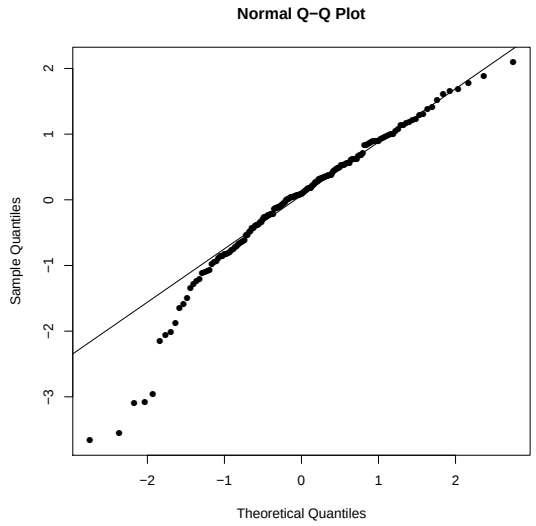

FIG. 3.1 – QQ-Plot des résidus studentisés

b) La figure 3.2 montre les résidus studentisés en fonction de chacune des variables exogènes et en fonction des valeurs prédites. Utiliser ces graphiques pour commenter sur la validité des postulats du modèle. Y en a-t-il qui ne sont pas respectés ? S'il y a lieu, expliquer l'impact de la violation de ce ou ces postulats.

3.20 La base de données `OutlierExample.csv` disponible sur le site du cours contient 19 observations de base, et trois observations supplémentaires, notées par les `CODES` 1, 2 et 3, qui sont aberrantes ou influentes.

a) Importez la base de données et tracez un nuage de points de `Y` en fonction de `x`.

b) Roulez les lignes de code suivantes pour observer le graphique avec les 3 points ajoutés

```
library(ggplot2)
ggplot(dat, aes(x= X, y= Y, label=CODES))+
  geom_point() +
  geom_text(aes(label=ifelse(CODES>0,CODES,'')),hjust=0,vjust=0)
```

c) Ajustez un modèle linéaire en incluant seulement les 19 points dont le code est 0. Regardez l'ajustement et commentez.

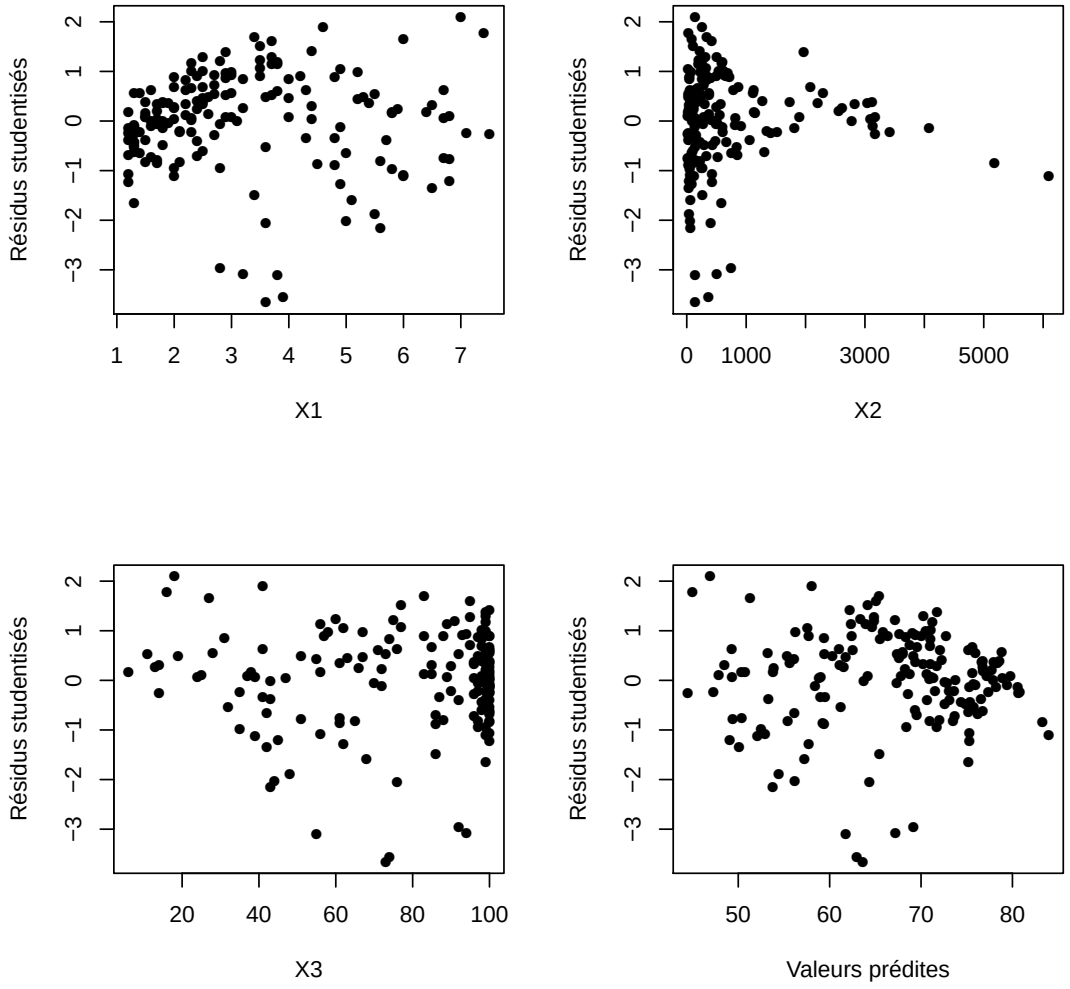

FIG. 3.2 – Nuage de points des résidus studentisés en fonction de chacune des variables exogènes et en fonction de la variable prédite

d) Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 1. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.

e) Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 2. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.

f) Ajustez un modèle linéaire en incluant les 19 points dont le code est 0 et le point de code 3. Quel est l'impact de l'inclusion de ce point sur le R^2 et sur les estimations des paramètres ? Étudiez le résultat de la fonction `influence.measures()`.

Réponses

3.2 a) $\hat{\beta}_0 = \bar{Y}$ b) $\hat{\beta}_1 = (\sum_{i=1}^n x_i Y_i)/(\sum_{i=1}^n x_i^2)$

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) $\hat{\beta}^* = \sum_{i=1}^n x_i Y_i / \sum_{i=1}^n x_i^2$, $\text{var}(\hat{\beta}^*) = \sigma^2 / \sum_{i=1}^n x_i^2$

b) $\hat{\beta}^* = \sum_{i=1}^n w_i x_i Y_i / \sum_{i=1}^n w_i x_i^2$, $\text{var}(\hat{\beta}^*) = \sigma^2 / \sum_{i=1}^n w_i x_i^2$

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

d) $\hat{\beta}^* = \sum_{i=1}^n Y_i/x_i$, $\text{var}(\hat{\beta}^*) = \sigma^2/n$

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

Modèles linéaires généralisés

Troisième partie

Annexes

A Révision de certains concepts de statistique et tables

A.1 Quelques distributions bien connues

Loi Normale : Si $Y \sim \mathcal{N}(\mu,\sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma^2 > 0$, alors sa densité est donnée, pour tout $y \in \mathbb{R}$, par

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

Dans ce cas, $Z = (Y-\mu)/\sigma$ suit une loi normale centrée réduite, notée $\mathcal{N}(0,1)$ ou $N(0,1)$.

Loi Khi-Carrée : Si $X \sim \chi^2_{(\nu)}$ avec $\nu > 0$, sa densité est donnée, pour tout $x \in (0,\infty)$, par

$$f_X(x) = \frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}.$$

On note que $\mathrm{E}[X] = \nu$. Les distributions normale et Khi-carrée sont reliées :

- Si $Z \sim \mathbb{N}(0,1)$, alors $Z^2 \sim \chi^2_{(1)}$.
- Si $Z_1,\ldots,Z_n$ sont mutuellement indépendantes et distribuées selon une loi $\mathcal{N}(0,1)$, alors

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}.$$

Loi Student t : Si $T \sim t_{(\nu)}$ avec $\nu > 0$, sa densité est donnée, pour tout $t \in \mathbb{R}$, par

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Quand $\nu \to \infty$, la loi t tend vers la loi normale centrée réduite. Aussi, si $Z \sim \mathbb{N}(0,1)$ et $X \sim \chi^2_{(\nu)}$ sont indépendantes, alors on a la représentation stochastique suivante :

$$\frac{Z}{\sqrt{X/\nu}} \sim t_{(\nu)}.$$

Loi de Fisher : Si $X_1 \sim \chi^2_{(\nu_1)}$ et $X_2 \sim \chi^2_{(\nu_2)}$ sont indépendantes et $\nu_1,\nu_2 > 0$, alors

$$\frac{X_1/\nu_1}{X_2/\nu_2} \sim F(\nu_1,\nu_2).$$

Aussi, on peut facilement montrer avec les relations précédentes que si $T \sim t_{(\nu)}$, alors $T^2 \sim F(1,\nu)$. La densité de la loi de Fisher est complexe et rarement utilisée. Cette distribution a un support positif.

A.2 Maximum de vraisemblance

Soient les observations $y_1,\ldots,y_n$, provenant de variables aléatoires indépendantes avec densités $f_{Y_i}(y_i;\theta)$, pour $i = 1,\ldots,n$. La fonction de vraisemblance est

$$L(\theta;y_1,\ldots,y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta).$$

La fonction de log-vraisemblance est

$$\ell(\theta;y_1,\ldots,y_n) = \ln L(\theta;y_1,\ldots,y_n) = \sum_{i=1}^{n} \ln f_{Y_i}(y_i;\theta).$$

La méthode du maximum de vraisemblance permet de trouver l'estimateur $\hat{\theta}_n$ qui maximise $L(\theta;y_1,\ldots,y_n)$, et par conséquent $\ell(\theta;y_1,\ldots,y_n)$. On travaille habituellement avec le logarithme naturel pour simplifier les calculs. La fonction de score est

$$\dot{\ell}_\theta(\theta;y_1,\ldots,y_n) = \frac{\partial}{\partial\theta}\ell(\theta;y_1,\ldots,y_n).$$

L'estimateur du maximum de vraisemblance (EMV, ou MLE pour *maximum likelihood estimator*) est $\hat{\theta}_n$ tel que

$$\dot{\ell}_\theta(\theta;y_1,\ldots,y_n)\big|_{\hat{\theta}_n} = 0.$$

A.3 Estimateur sans biais

Soit un échantillon aléatoire $Y_1,\ldots,Y_n$, avec densité $f(y;\theta)$. Soit l'estimateur de θ suivant : $\hat{\theta}_n = \hat{\theta}(Y_1,\ldots,Y_n)$. On dit de $\hat{\theta}_n$ qu'il est sans biais si $\mathrm{E}[\hat{\theta}_n] = \theta$. Dans ce cas, le biais est zéro,

$$\mathrm{biais}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n] - \theta = 0,$$

ce qui réduit l'erreur quadratique moyenne (EQM ou *MSE* pour *mean squared error*) de l'estimateur :

$$\mathrm{EQM}(\hat{\theta}_n) = \mathrm{E}[(\hat{\theta}_n - \theta)^2] = \mathrm{var}(\hat{\theta}_n) + \mathrm{biais}(\hat{\theta}_n).$$

A.4 Table de quantiles de la loi khi carré

Le tableau donne $\chi^2_\alpha(\nu)$, le quantile supérieur de niveau α de la loi khi carré avec ν degrés de liberté, α est donné dans les colonnes, et ν est donné dans les lignes. Précision : Si $X \sim \chi^2(\nu)$, alors $\Pr\{X > \chi^2_\alpha(\nu)\} = \alpha$.

	Queue de gauche					Queue de droite					
ν	0.99500	0.99000	0.97500	0.95000	0.90000	0.10000	0.05000	0.02500	0.01000	0.00500
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.70654	22.16426	24.43304	26.50930	29.05052	51.80506	55.75848	59.34171	63.69074	66.76596
50	27.99075	29.70668	32.35736	34.76425	37.68865	63.16712	67.50481	71.42020	76.15389	79.48998
60	35.53449	37.48485	40.48175	43.18796	46.45889	74.39701	79.08194	83.29767	88.37942	91.95170
70	43.27518	45.44172	48.75756	51.73928	55.32894	85.52704	90.53123	95.02318	100.42518	104.21490
80	51.17193	53.54008	57.15317	60.39148	64.27784	96.57820	101.87947	106.62857	112.32879	116.32106
90	59.19630	61.75408	65.64662	69.12603	73.29109	107.56501	113.14527	118.13589	124.11632	128.29894
100	67.32756	70.06489	74.22193	77.92947	82.35814	118.49800	124.34211	129.56120	135.80672	140.16949

A.5 Table de quantiles de la loi t

Le tableau donne $t_{\nu,\alpha}$, le quantile supérieur de niveau α de la loi de Student avec ν degrés de liberté, α est donné dans les colonnes, ν est donné dans les lignes. Précision : Si $T \sim t_{(\nu)}$, alors $\Pr\{T > t_{\nu,\alpha}\} = \alpha$.

	α				
ν	0.100	0.050	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

B La loi normale multivariée

En dimension $p = 1$, on considère d'abord le cas d'une variable aléatoire Z distribuée selon une loi normale centrée réduite $Z \sim \mathcal{N}(0,1)$. La densité de Z s'écrit alors, pour tout $z \in \mathbb{R}$, comme

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2).$$

Une variable X de distribution normale avec moyenne μ et variance σ^2, noté $X \sim \mathcal{N}(\mu,\sigma^2)$, peut être écrite selon la représentation stochastique $X = \sigma Z + \mu$. La densité de X est donnée, pour $x \in \mathbb{R}$, par

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}.$$

On considère maintenant la généralisation en dimension $p \geq 1$ en supposant $(Z_1,\ldots,Z_p)$ des variables aléatoires indépendantes et identiquement distribuées suivant une loi normale centrée réduite $\mathbb{N}(0,1)$. On a alors que

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p),$$

où p désigne le nombre de variables aléatoires, $\mathbf{0}$ est un vecteur de p zéros et $\mathbf{I}_p$ note la matrice identité de dimension $p \times p$. La densité conjointe de $\mathbf{Z} = (Z_1,\ldots,Z_p)^\top$, peut alors s'écrire comme

$$f(z_1,\ldots,z_p) = \prod_{i=1}^{p} f(z_i) = \left(\frac{1}{\sqrt{2\pi}} \right)^p \exp\left(-\sum_{i=1}^{p} z_i^2/2 \right) = \left(\frac{1}{\sqrt{2\pi}} \right)^p \exp\left(-\frac{1}{2} \mathbf{z}^\top \mathbf{z} \right),$$

où $\mathbf{z} = (z_1,\ldots,z_p)^\top \in \mathbb{R}^p$.

Soient $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)^\top \in \mathbb{R}^p$ un vecteur de p moyennes et $\boldsymbol{\Sigma}$ une matrice positive définie, de variance-covariance. En considérant la transformation de variables $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{(1/2)}\mathbf{Z}$, on obtient que

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

La densité conjointe de $(X_1,\ldots,X_p)$ est alors donnée par

$$\begin{aligned} f(x_1,\ldots,x_p) &= \prod_{i=1}^{p} f(x_i) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1/2})^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{B.1}$$

où $\boldsymbol{\mu}$ désigne le vecteur contenant la moyenne des p variables, $\boldsymbol{\Sigma}$ dénote la matrice des variances-covariances de $\mathbf{X}$ (de dimension $p \times p$), p est le nombre de variables. On a dans ce cas que le vecteur $\mathbf{Z} = \boldsymbol{\Sigma}^{(-1/2)}(\mathbf{X}-\boldsymbol{\mu})$ suit une normale multivariée centrée réduite.

B.1 Espérance et variance

On détermine ci-dessous l'espérance et la variance de $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. On a

$$\mathrm{E}[\mathbf{X}] = \mathrm{E}[\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}] = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathrm{E}[\mathbf{Z}] = \boldsymbol{\mu},$$

et

$$\begin{aligned} \mathrm{var}(\mathbf{X}) &= \mathrm{var}(\boldsymbol{\Sigma}^{1/2}\mathbf{Z}) \\ &= (\boldsymbol{\Sigma}^{1/2})^\top \mathrm{var}(\mathbf{Z}) \boldsymbol{\Sigma}^{1/2} \\ &= (\boldsymbol{\Sigma}^{1/2})^\top \mathbf{I}_p \boldsymbol{\Sigma}^{1/2} \\ &= \boldsymbol{\Sigma}. \end{aligned}$$

B.2 La matrice de variance-covariance

La matrice de variance-covariance s'écrit comme

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_p) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2,X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(X_p,X_1) & \mathrm{cov}(X_p,X_2) & \cdots & \mathrm{var}(X_p) \end{pmatrix}$$

Dans le cas centré-réduit, cette matrice peut s'écrire comme

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{Z}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I}_p.$$

On note $\sigma_{i,j}$ la valeur correspondant à $\mathrm{cov}(X_i,X_j)$. La matrice de variances-covariances possède les deux propriétés suivantes :

1. Elle est symétrique : $\sigma_{i,j} = \sigma_{j,i}$

2. Elle est semi-définie positive :

$$\forall \mathbf{a} \in \mathbb{R}^p, \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} \geq 0$$
$$\text{var}(\mathbf{a}^\top \mathbf{\Sigma}) \geq 0$$

Elle possède donc p valeurs propres.

Donc si

- $\mathbf{\Sigma}$ est la matrice de variances-covariances
- $\mathbf{\Lambda}$ est la matrice diagonale des p valeurs propres de $\mathbf{\Sigma}$
- et $\mathbf{P}$ est la matrice dont les colonnes sont les p vecteurs propres de $\mathbf{\Sigma}$,

alors $\mathbf{\Sigma} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$ et

$$|\mathbf{\Sigma}| = \left|\mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top\right| = |\mathbf{P}|\,|\mathbf{\Lambda}|\left|\mathbf{P}^\top\right| = \left|\mathbf{P}\mathbf{P}^\top\right||\mathbf{\Lambda}| = \prod_{i=1}^{p}\lambda_i.$$

On a donc que $|\mathbf{\Sigma}| \neq 0$ ce qui est équivalent à $\lambda_1 > \cdots > \lambda_p > 0$ et que $\mathbf{\Sigma}^{-1}$ existe.

Considérons maintenant l'exemple du cas bivarié ($p = 2$). La matrice de variances-covariances s'écrit

$$\mathbf{\Sigma} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) \end{pmatrix}.$$

Si

$$r = \text{corr}(X_1,X_2) = \frac{\text{cov}(X_1,X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}$$

et $\text{var}(X_1) = \sigma_1^2$ et $\text{var}(X_2) = \sigma_2^2$, alors $\text{cov}(X_1,X_2) = r\sigma_1\sigma_2$. On a donc

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix},$$

et son déterminant est $|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2 - r^2\sigma_1^2\sigma_2^2$.

On remarque de l'équation précédente que $|\mathbf{\Sigma}| \neq 0$ si et seulement si $r \neq \pm 1$. Autrement dit, si $r = \pm 1$, $\mathbf{\Sigma}$ n'est pas inversible, auquel cas la loi normale multivariée (équation B.1) n'a plus de sens. Ce sujet est abordé en détails dans la section sur la multicolinéarité.

C Éléments d'algèbre matricielle

C.1 Opérations de base sur les matrices

Soit une matrice finie $n \times p$

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,p} \\ a_{2,1} & \dots & \dots & a_{2,p} \\ \vdots & \vdots & \dots & \vdots \\ a_{n,1} & \dots & \dots & a_{n,p} \end{pmatrix} = (a_{i,j}).$$

Addition : Soit la matrice $\mathbf{B} = (b_{i,j})$, $n \times p$. Les opérations d'addition $\mathbf{A} + \mathbf{B}$ et $\mathbf{B} + \mathbf{A}$ sont définies :

$$(a_{i,j}) + (b_{i,j}) = (c_{i,j}),$$

où $c_{i,j} = a_{i,j} + b_{i,j}$.

Soustraction : Soit la matrice $\mathbf{B} = (b_{i,j})$, $n \times p$. Les opérations de soustraction $\mathbf{A} - \mathbf{B}$ et $\mathbf{B} - \mathbf{A}$ sont définies par

$$(a_{i,j}) - (b_{i,j}) = (c_{i,j}),$$

où $c_{i,j} = a_{i,j} - b_{i,j}$.

Multiplication : Soit la matrice $\mathbf{D} = (d_{i,j})$, $m \times n$. L'opération $\mathbf{DA}$ est définie. Le résultat est une matrice $m \times p$, dont l'élément de la ligne i et la colonne j est égal à

$$\sum_{k=1}^{n} d_{i,k} a_{k,j}.$$

Soit la matrice $\mathbf{E} = (e_{i,j})$, $p \times m$. L'opération $\mathbf{AE}$ est définie. Le résultat est une matrice $n \times m$, dont l'élément de la ligne i et la colonne j est égal à

$$\sum_{k=1}^{n} a_{i,k} e_{k,j}.$$

Exemple : On a

$$\begin{pmatrix} 2 & 3 & 0 \\ 7 & 1 & 8 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 4 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 4 \\ 9 & 2 & 11 \end{pmatrix},$$

$$\begin{pmatrix} 2 & 3 & 0 \\ 7 & 1 & 8 \end{pmatrix} - \begin{pmatrix} 1 & 1 & 4 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -4 \\ 5 & 0 & 5 \end{pmatrix},$$

et

$$\begin{pmatrix} 2 & 3 & 0 \\ 7 & 2 & 8 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 1 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} 8 & 9 \\ 51 & 39 \end{pmatrix}.$$

□

C.2 Propriétés de base des matrices

Rang : Le rang d'une matrice est le nombre de colonnes (ou de lignes) qui sont linéairement indépendantes.

Matrice Identité : La matrice identité d'ordre p, notée $\mathbf{I}_p$ ou $\mathbf{I}_p$, est une matrice $p \times p$ composée de 1 sur la diagonale et de zéros ailleurs.

Inverse d'une matrice : Soit une matrice $\mathbf{A}$, $p \times p$. L'inverse de la matrice $\mathbf{A}$ est notée $\mathbf{A}^{-1}$ et est telle que

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_p.$$

Exemple : Pour une matrice 2×2, on trouve

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

□

On peut facilement obtenir l'inverse d'une matrice de plus grande dimension à l'aide de la fonction `solve` en `R`.

Transposition : On note $\mathbf{A}^\top$ ou $\mathbf{A}'$ la transposée de $\mathbf{A}$. Cette opération consiste en échangeant les lignes et les colonnes, ce qui implique que l'élément de la ligne i et la colonne j de la matrice $\mathbf{A}^\top$ est $a_{j,i}$.
Exemple : On a

$$\begin{pmatrix} 2 & 3 & 0 \\ 7 & 1 & 8 \end{pmatrix}^\top = \begin{pmatrix} 2 & 7 \\ 3 & 1 \\ 0 & 8 \end{pmatrix}.$$

□

Propriétés des transpositions de matrices :

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
- $(k\mathbf{A})^\top = k\mathbf{A}^\top$, si k est un scalaire.
- $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top\mathbf{A}^\top$

Matrice symmétrique : On dit d'une matrice carrée $\mathbf{A}$ qu'elle est symmétrique si $\mathbf{A} = \mathbf{A}^\top$.

Matrice idempotente : On dit d'une matrice carrée $\mathbf{A}$ qu'elle est idempotente si $\mathbf{A} = \mathbf{A}\mathbf{A}$. Si $\mathbf{A}$ est aussi symétrique, on a aussi que $\mathbf{I} - \mathbf{A}$ est symétrique idempotente.

Trace : Soit une matrice carrée $\mathbf{A}$, $p \times p$. La trace d'une matrice carrée est un scalaire égal à la somme des éléments sur sa diagonale :

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{m} a_{i,i}.$$

Propriétés de la trace :

- Si $\mathbf{A}$ et $\mathbf{B}$ sont des matrices carrées $p \times p$, alors $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.
- Soit $\mathbf{A}$, $n \times p$ et $\mathbf{B}$, $p \times n$, alors $\text{tr}(\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{A})$.

Forme quadratique : Soit $\mathbf{A} = (a_{i,j})$ une matrice $p \times p$ symétrique et $\mathbf{b} = (b_1, \ldots, b_p)^\top$ un vecteur $p \times 1$. Alors

$$\mathbf{b}^\top\mathbf{A}\mathbf{b} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{i,j} b_i b_j$$

est une forme quadratique. Par exemple, si $p = 2$, alors

$$\mathbf{b}^\top \mathbf{A}\mathbf{b} = \sum_{i=1}^{2}\sum_{j=1}^{2} a_{i,j} b_i b_j = a_{11}b_1^2 + 2a_{12}b_1 b_2 + a_{22}b_2^2.$$

Aussi, si $\mathbf{A}$ est diagonale, $\mathbf{b}^\top \mathbf{A}\mathbf{b} = \sum_{i=1}^{p} a_{i,i} b_i^2$.

C.3 Dérivées

Dérivée d'un produit scalaire : Soient les vecteurs $\mathbf{a} = (a_1, \ldots, a_p)^\top$ et $\mathbf{b} = (b_1, \ldots, b_p)^\top$. Leur produit scalaire est $\mathbf{b}^\top \mathbf{a} = a_1 b_1 + \cdots + a_p b_p = \sum_{i=1}^{p} a_i x_i$. La dérivée de $\mathbf{b}^\top \mathbf{a}$ par rapport à $\mathbf{b}$ est

$$\frac{\partial}{\partial \mathbf{b}} \mathbf{b}^\top \mathbf{a} = \frac{\partial}{\partial \mathbf{b}} \sum_{i=1}^{k} a_i x_i = \begin{bmatrix} \frac{\partial}{\partial b_1} \sum_{i=1}^{k} a_i x_i \\ \vdots \\ \frac{\partial}{\partial b_p} \sum_{i=1}^{k} a_i x_i \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \mathbf{a}.$$

Dérivée d'une forme quadratique : Soit $\mathbf{A}_{p\times p}$ une matrice symétrique. Alors

$$\frac{\partial}{\partial \mathbf{b}} \mathbf{b}^\top \mathbf{A}\mathbf{b} = 2\mathbf{A}\mathbf{b}.$$

Preuve : On a

$$\mathbf{b}^\top \mathbf{A}\mathbf{b} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{i,j} b_i b_j = \sum_{i=1}^{p} a_{i,i} b_i^2 + \sum_{i=1}^{p} \sum_{j\neq i \in \{1,\ldots,p\}} a_{i,j} b_i b_j.$$

Pour $t = 1, \ldots, p$ et puisque $a_{i,j} = a_{j,i}$, par symmétrie, on trouve

$$\frac{\partial}{\partial b_t} \mathbf{b}^\top \mathbf{A}\mathbf{b} = 2a_{t,t} b_t + \sum_{i\neq t \in \{1,\ldots,p\}} a_{i,t} b_i + \sum_{j\neq t \in \{1,\ldots,p\}} a_{t,j} b_j = 2\sum_{i=1}^{p} a_{i,t} b_t.$$

On retrouve donc l'espression désirée. □

Dérivée d'une fonction : Si $f(\mathbf{b})$ est une fonction dérivable du vecteur $\mathbf{b}$, alors

$$\frac{\partial}{\partial \mathbf{b}} f(\mathbf{b})^\top \mathbf{A} f(\mathbf{b}) = 2\left\{ \frac{\partial}{\partial \mathbf{b}} f(\mathbf{b}) \right\}^\top \mathbf{A} f(\mathbf{b}).$$

C.4 Moments de vecteurs aléatoires

Espérance d'un vecteur aléatoire : Soit $\mathbf{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}^\top$ un vecteur aléatoire. Alors, l'espérance de $\mathbf{X}$ est

$$\mathrm{E}[\mathbf{X}] = \begin{pmatrix} \mathrm{E}[X_1] & \mathrm{E}[X_2] & \cdots & \mathrm{E}[X_n] \end{pmatrix}^\top.$$

Variance d'un vecteur aléatoire : Soit $\mathbf{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}^\top$ un vecteur aléatoire. Alors, la variance de $\mathbf{X}$ est

$$\mathrm{var}(\mathbf{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \ldots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \ldots & \mathrm{var}(X_n) \end{pmatrix}.$$

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

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 . \end{aligned}$$

```
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)
```

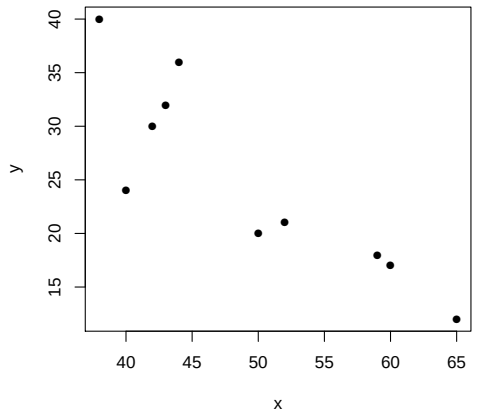

FIG. D.1 – Relation entre les données de l'exercice 2.1

Or,

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i,$$

d'où les équations normales sont

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0.$$

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

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0,$$

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

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

De la seconde équation normale, on obtient

$$\sum_{i=1}^{n} x_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

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

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}.$$

Par conséquent,

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

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 $\hat{Y}_i = 66{,}4488 - 0{,}8407 x_i$, $i = 1, 2, \ldots, 10$. Avec R, on crée un objet de type modèle de régression avec `lm` et on en extrait les valeurs ajustées avec `fitted` :

```
abline(fit)
```

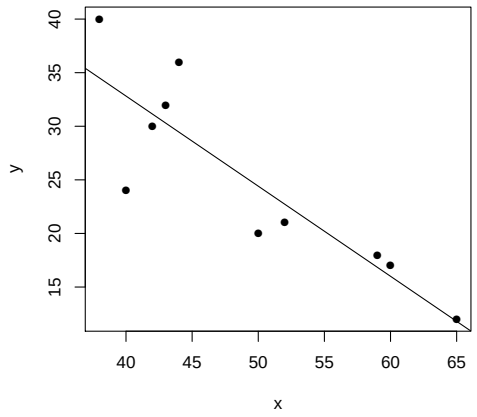

FIG. D.2 – Relation entre les données de l'exercice 2.1 et la droite de régression

```
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 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 $\varepsilon_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, 10$. 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
```

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

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

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

$$\begin{aligned}\text{SST} &= \sum_{i=1}^{8} (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{8} Y_i^2 - n\bar{Y}^2 \\ &= 214 - (8)(40/8)^2 \\ &= 14, \\ \text{SSR} &= \sum_{i=1}^{8} (\hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{8} \hat{\beta}_1^2 (x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 (\sum_{i=1}^{8} x_i^2 - n\bar{x}^2) \\ &= (-1/2)^2 (156 - (8)(32/8)^2) \\ &= 7.\end{aligned}$$

et SSE = SST − SSR = 14 − 7 = 7. Par conséquent, $R^2 = SSR/SST = 7/14 = 0{,}5$, donc la régression explique 50 % de la variation des Y_i 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` :

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

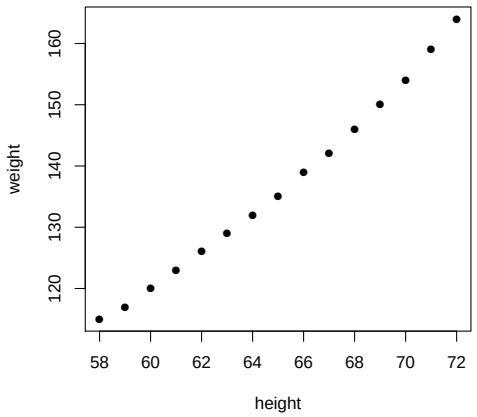

FIG. D.3 – Relation entre la taille et le poids moyen de femmes américaines âgées de 30 à 39 ans (données `women`)

```r
(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 :

```r
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 ***
## ---
```

```
abline(fit)
```

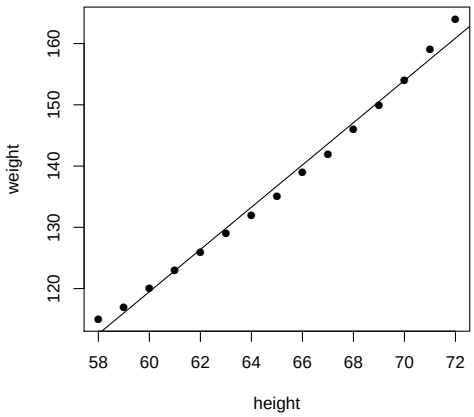

FIG. D.4 – Relation entre les données `women` et droite de régression linéaire simple

```
## 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 $\hat{Y}_i = (\bar{Y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1(x_i - \bar{x})$ et que $\varepsilon_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - \hat{\beta}_1(x_i - \bar{x})$, alors

$$\begin{aligned}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\varepsilon_i &= \hat{\beta}_1\left(\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\\ &= \hat{\beta}_1\left(S_{xy} - \frac{S_{xy}}{S_{xx}}S_{xx}\right)\\ &= 0.\end{aligned}$$

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

$$\begin{aligned}S(\beta) &= \sum_{i=1}^n \varepsilon_i^2 \\ &= \sum_{i=1}^n (Y_i - \beta x_i)^2.\end{aligned}$$

Or,

$$\frac{d}{d\beta} S(\beta) = -2\sum_{i=1}^n (Y_i - \hat{\beta}x_i)x_i,$$

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

$$\sum_{i=1}^n x_i Y_i - \hat{\beta}\sum_{i=1}^n x_i^2 = 0.$$

L'estimateur des moindres carrés de β est donc

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

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

$$\begin{aligned}
\mathrm{E}[\hat{\beta}] &= \mathrm{E}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] \\
&= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \mathrm{E}[Y_i] \\
&= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \beta x_i \\
&= \beta \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} \\
&= \beta.
\end{aligned}$$

c) Des hypothèses du modèle, on a

$$\begin{aligned}
\mathrm{var}(\hat{\beta}) &= \mathrm{var}\left(\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right) \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \mathrm{var}(Y_i) \\
&= \frac{\sigma^2}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.
\end{aligned}$$

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

$$\begin{aligned}
f(c_1, \ldots, c_n) &= \mathrm{var}(\beta^*) \\
&= \sum_{i=1}^n c_i^2 \mathrm{var}(Y_i) \\
&= \sigma^2 \sum_{i=1}^n c_i^2
\end{aligned}$$

sous la contrainte $\mathrm{E}[\beta^*] = \sum_{i=1}^n c_i \mathrm{E}[Y_i] = \sum_{i=1}^n c_i \beta x_i = \beta \sum_{i=1}^n c_i x_i = \beta$, soit $\sum_{i=1}^n c_i x_i = 1$ ou $g(c_1, \ldots, c_n) = 0$ avec

$$g(c_1, \ldots, c_n) = \sum_{i=1}^n c_i x_i - 1.$$

Pour utiliser la méthode des multiplicateurs de Lagrange, on pose

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

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

$$\frac{\partial \mathcal{L}}{\partial c_u} = 2\sigma^2 c_u - \lambda x_u, \quad u = 1,\ldots,n$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\sum_{i=1}^{n} c_i x_i + 1.$$

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

$$c_i = \frac{\lambda x_i}{2\sigma^2}.$$

Or, de la contrainte,

$$\sum_{i=1}^{n} c_i x_i = \frac{\lambda}{2\sigma^2} \sum_{i=1}^{n} x_i^2 = 1,$$

d'où

$$\frac{\lambda}{2\sigma^2} = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

et, donc,

$$c_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}.$$

Finalement,

$$\begin{aligned}\beta^* &= \sum_{i=1}^{n} c_i Y_i \\ &= \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \\ &= \hat{\beta}.\end{aligned}$$

2.8 a) Tout d'abord, puisque $\text{MSE} = \text{SSE}/(n-2) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n-2)$ et que $\text{E}[Y_i] = \text{E}[\hat{Y}_i]$, alors

$$\begin{aligned}\text{E}[MSE] &= \frac{1}{n-2}\text{E}\left[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\text{E}[(Y_i - \hat{Y}_i)^2] \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\text{E}[((Y_i - \text{E}[Y_i]) - (\hat{Y}_i - \text{E}[\hat{Y}_i]))^2] \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\left(\text{var}(Y_i) + \text{var}(\hat{Y}_i) - 2\,\text{cov}(Y_i, \hat{Y}_i)\right).\end{aligned}$$

Or, on a par hypothèse du modèle que $\text{cov}(Y_i, Y_j) = \text{cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, d'où $\text{var}(Y_i) = \sigma^2$ et $\text{var}(\bar{Y}) = \sigma^2/n$. D'autre part,

$$\begin{aligned}\text{var}(\hat{Y}_i) &= \text{var}(\bar{Y} + \hat{\beta}_1(x_i - \bar{x})) \\ &= \text{var}(\bar{Y}) + (x_i - \bar{x})^2\text{var}(\hat{\beta}_1) + 2(x_i - \bar{x})\text{cov}(\bar{Y}, \hat{\beta}_1)\end{aligned}$$

et l'on sait que

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et que

$$\begin{aligned}
\text{cov}(\bar{Y}, \hat{\beta}_1) &= \text{cov}\left(\frac{\sum_{i=1}^n Y_i}{n}, \frac{\sum_{j=1}^n (x_j - \bar{x}) Y_j}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) \\
&= \frac{1}{n \sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}(Y_i, (x_j - \bar{x}) Y_j) \\
&= \frac{1}{n \sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_j - \bar{x}) \text{var}(Y_i) \\
&= \frac{\sigma^2}{n \sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) \\
&= 0,
\end{aligned}$$

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

$$\text{var}(\hat{Y}_i) = \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

De manière similaire, on détermine que

$$\begin{aligned}
\text{cov}(Y_i, \hat{Y}_i) &= \text{cov}(Y_i, \bar{Y} + \hat{\beta}_1 (x_i - \bar{x})) \\
&= \text{cov}(Y_i, \bar{Y}) + (x_i - \bar{x}) \text{cov}(Y_i, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

Par conséquent,

$$\text{E}[(Y_i - \hat{Y}_i)^2] = \frac{n-1}{n} \sigma^2 - \frac{(x_i - \bar{x})^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et

$$\sum_{i=1}^n \text{E}[(Y_i - \hat{Y}_i)^2] = (n-2)\sigma^2,$$

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

b) On a

$$\begin{aligned}
\mathrm{E}[\mathrm{MSR}] &= \mathrm{E}[\mathrm{SSR}] \\
&= \mathrm{E}\left[\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2\right] \\
&= \sum_{i=1}^{n}\mathrm{E}[\hat{\beta}_1^2(x_i - \bar{x})^2] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{E}[\hat{\beta}_1^2] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2(\mathrm{var}(\hat{\beta}_1) + \mathrm{E}[\hat{\beta}_1]^2) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2\left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \beta_1^2\right) \\
&= \sigma^2 + \beta_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}$$

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,

$$\begin{aligned}
\bar{x}' &= \frac{1}{n}\sum_{i=1}^{n}x_i' \\
&= \frac{1}{n}\sum_{i=1}^{n}(c + dx_i) \\
&= c + d\bar{x},
\end{aligned}$$

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

$$\begin{aligned}
S_{xx}' &= \sum_{i=1}^{n}(x_i' - \bar{x}')^2 \\
&= \sum_{i=1}^{n}(c + dx_i - c - d\bar{x})^2 \\
&= d^2S_{xx}
\end{aligned}$$

et $S_{yy}' = b^2S_{yy}$, $S_{xy}' = bdS_{xy}$. Par conséquent,

$$\begin{aligned}
\hat{\beta}_1' &= \frac{S_{xy}'}{S_{xx}'} \\
&= \frac{bdS_{xy}}{d^2S_{xx}} \\
&= \frac{b}{d}\hat{\beta}_1
\end{aligned}$$

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

$$\begin{aligned}R^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}.\end{aligned}$$

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 $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, $m = n/a$ et $a = \sqrt{n+1} - 1$. On définit

$$\begin{aligned}\bar{x}' &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i \\ &= \frac{1}{n+1}\sum_{i=1}^{n} x_i + \frac{m}{n+1}\bar{x} \\ &= k\bar{x}\end{aligned}$$

et, de manière similaire,

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

où

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

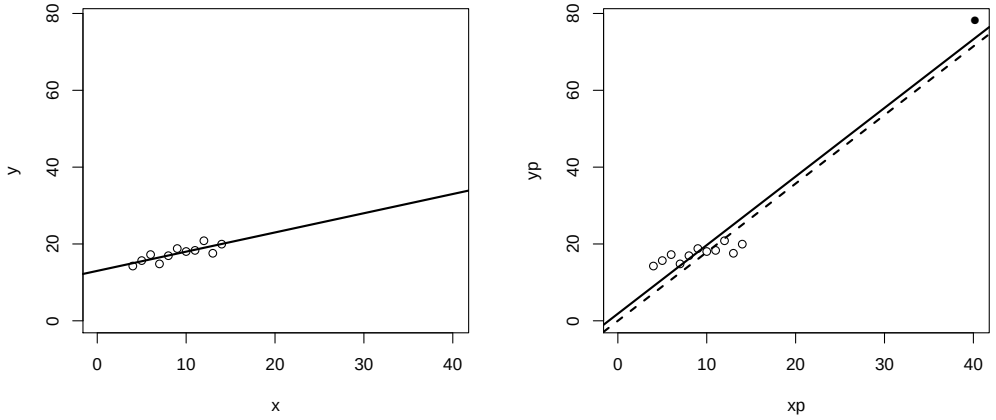

FIG. D.5 – Illustration de l'effet de l'ajout d'un point spécial à un ensemble de données. À gauche, la droite de régression usuelle. À droite, le même ensemble de points avec le point spécial ajouté (cercle plein), la droite de régression avec ce nouveau point (ligne pleine) et la droite de régression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

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

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n+1} x_i Y_i - (n+1)\bar{x}'\bar{Y}'}{\sum_{i=1}^{n+1} x_i^2 - (n+1)(\bar{x}')^2} \\ &= \frac{\sum_{i=1}^{n} x_i Y_i + m^2\bar{x}\bar{Y} - (n+1)k^2\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 + m^2\bar{x}^2 - (n+1)k^2\bar{x}^2}.\end{aligned}$$

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,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \\ &= \hat{\beta}.\end{aligned}$$

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, $\varepsilon_i \sim N(0,\sigma^2)$ et que $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, alors $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. De plus, on sait que

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},\end{aligned}$$

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

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

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 $\mathrm{SST} = S_{yy} = 20\,838$ et $S_{xx} = 10\,668$. Par conséquent,

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

et

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

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 :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= 1{,}436 \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} \\ &= 9{,}273.\end{aligned}$$

b) Les sommes de carrés sont

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

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 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 statis-

tique F. Or, à partir de l'information donnée dans l'énoncé, on peut calculer

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

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 $\text{E}[\hat{\beta}] = \beta$, d'où $\text{E}[Y_0 - \hat{Y}_0] = \text{E}[\beta x_0 + \varepsilon_0 - \hat{\beta} x_0] = \beta x_0 - \beta x_0 = 0$ 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,

$$\begin{aligned}
\text{var}(\hat{\beta}) &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} x_i^2 \,\text{var}(Y_i) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

d'où, finalement,

$$\text{var}(Y_0 - \hat{Y}_0) = \sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right).$$

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

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)\right)$$

ou, de manière équivalente,

$$\frac{Y_0 - \hat{Y}_0}{\sigma\sqrt{1 + x_0^2 / \sum_{i=1}^n x_i^2}} \sim N(0,1).$$

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

$$\frac{Y_0 - \hat{Y}_0}{s\sqrt{1 + x_0^2 / \sum_{i=1}^n x_i^2}} \sim t(n-1).$$

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

$$\hat{Y}_0 \pm t_{\alpha/2}(n-1)\, s \sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}.$$

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 $\bar{x} = 3{,}72$, $\bar{Y} = 7{,}55$, $\sum_{i=1}^{10} x_i^2 = 147{,}18$, $\sum_{t=1}^{10} Y_i^2 = 597{,}03$ et $\sum_{i=1}^{10} x_i Y_i = 295{,}95$. Par conséquent,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{10} x_i Y_i - 10\bar{x}\bar{Y}}{\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2} \\ &= 1{,}716\end{aligned}$$

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 :

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

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 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.

```r
plot(medv ~ rm + age + lstat + tax, data = house, ask = FALSE)
```

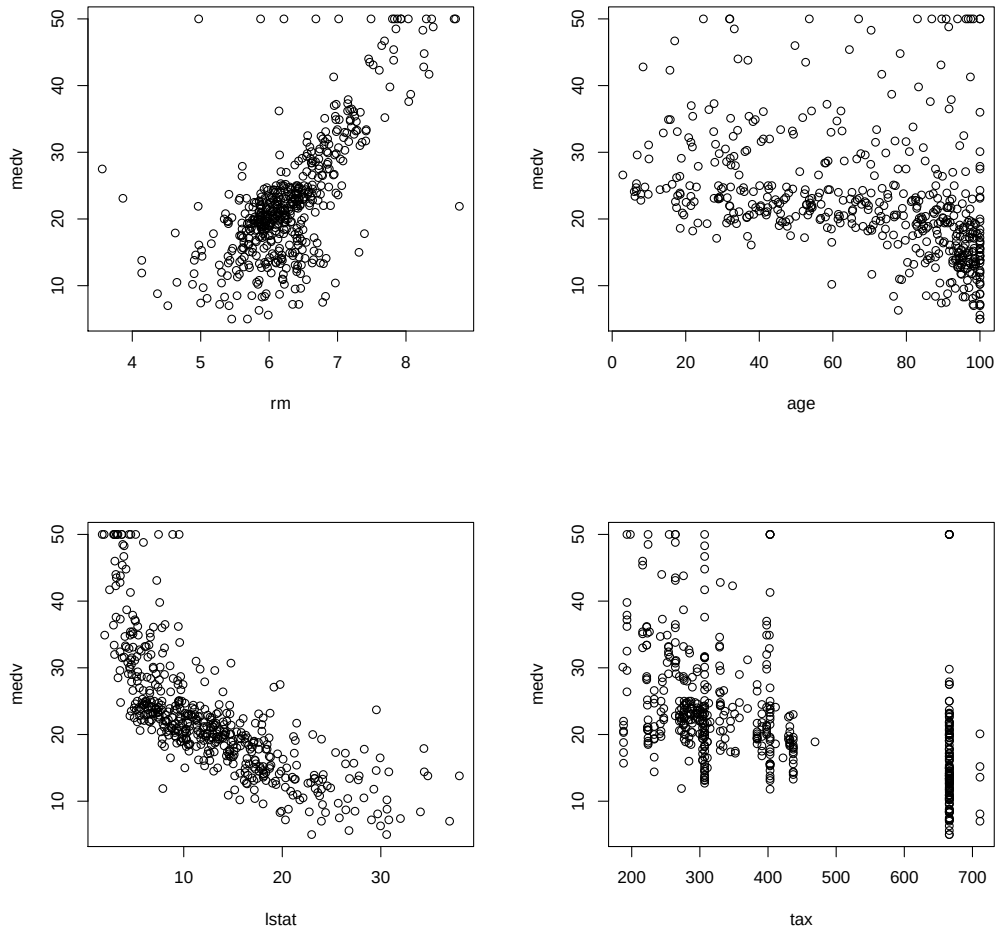

FIG. D.6 – Relation entre la variable `medv` et les variables `rm`, `age`, `lstat` et `tax` des données `house.dat`

```
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 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)
##
## 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)
```

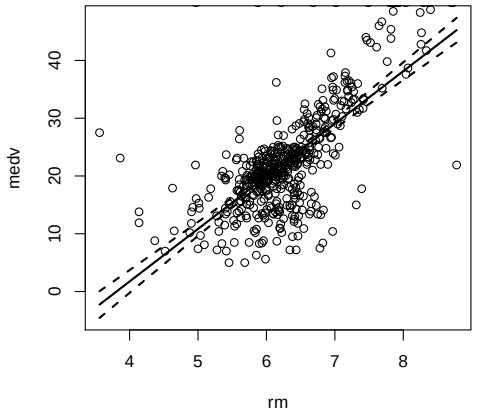

FIG. D.7 – Résultat de la régression de la variable `rm` sur la variable `medv` des données `house.dat`

```
## 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 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 :

```r
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)
```

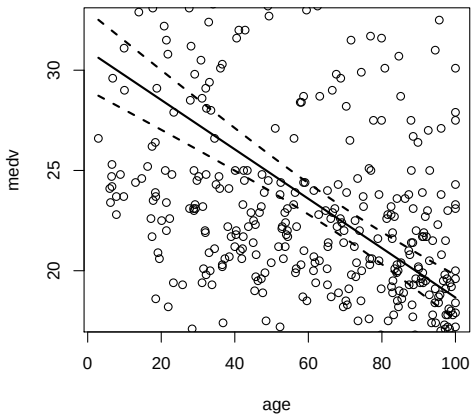

FIG. D.8 – Résultat de la régression de la variable `age` sur la variable `medv` des données `house.dat`

```r
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 : $Y_i = -0{,}01465 + 0{,}007838 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, 1{,}039^2)$, où Y_i est la consommation en litres aux 100 kilomètres et x_i 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 10.5669 8.432089 12.7017
```

2.19 a) On a

$$\bar{Y} = \frac{\sum_{i=1}^{500} Y_i}{500} = \frac{300\bar{Y}_F + 200\bar{Y}_H}{500}.$$

Aussi,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{500} x_i Y_i - 500\bar{x}\bar{Y}}{\sum_{i=1}^{500} x_i^2 - 500\bar{x}^2}.$$

Or,

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{500} x_i}{500} = \frac{300}{500}, \\ \sum_{i=1}^{500} x_i^2 &= 300, \\ \sum_{i=1}^{500} x_i Y_i &= 300\bar{Y}_F \end{aligned}$$

Donc,

$$\begin{aligned} \hat{\beta}_1 &= \frac{300\bar{Y}_F - 500 \times \frac{300}{500} \times \frac{300\bar{Y}_F + 200\bar{Y}_H}{500}}{300 - 500\left(\frac{300}{500}\right)^2} \\ &= \frac{500\bar{Y}_F - 300\bar{Y}_F - 200\bar{Y}_H}{500 - 300} \\ &= \bar{Y}_F - \bar{Y}_H. \end{aligned}$$

b) Oui, le coefficient relié à la variable indicatrice qui vaut 1 si le sexe est F représente la différence etre la moyenne de l'espérance de vie pour les femmes et la moyenne de l'espérace de vie pour les hommes.

c)

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} = \bar{Y} - (\bar{Y}_F - \bar{Y}_H)\frac{300}{500} = \bar{Y}_H.$$

$\Rightarrow \hat{\beta}_0$ est la moyenne de l'espérance de vie pour les hommes.

2.20 a)

$$\begin{aligned}
\text{cov}(Y_i,\hat{Y}_j) &= \text{cov}(Y_i,\hat{\beta}_0 + \hat{\beta}_1 x_j) \\
&= \text{cov}(Y_i,\bar{Y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_j) \\
&= \text{cov}(Y_i,\bar{Y}) + (x_j - \bar{x})\text{cov}(Y_i,\hat{\beta}_1) \text{ par indépendance des observations} \\
&= \frac{\sigma^2}{n} + \frac{(x_j - \bar{x})}{S_{xx}} \sum_{l=1}^{n} (x_l - \bar{x})\text{cov}(Y_i,Y_l) \\
&= \frac{\sigma^2}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}\sigma^2 \text{ par indépendance des observations.}
\end{aligned}$$

b)

$$\begin{aligned}
\text{cov}(\hat{Y}_i,\hat{Y}_j) &= \text{cov}(\hat{\beta}_0 + \hat{\beta}_1 x_i,\hat{\beta}_0 + \hat{\beta}_1 x_j) \\
&= \text{var}(\hat{\beta}_0) + (x_i + x_j)\text{cov}(\hat{\beta}_0,\hat{\beta}_1) + x_i x_j \text{var}(\hat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) - (x_i + x_j)\frac{\bar{x}\sigma^2}{S_{xx}} + x_i x_j \frac{\sigma^2}{S_{xx}} \\
&= \ldots \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}\right).
\end{aligned}$$

c)

$$\begin{aligned}
\text{cov}(\hat{\varepsilon}_i,\hat{\varepsilon}_j) &= \text{cov}(Y_i - \hat{Y}_i, Y_j - \hat{Y}_j) \\
&= \text{cov}(Y_i,Y_j) - \text{cov}(Y_i,\hat{Y}_j) - \text{cov}(\hat{Y}_i,Y_j) + \text{cov}(\hat{Y}_i,\hat{Y}_j) \\
&= 0 - 2\sigma^2\left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}\right) + \left(\frac{1}{n} + \frac{(x_j - \bar{x})(x_i - \bar{x})}{S_{xx}}\right) \\
&= -\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}\right).
\end{aligned}$$

2.21 Utiliser l'approximation de Taylor de premier ordre pour montrer que la variance de $g(Y) = 1/Y$ est approximativement constante.

2.22 a) Figure D.9 shows a scatter plot of the number of bacteria versus the minutes of exposure. The plot shows a straight line would be a reasonable model, but an even better model would capture the curvature. In fact, the plot shows that when the canned food is exposed to 300^o F for a long time, there is ultimately no bacteria left. This suggests a model that would capture the asymptotic behavior of the number of bacteria when the number of minutes of exposure increases. A linear model would continue to drive down the number of bacteria, eventually leading to negative values, which is nonsensical in this context.

b) A simple linear model is fitted to the data using R. Here is a summary of the model :

```
fit1 <- lm(bact~min)
summary(fit1)

##
```

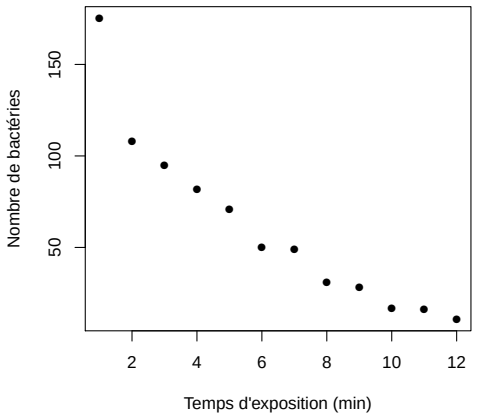

FIG. D.9 – Scatter Plot of the Number of Bacteria versus the Minutes of Exposure to 300^o F

```
## Call:
## lm(formula = bact ~ min)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.323  -9.890  -7.323   2.463  45.282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   142.20      11.26  12.627 1.81e-07 ***
## min           -12.48       1.53  -8.155 9.94e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.3 on 10 degrees of freedom
## Multiple R-squared:  0.8693,Adjusted R-squared:  0.8562
## F-statistic: 66.51 on 1 and 10 DF,  p-value: 9.944e-06
```

The fitted model is

$$\hat{y} = 142.20 - 12.48x,$$

where the parameters of the model are estimated by the best linear unbiased estimators. The ANOVA table is obtained using R :

```
anova(fit1)

## Analysis of Variance Table
##
## Response: bact
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## min        1 22268.8 22268.8  66.512 9.944e-06 ***
```

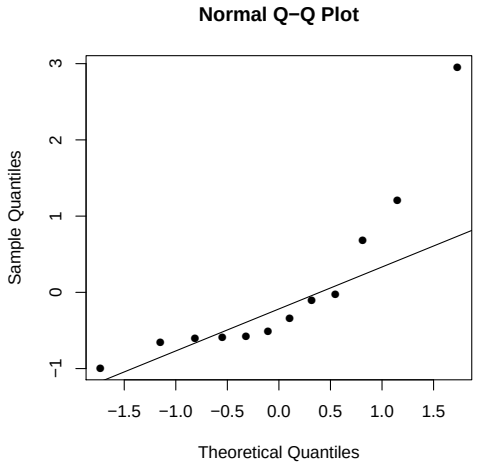

FIG. D.10 – Q-Q Plot for Simple Linear Model in Problem 2.22

```
## Residuals 10  3348.1   334.8
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to test for the significance of regression, we use the F-statistic. The F-statistic is 66.512, and it has 1 and 10 degrees of freedom, so the p-value is

$$P[F_{(1,10)} > 66.512] = 9.944 \times 10^{-6}.$$

Since the p-value is much smaller than 1%, there is enough evidence to reject the null hypothesis that $\beta_1 = 0$ at the 1% level. The simple linear model is significant.

The value of R^2 is 86.93%. This is a high coefficient of correlation, it means that about 87% of the variation in the number of bacteria in the canned food is explained by the minutes of exposure to 300^oF. The model seems to perform well.

The Q-Q Plot of the studentized residuals is shown in Figure D.10. The line represents when the empirical quantiles are exactly equal to the standard normal quantiles. The normality assumption is seriously violated as the dots are clearly not on a straight line. This means there are serious flaws in the model, including the fact that the hypothesis tests are not reliable.

Figure D.11 shows a plot of the studentized residuals versus the fitted values. The plot suggests a clear curve, which is usually an indicator of non-linearity. This is in line with the previous comments.

Finally, this model is inadequate and transformations on the response variables are required.

c) The Box-Cox method is used to determine which transformation is optimal. Figure D.12 shows the plot of the log-likelihood function in terms of λ, for two different ranges of λ. It was obtained with the R commands :

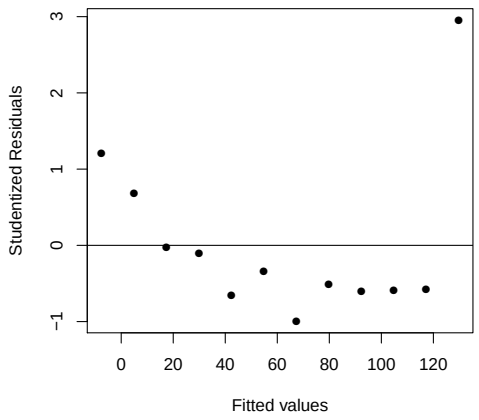

FIG. D.11 – Residuals versus the Fitted Values for Simple Linear Model in Problem 2.22

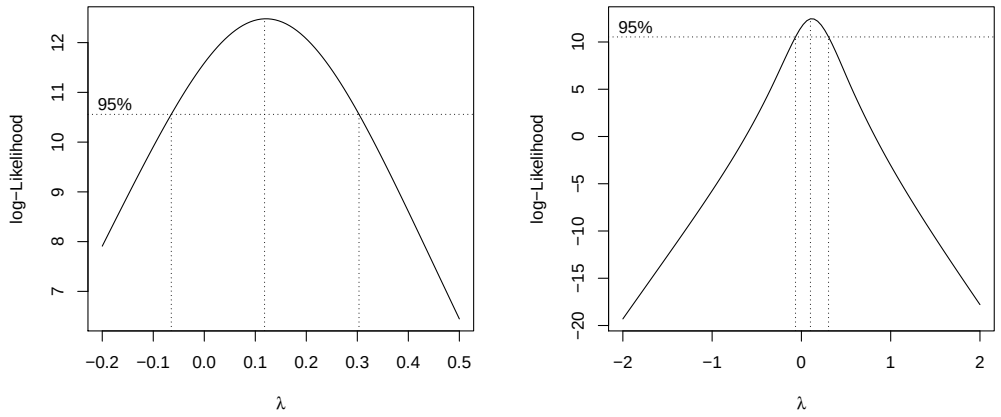

FIG. D.12 – Log-likelihood versus λ in the Box-Cox method for Problem 2.22

```
boxcox(bact~min, lambda = seq(-2, 2, len = 20), plotit = TRUE)
boxcox(bact~min, lambda = seq(-0.2, 0.5, len = 20), plotit = TRUE)
```

Note that the maximum is around 0.1 and 0 is included in the 95% confidence interval for λ. Therefore, it is preferable to use 0 as this is a common transformation, it represents the logarithm transformation. Let $y^* = \ln(y)$. A simple linear model is fitted to the transformed data. The output is the following :

```
logbact <- log(bact)
fit2 <- lm(logbact~min)
summary(fit2)

##
```

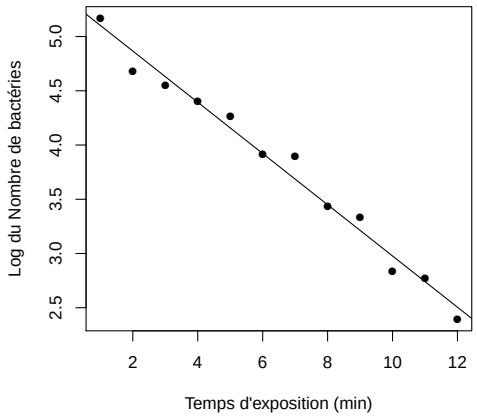

FIG. D.13 – Scatter Plot of the Logarithm of the Number of Bacteria versus the Minutes of Exposure to 300^o F

```
## Call:
## lm(formula = logbact ~ min)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.184303 -0.083994  0.001453  0.072825  0.206246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.33878    0.07409   72.05 6.47e-15 ***
## min         -0.23617    0.01007  -23.46 4.49e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1204 on 10 degrees of freedom
## Multiple R-squared:  0.9822,Adjusted R-squared:  0.9804
## F-statistic: 550.3 on 1 and 10 DF,  p-value: 4.489e-10
```

The fitted model is

$$\hat{y}^* = 5.33878 - 0.23617x,$$

where the parameters of the model are estimated by the best linear unbiased estimators. Figure D.13 is a scatter plot of the transformed response variable versus the covariate, along with the fitted line. The scatter plot looks much more linear now than in (a).

The ANOVA table is obtained using R :

```
anova(fit2)

## Analysis of Variance Table
##
```

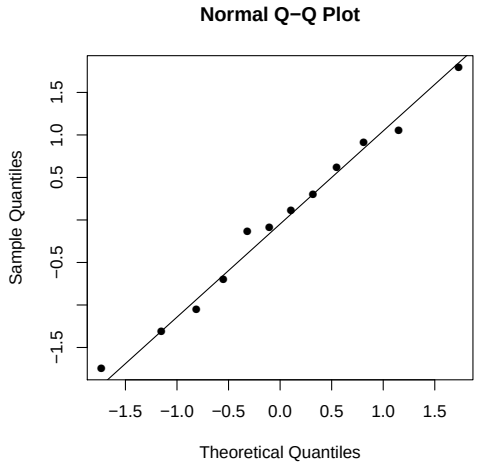

FIG. D.14 – Q-Q Plot of Model for the Logarithm of the Number of Bacteria in Problem 2.22

```
## Response: logbact
##           Df Sum Sq Mean Sq F value    Pr(>F)
## min        1 7.9761  7.9761  550.33 4.489e-10 ***
## Residuals 10 0.1449  0.0145
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic for the test of significance of regression is 550.33, and it has 1 and 10 degrees of freedom, so the p-value is

$$P[F_{(1,10)} > 550.33] = 4.489 \times 10^{-10}.$$

Since the p-value is much smaller than 1%, there is enough evidence to reject the null hypothesis that $\beta_1 = 0$ at the 1% level. This model is significant.

The value of R^2 is very high at 98.22%. This means that about 98% of the variation in the log of the number of bacteria in the canned food is explained by the minutes of exposure to 300^oF. The model seems to perform very well, better than the model proposed in (b).

The Q-Q Plot of the studentized residuals is shown in Figure D.14. The dots are beautifully aligned with the standard normal quantiles. The normality assumption is appropriate. Figure D.15 shows a plot of the studentized residuals versus the fitted values. The dots can be contained in horizontal bands and looks randomly scattered.

Finally, this model is adequate and the transformation used on the response variables fixed the problems in the model.

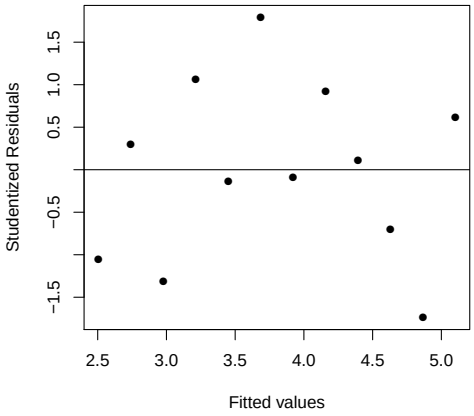

FIG. D.15 – Residuals versus the Fitted Values for Model for the Logarithm of the Number of Bacteria in Problem 2.22