Modèles linéaires en actuariat

Exercices et solutions

Modèles linéaires en actuariat

Exercices et solutions

Marie-Pier Côté

Vincent Mercier

École d'actuariat, Université Laval

Seconde édition

Historique de publication
Septembre 2019 : Première édition

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

Introduction

Ce document contient les exercices proposés par Marie-Pier Côté pour le cours ACT-2003 Modèles linéaires en actuariat, donné à l'École d'actuariat de l'Université Laval. Certains exercices sont le fruit de l'imagination des auteurs ou de ceux des versions précédentes, alors que plusieurs autres sont des adaptations d'exercices tirés des ouvrages cités dans la bibliographie.

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

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

L'estimation des paramètres, le calcul de prévisions et l'analyse des résultats sont toutes des procédures à forte composante numérique. Il serait tout à fait artificiel de se restreindre, dans les exercices, à de petits ensembles de données se prêtant au calcul manuel. Dans cette optique, plusieurs des exercices de ce recueil requièrent l'utilisation du logiciel statistique R. D'ailleurs, l'annexe **??** présente les principales fonctions de R pour la régression.

Le format de cet annexe est inspiré de **?**] : la présentation des fonctions compte peu d'exemples. Par contre, le lecteur est invité à lire et exécuter le code informatique des sections d'exemples **??**.

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

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

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

???? à régler

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

```
> data()
```

en fournit une liste complète.

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

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

Table des matières

Première partie

Régression linéaire

2 Régression linéaire simple

2.1 Considérer les données suivantes et le modèle de régression linéaire $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$:

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

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

$$\text{var}[\beta^*] = \text{var}\left[\sum_{i=1}^{n} c_i Y_i\right]$$

sous la contrainte que

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

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

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

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

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

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 $\mathbb{C}\text{ov}(Y_i, \hat{Y}_j)$.

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

c) Déduire de a) et b) que

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

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

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

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

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

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

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

a) $Y_t = \beta_0 + \varepsilon_t$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.6 En ajustant le modèle

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

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

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

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

3.7 On vous donne les observations suivantes :

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

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

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

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

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

et

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

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

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

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

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

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

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

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

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

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

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

où $E[\varepsilon_t] = 0$ et $\text{cov}(\varepsilon_t, \varepsilon_s) = \delta_{ts}\sigma^2$. 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.

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

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

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

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

Calculer la statistique appropriée pour le test

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

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

a) Vérifier que

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

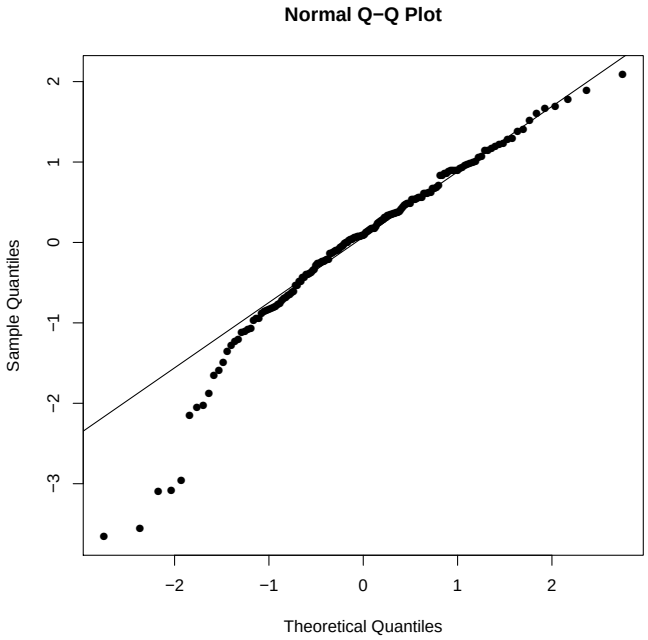
b) La figure 3.2 montre les résidus studentisés en fonction de chacune des variables exogènes et en fonction des valeurs prédites. Utiliser ces graphiques pour commenter sur la validité des postulats du modèle. Y en a-t-il qui ne sont pas respectés ? S'il y a lieu, expliquer l'impact de la violation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

FIG. 3.1 – QQ-Plot des résidus studentisés




```
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.

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

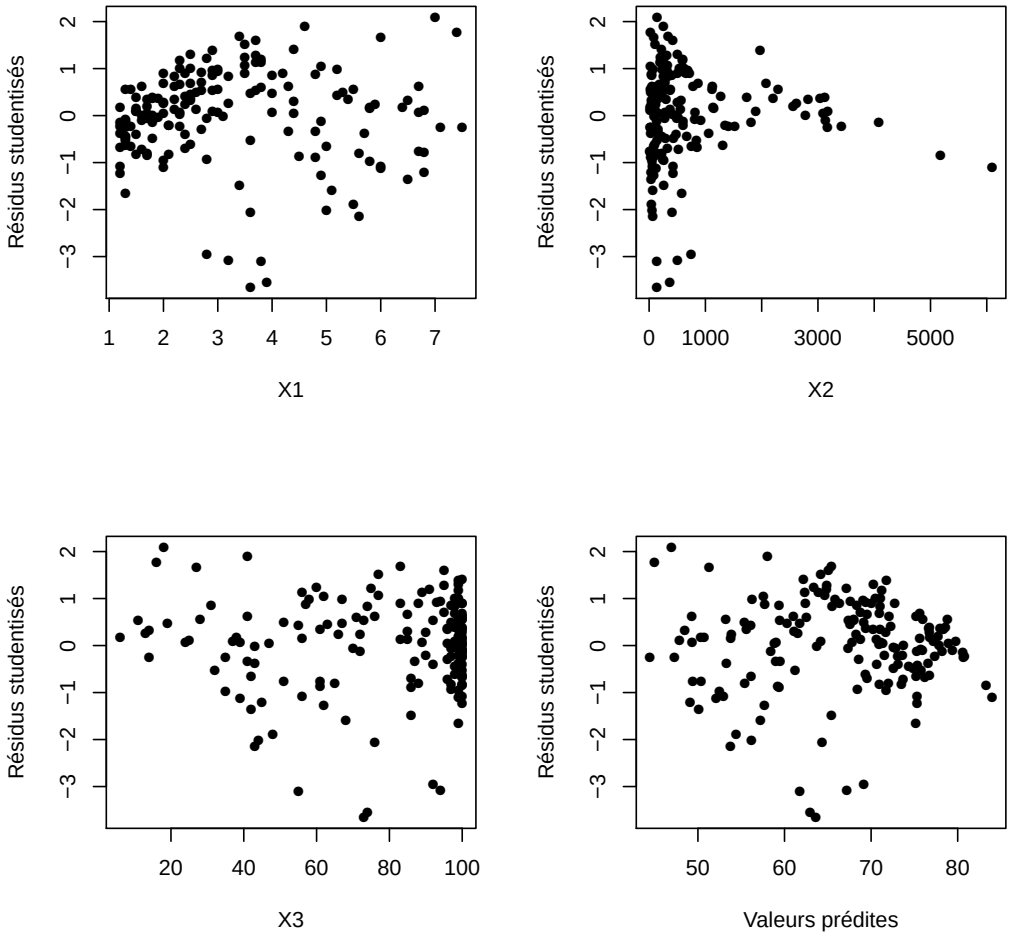
3.2 a) $\hat{\beta}_0 = \bar{Y}$ b) $\hat{\beta}_1 = (\sum_{t=1}^{n} X_t Y_t)/(\sum_{t=1}^{n} X_t^2)$

3.6 $p \approx 0{,}01$

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

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

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



3.10 a) $\mathbf{y}_{20\times1}$, $\mathbf{X}_{20\times4}$, $\boldsymbol{\beta}_{4\times1}$ et $\boldsymbol{\varepsilon}_{20\times1}$

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

3.12 103,67

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

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

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

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

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

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

Deuxième partie

Annexes

A Solutions

Chapitre 2

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

b) Les équations normales sont les équations à résoudre pour trouver les estimateurs de 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)
```

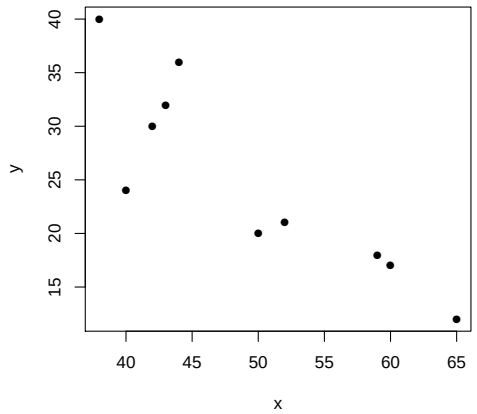


FIG. A.1 – Relation entre les données de l'exercice 2.1

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)
```

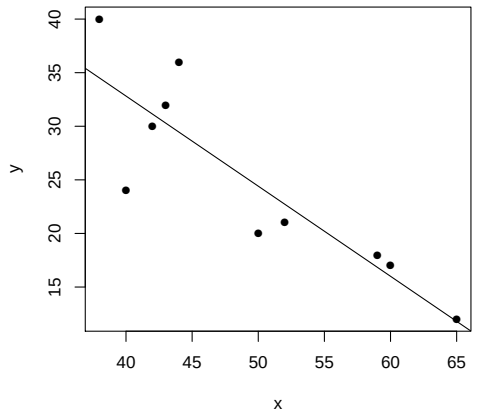


FIG. A.2 – Relation entre les données de l'exercice 2.1 et la droite de régression

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

e) Les résidus de la régression sont $\varepsilon_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, 10$. 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 A.3.

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

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

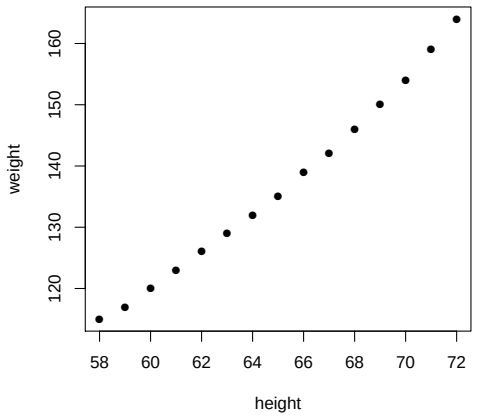


FIG. A.3 – Relation entre la taille et le poids moyen de femmes américaines âgées de 30 à 39 ans (données `women`)

```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 A.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 ***
## ---
```

```r
abline(fit)
```

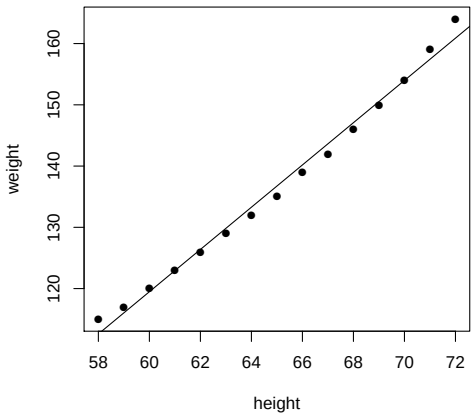


FIG. A.4 – Relation entre les données `women` et droite de régression linéaire simple

```
## 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

```r
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 $E[\hat{\beta}] = \beta$. On a

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

c) Des hypothèses du modèle, on a

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

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

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

sous la contrainte $E[\beta^*] = \sum_{i=1}^n c_i E[Y_i] = \sum_{i=1}^n c_i \beta x_i = \beta \sum_{i=1}^n c_i x_i = \beta$, 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 $E[Y_i] = E[\hat{Y}_i]$, alors

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

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,

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

et

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

d'où $E[\text{MSE}] = \sigma^2$.

b) On a

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

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,

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

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

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

et

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

b) Tout d'abord, on établit que

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

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

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

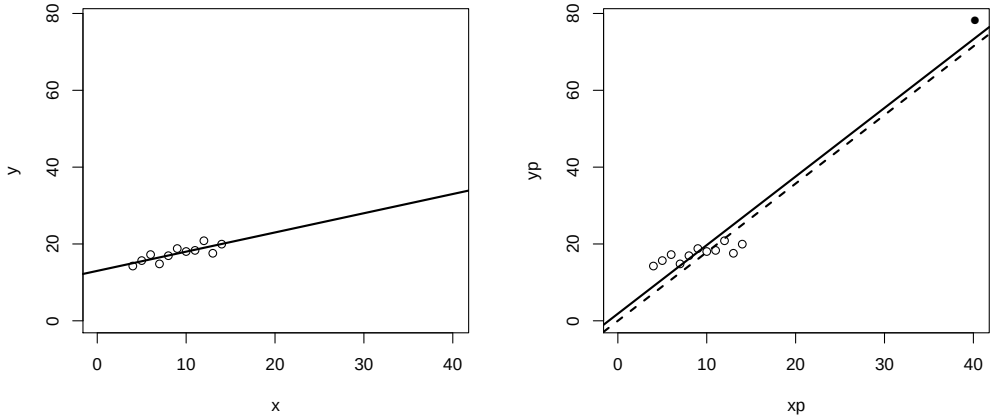


FIG. A.5 – Illustration de l'effet de l'ajout d'un point spécial à un ensemble de données. À gauche, la droite de régression usuelle. À droite, le même ensemble de points avec le point spécial ajouté (cercle plein), la droite de régression avec ce nouveau point (ligne pleine) et la droite de régression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

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 A.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, $\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \text{var}[\hat{\beta}_1])$, où $E[\hat{\beta}_1] = \beta_1$ et $\text{var}[\hat{\beta}_1] = \sigma^2 / S_{xx}$ et, donc,

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

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

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

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

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

et

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

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

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

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

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

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

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

Or, $\text{cov}(Y_0, \hat{Y}_0) = 0$ par l'hypothèse ii) de l'énoncé, $\text{var}[Y_0] = \sigma^2$ et $\text{var}[\hat{Y}_0] = x_0^2 \text{var}[\hat{\beta}]$. De plus,

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

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

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

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

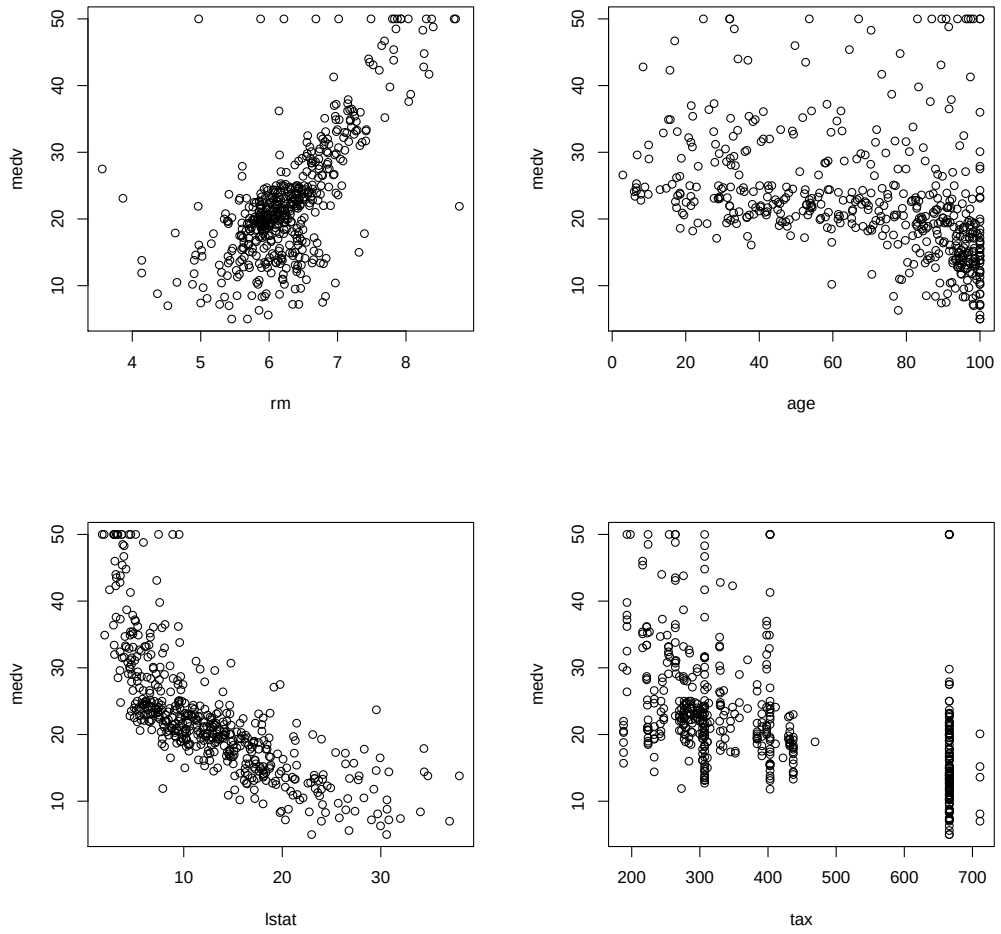


FIG. A.6 – Relation entre la variable `medv` et les variables `rm`, `age`, `lstat` et `tax` des données `house.dat`

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

```
##
## Call:
## lm(formula = medv ~ rm, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835,Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

On peut voir que tant l'ordonnée à l'origine que la pente sont très significativement différentes de zéro. La régression est donc elle-même significative. Cependant, le coefficient de détermination n'est que de $R^2 = 0{,}4835$, ce qui indique que d'autres facteurs pourraient expliquer la variation dans `medv`.

On calcule les bornes de l'intervalle de confiance de la régression avec la fonction `predict` :

```
pred.ci <- predict(fit1, interval = "confidence", level = 0.95)
```

La droite de régression et ses bornes d'intervalle de confiance inférieure et supérieure sont illustrée à la figure A.7.

c) On reprend la même démarche, mais cette fois avec la variable `age` :

```
fit2 <- lm(medv ~ age, data = house)
summary(fit2)
```

```
##
## Call:
## lm(formula = medv ~ age, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
```

```
ord <- order(house$rm)
plot(medv ~ rm, data = house, ylim = range(pred.ci))
matplot(house$rm[ord], pred.ci[ord,],
        type = "l", lty = c(1, 2, 2), lwd= 2,
        col = "black", add = TRUE)
```

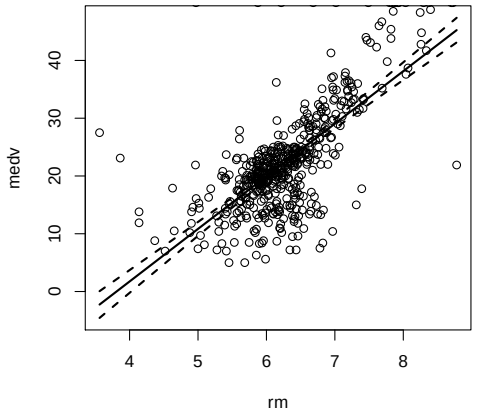


FIG. A.7 – Résultat de la régression de la variable `rm` sur la variable `medv` des données `house.dat`

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421,Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16

pred.ci <- predict(fit2, interval = "confidence", level = 0.95)
```

La régression est encore une fois très significative. Cependant, le R^2 est encore plus faible qu'avec la variable `rm`. Les variables `rm` et `age` contribuent donc chacune à expliquer les variations de la variable `medv` (et `rm` mieux que `age`), mais aucune ne sait le faire seule de manière satisfaisante. La droite de régression et l'intervalle de confiance de celle-ci sont reproduits à la figure A.8. On constate que l'intervalle de confiance est plus large qu'en b).

2.18 a) On importe les données dans R, puis on effectue les conversions demandées. La variable `consommation` contient la consommation des voitures en $\ell/100$ km et la variable `poids` le poids en kilogrammes.

```
carburant <- read.table("carburant.dat", header = TRUE)
consommation <- 235.1954/carburant$mpg
poids <- carburant$poids * 0.45455 * 1000
```

b) La fonction `summary` fournit l'information essentielle pour juger de la validité et de la qualité du modèle :

```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)
```

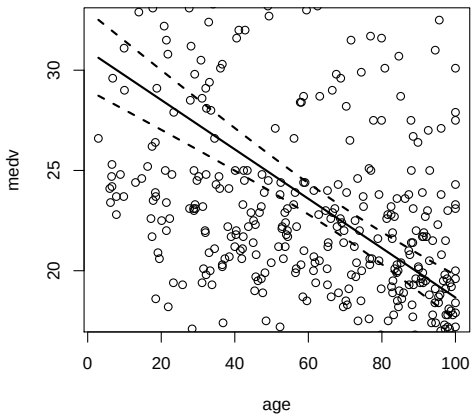


FIG. A.8 – Résultat de la régression de la variable `age` sur la variable `medv` des données `house.dat`

```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)

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

b)

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

c)

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

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 A.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)

##
```

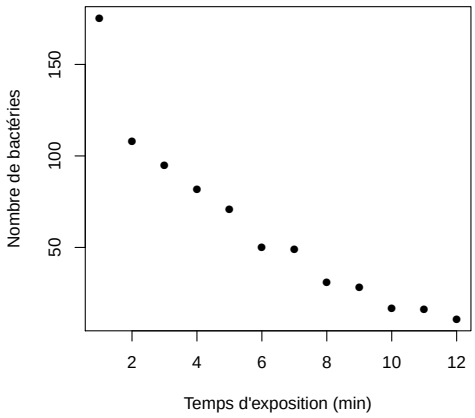


FIG. A.9 – Scatter Plot of the Number of Bacteria versus the Minutes of Exposure to 300^o F

```
## 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 ***
```

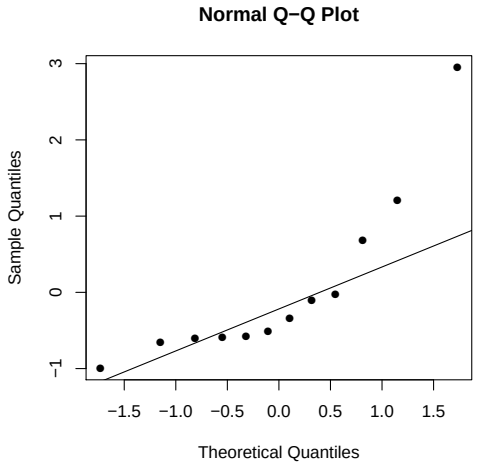


FIG. A.10 – Q-Q Plot for Simple Linear Model in Problem 2.22

```
## 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 A.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 A.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 A.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 :

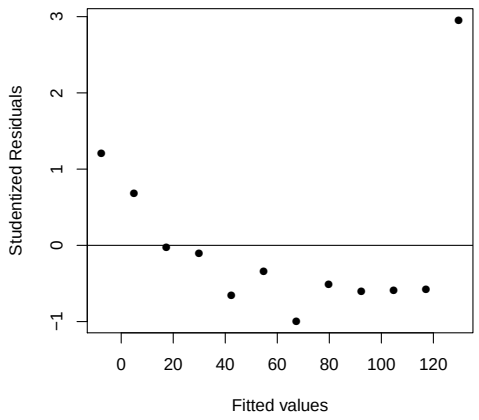


FIG. A.11 – Residuals versus the Fitted Values for Simple Linear Model in Problem 2.22

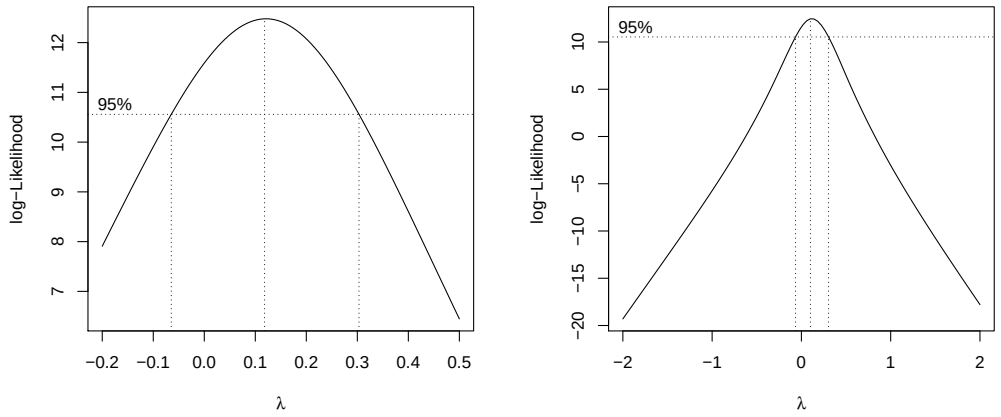


FIG. A.12 – Log-likelihood versus λ in the Box-Cox method for Problem 2.22

```
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)

##
```

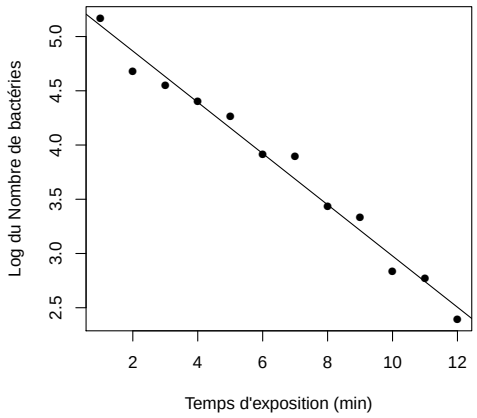


FIG. A.13 – Scatter Plot of the Logarithm of the Number of Bacteria versus the Minutes of Exposure to 300^o F

```
## 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 A.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
##
```

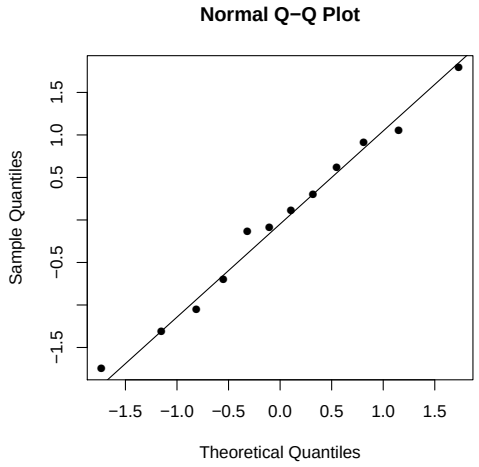


FIG. A.14 – Q-Q Plot of Model for the Logarithm of the Number of Bacteria in Problem 2.22

```
## 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 A.14. The dots are beautifully aligned with the standard normal quantiles. The normality assumption is appropriate. Figure A.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.

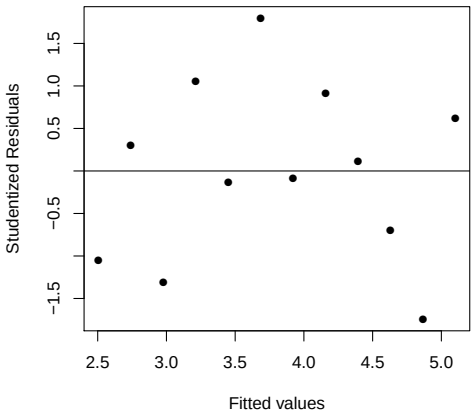


FIG. A.15 – Residuals versus the Fitted Values for Model for the Logarithm of the Number of Bacteria in Problem 2.22

Chapitre 3

3.1 Tout d'abord, selon le théorème **??** de l'annexe **??**,

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

Il suffit, pour faire la démonstration, d'appliquer directement ce résultat à la forme quadratique

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

avec $f(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ et $\mathbf{A} = \mathbf{I}$, la matrice identité. On a alors

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2 \left(\frac{d}{d\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right)' \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ &= 2(-\mathbf{X})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

En posant ces dérivées exprimées sous forme matricielle simultanément égales à zéro, on obtient les équations normales à résoudre pour calculer l'estimateur des moindres carrés du vecteur $\boldsymbol{\beta}$, soit

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

En isolant $\hat{\boldsymbol{\beta}}$ dans l'équation ci-dessus, on obtient, finalement, l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

3.2 a) On a un modèle sans variable explicative. Intuitivement, la meilleure prévision de Y_t sera alors $\bar{Y}$. En effet, pour ce modèle,

$$\mathbf{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$$

et

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= n^{-1} \sum_{t=1}^{n} Y_t \\ &= \bar{Y}.\end{aligned}$$

b) Il s'agit du modèle de régression linéaire simple passant par l'origine, pour lequel la matrice de schéma est

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}_{n \times 1}.$$

Par conséquent,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \left(\sum_{t=1}^{n} X_t^2 \right)^{-1} \sum_{t=1}^{n} X_t Y_t \\ &= \frac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2},\end{aligned}$$

tel qu'obtenu à l'exercice 2.6.

c) On est ici en présence d'un modèle de régression multiple ne passant pas par l'origine et ayant deux variables explicatives. La matrice de schéma est alors

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}_{n \times 3}.$$

Par conséquent,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ X_{12} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ X_{12} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \begin{bmatrix} n & n\bar{X}_1 & n\bar{X}_2 \\ n\bar{X}_1 & \sum_{t=1}^{n} X_{t1}^2 & \sum_{t=1}^{n} X_{t1} X_{t2} \\ n\bar{X}_2 & \sum_{t=1}^{n} X_{t1} X_{t2} & \sum_{t=1}^{n} X_{t2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{n} Y_t \\ \sum_{t=1}^{n} X_{t1} Y_t \\ \sum_{t=1}^{n} X_{t2} Y_t \end{bmatrix}.\end{aligned}$$

L'inversion de la première matrice et le produit par la seconde sont laissés aux bons soins du lecteur plus patient que les rédacteurs de ces solutions.

3.3 Dans le modèle de régression linéaire simple, la matrice schéma est

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

Par conséquent,

$$\begin{aligned} \text{var}[\hat{\boldsymbol{\beta}}] &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 \left(\begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \right)^{-1} \\ &= \sigma^2 \begin{bmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{t=1}^n X_t^2 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{n\sum_{t=1}^n X_t^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{t=1}^n X_t^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix} \\ &= \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X}^2)} \begin{bmatrix} n^{-1}\sum_{t=1}^n X_t^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}, \end{aligned}$$

d'où

$$\begin{aligned} \text{var}[\hat{\beta}_0] &= \sigma^2 \frac{\sum_{t=1}^n X_t^2}{n\sum_{t=1}^n (X_t - \bar{X})^2} \\ &= \sigma^2 \frac{\sum_{t=1}^n (X_t - \bar{X}^2) + n\bar{X}^2}{n\sum_{t=1}^n (X_t - \bar{X})^2} \end{aligned}$$

et

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

Ceci correspond aux résultats antérieurs.

3.4 Dans les démonstrations qui suivent, trois relations de base seront utilisées : $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

a) On a

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} \\ &= \mathbf{0}. \end{aligned}$$

En régression linéaire simple, cela donne

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{bmatrix} \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_{t=1}^n e_t \\ \sum_{t=1}^n X_t e_t \end{bmatrix}. \end{aligned}$$

Par conséquent, $\mathbf{X}'\mathbf{e} = \mathbf{0}$ se simplifie en $\sum_{t=1}^{n} e_t = 0$ et $\sum_{t=1}^{n} X_t e_t = 0$ soit, respectivement, la condition pour que l'estimateur des moindres carrés soit sans biais et la seconde équation normale obtenue à la partie b) de l'exercice 2.1.

b) On a

$$\begin{aligned}\hat{\mathbf{y}}'\mathbf{e} &= (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= 0.\end{aligned}$$

Pour tout modèle de régression cette équation peut aussi s'écrire sous la forme plus conventionnelle $\sum_{t=1}^{n} \hat{Y}_t e_t = 0$. Cela signifie que le produit scalaire entre le vecteur des prévisions et celui des erreurs doit être nul ou, autrement dit, que les vecteurs doivent être orthogonaux. C'est là une condition essentielle pour que l'erreur quadratique moyenne entre les vecteurs $\mathbf{y}$ et $\hat{\mathbf{y}}$ soit minimale. (Pour de plus amples détails sur l'interprétation géométrique du modèle de régression, consulter **?** , chapitres 20 et 21].) D'ailleurs, on constate que $\hat{\mathbf{y}}'\mathbf{e} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e}$ et donc, en supposant sans perte de généralité que $\hat{\boldsymbol{\beta}} \neq \mathbf{0}$, que $\hat{\mathbf{y}}'\mathbf{e} = 0$ et $\mathbf{X}'\mathbf{e} = \mathbf{0}$ sont des conditions en tous points équivalentes.

c) On a

$$\begin{aligned}\hat{\mathbf{y}}'\hat{\mathbf{y}} &= (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.\end{aligned}$$

Cette équation est l'équivalent matriciel de l'identité

$$\begin{aligned}\text{SSR} &= \hat{\beta}_1^2 \sum_{t=1}^{n} (X_t - \bar{X})^2 \\ &= \frac{S_{XY}^2}{S_{XX}}\end{aligned}$$

utilisée à plusieurs reprises dans les solutions du chapitre 2. En effet, en régression linéaire simple, $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \sum_{t=1}^{n} \hat{Y}_t^2 = \sum_{t=1}^{n} (\hat{Y} - \bar{Y})^2 + n\bar{Y}^2 = \text{SSR} + n\bar{Y}^2$ et

$$\begin{aligned}\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} &= \hat{\beta}_0 n\bar{Y} + \hat{\beta}_1 \sum_{t=1}^{n} X_t Y_t \\ &= (\bar{Y} - \hat{\beta}_1 \bar{X}) n\bar{Y} + \hat{\beta}_1 \sum_{t=1}^{n} X_t Y_t \\ &= \hat{\beta}_1 \sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) + n\bar{Y}^2 \\ &= \frac{S_{XY}^2}{S_{XX}} + n\bar{Y}^2,\end{aligned}$$

d'où $\text{SSR} = S_{XY}^2 / S_{XX}$.

3.5 a) Premièrement, $Y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$ avec $E[\varepsilon_0] = 0$. Par conséquent, $E[Y_0] = E[x_0\boldsymbol{\beta} + \varepsilon_0] = \mathbf{x}_0\boldsymbol{\beta}$. Deuxièmement, $E[\hat{Y}_0] = E[\mathbf{x}_0\hat{\boldsymbol{\beta}}] = \mathbf{x}_0 E[\hat{\boldsymbol{\beta}}] = x_0\boldsymbol{\beta}$ puisque l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est sans biais. Ceci complète la preuve.

b) Tout d'abord, $E[(\hat{Y}_0 - E[Y_0])^2] = \mathbf{V}[\hat{Y}_0] = \text{var}[\hat{Y}_0]$ puisque la matrice de variance-covariance du vecteur aléatoire $\hat{Y}_0$ ne contient, ici, qu'une seule valeur. Or, par le théorème **??**,

$$\begin{aligned}\text{var}[\hat{Y}_0] &= \mathbf{V}[\mathbf{x}_0\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0\mathbf{V}[\hat{\boldsymbol{\beta}}]\mathbf{x}_0' \\ &= \sigma^2\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'.\end{aligned}$$

Afin de construire un intervalle de confiance pour $E[Y_0]$, on ajoute au modèle l'hypothèse $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Par linéarité de l'estimateur des moindres carrés, on a alors $\hat{Y}_0 \sim N(E[Y_0], \text{var}[\hat{Y}_0])$. Par conséquent,

$$\Pr\left[-z_{\alpha/2} \leq \frac{\hat{Y} - E[\hat{Y}_0]}{\sqrt{\text{var}[\hat{Y}_0]}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

d'où un intervalle de confiance de niveau $1 - \alpha$ pour $E[Y_0]$ est

$$E[Y_0] \in \hat{Y}_0 \pm z_{\alpha/2}\sigma\sqrt{\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

Si la variance σ^2 est inconnue et estimée par s^2, alors la distribution normale est remplacée par une distribution de Student avec $n - p - 1$ degrés de liberté. L'intervalle de confiance devient alors

$$E[Y_0] \in \hat{Y}_0 \pm t_{\alpha/2}(n - p - 1)\, s\sqrt{\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

c) Par le résultat obtenu en a) et en supposant que $\text{cov}(\varepsilon_0, \varepsilon_t) = 0$ pour tout $t = 1, \ldots, n$, on a

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

Ainsi, avec l'hypothèse sur le terme d'erreur énoncée en b), $Y_0 - \hat{Y}_0 \sim N(0, \text{var}[Y_0 - \hat{Y}_0])$. En suivant le même cheminement qu'en b), on détermine qu'un intervalle de confiance de niveau $1 - \alpha$ pour Y_0 est

$$Y_0 \in \hat{Y}_0 \pm z_{\alpha/2}\sigma\sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

ou, si la variance σ^2 est inconnue et estimée par s^2,

$$Y_0 \in \hat{Y}_0 \pm t_{\alpha/2}(n - p - 1)\, s\sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'}.$$

3.6 On a la relation suivante liant la statistique F et le coefficient de détermination R^2 :

$$F = \frac{R^2}{1 - R^2}\,\frac{n - p - 1}{p}$$

La principale inconnue dans le problème est n, le nombre de données. Or,

$$\begin{aligned} n &= pF\left(\frac{1-R^2}{R^2}\right) + p + 1 \\ &= 3(5{,}438)\left(\frac{1-0{,}521}{0{,}521}\right) + 3 + 1 \\ &= 19. \end{aligned}$$

Soit F une variable aléatoire dont la distribution est une loi de Fisher avec 3 et $19-3-1=15$ degrés de liberté, soit la même distribution que la statistique F du modèle. On obtient la valeur p du test global de validité du modèle dans un tableau de quantiles de la distribution F ou avec la fonction `pf` dans R :

$$\Pr[F > 5{,}438] = 0{,}0099$$

3.7 a) On a

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \frac{1}{2}\begin{bmatrix} -6 & 34 & -13 & -13 \\ 2 & -4 & 1 & 1 \\ 0 & -2 & 1 & 1 \end{bmatrix}\begin{bmatrix} 17 \\ 12 \\ 14 \\ 13 \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} -45 \\ 13 \\ 3 \end{bmatrix} = \begin{bmatrix} -22{,}5 \\ 6{,}5 \\ 1{,}5 \end{bmatrix} \end{aligned}$$

b) Avec les résultats de la partie a), on a

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 17 \\ 12 \\ 13{,}5 \\ 13{,}5 \end{bmatrix},$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 0 \\ 0 \\ 0{,}5 \\ -0{,}5 \end{bmatrix}$$

et $\bar{Y} = 14$. Par conséquent,

$$\begin{aligned} \text{SST} &= \mathbf{y}'\mathbf{y} - n\bar{Y}^2 = 14 \\ \text{SSE} &= \mathbf{e}'\mathbf{e} = 0{,}5 \\ \text{SSR} &= \text{SST} - \text{SSR} = 13{,}5, \end{aligned}$$

d'où le tableau d'analyse de variance est le suivant :

Source	SS	d.l.	MS	F
Régression	13,5	2	6,75	13,5
Erreur	0,5	1	0,5	
Total	14			

Le coefficient de détermination est

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 0{,}9643.$$

c) On sait que $\text{var}[\hat{\beta}_i] = \sigma^2 c_{ii}$, où c_{ii} est l'élément en position $(i+1, i+1)$ de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Or, $\hat{\sigma}^2 = s^2 = \text{MSE} = 0{,}5$, tel que calculé en b). Par conséquent, la statistique t du test $H_0 : \beta_1 = 0$ est

$$t = \frac{\hat{\beta}_1}{s\sqrt{c_{11}}} = \frac{6{,}5}{\sqrt{0{,}5(\frac{11}{2})}} = 3{,}920,$$

alors que celle du test $H_0 : \beta_2 = 0$ est

$$t = \frac{\hat{\beta}_2}{s\sqrt{c_{22}}} = \frac{1{,}5}{\sqrt{0{,}5(\frac{3}{2})}} = 1{,}732.$$

À un niveau de signification de 5 %, la valeur critique de ces tests est $t_{0{,}025}(1) = 12{,}706$. Dans les deux cas, on ne rejette donc pas H_0, les variables X_1 et X_2 ne sont pas significatives dans le modèle.

d) Soit $\mathbf{x}_0 = \begin{bmatrix} 1 & 3{,}5 & 9 \end{bmatrix}$ et Y_0 la valeur de la variable dépendante correspondant à $\mathbf{x}_0$. La prévision de Y_0 donnée par le modèle trouvé en a) est

$$\begin{aligned} \hat{Y}_0 &= \mathbf{x}_0 \hat{\boldsymbol{\beta}} \\ &= -22{,}5 + 6{,}5(3{,}5) + 1{,}5(9) \\ &= 13{,}75. \end{aligned}$$

D'autre part,

$$\begin{aligned} \widehat{\text{Var}}[Y_0 - \hat{Y}_0] &= s^2(1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0') \\ &= 1{,}1875. \end{aligned}$$

Par conséquent, un intervalle de confiance à 95 % pour Y_0 est

$$\begin{aligned} E[Y_0] &\in \hat{Y}_0 \pm t_{0{,}025}(1) s \sqrt{1 + \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0'} \\ &\in 13{,}75 \pm 12{,}706\sqrt{1{,}1875} \\ &\in (-0{,}096, 27{,}596). \end{aligned}$$

3.8 a) On importe les données dans R, puis on effectue les conversions nécessaires. Comme précédemment, la variable `consommation` contient la consommation des voitures en $\ell/100$ km et la variable `poids` le poids en kilogrammes. On ajoute la variable `cylindree`, qui contient la cylindrée des voitures en litres.

```
carburant <- read.table("carburant.dat", header = TRUE)
consommation <- 235.1954/carburant$mpg
poids <- carburant$poids * 0.45455 * 1000
cylindree <- carburant$cylindree * 2.54^3/1000
```

b) La fonction `summary` fournit l'information essentielle pour juger de la validité et de la qualité du modèle :

```
fit <- lm(consommation ~ poids + cylindree)
summary(fit)

##
## Call:
## lm(formula = consommation ~ poids + cylindree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8799 -0.5595  0.1577  0.6051  1.7900
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.049304   1.098281  -2.776  0.00877 **
## poids        0.012677   0.001512   8.386 6.85e-10 ***
## cylindree   -1.122696   0.333479  -3.367  0.00186 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9156 on 35 degrees of freedom
## Multiple R-squared:  0.8927,Adjusted R-squared:  0.8866
## F-statistic: 145.6 on 2 and 35 DF,  p-value: < 2.2e-16
```

Le modèle est donc le suivant :

$$Y_t = -3{,}049 + 0{,}01268 X_{t1} + -1{,}123 X_{t2} + \varepsilon_t, \quad \boldsymbol{\varepsilon}_t \sim N(0, 0{,}9156^2 \mathbf{I})$$

où Y_t est la consommation en litres aux 100 kilomètres, X_{t1} le poids en kilogrammes et X_{t2} la cylindrée en litres. La faible valeur p du test F indique une régression globalement très significative. Les tests t des paramètres individuels indiquent également que les deux variables du modèle sont significatives. Enfin, le R^2 de 0,8927 confirme que l'ajustement du modèle est toujours bon.

c) On veut calculer un intervalle de confiance pour la consommation prévue d'une voiture de 1 350 kg ayant un moteur d'une cylindrée de 1,8 litres. On obtient, avec la fonction `predict` :

```
predict(fit, newdata = data.frame(poids = 1350, cylindree = 1.8),
        interval = "prediction")

##        fit      lwr      upr
## 1 12.04325 9.959855 14.12665
```

3.9 Il y a plusieurs réponses possibles pour cet exercice. Si l'on cherche, tel que suggéré dans l'énoncé, à distinguer les voitures sport des minifourgonnettes (en supposant que ces dernières ont moins d'accidents que les premières), alors on pourrait s'intéresser, en premier lieu, à la variable `peak.rpm`. Il s'agit du régime moteur maximal, qui est en général beaucoup plus élevé sur les voitures sport. Puisque l'on souhaite expliquer le montant total des sinistres de différents types de voitures, il devient assez naturel de sélectionner également la variable `price`, soit le prix du véhicule. Un véhicule plus luxueux coûte en général plus cher à faire réparer à dommages égaux. Voyons l'effet de l'ajout, pas à pas, de ces deux variables au modèle précédent ne comportant que la variable `horsepower` :

```
autoprice <- read.table("data/auto-price.dat", header = TRUE)
fit1 <- lm(losses ~ horsepower + peak.rpm, data = autoprice)
summary(fit1)
```

```
##
## Call:
## lm(formula = losses ~ horsepower + peak.rpm, data = autoprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.973 -24.074  -6.373  18.049 130.301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.521414  29.967570   0.184 0.854060
## horsepower   0.318477   0.086840   3.667 0.000336 ***
## peak.rpm     0.016639   0.005727   2.905 0.004205 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.44 on 156 degrees of freedom
## Multiple R-squared:  0.1314,Adjusted R-squared:  0.1203
## F-statistic:  11.8 on 2 and 156 DF,  p-value: 1.692e-05
```

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: losses
##             Df Sum Sq Mean Sq F value    Pr(>F)
## horsepower   1  16949 16948.5 15.1573 0.0001463 ***
## peak.rpm     1   9437  9437.0  8.4397 0.0042049 **
## Residuals  156 174435  1118.2
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La variable `peak.rpm` est significative, mais le R^2 demeure faible. Ajoutons maintenant la variable `price` au modèle :

```
fit2 <- lm(losses ~ horsepower + peak.rpm + price, data = autoprice)
summary(fit2)
```

```
##
## Call:
## lm(formula = losses ~ horsepower + peak.rpm + price, data = autoprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.745 -25.214  -5.867  18.407 130.032
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6972172 31.3221462  -0.022  0.98227
## horsepower   0.2414922  0.1408272   1.715  0.08838 .
## peak.rpm     0.0181386  0.0061292   2.959  0.00357 **
## price        0.0005179  0.0007451   0.695  0.48803
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.49 on 155 degrees of freedom
## Multiple R-squared:  0.1341,Adjusted R-squared:  0.1173
## F-statistic: 8.001 on 3 and 155 DF,  p-value: 5.42e-05
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: losses
##             Df Sum Sq Mean Sq F value    Pr(>F)
## horsepower   1  16949 16948.5 15.1071 0.0001502 ***
## peak.rpm     1   9437  9437.0  8.4118 0.0042702 **
## price        1    542   542.1  0.4832 0.4880298
## Residuals  155 173893  1121.9
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Du moins avec les variables `horsepower` et `peak.rpm`, la variable `price` n'est pas significative. D'ailleurs, l'augmentation du R^2 suite à l'ajout de cette variable est minime. À ce stade de l'analyse, il vaudrait sans doute mieux reprendre tout depuis le début avec d'autres variables. Des méthodes de sélection des variables seront étudiées plus avant dans le chapitre.

3.10 a) On a $p = 3$ variables explicatives et, du nombre de degrés de liberté de la statistique F, on apprend que $n - p - 1 = 16$. Par conséquent, $n = 16 + 3 + 1 = 20$. Les dimensions des vecteurs et de la matrice de schéma dans la représentation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ sont donc : $n \times 1 = 20 \times 1$ pour les vecteurs $\mathbf{y}$ et $\boldsymbol{\varepsilon}$, $n \times (p+1) = 20 \times 4$ pour la matrice $\mathbf{X}$, $(p+1) \times 1$ pour le vecteur $\boldsymbol{\beta}$.

b) La valeur p associée à la statistique F est, à toute fin pratique, nulle. Cela permet de rejeter facilement l'hypothèse nulle selon laquelle la régression n'est pas significative.

c) On doit se fier ici au résultat du test t associé à la variable X_2. Dans les résultats obtenus avec R, on voit que la valeur p de la statistique t du paramètre β_2 est 0,0916. Cela signifie que jusqu'à un seuil de signification de 9,16 % (ou un niveau de confiance supérieur à 90,84 %), on ne peut rejeter l'hypothèse $H_0 : \beta_2 = 0$ en faveur de $H_1 : \beta_2 \neq 0$. Il s'agit néanmoins d'un cas limite et il est alors du ressort de l'analyste de décider d'inclure ou non le revenu disponible dans le modèle.

d) Le coefficient de détermination est de $R^2 = 0{,}981$. Cela signifie que le prix de la bière, le revenu disponible et la demande de l'année précédente expliquent plus de 98 % de la variation de la demande en bière. L'ajustement du modèle aux données est donc particulièrement bon. Il est tout à fait possible d'obtenir un R^2 élevé et, simultanément, toutes les statistiques t non significatives : comme chaque test t mesure l'impact d'une variable sur la régression étant donné la présence des autres variables, il suffit d'avoir une bonne

variable dans un modèle pour obtenir un R^2 élevé et une ou plusieurs autres variables redondantes avec la première pour rendre les tests t non significatifs.

3.11 a) L'information demandée doit évidemment être extraite des deux tableaux d'analyse de variance fournis dans l'énoncé. Il importe, ici, de savoir que le résultat de la fonction `anova` de R est un tableau d'analyse de variance séquentiel, où chaque ligne identifiée par le nom d'une variable correspond au test F partiel résultant de l'ajout de cette variable au modèle. Ainsi, du premier tableau on obtient les sommes de carrés

$$\begin{aligned} \text{SSR}(X_2) &= 45{,}59085 \\ \text{SSR}(X_3|X_2) &= 8{,}76355 \end{aligned}$$

alors que du second tableau on a

$$\begin{aligned} \text{SSR}(X_1) &= 45{,}59240 \\ \text{SSR}(X_2|X_1) &= 0{,}01842 \\ \text{SSR}(X_3|X_1, X_2) &= 8{,}78766, \end{aligned}$$

ainsi que

$$\begin{aligned} \text{MSE} &= \frac{SSE(X_1, X_2, X_3)}{n-p-1} \\ &= 0{,}44844. \end{aligned}$$

i) Le test d'hypothèse $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ est le test global de validité du modèle. La statistique F pour ce test est

$$\begin{aligned} F &= \frac{\text{SSR}(X_1, X_2, X_3)/3}{\text{MSE}} \\ &= \frac{(\text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1,X_2))/3}{\text{MSE}} \\ &= \frac{(45{,}5924 + 0{,}01842 + 8{,}78766)/3}{0{,}44844} \\ &= 40{,}44. \end{aligned}$$

Puisque la statistique MSE a 21 degrés de liberté, la statistique F en a 3 et 21.

ii) Pour tester cette hypothèse, il faut utiliser un test F partiel. On teste si la variable X_1 est significative dans la régression globale. La statistique du test est alors

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_1|X_2,X_3)/1}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_2, X_3)}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_2) - \text{SSR}(X_3|X_2)}{\text{MSE}} \\ &= \frac{54{,}39848 - 45{,}59085 - 8{,}76355}{0{,}44844} \\ &= 0{,}098, \end{aligned}$$

avec 1 et 21 degrés de liberté.

iii) Cette fois, on teste si les variables X_2 et X_3 (les deux ensemble) sont significatives dans la régression globale. On effectue donc encore un test F partiel avec la statistique

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_2, X_3|X_1)/2}{\text{MSE}} \\ &= \frac{(\text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1))/2}{\text{MSE}} \\ &= \frac{(54{,}39848 - 45{,}5924)/2}{0{,}44844} \\ &= 9{,}819, \end{aligned}$$

avec 2 et 21 degrés de liberté.

b) À la lecture du premier tableau d'analyse de variance que tant les variables X_2 que X_3 sont significatives dans le modèle. Par contre, comme on le voit dans le second tableau, la variable X_2 devient non significative dès lors que la variable X_1 est ajoutée au modèle. (L'impact de la variable X_3 demeure, lui, inchangé.) Cela signifie que les variables X_1 et X_2 sont redondantes et qu'il faut choisir l'une ou l'autre, mais pas les deux. Par conséquent, les choix de modèle possibles sont X_1 et X_3, ou X_2 et X_3.

3.12 La statistique à utiliser pour faire ce test F partiel est

$$\begin{aligned} F^* &= \frac{\text{SSR}(X_2, X_3|X_1, X_4)/2}{\text{MSE}} \\ &= \frac{\text{SSR}(X_1, X_2, X_3, X_4) - \text{SSR}(X_1, X_4)}{2\text{MSE}} \\ &= \frac{\text{SSR} - \text{SSR}(X_4) - \text{SSR}(X_1|X_4)}{2s^2} \end{aligned}$$

où $\text{SSR} = \text{SSR}(X_1,X_2,X_3,X_4)$. Or,

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{\text{SSR}}{\text{SSR} + \text{SSE}}, \end{aligned}$$

d'où

$$\begin{aligned} \text{SSR} &= \frac{R^2}{1 - R^2}\,\text{SSE} \\ &= \frac{R^2}{1 - R^2}\,\text{MSE}(n - p - 1) \\ &= \frac{0{,}6903}{1 - 0{,}6903}\,(26{,}41)(506 - 4 - 1) \\ &= 29\,492. \end{aligned}$$

Par conséquent,

$$\begin{aligned} F^* &= \frac{29\,492 - 2\,668 - 21\,348}{(2)(26{,}41)} \\ &= 103{,}67. \end{aligned}$$

3.13 a) Tout d'abord, si $Z \sim N(0,1)$ et $V \sim \chi^2(r)$ alors, par définition,

$$\frac{Z}{\sqrt{V/r}} \sim t(r).$$

Tel que mentionné dans l'énoncé, $\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$ ou, de manière équivalente,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1).$$

Par conséquent,

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}}{\sqrt{\frac{\text{SSE}}{\sigma^2(n-p-1)}}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t(n-p-1).$$

b) En régression linéaire simple, $c_{11} = 1/\sum_{t=1}^{n}(X_t - \bar{X})^2 = 1/S_{XX}$ et $\sigma^2 c_{11} = \text{var}[\hat{\beta}_1]$. Le résultat général en a) se réduit donc, en régression linéaire simple, au résultat bien connu du test t sur le paramètre β_1

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{1/S_{XX}}} \sim t(n-1-1).$$

3.14 En suivant les indications donnée dans l'énoncé, on obtient aisément

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2\left(\frac{d}{d\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)' \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2(\mathbf{X}'\mathbf{W}\mathbf{y} - \mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

Par conséquent, les équations normales à résoudre pour trouver l'estimateur $\hat{\boldsymbol{\beta}}^*$ minimisant la somme de carrés pondérés $S(\boldsymbol{\beta})$ sont $(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}}^* = \mathbf{X}'\mathbf{W}\mathbf{y}$ et l'estimateur des moindres carrés pondérés est

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

3.15 De manière tout à fait générale, l'estimateur linéaire sans biais à variance minimale dans le modèle de régression linéaire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\text{var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{W}^{-1}$ est

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

et sa variance est, par le théorème **??**,

$$\begin{aligned}\mathbf{V}[\hat{\boldsymbol{\beta}}^*] &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{V}[\mathbf{y}]\mathbf{W}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{W}^{-1}\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\end{aligned}$$

puisque les matrices $\mathbf{W}$ et $\mathbf{X}'\mathbf{W}\mathbf{X}$ sont symétriques. Dans le cas de la régression linéaire simple passant par l'origine et en supposant que $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$, ces formules se réduisent en

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

et

$$\text{var}[\hat{\beta}^*] = \frac{\sigma^2}{\sum_{t=1}^{n} w_t X_t^2}.$$

a) Cas déjà traité à l'exercice 2.6 où $\mathbf{W} = \mathbf{I}$ et, donc,

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

et

$$\text{var}[\hat{\beta}^*] = \frac{\sigma^2}{\sum_{t=1}^n w_t X_t^2}.$$

b) Cas général traité ci-dessus.

c) Si $\text{var}[\varepsilon_t] = \sigma^2 X_t$, alors $w_t = X_t^{-1}$. Le cas général se simplifie donc en

$$\begin{aligned}\hat{\beta}^* &= \frac{\sum_{t=1}^n Y_t}{\sum_{t=1}^n X_t} \\ &= \frac{\bar{Y}}{\bar{X}}, \\ \text{var}[\hat{\beta}^*] &= \frac{\sigma^2}{\sum_{t=1}^n X_t} \\ &= \frac{\sigma^2}{n\bar{X}}.\end{aligned}$$

d) Si $\text{var}[\varepsilon_t] = \sigma^2 X_t^2$, alors $w_t = X_t^{-2}$. On a donc

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

3.16 Le graphique des valeurs de Y en fonction de celles de X, à la figure A.16, montre clairement une relation quadratique. On postule donc le modèle

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2).$$

Par la suite, on peut estimer les paramètres de ce modèle avec la fonction `lm` de R :

```
fit <- lm(Y ~ poly(X, 2), data = donnees)
summary(fit)

##
## Call:
## lm(formula = Y ~ poly(X, 2), data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9123 -0.6150 -0.1905  0.6367  1.6921
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.1240     0.3025   59.91 3.10e-16 ***
```

```
plot(Y ~ X, data = donnees)
```

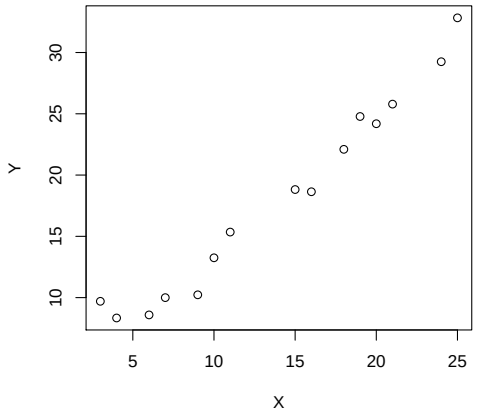


FIG. A.16 – Graphique des données de l'exercice 3.16

```
## poly(X, 2)1  29.6754     1.1717   25.33 8.72e-12 ***
## poly(X, 2)2   4.0899     1.1717    3.49  0.00446 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.172 on 12 degrees of freedom
## Multiple R-squared:  0.982,Adjusted R-squared:  0.979
## F-statistic: 326.8 on 2 and 12 DF,  p-value: 3.434e-11
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## poly(X, 2)  2 897.36  448.68  326.79 3.434e-11 ***
## Residuals  12  16.48    1.37
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tant le test F global que les tests t individuels sont concluants, le coefficient de détermination est élevé et l'on peut constater à la figure A.17 que l'ajustement du modèle est bon. On conclut donc qu'un modèle adéquat pour cet ensemble de données est

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

3.17 Comme on peut le constater à la figure A.18, le point (X_{16}, Y_{16}) est plus éloigné des autres. En b) et c), on diminue son poids dans la régression.

```
plot(Y ~ X, data = donnees)
x <- seq(min(donnees$X), max(donnees$X), length = 200)
lines(x, predict(fit, data.frame(X = x), lwd = 2))
```

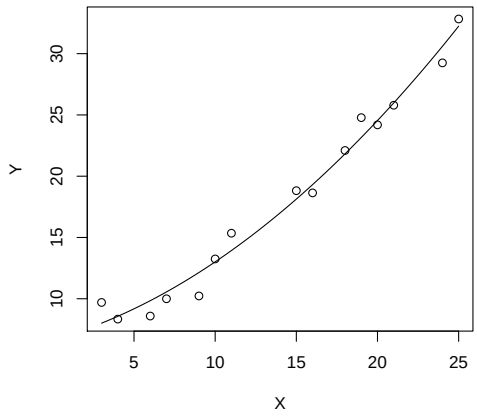


FIG. A.17 – Graphique des données de l'exercice 3.16 et courbe obtenue par régression

```
plot(Y ~ X, data = donnees)
points(donnees$X[16], donnees$Y[16], pch = 16)
```

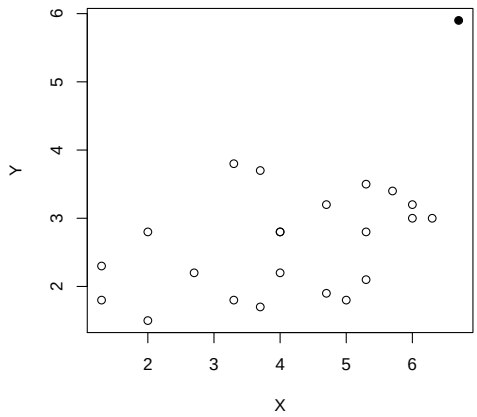


FIG. A.18 – Graphique des données de l'exercice 3.17. Le cercle plein représente la donnée (X_{16}, Y_{16}).

a) On calcule d'abord l'estimateur des moindres carrés ordinaires :

```
(fit1 <- lm(Y ~ X, data = donnees))

##
## Call:
## lm(formula = Y ~ X, data = donnees)
##
## Coefficients:
## (Intercept)            X
##      1.4256       0.3158
```

b) Si l'on suppose que la variance de la données (X_{16}, Y_{16}) est quatre fois plus élevée que la variance des autres données, alors il convient d'accorder un point quatre fois moins grand à cette donnée dans la régression. Cela requiert les moindres carrés pondérés. Pour calculer les estimateurs avec `lm` dans R, on utilise l'argument `weights` :

```
w <- rep(1, nrow(donnees))
w[16] <- 0.25
(fit2 <- update(fit1, weights = w))

##
## Call:
## lm(formula = Y ~ X, data = donnees, weights = w)
##
## Coefficients:
## (Intercept)            X
##      1.7213       0.2243
```

c) On répète la procédure en b) avec un poids de encore plus petit pour la donnée (X_{16}, Y_{16}) :

```
w[16] <- 0.0625
(fit3 <- update(fit1, weights = w))

##
## Call:
## lm(formula = Y ~ X, data = donnees, weights = w)
##
## Coefficients:
## (Intercept)            X
##      1.8080       0.1975
```

Plus le poids accordé à la donnée (X_{16}, Y_{16}) est faible, moins la droite de régression est attirée vers ce point (voir la figure A.19).

3.18 a) Voir la figure A.20 pour le graphique. Il y a effectivement une différence entre la consommation de carburant des hommes et des femmes : ces dernières font plus de milles avec un gallon d'essence.

b) Remarquer que la variable `sexe` est un facteur et peut être utilisée telle quelle dans `lm` :

```
(fit <- lm(mpg ~ age + sexe, data = donnees))

##
## Call:
## lm(formula = mpg ~ age + sexe, data = donnees)
```

```
plot(Y ~ X, data = donnees)
points(donnees$X[16], donnees$Y[16], pch = 16)
abline(fit1, lwd = 2, lty = 1)
abline(fit2, lwd = 2, lty = 2)
abline(fit3, lwd = 2, lty = 3)
legend(1.2, 6, legend = c("Modèle a)", "Modèle b)", "Modèle c)"),
       lwd = 2, lty = 1:3)
```

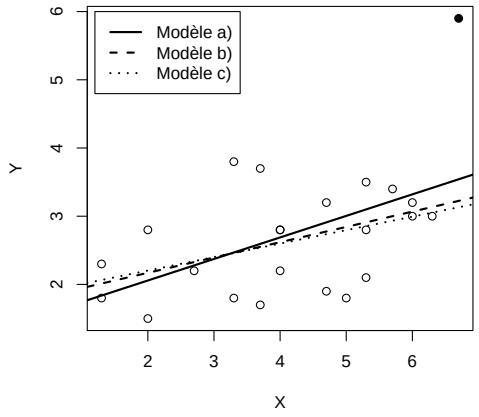


FIG. A.19 – Graphique des données de l'exercice 3.17 avec les droites de régression obtenues à l'aide des moindres carrés pondérés.

```
##
## Coefficients:
## (Intercept)          age         sexeM
##      16.687       -1.040       -1.206
```

c) Calcul d'une prévision pour la valeur moyenne de la variable `mpg` :

```
predict(fit, newdata = data.frame(age = 4, sexe = "F"),
        interval = "confidence", level = 0.90)

##        fit      lwr      upr
## 1 12.52876 11.94584 13.11168
```

3.19 a) Le postulat de normalité semble violé.
La distribution des résidus a une queue inférieure plus épaisse que la loi normale, ce que l'on voit à gauche du Q-Q plot, puisque les poits ne sont pas alignés.
Le postulat de normalité n'est pas critique, parce que les estimateurs des moindres carrés ont un sens quand même. Toutefois, les tests d'hypothèses et les intervalles de confiance ne sont pas valides.

b) Le graphique des résidus en fonction de x_2 montre que le postulat de linéarité semble violé. Cela implique que le modèle n'est pas valide.
On observe de l'hétéroscédasticité (par exemple, dans les graphiques 1, 3 ou 4) puisque les résidus ne semblent pas avoir une variance constante.

```
hommes <- subset(donnees, sexe == "M")
femmes <- subset(donnees, sexe == "F")
plot(mpg ~ age, data = hommes,
     xlim = range(donnees$age), ylim = range(donnees$mpg))
points(mpg ~ age, data = femmes, pch = 16)
legend(4, 16, legend = c("Hommes", "Femmes"), pch = c(1, 16))
```

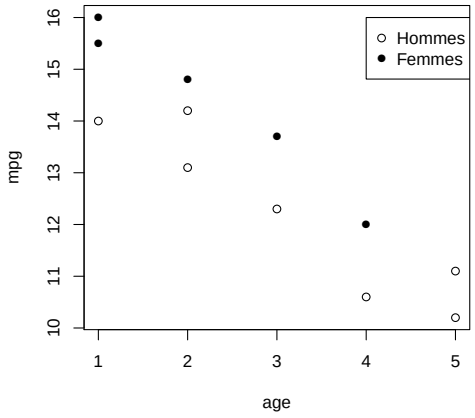


FIG. A.20 – Graphique des données de l'exercice 3.18

Cela signifie que les variances des paramètres ne sont pas calculées de façon appropriée OU il faudrait effectuer une transformation sur les variables pour régler ces problèmes.

3.20 On pourrait croire qu'un point sur 20, ça ne change rien, mais ce n'est pas le cas ! Le point 1 a un impact sur la pente et la qualité de l'ajustement. Le point 2 a un grand levier mais n'affecte pas beaucoup les estimations, le point 3 a un grand levier et un gros impact.

```
dat <- read.csv("OutlierExample.csv")

dim(dat)

summary(dat)

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)

fit0 <- lm(Y~X,dat,subset=(CODES==0))
summary(fit0)
plot(dat[,1:2],pch=16)
points(dat[match(1:3,dat$CODES),1:2],col=2:4,pch=16:18,cex=1.2)
abline(fit0)

fit1 <- lm(Y~X,dat,subset=(CODES<=1))
```

```
summary(fit1)
abline(fit1,col=2,lty=2)

fit2 <- lm(Y~X,dat,subset=(CODES%in%c(0,2)))
summary(fit2)
abline(fit2,col=3,lty=3)

fit3 <- lm(Y~X,dat,subset=(CODES%in%c(0,3)))
summary(fit3)
abline(fit3,col=4,lty=4)

influence.measures(fit0)
influence.measures(fit1)
influence.measures(fit2)
influence.measures(fit3)
```

Chapitre ??

4.1 a) i) modèle D
ii) modèle D
iii) modèle G
iv) modèle G
v) modèle C
vi) modèle H

b) Il y a un très gros problème de multicolinéarité pour les modèle F, G et H, car certains VIFs sont beaucoup plus grands que 10. Ce problème augmente inutilement la variance des paramètres estimés.

c) On évite les modèles F G et H pour ne pas avoir de problème de multicolinéarité. Le modèle D est préférable selon les critères PRESS et R^2_p. De plus, ses critères AIC et BIC sont les deuxièmes plus petits. Le C_p est 8, donc 8-5=3. Ce n'est pas parfait, mais ce n'est pas si mal, etc.

4.2 a) Puisque $n = p$, $\beta_0 = 0$ et que la matrice d'incidence est diagonale, on a $\hat{y}_i = \hat{\beta}_i$ pour $i = 1,\ldots,n$. On minimise $S(\beta) = \sum_{i=1}^n (y_i - \beta_i)^2$ et on trouve pour $i \in \{1,\ldots,n\}$,

$$\frac{\partial}{\partial \beta_i} S(\beta)\bigg|_{\hat{\beta}_i} = -2(y_i - \hat{\beta}_i) = 0 \quad \Rightarrow \quad \hat{\beta}_i = y_i.$$

b) On minimise, pour une valeur $\lambda > 0$,

$$S^{\text{ridge}}(\beta) = \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^n \beta_i^2.$$

c) On a

$$\frac{\partial}{\partial \beta_i} S^{\text{ridge}}(\beta) = -2(y_i - \beta_i) + 2\lambda\beta_i.$$

On pose égal à 0 et on trouve

$$y_i - \hat{\beta}_i^{\text{ridge}} = \lambda \hat{\beta}_i^{\text{ridge}} \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = \frac{y_i}{1+\lambda}.$$

d) On minimise, pour une valeur $\lambda > 0$,

$$S^{\text{lasso}}(\beta) = \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^n |\beta_i|.$$

e) On a

$$\frac{\partial}{\partial \beta_i} S^{\text{lasso}}(\beta) = -2(y_i - \beta_i) + \lambda \,\text{signe}(\beta_i).$$

On utilise les EMV trouvés en a) pour définir le signe. Supposons d'abord que $\hat{\beta}_i = y_i > 0$. Alors, on a aussi $\hat{\beta}_i^{\text{lasso}} > 0$ (sinon, changer le signe donnera une valeur plus petite de l'équation à minimiser). On pose la dérivée égale à 0 et on trouve

$$2(y_i - \hat{\beta}_i^{\text{lasso}}) = \lambda \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = y_i - \lambda/2,$$

ce qui tient seulement si $\hat{\beta}_i^{\text{lasso}} > 0$, alors on a $\hat{\beta}_i^{\text{ridge}} = \max(0, y_i - \lambda/2)$. Supposons ensuite que $\hat{\beta}_i = y_i < 0$. Alors, on a aussi $\hat{\beta}_i^{\text{lasso}} < 0$. On pose la dérivée égale à 0 et on trouve

$$2(y_i - \hat{\beta}_i^{\text{lasso}}) = -\lambda \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = y_i + \lambda/2,$$

sous la contrainte que ce soit négatif, donc dans ce cas, $\hat{\beta}_i^{\text{ridge}} = \min(0, y_i + \lambda/2)$. On combine les deux cas et on obtient l'équation donnée.

f) On peut voir que la façon de rapetisser les paramètres est bien différente pour les deux méthodes. Avec ridge, chaque coefficient des moindres carrés est réduit par la même proportion. Avec lasso, chaque coefficient des moindres carrés est réduit vers 0 d'un montant constant $\lambda/2$; ceux qui sont plus petits que $\lambda/2$ en valeur absolue sont mis exactement égaux à 0. C'est de cette façon que le lasso permet de faire la sélection des variables explicatives.

Chapitre ??

5.1 a) Normale(μ,σ^2) : oui,

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi\sigma^2)^1/2}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), y \in \mathbb{R} \\ &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2}\right), y \in \mathbb{R}. \end{aligned}$$

- Paramètre canonique : $\theta = \mu$
- Paramètre de dispersion : $\phi = \sigma^2$
- $b(\theta) = \frac{\theta^2}{2}$
- $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\frac{\theta^2}{2} = \theta = \mu$
- var[(]Y) $= \phi\ddot{b}(\theta) = \sigma^2\frac{\partial}{\partial\theta}\theta = \sigma^2$
- $V(\mu) = 1$.

b) Uniforme(0,β) : non. Le domaine dépend du paramètre β.

c) Poisson(λ) :

$$\begin{aligned} f_Y(y;\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!}, \text{ pour } y \in \mathbb{N}^+ \\ &= \exp\{y\ln\lambda - \lambda - \ln y!\} \\ f_Y(y;\theta,\phi) &= \exp\left\{\frac{y\theta - e^\theta}{\phi} - \ln y!\right\}. \end{aligned}$$

- Paramètre canonique : $\theta = \ln\lambda$
- Paramètre de dispersion : $\phi = 1$
- $b(\theta) = e^\theta$
- $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}e^\theta = e^\theta = \lambda$
- var[(]Y) $= \phi\ddot{b}(\theta) = \frac{\partial}{\partial\theta}e^\theta = e^\theta = \lambda$
- $V(\mu) = \mu$.

d) Bernoulli(π)

$$\begin{aligned} f_Y(y;\pi) &= \pi^y(1-\pi)^{1-y}1(y \in \{0,1\}) \\ &= \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\}1(y \in \{0,1\}). \end{aligned}$$

- Paramètre canonique : $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
- Paramètre de dispersion : $\phi = 1$
- $b(\theta) = \ln(1+e^\theta)$
- $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\ln(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi$
- var[(]Y) $= \phi\ddot{b}(\theta) = \frac{\partial}{\partial\theta}\frac{e^\theta}{1+e^\theta} = \frac{e^\theta}{(1+e^\theta)^2} = \pi(1-\pi)$
- $V(\mu) = \mu(1-\mu)$.

e) Binomiale(m,π), $m > 0$ est un entier et est connu.

$$\begin{aligned} f_Y(y;\pi) &= \binom{m}{y}\pi^y(1-\pi)^{m-y}1(y \in \{0,1,...,m\}) \\ &= \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{y}\right\}1(y \in \{0,1,...,m\}). \end{aligned}$$

Dans cette représentation, on a

$$E[Y] = m\pi \text{ et } Var(Y) = m\pi(1-\pi).$$

Cette forme est moins utilisée car l'espérance de Y dépend de m, le paramètre de dispersion. Souvent, on transforme les données. On utilise plutôt $Y^* = Y/m$. Alors, pour ces données transformées,

$$\begin{aligned} f_{Y^*}(y;\pi) &= \exp\left\{my\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{my}\right\}, y \in \{0,1/m,...,1\} \\ &= \exp\left\{\frac{y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)}{1/m} + \ln\binom{m}{my}\right\}, y \in \{0,1/m,...,1\}. \end{aligned}$$

- Paramètre canonique : $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
- Paramètre de dispersion : $\phi = 1/m$
- $b(\theta) = \ln(1+e^\theta)$
- $E[Y^*] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\ln(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi$
- $\text{var}[(]Y^*) = \phi\ddot{b}(\theta) = \frac{1}{m}\frac{\partial}{\partial\theta}\frac{e^\theta}{1+e^\theta} = \frac{e^\theta}{m(1+e^\theta)^2} = \frac{\pi(1-\pi)}{m}$
- $V(\mu) = \mu(1-\mu)$.

f) Pareto(α,λ) : non.

g) Gamma(α,β) Soit $Y \sim Gamma(\alpha,\beta)$. Alors, avec un peu de travail, la densité peut être écrite sous la forme exponentielle linéaire.

$$f_Y(y;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y},$$

pour $y > 0$. On reparamétrise : $\mu = \alpha/\beta = E[Y]$ et α, on a donc $\beta = \alpha/\mu$ et

$$f_Y(y;\alpha,\mu) = \frac{1}{y\Gamma(\alpha)}\left(\frac{\alpha y}{\mu}\right)^\alpha \exp\left\{-\frac{\alpha y}{\mu}\right\}.$$

Posons $\theta = -1/\mu$, et $a(\phi) = 1/\alpha$, alors on trouve

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta + \ln(-\theta)}{\phi} + \alpha\ln\alpha + (\alpha-1)\ln y - \ln\Gamma(\alpha)\right\}.$$

Donc, $b(\theta) = -\ln(-\theta)$ et $a(\phi) = 1/\alpha \Rightarrow \dot{b}(\theta) = \frac{-1}{\theta} = \mu$ et $\ddot{b}(\theta) = \frac{1}{\theta^2} = \mu^2$. Finalement,

$$E[Y] = \frac{-1}{\theta} = \mu \text{ et } Var(Y) = \frac{1}{\alpha}\mu^2.$$

h) Binomiale négative(r,π) avec r connu. On considère $Y^* = Y/r$:

$$\begin{aligned} f_Y^*(y) &= \binom{r+ry-1}{ry} \pi^r (1-\pi)^{ry}, \text{ pour } y \in \{0, \frac{1}{r}, \frac{2}{r}, \ldots\} \\ &= \exp\left(ry \ln(1-\pi) + r\ln\pi + \ln\binom{r+ry-1}{ry} \right). \end{aligned}$$

- Paramètre canonique : $\theta = \ln(1-\pi)$
- Paramètre de dispersion : $\phi = 1/r$
- $b(\theta) = -\ln(1-e^\theta)$
- $E[Y^*] = \dot{b}(\theta) = \frac{\partial}{\partial\theta} - \ln(1-e^\theta) = \frac{e^\theta}{1-e^\theta} = \frac{1-\pi}{\pi}$
- $\text{var}[(]Y^*) = \phi\ddot{b}(\theta) = \frac{1}{r}\frac{\partial}{\partial\theta}\frac{e^\theta}{1-e^\theta} = \frac{e^\theta}{r(1-e^\theta)^2} = \frac{(1-\pi)}{r\pi^2}$
- $V(\mu) = \mu(\mu+1)$.

5.2 Le lien canonique est le lien log : $\eta = \ln(\mu)$. On pourrait aussi utiliser d'autres fonctions de lien, telle que le lien identité $\eta = \mu$, le lien inverse $\eta = \frac{1}{\mu}$, mais le lien log est le plus approprié parce que son utilisation garantit une moyenne μ positive, ce qui est nécessaire pour la loi de Poisson.

5.3 Le lien canonique pour la loi Gamma est le lien inverse $\eta = 1/\mu$. Comme la moyenne d'une loi Gamma est toujours positive, ce lien n'est pas toujours approprié parce qu'il ne restreint pas le domaine de μ aux réels positifs. Le lien log serait plus approprié dans certains cas.

5.4 a) $\eta = g(\mu) = \ln(\mu)$

b) On a

$$\ln(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i.$$

La densité de la loi Poisson est

$$\begin{aligned} f_{Y_i}(y_i;\mu_i) &= \exp\left(y_i \ln\mu_i - \mu_i - \ln y_i!\right) \\ f_{Y_i}(y_i;\beta_0,\beta_1) &= \exp\left(y_i(\beta_0+\beta_1 x_i) - e^{\beta_0+\beta_1 x_i} - \ln y_i!\right). \end{aligned}$$

La fonction de vraisemblance et la log-vraisemblance sont donc :

$$\begin{aligned} \mathcal{L}(\beta_0,\beta_1) &= \prod_{i=1}^n f_{Y_i}(y_i;\beta_0,\beta_1) = \prod_{i=1}^n \exp\left(y_i(\beta_0+\beta_1 x_i) - e^{\beta_0+\beta_1 x_i} - \ln y_i!\right) \\ \ell(\beta_0,\beta_1) &= \sum_{i=1}^n y_i(\beta_0+\beta_1 x_i) - e^{\beta_0+\beta_1 x_i} + \text{constante}. \end{aligned}$$

On maximise la log-vraisemblance :

$$\begin{aligned} \frac{\partial}{\partial\beta_0}\ell(\beta_0,\beta_1) &= \sum_{i=1}^n y_i - e^{\beta_0+\beta_1 x_i} \\ \frac{\partial}{\partial\beta_1}\ell(\beta_0,\beta_1) &= \sum_{i=1}^n y_i x_i - x_i e^{\beta_0+\beta_1 x_i} \end{aligned}$$

Donc, les équations à résoudre sont

$$\sum_{i=1}^{n} y_i - e^{\beta_0+\beta_1 x_i} = 0$$
$$\sum_{i=1}^{n} x_i(y_i - e^{\beta_0+\beta_1 x_i}) = 0.$$

5.5 La déviance est

$$D(y;\hat{\mu}) = 2(\ell_n(\tilde{\theta}) - \ell_n(\hat{\theta})).$$

Pour le modèle Binomial, on a que

$$\ell_n(\theta) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)}{1/m_i}.$$

Alors, dans le modèle complet, $\mu_i = y_i$ et on trouve

$$\ell_n(\tilde{\theta}) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{y_i}{1-y_i}\right) + \ln(1-y_i)}{1/m_i}.$$

Dans le modèle développé avec le lien log, $\mu_i = \hat{\mu}_i$ et on trouve

$$\ell_n(\hat{\theta}) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + \ln(1-\hat{\mu}_i)}{1/m_i}.$$

Finalement, la déviance est

$$\begin{aligned} D(y;\hat{\mu}) &= \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{y_i}{1-y_i}\right) + \ln(1-y_i)}{1/m_i} - \frac{y_i \ln\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + \ln(1-\hat{\mu}_i)}{1/m_i} \\ &= \sum_{i=1}^{n} m_i \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right]. \end{aligned}$$

5.6 Pour la distribution Gamma, on a $V(t) = t^2$ et $b(t) = -\ln(-t)$.
Résidus de Pearson :

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i^2}} = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

Résidus d'Anscombe :

$$\begin{aligned} A(t) &= \int_0^t \frac{\mathrm{d}s}{s^{2/3}} = 3t^{1/3} \\ \dot{A}(t) &= \frac{1}{s^{2/3}} \\ r_{A_i} &= \frac{A(y_i) - A(\hat{\mu}_i)}{\dot{A}(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}} = \frac{3(y_i^{1/3} - \hat{\mu}_i^{1/3})}{\hat{\mu}_i^{1/3}}. \end{aligned}$$

Résidus de déviance :

$$\begin{aligned} D_i &= 2\left(-\frac{y_i}{y_i} - \ln(y_i) + \frac{y_i}{\hat{\mu}_i} + \ln(\hat{\mu}_i)\right) \\ &= 2\left(\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) \\ r_{D_i} &= sign(y_i - \hat{\mu}_i)\sqrt{2\left(\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)}. \end{aligned}$$

5.7 Cette solution est en anglais, vous pouvez poser vos questions sur le forum, s'il y a lieu.
This is a two-factor model, «Device» takes three levels (M1, M2 and M3) and «Stress» takes 4 levels. The baseline group is M1 device at stress level I. An analysis of deviance is carried out to assess if the parameters for the devices are significant.

```
glm <- glm(Failures~Level*Machine,family=poisson,data=stresstest)
anova(glm)

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Failures
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev
## NULL                             11     35.844
## Level          3  20.8567         8     14.987
## Machine        2  12.2154         6      2.772
## Level:Machine  6   2.7719         0      0.000

qchisq(0.95, 6)

## [1] 12.59159

qchisq(0.95, 2)

## [1] 5.991465
```

The model

Stress+Device+Stress.Device

is fitted first. The change in deviance from the simpler model Stress+Device is 2.7719 on 6 degrees of freedom, which is not significant when compared to $\chi^2_{(6,0.95)} = 12.59$. Hence, the model Stress+Device is an adequate simplification of the more complex model. If we then test for the significance of the Device parameters, we find that the change in deviance from the simpler model Stress is 12.2154 on 2 degrees of freedom, which is significant because $\chi^2_{(2,0.95)} = 5.99$. From this analysis, we can conclude that there is a significant difference between the failure rates of the different devices.

5.8 a) En R, on obtient

```
modinv <- glm(AvCost~OwnerAge+Model+CarAge,family=Gamma,data=Bcar)
summary(modinv)
```

```
##
## Call:
## glm(formula = AvCost ~ OwnerAge + Model + CarAge, family = Gamma,
##     data = Bcar)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.85536  -0.13930  -0.00821   0.07444   1.49969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0033233  0.0004038   8.230 4.42e-13
## OwnerAge21-24  0.0006043  0.0004159   1.453  0.14908
## OwnerAge25-29  0.0003529  0.0003933   0.897  0.37163
## OwnerAge30-34  0.0011783  0.0004572   2.577  0.01130
## OwnerAge35-39  0.0016372  0.0004990   3.281  0.00139
## OwnerAge40-49  0.0012039  0.0004592   2.622  0.01000
## OwnerAge50-59  0.0010998  0.0004511   2.438  0.01638
## OwnerAge60+    0.0012390  0.0004619   2.682  0.00845
## ModelB        -0.0002817  0.0004049  -0.696  0.48806
## ModelC        -0.0006502  0.0003906  -1.664  0.09893
## ModelD        -0.0018235  0.0003481  -5.239 7.96e-07
## CarAge10+      0.0033776  0.0004747   7.115 1.24e-10
## CarAge4-7      0.0003393  0.0002723   1.246  0.21539
## CarAge8-9      0.0017423  0.0003575   4.873 3.75e-06
##
## (Intercept)   ***
## OwnerAge21-24
## OwnerAge25-29
## OwnerAge30-34 *
## OwnerAge35-39 **
## OwnerAge40-49 **
## OwnerAge50-59 *
## OwnerAge60+   **
## ModelB
## ModelC        .
## ModelD        ***
## CarAge10+     ***
## CarAge4-7
## CarAge8-9     ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1074529)
##
##     Null deviance: 27.841  on 122  degrees of freedom
## Residual deviance: 11.511  on 109  degrees of freedom
```

```
##   (5 observations deleted due to missingness)
## AIC: 1400.7
##
## Number of Fisher Scoring iterations: 5
```

b) On a utilisé un lien inverse, alors $E[Y_i] = \frac{1}{\eta_i}$. Puisque les variables explicatives prennent toutes leur niveau de base, on a que $\hat{\eta}_i = \hat{\beta}_0 = 0.0033233$ et

$$\widehat{E[Y_i]} = 0.0033233^{-1} = 300.91.$$

c) Puisqu'on a utilisé un lien inverse, un coefficient plus élevé implique une diminution de l'espérance du coût de la réclamation, alors qu'un coefficient négatif signifie une augmentation de cette espérance. Ici on observe que les sept coefficients sont positifs, alors la catégorie d'âge ayant une espérance de coût la plus élevée est la catégorie de base, 17-20 ans. Le coût moyen semble ensuite relativement élevé pour les jeunes entre 21 et 29 ans. La catégorie d'âge avec coût de réclamation minimal est 35-39 ans, puis la moyenne semble relativement stable pour les détenteurs de police plus âgés.

d) Les trois coefficients pour la variable modèle sont négatifs, ce qui signifie que les réclamations pour les véhicules de type A (niveau de base) sont moins élevées en moyenne que celles pour les autres types de véhicule. Les réclamatios pour les véhicules du modèle D semblent particulièrement coûteuse car le coefficient est beaucoup plus grand en valeur absolue que les autres.

e) De la même façon, on observe que d'augmenter l'âge du véhicule diminue le coût moyen des réclamations.

f) Pour un détenteur de police entre 17 et 20 ans, avec un véhicule de type D âgé de un à 3 ans, on trouve que

$$\widehat{E[Y_i]} = \frac{1}{\hat{\beta}_0 + \hat{\beta}_D^{MODEL}} = \frac{1}{0.0033233 - 0.0018235} = 666.76.$$

g) Pour un détenteur de police entre 35 et 39 ans, avec un véhicule de type A âgé de plus de 10 ans, on trouve que

$$\widehat{E[Y_i]} = \frac{1}{\hat{\beta}_0 + \hat{\beta}_{35-39}^{OWNERAGE} + \hat{\beta}_{10+}^{CARAGE}} = \frac{1}{0.0033233 + 0.0016372 + 0.0033776} = 119.93.$$

h) La déviance $D(y,\hat{\mu}) = 11.511$ est donnée dans la sortie `R` pour la sous-question a). On a que

$$\frac{D(y,\hat{\mu})}{\hat{\phi}} = \frac{11.511}{0.1074529} = 107.126,$$

ce qui est très près de $n - p' = 109$. Le modèle semble donc adéquat.

i) Les résidus sont calculés avec les formules trouvées à la question 6. Il faut d'abord enlever les données manquantes du vecteur contenant les coûts moyens. On obtient les graphiques de la Figure A.21.

j) a. Le modèle avec le lien logarithmique est

```
modlog <- glm(AvCost~OwnerAge+Model+CarAge,family=Gamma(link=log),data=Bcar)
summary(modlog)
```

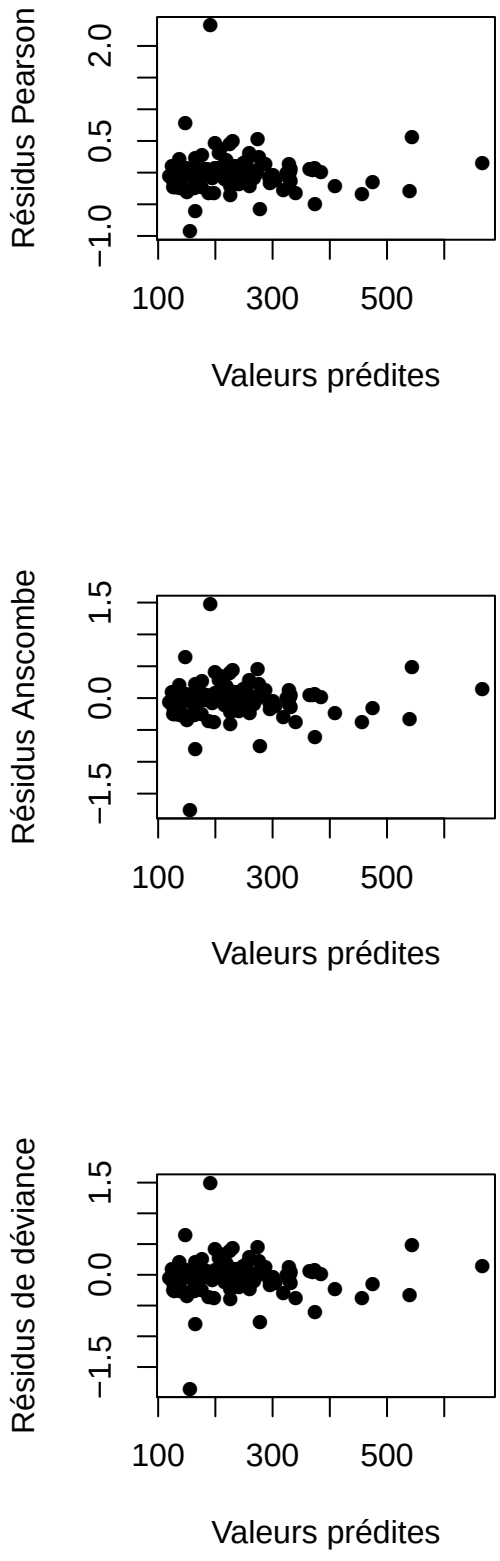


FIG. A.21 – Résidus pour GLM Gamma

```
##
## Call:
## glm(formula = AvCost ~ OwnerAge + Model + CarAge, family = Gamma(link = log),
##     data = Bcar)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84819  -0.12796  -0.00834   0.08552   1.20066
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.711739   0.103835  55.008  < 2e-16
## OwnerAge21-24 -0.108159   0.114547  -0.944   0.3471
## OwnerAge25-29  0.005223   0.113170   0.046   0.9633
## OwnerAge30-34 -0.288090   0.113170  -2.546   0.0123
## OwnerAge35-39 -0.331420   0.114547  -2.893   0.0046
## OwnerAge40-49 -0.280775   0.113170  -2.481   0.0146
## OwnerAge50-59 -0.238136   0.113170  -2.104   0.0377
## OwnerAge60+   -0.283521   0.113170  -2.505   0.0137
## ModelB         0.057951   0.075447   0.768   0.4441
## ModelC         0.154588   0.076115   2.031   0.0447
## ModelD         0.472290   0.078497   6.017 2.43e-08
## CarAge10+     -0.735513   0.078497  -9.370 1.17e-15
## CarAge4-7     -0.111412   0.075447  -1.477   0.1426
## CarAge8-9     -0.422538   0.076115  -5.551 2.02e-07
##
## (Intercept)   ***
## OwnerAge21-24
## OwnerAge25-29
## OwnerAge30-34 *
## OwnerAge35-39 **
## OwnerAge40-49 *
## OwnerAge50-59 *
## OwnerAge60+   *
## ModelB
## ModelC        *
## ModelD        ***
## CarAge10+     ***
## CarAge4-7
## CarAge8-9     ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0910768)
##
##     Null deviance: 27.841  on 122  degrees of freedom
## Residual deviance: 11.263  on 109  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1398
```

```
##
## Number of Fisher Scoring iterations: 7
```

b. Avec ce modèle $E[Y_i] = e^{\eta_i}$. Puisque les variables explicatives prennent toutes leur niveau de base, on a que $\hat{\eta}_i = \hat{\beta}_0 = 5.711739$ et

$$\widehat{E[Y_i]} = e^5.711739 = 302.39.$$

Cela ne diffère pas beaucoup du résultat trouvé en b).

c-d-e. Puisqu'on a utilisé un lien logarithmique, on a un modèle multiplicatif. Si $e^{\beta} > 1$, alors l'espérance du coût augmente, alors que si $e^{\beta} < 1$ alors l'espérance du coût diminue. On peut donc tirer des conclusions similaires à celles en c), d) et e).

f. Pour un détenteur de police entre 17 et 20 ans, avec un véhicule de type D âgé de un à 3 ans, on trouve que

$$\widehat{E[Y_i]} = \exp \hat{\beta}_0 + \hat{\beta}_D^{MODEL} = \exp 5.711739 + 0.472290 = 484.94.$$

On note que cette valeur est beaucoup moins élevée que celle obtenue en f).

g. Pour un détenteur de police entre 35 et 39 ans, avec un véhicule de type A âgé de plus de 10 ans, on trouve que

$$\widehat{E[Y_i]} = \exp \hat{\beta}_0 + \hat{\beta}_{35-39}^{OWNERAGE} + \hat{\beta}_{10+}^{CARAGE} = \exp 5.711739 - 0.331420 - 0.735513 = 104.04.$$

h. La déviance $D(y,\hat{\mu}) = 11.263$ est donnée dans la sortie `R` pour la sous-question a). On a que

$$\frac{D(y,\hat{\mu})}{\hat{\phi}} = \frac{11.263}{0.0910768} = 123.66,$$

ce qui est moins près de $n - p' = 109$ que pour le modèle avec le lien inverse. Le modèle semble donc moins adéquat.

5.9 a) If x_i is treated as a factor predictor with 11 levels, the linear predictor is written as

$$\eta_i = \beta_0 + \beta_i,\ i = 1,...,11$$

and $\beta_1 = 0$. The binomial density is the following :

$$f_Y(y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i},$$

which can be rewritten in a exponential family representation as :

$$f_Y(y_i) = \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + m \ln(1-\pi_i) + \ln\binom{m_i}{y_i} \right].$$

Hence, the canonical parameter is $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ and the canonical link is the logit link. Thus,

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_i$$
$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}}$$

The expression of the density in the reparametrization is then

$$\begin{aligned} f_Y(y_i) &= \binom{m_i}{y_i} \left(\frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_i}} \right)^{m_i-y_i} \\ &= \binom{m_i}{y_i} \frac{e^{y_i(\beta_0+\beta_i)}}{(1+e^{\beta_0+\beta_i})^{m_i}} \end{aligned}$$

The likelihood L and the log-likelihood l are shown below :

$$\begin{aligned} L(\beta_0,...,\beta_{11};y_1,...,y_{11}) &= \prod_{i=1}^{11} \binom{m_i}{y_i} \frac{e^{y_i(\beta_0+\beta_i)}}{(1+e^{\beta_0+\beta_i})^{m_i}} \\ \ell(\beta_0,...,\beta_{11};y_1,...,y_{11}) &= \sum_{i=1}^{11} \left[\ln \binom{m_i}{y_i} + y_i(\beta_0+\beta_i) - m_i \ln(1+e^{\beta_0+\beta_i}) \right] \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \sum_{i=1}^{11} \left[y_i - m_i \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}} \right] \\ \frac{\partial \ell}{\partial \beta_i} &= y_i - m_i \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}},\ i = 2,...,11, \end{aligned}$$

and $\beta_1 = 0$ by constraint of the model. The maximum likelihood estimators for the parameters are derived by solving the system of equations $\frac{\partial \ell}{\partial \beta_i} = 0$, $i = 0,...,11$:

$$\begin{aligned} \sum_{i=1}^{11} \left[y_i - m_i \frac{e^{\hat{\beta}_0+\hat{\beta}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_i}} \right] &= 0 \\ y_i - m_i \frac{e^{\hat{\beta}_0+\hat{\beta}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_i}} &= 0,\ i = 2,...,11, \\ \Rightarrow \hat{\beta}_0 + \hat{\beta}_i = \ln\left(\frac{y_i}{m_i - y_i} \right),\ i &= 2,...,11, \end{aligned}$$

Using the first equation and replacing $\hat{\beta}_0 + \hat{\beta}_i$ by $\ln\left(\frac{y_i}{m_i - y_i}\right)$,

$$
\begin{aligned}
y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} + \sum_{i=2}^{11} \left[y_i - m_i \frac{\left(\frac{y_i}{m_i - y_i}\right)}{1 + \left(\frac{y_i}{m_i - y_i}\right)} \right] &= 0 \\
y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} + \sum_{i=2}^{11} \left[y_i - m_i \frac{y_i}{m_i} \right] &= 0 \\
y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} &= 0 \\
\hat{\beta}_0 = \ln\left(\frac{y_1}{m_1 - y_1}\right) & \\
\hat{\beta}_i = \ln\left(\frac{y_i}{m_i - y_i}\right) - \hat{\beta}_0 = \ln\left(\frac{y_i/(m_i - y_i)}{y_1/(m_1 - y_1)}\right), & \; i = 2,...,11.
\end{aligned}
$$

The estimates of the model parameters are easily found in R as follows :

```
(beta0 <- log(y[1]/(m[1]-y[1])))

## [1] -2.917771

(beta <- c(0,log(y[-1]/(m[-1]-y[-1]))-beta0))

##  [1] 0.0000000 0.4122448 0.6151856 0.6740261 1.3083328
##  [6] 1.7137979 1.6184877 2.7636201 3.5239065 3.0178542
## [11] 3.2542430
```

Hence, here

$$\hat{\beta} = (-2.9178, 0, 0.4122, 0.6152, 0.6740, 1.3083, 1.7138, 1.6185, 2.7636, 3.5239, 3.0179, 3.2542)^T.$$

As a consistency check following from the invariance property of maximum likelihood estimation, we can verify that the estimates of π_i using the expit function are equal to the MLE estimates $\hat{\pi}_i = \frac{y_i}{m_i}$:

```
(pi <- exp(beta0+beta)/(1+exp(beta0+beta)))

##  [1] 0.05128205 0.07547170 0.09090909 0.09589041
##  [5] 0.16666667 0.23076923 0.21428571 0.46153846
##  [9] 0.64705882 0.52500000 0.58333333

y/m

##  [1] 0.05128205 0.07547170 0.09090909 0.09589041
##  [5] 0.16666667 0.23076923 0.21428571 0.46153846
##  [9] 0.64705882 0.52500000 0.58333333
```

b) The Binomial GLM model with logit link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i = 1,...,11$ is fitted to the data using R and the command :

```
glm(cbind(y,m-y)~x,family=binomial)
```

The estimates of the parameters are $\hat{\beta}_0 = -3.6070615$ and $\hat{\beta}_1 = 0.0009121$, with standard error $SE(\hat{\beta}_0) = 0.3533875$ and $SE(\hat{\beta}_1) = 0.0001084$.

c) The Binomial GLM model with probit link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i =$ 1,...,11 is fitted to the data using R and the command :

```
glm(cbind(y,m-y)~x,family=binomial(link=probit))
```

The estimates of the parameters are $\hat{\beta}_0 = -2.080$ and $\hat{\beta}_1 = 5.230 \times 10^{-4}$, with standard error $SE(\hat{\beta}_0) = 0.1852$ and $SE(\hat{\beta}_1) = 5.973 \times 10^{-5}$.

d) The Binomial GLM model with complementary log-log link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i = 1,...,11$ is fitted to the data using R and the command :

```
glm(cbind(y,m-y)~x,family=binomial(link=cloglog))
```

The estimates of the parameters are $\hat{\beta}_0 = -3.360$ and $\hat{\beta}_1 = 7.480 \times 10^{-4}$, with standard error $SE(\hat{\beta}_0) = 0.3061$ and $SE(\hat{\beta}_1) = 8.622 \times 10^{-5}$.

e) Predictions can be found using the inverse of the link function. For the model with canonical link (model from b), we find that

$$\hat{y}_{2000} = \frac{e^{\hat{\beta}_0 + 2000\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + 2000\hat{\beta}_1}} = 0.1439472.$$

Alternatively, the command `predict(modelb,data.frame(x=2000),type="response",se.fit=TRUE)` can be used to calculate the predictions and associated standard errors. The resulting predictions and standard errors are presented in Table A.1. The probability of developing fissures after 2000 hours of operations is 14.39% according to model b, and the standard deviation is 2.08%. This prediciton is quite comparable with the complementary log-log model (d), for which the estimated probability of developing fissures is 14.36%, with a standard error of 2.03%. The precision is slightly better in this model than the two others due to a smaller variance. The estimated probability with Model c, using the probit link, is higher at 15.06%, with standard error of 2.06%.

	Model b (logit link)	Model c (probit link)	Model d (compl. log-log link)
$\hat{y}_{2000}$	0.1439472	0.150615	0.1436447
$SE(\hat{y}_{2000})$	0.02080256	0.02062154	0.02030803

TAB. A.1 – Predictions and Standard Errors for Predictions for the 3 Models

f) Figure A.22 shows a plot of the data points along with the three fitted lines. It was obtained using the following code in R, where `ilogit`, `iprobit` and `icloglog` are the inverse of the corresponding link functions :

```
plot(x,y/m,pch=19,xlab="Number of Hours of Operation (x)",
     ylab="Prob of Developing Fissures")
j <- seq(0,4800,1)
lines(j,ilogit(coef(modelb)[1]+coef(modelb)[2]*j),lwd=2)
lines(j,iprobit(coef(modelc)[1]+coef(modelc)[2]*j),lty=2,col=2,lwd=2)
lines(j,icloglog(coef(modeld)[1]+coef(modeld)[2]*j),lty=3,col=4,lwd=2)
legend("topleft",legend=c("Logit","Probit","Complementary log-log"),
       lty=c(1,2,3),col=c(1,2,4),lwd=rep(2,3))
```

It is easy to observe that the fit is better when the number of hours of operations is lower, it seems that the variance of the observations is increasing with the predictor. This is expected in a generalized linear model framework. The three fitted lines are slightly

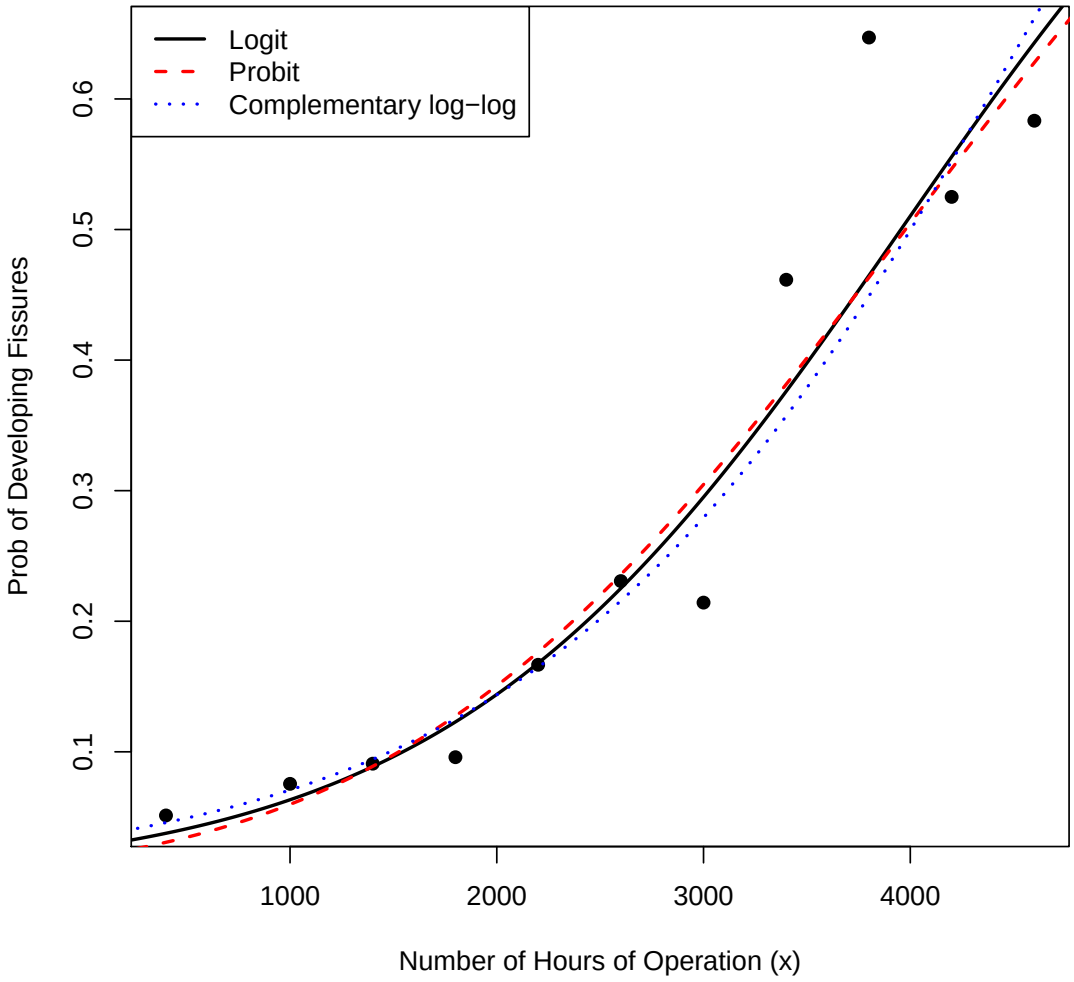


FIG. A.22 – Logistic, Probit and Complementary Log-Log Model Fit

different. The probit link produces lower estimates in the tails and higher estimates in the middle of the range of the predictors. It seems like this model is less representative of the data than the others. The complementary log-log model (d) predicts higher probabilities of failures in the extremes of the range of the predictors. This seems to fit the data well, and recall that the variance of the predictions where also smaller than other models in this case, which is a desirable property. The line for the model with canonical link is between the two others. It could also be a reasonable model for the data.

Chapitre ??

6.1 On ajuste d'abord le modèle avec les effets principaux et les interactions.

```
library(datasets)
fit1 <- glm(ncases~factor(agegp)*(factor(alcgp)+factor(tobgp))+factor(alcgp):factor(tobgp),f
```

```
## Warning: glm.fit: fitted rates numerically 0 occurred
```

```
anova(fit1)
```

```
## Warning: glm.fit: fitted rates numerically 0 occurred
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: ncases
##
## Terms added sequentially (first to last)
##
##
##                             Df Deviance Resid. Df
## NULL                                           87
## factor(agegp)                5  138.256        82
## factor(alcgp)                3   24.106        79
## factor(tobgp)                3   22.169        76
## factor(agegp):factor(alcgp) 15   32.417        61
## factor(agegp):factor(tobgp) 15   18.109        46
## factor(alcgp):factor(tobgp)  9    7.658        37
##                             Resid. Dev
## NULL                           262.926
## factor(agegp)                  124.670
## factor(alcgp)                  100.564
## factor(tobgp)                   78.395
## factor(agegp):factor(alcgp)     45.979
## factor(agegp):factor(tobgp)     27.870
## factor(alcgp):factor(tobgp)     20.212
```

```
qchisq(0.95,9) ## rejette alcgp:tobgp
```

```
## [1] 16.91898
```

```
qchisq(0.95,15) ## rejette agegp:tobgp mais conserve agegp:alcgp
```

```
## [1] 24.99579
```

On trouve donc que l'interaction entre la consommation d'alcool et de tabac n'est pas significative parce que

$$\Delta Deviance = 7.658 < \chi^2_{(9;0.95)} = 16.92.$$

Cela signifie que le modèle `agegp*(alcgp+tobgp)` est une simplification adéquate du modèle `agegp+alcgp+tobgp+agegp.alcgp+agegp.tobgp+alcgp.tobgp`. De plus, on peut enlever

l'interaction entre l'âge et la consommation de tabac :

$$\Delta Deviance = 18.109 < \chi^2_{(15;0.95)} = 25.$$

Cela signifie que le modèle `agegp*alcgp+tobgp` est une simplification adéquate du modèle `agegp*(alcgp+tobgp)`. Toutefois, on ne peut pas enlever l'autre terme d'interaction car

$$\Delta Deviance = 32.417 > \chi^2_{(15;0.95)} = 25.$$

Si on tente de remettre l'interaction entre la consommation d'alcool et de tabac dans le modèle, on trouve qu'elle n'est toujours pas significative :

```
fit2 <- glm(ncases ~ agegp * alcgp + tobgp,family=poisson,data=esoph)
fit3 <- update(fit1,~.-factor(agegp):factor(tobgp))
anova(fit2,fit3)

## Analysis of Deviance Table
##
## Model 1: ncases ~ agegp * alcgp + tobgp
## Model 2: ncases ~ factor(agegp) + factor(alcgp) + factor(tobgp) + factor(agegp):factor(alcgp) +
##     factor(alcgp):factor(tobgp)
##   Resid. Df Resid. Dev Df Deviance
## 1        61     45.979
## 2        52     38.973  9    7.006
```

Par conséquent, le modèle final est `agegp*alcgp+tobgp`. Cela signifie que l'effet de consommer de l'alcool sur l'occurence du cancer de l'oesophage est différent pour chaque groupe d'âge.

6.2 Intégrer la densité conditionnelle Poisson sur z. Comme c'est plus agréable à faire à la main qu'à taper, je vous laisse le soin de réussir par vous-même.

6.3 On note $n = n_A + n_B$. La vraisemblance pour ce GLM est

$$\begin{aligned}\mathcal{L}(\beta_0,\beta_1) &= \prod_{i=1}^{n} \exp(y_i \log(\mu_i) - \mu_i + \text{cte}) \\ &= \prod_{i=1}^{n} \exp(y_i \log(g^{-1}(\eta_i)) - g^{-1}(\eta_i) + \text{cte}).\end{aligned}$$

La log-vraisemblance est donc :

$$\ell(\beta_0,\beta_1) = \sum_{i=1}^{n} (y_i \log(g^{-1}(\eta_i)) - g^{-1}(\eta_i) + \text{cte}).$$

On dérive par rapport à β_0 et β_1 :

$$\begin{aligned}\frac{\partial \ell(\beta_0,\beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} (y_i \frac{1}{g^{-1}(\eta_i) g'(g^{-1}(\eta_i))} - \frac{1}{g'(g^{-1}(\eta_i))}) \\ &= \sum_{i=1}^{n} \frac{(y_i - g^{-1}(\eta_i))}{g^{-1}(\eta_i) g'(g^{-1}(\eta_i))} \\ \frac{\partial \ell(\beta_0,\beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} (y_i \frac{x_i}{g^{-1}(\eta_i) g'(g^{-1}(\eta_i))} - \frac{x_i}{g'(g^{-1}(\eta_i))}) \\ &= \sum_{i=1}^{n} \frac{x_i (y_i - g^{-1}(\eta_i))}{g^{-1}(\eta_i) g'(g^{-1}(\eta_i))}.\end{aligned}$$

On égalise à 0 pour obtenir le système d'équations à résoudre.

$$0 = \sum_{i=1}^{n} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i)g'(g^{-1}(\hat{\eta}_i))}$$
$$0 = \sum_{i=1}^{n} \frac{x_i(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i)g'(g^{-1}(\hat{\eta}_i))}$$

On utilise que $x_i = 0 \forall i \in (n_A + 1,...,n_A + n_B)$:

$$0 = \sum_{i=1}^{n_A+n_B} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i)g'(g^{-1}(\hat{\eta}_i))}$$
$$0 = \sum_{i=1}^{n_A} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i)g'(g^{-1}(\hat{\eta}_i))}$$
$$\Rightarrow 0 = \sum_{i=n_A+1}^{n_A+n_B} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i)g'(g^{-1}(\hat{\eta}_i))}.$$

Aussi, $\forall i \in (1,...,n_A), \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1$, ce qui ne dépend pas de i. Le dénominateur ne dépend pas de i et peut sortir de la somme et s'annuler. De même, $\forall i \in (n_A + 1,...,n_A + n_B), \hat{\eta}_i = \hat{\beta}_0$, ce qui ne dépend pas de i. Le dénominateur ne dépend pas de i et peut sortir de la somme et s'annuler. On obtient donc les équations :

$$0 = \sum_{i=1}^{n_A} (y_i - g^{-1}(\hat{\eta}_i))$$
$$0 = \sum_{i=n_A+1}^{n_A+n_B} (y_i - g^{-1}(\hat{\eta}_i)).$$

Finalement, $g^{-1}(\hat{\eta}_i) = \hat{\mu}_i$ par définition. Alors

$$0 = \sum_{i=1}^{n_A} (y_i - \hat{\mu}_A) \Rightarrow \sum_{i=1}^{n_A} y_i = n_A \hat{\mu}_A \Rightarrow \frac{\sum_{i=1}^{n_A} y_i}{n_A} = \hat{\mu}_A$$
$$0 = \sum_{i=n_A+1}^{n_A+n_B} (y_i - \hat{\mu}_B) \Rightarrow \sum_{i=n_A+1}^{n_A+n_B} y_i = n_B \hat{\mu}_B \Rightarrow \frac{\sum_{i=n_A+1}^{n_A+n_B} y_i}{n_B} = \hat{\mu}_B.$$

6.4 a) Avec ce modèle, on a que

$$\mu_A = \exp(\beta_0)$$
$$\mu_B = \exp(\beta_0 + \beta_1) = \mu_A \exp(\beta_1),$$

ce qui implique que $\exp(\beta_1) = \mu_B / \mu_A$. On ajuste le modèle en R, et on vérifie que cela est bien vrai :

```
y <- c( 8,7,6,6,3,4,7,2,3,4,9,9,8,14,8,13,11,5,7,6)
x <- rep(0:1,each=10)
fit1 <- glm(y~x,family=poisson)
summary(fit1)

##
```

```
## Call:
## glm(formula = y ~ x, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5280  -0.7622  -0.1699   0.6938   1.5399
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
## x             0.5878     0.1764   3.332 0.000861 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 27.857  on 19  degrees of freedom
## Residual deviance: 16.268  on 18  degrees of freedom
## AIC: 94.349
##
## Number of Fisher Scoring iterations: 4
```

```
log(mean(y[which(x==1)])/mean(y[which(x==0)]))
```

```
## [1] 0.5877867
```

b) Puisque $\exp(\beta_1) = \mu_B/\mu_A$, alors si $H_0 : \mu_A = \mu_B$ est vraie, $\beta_1 = 0$. On peut utiliser la statistique de Wald directement, on trouve que le seuil observé du test est 0.000861. On rejette donc l'hypothèse nulle à un niveau de confiance de 99%, ce qui implique que les moyennes diffèrent de façon significative.

c) Un I.C. à 95% pour β_1 est

```
fit1$coef[2]+c(-1,1)*qnorm(0.975)*summary(fit1)$coefficients[2,2]
```

```
## [1] 0.2420820 0.9334913
```

Alors, un I.C. pour μ_B/μ_A est $(\exp(0.2421), \exp(0.93349)) = (1.273899, 2.543373)$.

d) Il n'y a pas d'indications de surdispersion, puisque la déviance est 16.26 sur 18 degrés de liberté, et $16.26/18 < 1$.

e) Quand on ajuste une binomiale négative à ces données, on trouve que θ_z tend vers l'infini, donc le modèle Poisson est une simplification adéquate du modèle NB. En fait, les estimations des paramètres β_0 et β_1 sont exactement les mêmes que celles obtenues dans le modèle Poisson.

```
library(MASS)
fit2 <- glm.nb(y~x)
```

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace
> : iteration limit reached
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace
> : iteration limit reached
```

```r
summary(fit2)
```

```
##
## Call:
## glm.nb(formula = y ~ x, init.theta = 113420.3107, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5280  -0.7622  -0.1699   0.6937   1.5398
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
## x             0.5878     0.1764   3.332 0.000861 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(113420.3) family taken to be 1)
##
##     Null deviance: 27.855  on 19  degrees of freedom
## Residual deviance: 16.267  on 18  degrees of freedom
## AIC: 96.349
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  113420
##           Std. Err.:  4076965
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -90.349
```

f) Dans ce cas, on remarque que, bien que l'estimation du paramètre est égale pour les deux modèles, l'écart-type diffère. Aussi, le modèle de Poisson ne semble plus adéquat, car $Deviance/dl = 27.857/19 > 1$, alors que le modèle NB s'ajuste bien aux données. Cela montre que lorsqu'une variable explicative importante n'est pas observée, le modèle de Poisson peut perdre sa validité pour des données de comptage. La variable explicative manquante introduit de la sur-dispersion dans les données, ce qui est capturé efficacement avec la loi NB.

```r
fit3 <- glm(y~1,family=poisson)
fit4 <- glm.nb(y~1)
summary(fit3)
```

```
##
## Call:
## glm(formula = y ~ 1, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2336  -0.9063   0.0000   0.4580   2.3255
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.94591    0.08451   23.02   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 27.857  on 19  degrees of freedom
## Residual deviance: 27.857  on 19  degrees of freedom
## AIC: 103.94
##
## Number of Fisher Scoring iterations: 4
```

```
summary(fit4)
```

```
##
## Call:
## glm.nb(formula = y ~ 1, init.theta = 18.2073559, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9810  -0.7836   0.0000   0.3859   1.9033
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.94591    0.09944   19.57   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(18.2074) family taken to be 1)
##
##     Null deviance: 20.279  on 19  degrees of freedom
## Residual deviance: 20.279  on 19  degrees of freedom
## AIC: 104.77
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  18.2
##           Std. Err.:  21.0
##
##  2 x log-likelihood:  -100.767
```

```
exp(fit3$coef[1]+c(-1,1)*qnorm(0.975)*summary(fit3)$coefficients[1,2])
```

```
## [1] 5.931421 8.261090
```

```
exp(fit4$coef[1]+c(-1,1)*qnorm(0.975)*summary(fit4)$coefficients[1,2])
```

```
## [1] 5.760386 8.506374
```

6.5 a) On y va

```
sex <- rep(0:1,each=6)
Dep <- rep(0:5,2)
y <- c(512,353,120,138,53,22,89,17,202,131,94,24)
no <- c(313,207,205,279,138,351,19,8,391,244,299,317)
nb <- y+no
fitpSex <- glm(y~factor(sex)+offset(log(nb)),family=poisson)
summary(fitpSex)
```

```
##
## Call:
## glm(formula = y ~ factor(sex) + offset(log(nb)), family = poisson)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.1129   -3.6826   -0.2719    3.7437    8.0834
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.80926    0.02889 -28.011  < 2e-16 ***
## factor(sex)1 -0.38298    0.05128  -7.468 8.15e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 551.69  on 11  degrees of freedom
## Residual deviance: 493.56  on 10  degrees of freedom
## AIC: 573.76
##
## Number of Fisher Scoring iterations: 4
```

On trouve donc que la valeur-p du test de Wald $H_0 : \beta^{SEX} = 0$ est 8.15×10^{-14} ce qui est hautement significatif. Puisque le coefficient est négatif et que le niveau de base utilisé est "hommes", cela signifie que les femmes ont moins de chance d'être acceptées aux études graduées que les hommes.

b) On ajoute le département :

```
fitp2 <- glm(y~factor(sex)+factor(Dep)+offset(log(nb)),family=poisson)
summary(fitp2)
```

```
##
## Call:
## glm(formula = y ~ factor(sex) + factor(Dep) + offset(log(nb)),
##     family = poisson)
##
## Deviance Residuals:
##        1         2         3         4         5
## -0.68882  -0.01474   0.96655   0.02569   0.97713
##        6         7         8         9        10
## -0.28371   1.77895   0.06756  -0.71131  -0.02632
```

```
##       11        12
## -0.68503   0.28254
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.44677    0.04148 -10.771   <2e-16 ***
## factor(sex)1  0.05859    0.06166   0.950    0.342
## factor(Dep)1 -0.01391    0.06625  -0.210    0.834
## factor(Dep)2 -0.63911    0.07660  -8.344   <2e-16 ***
## factor(Dep)3 -0.66125    0.07675  -8.615   <2e-16 ***
## factor(Dep)4 -0.97250    0.09836  -9.887   <2e-16 ***
## factor(Dep)5 -2.32388    0.15468 -15.024   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 551.6926  on 11  degrees of freedom
## Residual deviance:   6.6698  on  5  degrees of freedom
## AIC: 96.868
##
## Number of Fisher Scoring iterations: 3
```

Dans ce modèle, le résultat du test de Wald pour le coefficient de la variable Sexe est différent. Puisque le seuil observé du test est 34.5%, on ne peut pas rejeter l'hypothèse nulle que $\beta^{SEX} = 0$. Cela signifie que le sexe n'est pas un facteur qui influence le taux d'admission aux études graduées lorsqu'on prend en considération le département. Il en est ainsi car les femmes appliquent plus souvent que les hommes dans des départements où il est plus difficile d'être admis.

c) À l'aide de l'analyse de la déviance, on trouve que l'interaction n'est pas significative :

$$\Delta Deviance = 6.67 < \chi^2(0.95,5) = 11.07.$$

```
fitp <- glm(y~factor(sex)*factor(Dep)+offset(log(nb)),family=poisson)
anova(fitp)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                       Df Deviance Resid. Df
## NULL                                     11
## factor(sex)            1    58.13        10
## factor(Dep)            5   486.89         5
```

```
## factor(sex):factor(Dep)  5     6.67         0
##                         Resid. Dev
## NULL                        551.69
## factor(sex)                 493.56
## factor(Dep)                   6.67
## factor(sex):factor(Dep)       0.00
```

```
qchisq(0.95,5) ## reject interaction
```

```
## [1] 11.0705
```

d) Le modèle final est celui avec une seule variable explicative dichotomique : le département. La déviance pour ce modèle est 7.5706, ce qui est légèrement supérieur à 6, le nombre de degrés de liberté. Toutefois, puisque $Deviance/dl \approx 1.26$, cela n'est pas très alarmant, et il n'y a pas de raison de supposer que le modèle de Poisson est inadéquat. La statistique de Pearson est 8.03, ce qui est aussi une valeur attendue pour la loi chi-carrée avec 6 degrés de liberté.

```
fitpDep <- glm(y~factor(Dep)+offset(log(nb)),family=poisson)
summary(fitpDep)
```

```
##
## Call:
## glm(formula = y ~ factor(Dep) + offset(log(nb)), family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8481  -0.4187   0.1160   0.4595   2.2321
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.43981    0.04079 -10.782   <2e-16 ***
## factor(Dep)1 -0.01830    0.06608  -0.277    0.782
## factor(Dep)2 -0.60784    0.06906  -8.801   <2e-16 ***
## factor(Dep)3 -0.64004    0.07336  -8.725   <2e-16 ***
## factor(Dep)4 -0.93966    0.09201 -10.212   <2e-16 ***
## factor(Dep)5 -2.30243    0.15298 -15.050   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 551.6926  on 11  degrees of freedom
## Residual deviance:   7.5706  on  6  degrees of freedom
## AIC: 95.769
##
## Number of Fisher Scoring iterations: 4
```

```
sum((y-fitted(fitpDep))^2/fitted(fitpDep))
```

```
## [1] 8.025236
```

```
pchisq(8.025236,6)
```

```
## [1] 0.7637397
```

e) On recommence et on obtient exactement les mêmes conclusions :

```
fitbSex <- glm(cbind(y,nb-y)~factor(sex),family=binomial)
summary(fitbSex)
```

```
##
## Call:
## glm(formula = cbind(y, nb - y) ~ factor(sex), family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.7915   -4.7613   -0.4365    5.1025   11.2022
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.22013    0.03879  -5.675 1.38e-08 ***
## factor(sex)1 -0.61035    0.06389  -9.553  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.06  on 11  degrees of freedom
## Residual deviance: 783.61  on 10  degrees of freedom
## AIC: 856.55
##
## Number of Fisher Scoring iterations: 4
```

```
fitb2 <- glm(cbind(y,nb-y)~factor(sex)+factor(Dep),family=binomial)
summary(fitb2)
```

```
##
## Call:
## glm(formula = cbind(y, nb - y) ~ factor(sex) + factor(Dep), family = binomial)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -1.2487  -0.0560   1.2533   0.0826   1.2205  -0.2076
##       7        8        9       10       11       12
##  3.7189   0.2706  -0.9243  -0.0858  -0.8509   0.2052
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.58205    0.06899   8.436   <2e-16 ***
## factor(sex)1  0.09987    0.08085   1.235    0.217
## factor(Dep)1 -0.04340    0.10984  -0.395    0.693
## factor(Dep)2 -1.26260    0.10663 -11.841   <2e-16 ***
## factor(Dep)3 -1.29461    0.10582 -12.234   <2e-16 ***
## factor(Dep)4 -1.73931    0.12611 -13.792   <2e-16 ***
```

```
## factor(Dep)5 -3.30648    0.16998 -19.452   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.056  on 11  degrees of freedom
## Residual deviance:  20.204  on  5  degrees of freedom
## AIC: 103.14
##
## Number of Fisher Scoring iterations: 4
```

```
fitb <- glm(cbind(y,nb-y)~factor(sex)*factor(Dep),family=binomial)
anova(fitb)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(y, nb - y)
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df
## NULL                                        11
## factor(sex)               1    93.45        10
## factor(Dep)               5   763.40         5
## factor(sex):factor(Dep)   5    20.20         0
##                          Resid. Dev
## NULL                         877.06
## factor(sex)                  783.61
## factor(Dep)                   20.20
## factor(sex):factor(Dep)        0.00
```

```
qchisq(0.95,5)
```

```
## [1] 11.0705
```

6.6 If $Y_i \sim Poisson(E_i\lambda_i)$, then, using the canonical link,

$$\log(\mu_i) = log(E_i) + log(\lambda_i),$$

where λ_i is the mean SMR for observation i. $\log(E_i)$, the natural logarithm of the expected count of lung cancer based on the demographics of the county, is passed to the `glm` function as an offset factor.

The data for males and females are concatenated to create a model with one covariate, Radon exposure, and one factor predictor, Sex, which takes 2 levels (0 for males and 1 for females).

```
Ytot <- c(YM,YF)
Etot <- c(EM,EF)
Sex <- c(rep(0,length(YM)),rep(1,length(YF))) ## 1 if female
Radontot <- rep(Radon,2)

modsex <- glm(Ytot~Sex+offset(log(Etot)),family=poisson)
modsexrad <- glm(Ytot~Radontot+Sex+offset(log(Etot)),family=poisson)

anova(modsex,modsexrad)

## Analysis of Deviance Table
##
## Model 1: Ytot ~ Sex + offset(log(Etot))
## Model 2: Ytot ~ Radontot + Sex + offset(log(Etot))
##   Resid. Df Resid. Dev Df Deviance
## 1       172     410.27
## 2       171     364.05  1    46.22

qchisq(0.99,1)

## [1] 6.634897
```

As shown above, the analysis of deviance shows strong evidence that the radon exposure influences the number of lung cancer in a particular county :

$$\Delta Deviance = 46.22 > \chi^2_{(1;0.99)} = 6.6349.$$

The null model and the model including sex only are fitted.

a)
```
modtot <- glm(Ytot~1+offset(log(Etot)),family=poisson)
modsex <- glm(Ytot~Sex+offset(log(Etot)),family=poisson)
anova(modtot,modsex)

## Analysis of Deviance Table
##
## Model 1: Ytot ~ 1 + offset(log(Etot))
## Model 2: Ytot ~ Sex + offset(log(Etot))
##   Resid. Df Resid. Dev Df  Deviance
## 1       173     410.28
## 2       172     410.27  1 0.0093398

qchisq(0.95,1)

## [1] 3.841459
```

The analysis of deviance shows that $\Delta Deviance = 0.0093398 < \chi^2_{(1;0.95)} = 3.8415$. Hence, the null model is an appropriate simplification of the model including the factor Sex, so the factor is not significant. However, below is the R output for the analysis of deviance when the covariate Radon (known to be significant from a) is included in the model. If we first consider the model with main effects and interactions, we see that $\Delta Deviance = 8.823 > \chi^2_{(1;0.99)}$, meaning that the model with main effects only is not an adequate simplification of the model with main effects and interactions. Thus, the factor predictor Sex is significant in the model through its interaction with the covariate Radon. Note that even if the main effect of the Sex does not appear to be significant, it is kept in the model by convention.

```
modtotrad <- glm(Ytot~Radontot+offset(log(Etot)),family=poisson)
modsexrad <- glm(Ytot~Radontot+Sex+offset(log(Etot)),family=poisson)
modsexradINT <- glm(Ytot~Radontot*Sex+offset(log(Etot)),family=poisson)
anova(modtot,modtotrad,modsexrad,modsexradINT)

## Analysis of Deviance Table
##
## Model 1: Ytot ~ 1 + offset(log(Etot))
## Model 2: Ytot ~ Radontot + offset(log(Etot))
## Model 3: Ytot ~ Radontot + Sex + offset(log(Etot))
## Model 4: Ytot ~ Radontot * Sex + offset(log(Etot))
##   Resid. Df Resid. Dev Df Deviance
## 1       173     410.28
## 2       172     364.06  1   46.219
## 3       171     364.05  1    0.011
## 4       170     355.23  1    8.823
```

b) The predictions are obtained using the command

```
predict(modsexradINT,data.frame(Radontot=6,Sex=0,Etot=1),type="response",se.fit=TRUE)
```

If the model Sex*Radon is used, we find

$$\hat{SMR}_{Sex=0,Radon=6} = 0.9708183,$$

with a standard error of 0.01415307.

c) The model Sex*Radon has a deviance of 355.23 on 170 degrees of freedom. A heuristic check for the validity of the model is to calculate the estimated dispersion parameter

$$\hat{\phi} = \frac{355.23}{170} = 2.089$$

and to compare it with 1, the dispersion parameter implied in the Poisson model. This check suggests the presence of overdispersion in the data as $\hat{\phi}$ is greater than 1. Fitting the quasipoisson model also leads to the same conclusion : the estimated dispersion parameter is 1.98311, closer to 2. Thus, we can conclude that the Poisson model is not adequate, we might consider fitting a Negative Binomial model to capture the overdispersion.