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 Modèles linéaires généralisés (GLM)

2.1 Est-ce que les distributions suivantes font partie de la famille exponentielle linéaire ? Si oui, écrire la densité sous la forme exponentielle linéaire, donner le paramètre canonique, le paramètre de dispersion, l'espérance et la variance de Y en termes de la fonction $b()$ et la relation $V()$ entre la moyenne et la variance.

a) Normale(μ,σ^2)

b) Uniforme(0,β)

c) Poisson(λ)

d) Bernoulli(π)

e) Binomiale(m,π), $m > 0$ est un entier et est connu (On considère $Y^* = Y/m$).

f) Pareto(α,λ)

g) Gamma(α,β)

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

2.2 Quelles fonctions de lien peut-on utiliser pour un GLM avec une loi de Poisson ?

2.3 Quel est le lien canonique pour la loi gamma ? Est-ce que ce lien est toujours approprié ?

2.4 On suppose que Y_1,...,Y_n sont des v.a.s indépendantes et $Y_i \sim Poisson(\mu_i)$. Pour chaque observation, on a une seule variable explicative x_i.

a) Quel est le lien canonique ?

b) Trouver les fonctions de score (à résoudre pour l'estimation des paramètres par maximum de vraisemblance)

2.5 Montrer que la déviance pour le modèle binomial est

$$D(y,\hat{\mu}) = 2\sum_{i=1}^{n} m_i \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right].$$

2.6 Trouver les expressions des résidus de Pearson, d'Anscombe et de déviance pour la loi Gamma.

2.7 Les données suivantes représentent des données de comptage, du nombre d'échec pour trois appareils médicaux (M1, M2 et M3) lors de tests de résistance sur 1000 appareils de chaque type et pour quatre niveaux de résistance mécanique différents (I, II, III, IV).

Device\ Stress Level	I	II	III	IV
M1	6	8	18	10
M2	13	18	29	20
M3	9	8	21	19

À l'aide de la modélisation Poisson (lien canonique), évaluer s'il y a une différence significative entre les taux d'échec des appareils.

2.8 Les données pour cet exercice sot contenues dans le fichier `BritishCar.csv (sep=";")` disponible sur le site du cours. On y trouve les montants de réclamations moyens pour les dommages causés au véhicule du détenteur de la police pour les véhicules assurés au Royaume-Uni en 1975. Les moyennes sont en livres sterling ajustées pour l'inflation.

Variable	Description
`OwnerAge`	Âge du détenteur de la police (8 catégories)
`Model`	Type de voiture (4 groupes)
`CarAge`	Âge du véhicule, en années (4 catégories)
`NClaims`	Nombre de réclamations
`AvCost`	Coût moyen par réclamation, en livres sterling

On s'intéresse à la modélisation du coût moyen par réclamation.

a) Ajuster un modèle de régression Gamma avec lien inverse pour la variable endogène `AvCost`. Inclure les effets principaux `OwnerAge`, `Model` et `CarAge`.

b) Quelle est l'espérance du coût moyen de la réclamation pour un détenteur de police âgé entre 17 et 20 ans, avec une auto de type A âgée de moins de 3 ans ?

c) Interpréter brièvement les coefficients pour la variable exogène `OwnerAge`.

d) Interpréter brièvement les coefficients pour la variable exogène `Model`.

e) Interpréter brièvement les coefficients pour la variable exogène `CarAge`.

f) Pour quelle combinaison de variables exogènes l'espérance du coût de réclamation est-elle la plus élevée ? Calculer sa valeur.

g) Pour quelle combinaison de variables exogènes l'espérance du coût de réclamation est-elle la plus faible ? Calculer sa valeur.

h) Quelle est la déviance pour ce modèle ? Est-ce que le modèle semble adéquat ?

i) Tracer le graphique des résidus de Pearson en fonction des valeurs prédites, des résidus d'Ascombe en fonction des valeurs prédites et des résidus de déviance en fonction des valeurs prédites.

j) Obtient-on les mêmes conclusions aux sous-questions a) à h) si on utilise un lien logarithmique plutôt que le lien inverse ?

2.9 On considère les données suivantes, qui contiennent le nombre Y_i de turbines sur m_i qui ont été fissurées après x_i heures d'opération.

x_i	m_i	Y_i
400	39	2
1000	53	4
1400	33	3
1800	73	7
2200	30	5
2600	39	9
3000	42	9
3400	13	6
3800	34	22
4200	40	21
4600	36	21

a) En utilisant un GLM binomial avec lien canonique, dériver les estimateurs des paramètres lorsque x_i est traité comme une variable exogène dichotomique avec 11 niveaux, et lorsque le prédicteur linéaire pour la donnée i est

$$\eta_i = \beta_0 + \beta_i, \text{ pour } i = 1,...,11,$$

avec la contrainte d'identifiabilité que $\beta_1 = 0$.

b) En utilisant R et un GLM binomial avec lien canonique, ajuster le modèle où le prédicteur linéaire est

$$\eta_i = \beta_0 + \beta_1 x_i, \text{ pour } i = 1,...,11.$$

Donner les estimations des paramètres et leur écart-type.

c) Refaire (b) en utilisant un lien probit. Donner les estimations des paramètres et leur écart-type.

d) Refaire (b) en utilisant un lien log-log complémentaire. Donner les estimations des paramètres et leur écart-type.

e) Comparer les prévisions (et leurs mesures d'incertitude) sous les trois modèles ajustés en (b), (c) et (d) pour une turbine qui était en opération pour 2000 heures.

f) Tracer un graphique pour montrer si les modèles en (b), (c) et (d) ajustent bien (ou non) les données. Commenter.

Réponses

A Solutions

Chapitre ??

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

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

Or,

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

d'où les équations normales sont

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

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

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

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

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

De la seconde équation normale, on obtient

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

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

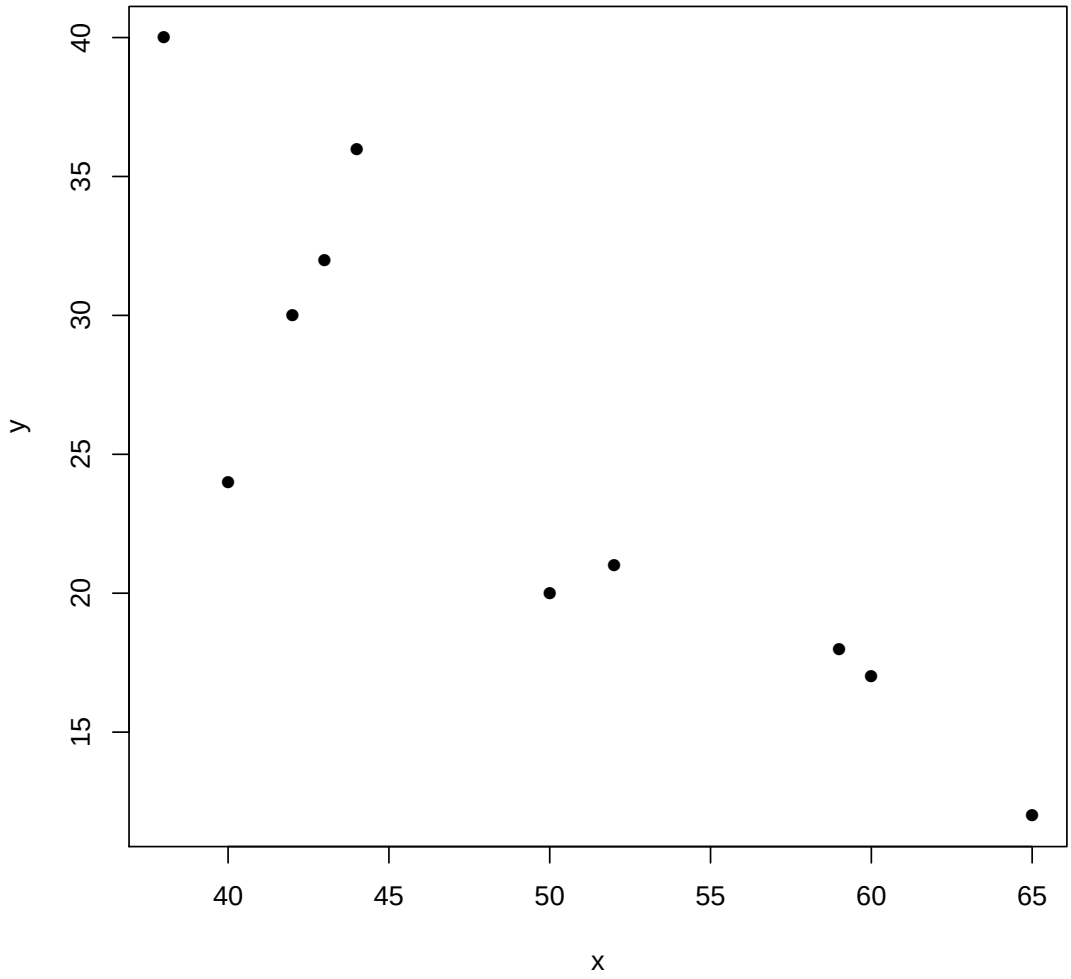

FIG. A.1 – Relation entre les données de l'exercice **??.??**

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

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

Par conséquent,

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

et

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

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

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

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

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

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

```
residuals(fit)
```

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

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

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

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

```
abline(fit)
```

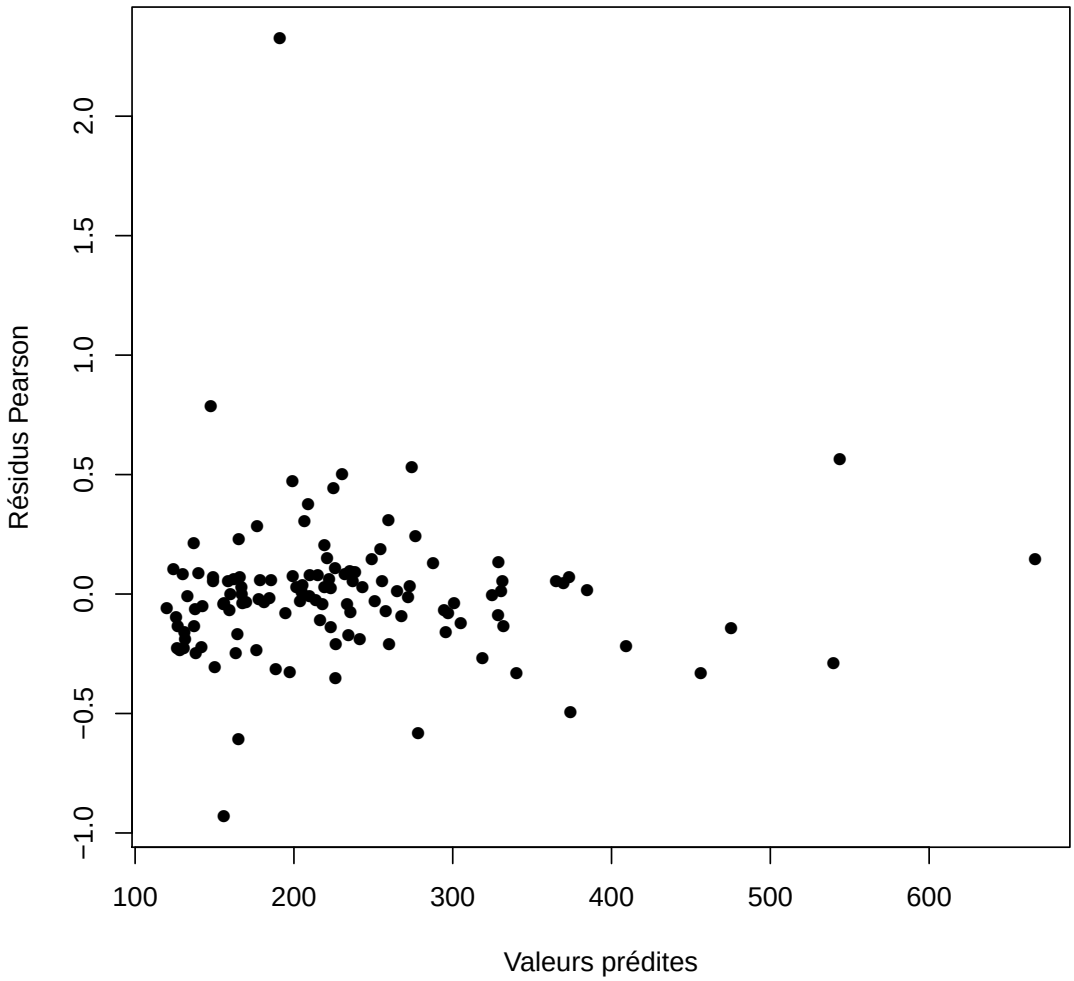

FIG. A.2 – Relation entre les données de l'exercice **??.??** et la droite de régression

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

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

et

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

b) Les sommes de carrés sont

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

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

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

2.3 a) Voir la figure A.3.

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

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

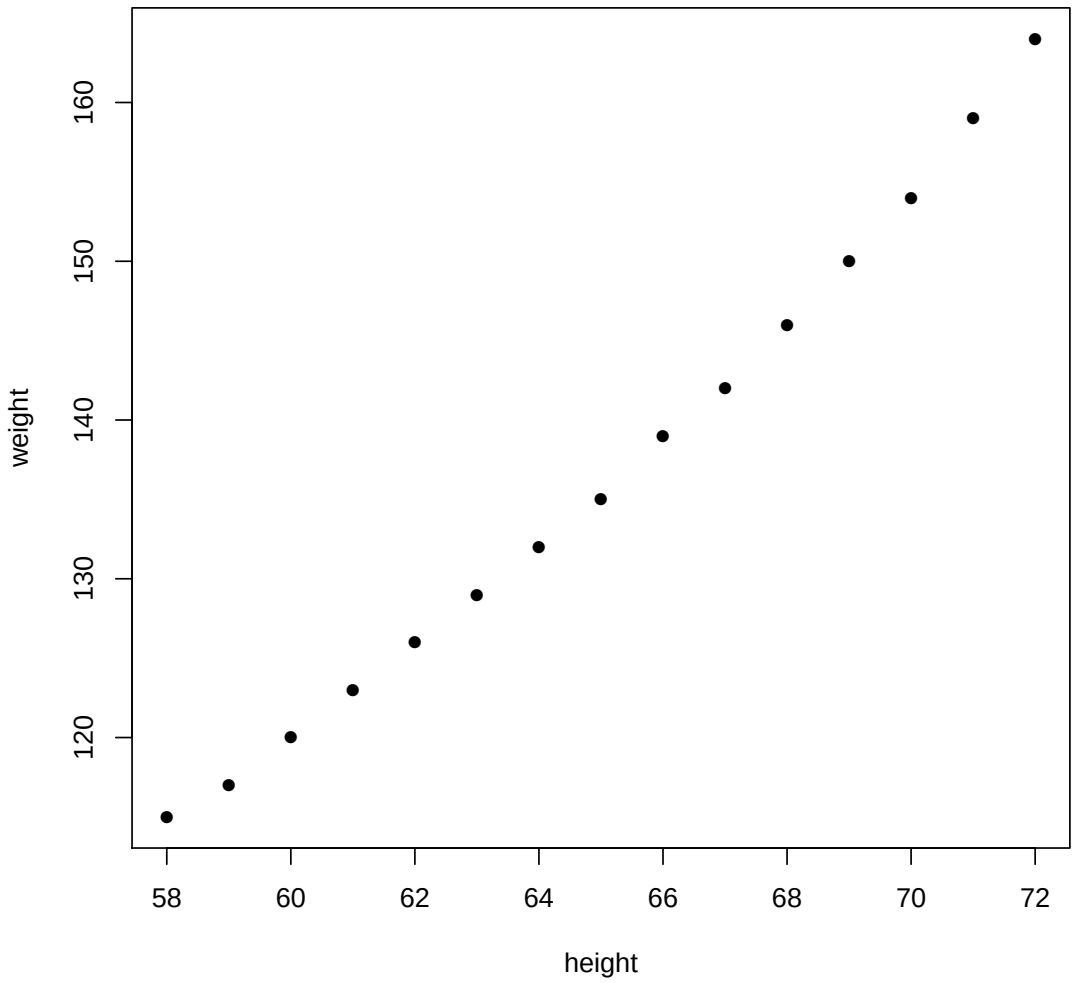

FIG. A.3 – Relation entre la taille et le poids moyen de femmes américaines âgées de 30 à 39 ans (données `women`)

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

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

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

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

```
summary(fit)

##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

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

e) On a

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

## [1] TRUE

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

## [1] TRUE
```

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

```
abline(fit)
```

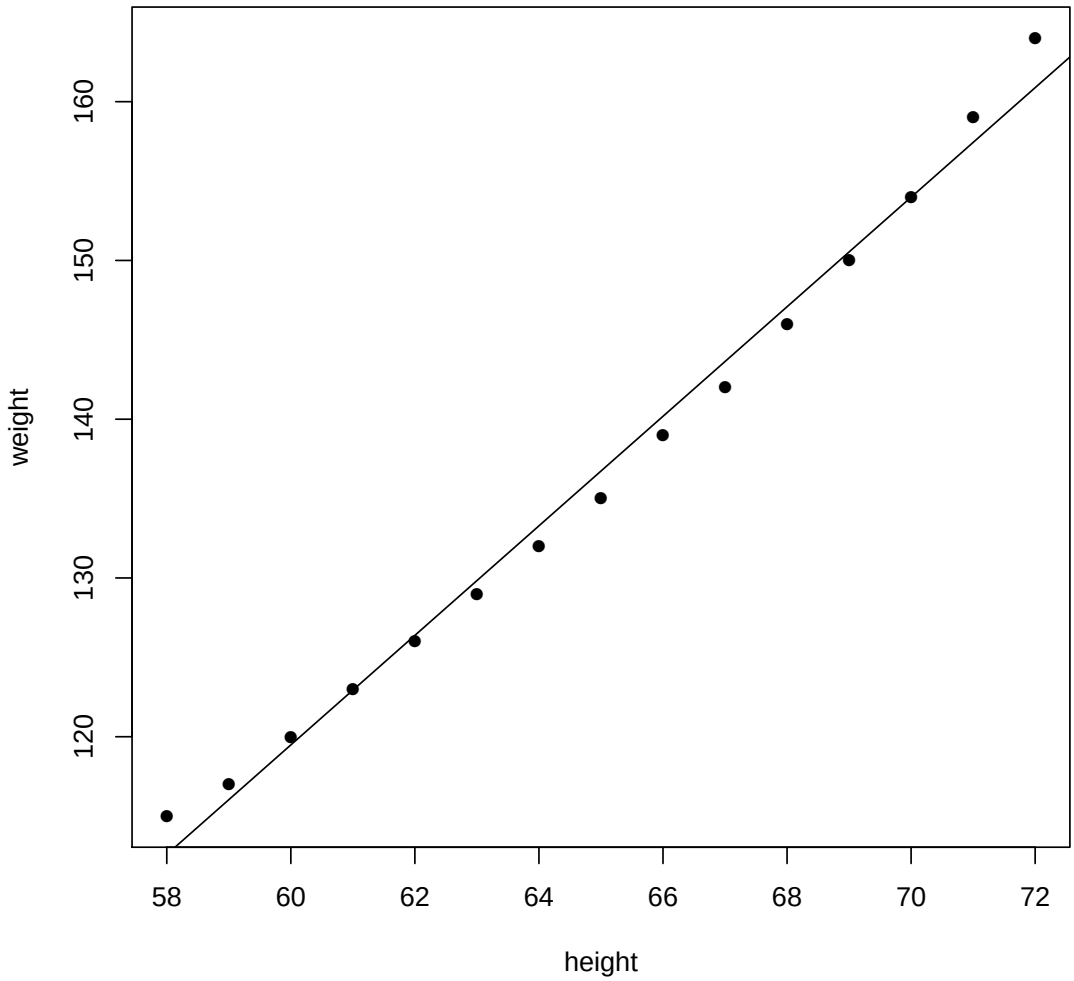

FIG. A.4 – Relation entre les données `women` et droite de régression linéaire simple

alors

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

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

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

et

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

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

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

et

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

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

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

Or,

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

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

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

L'estimateur des moindres carrés de β est donc

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

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

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

c) Des hypothèses du modèle, on a

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

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

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

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

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

Pour utiliser la méthode des multiplicateurs de Lagrange, on pose

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

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

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

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

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

Or, de la contrainte,

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

d'où

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

et, donc,

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

Finalement,

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

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

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

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

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

et l'on sait que

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

et que

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

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

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

De manière similaire, on détermine que

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

Par conséquent,

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

et

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

d'où $E[\text{MSE}] = \sigma^2$.

b) On a

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

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

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

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

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

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

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

et

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

b) Tout d'abord, on établit que

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

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

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

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

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

et, de manière similaire,

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

où

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

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

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

Or,

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

Par conséquent,

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

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

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

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

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

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

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

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

2.12 L'intervalle de confiance pour β_1 est

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

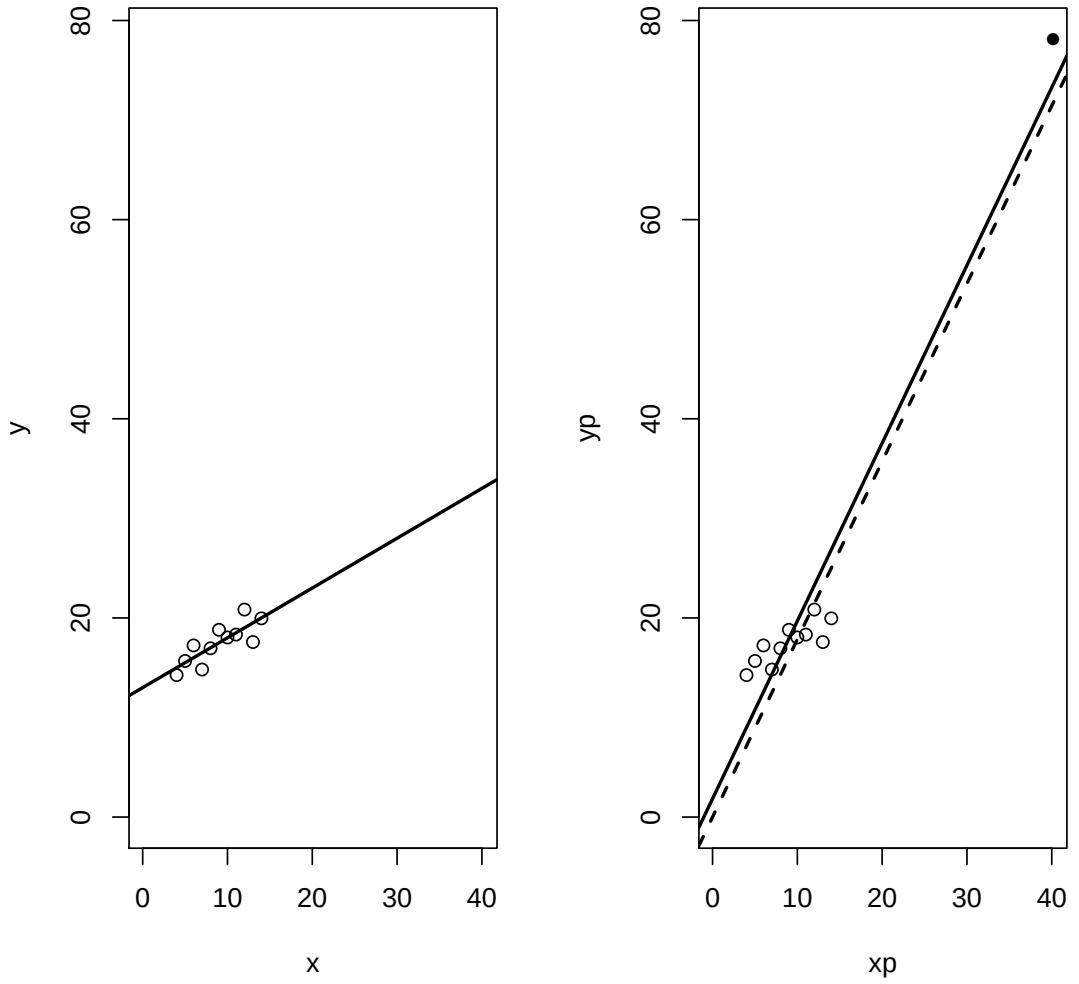

FIG. A.5 – Illustration de l'effet de l'ajout d'un point spécial à un ensemble de données. À gauche, la droite de régression usuelle. À droite, le même ensemble de points avec le point spécial ajouté (cercle plein), la droite de régression avec ce nouveau point (ligne pleine) et la droite de régression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

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

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

et

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

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

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

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

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

b) Les sommes de carrés sont

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

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

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

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

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

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

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

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

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

d'où

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

et, enfin,

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

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

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

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

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

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

d'où, finalement,

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

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

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

ou, de manière équivalente,

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

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

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

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

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

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

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

et

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

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

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

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

puis $\text{SSE} = \text{SST} - \text{SSR} = 1{,}104$ 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

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

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

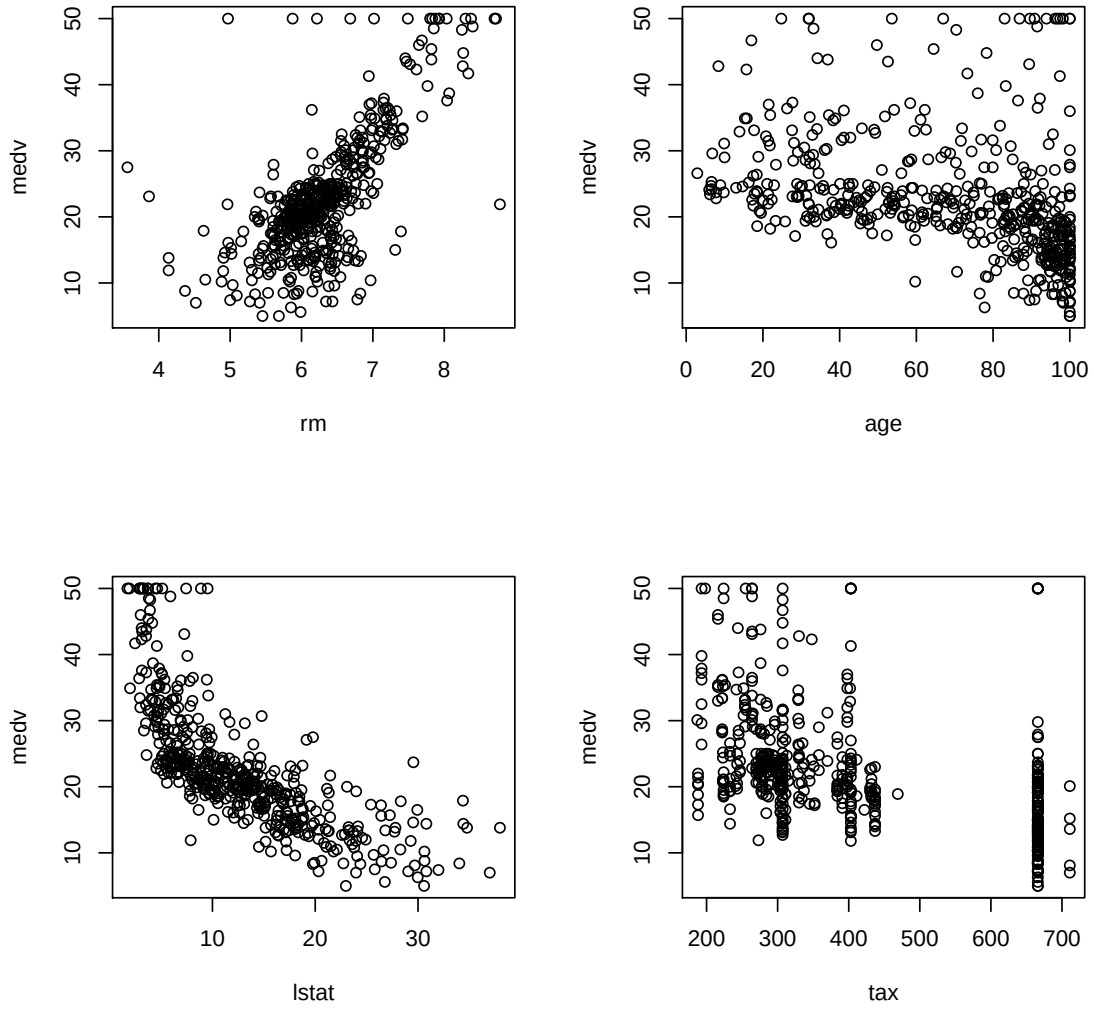

FIG. A.6 – Relation entre la variable `medv` et les variables `rm`, `age`, `lstat` et `tax` des données `house.dat`

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

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

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

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

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

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

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

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

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

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

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

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

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

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

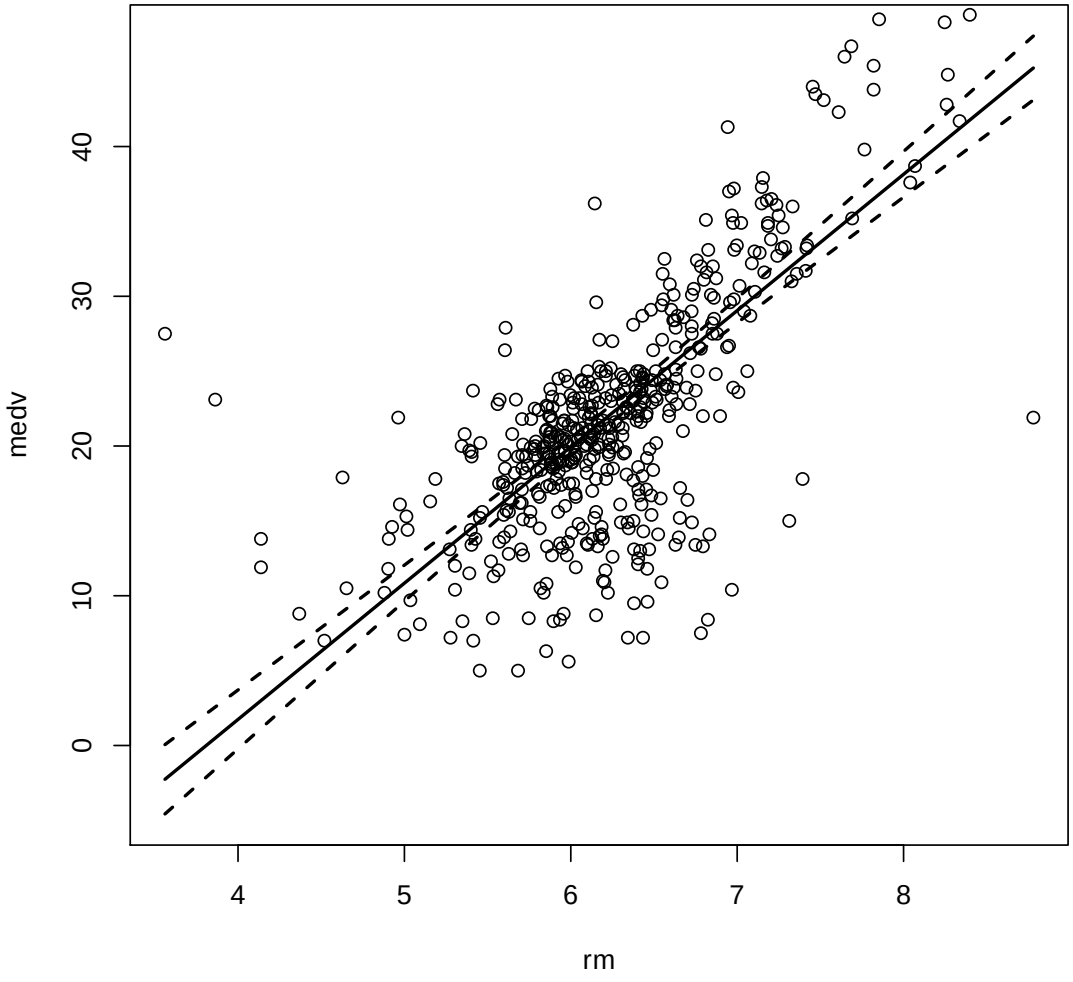

FIG. A.7 – Résultat de la régression de la variable `rm` sur la variable `medv` des données `house.dat`

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

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

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

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

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

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

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

##
## Call:
## lm(formula = consommation ~ poids)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07123 -0.68380  0.01488  0.44802  2.66234
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0146530  0.7118445  -0.021    0.984
## poids        0.0078382  0.0005315  14.748   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 36 degrees of freedom
## Multiple R-squared:  0.858,Adjusted R-squared:  0.854
## F-statistic: 217.5 on 1 and 36 DF,  p-value: < 2.2e-16
```

Le modèle est donc le suivant : $Y_t = -0{,}01465 + 0{,}007838 X_t + \varepsilon_t$, $\varepsilon_t \sim N(0, 1{,}039^2)$, où Y_t est la consommation en litres aux 100 kilomètres et X_t le poids en kilogrammes. La faible valeur p du test F indique une régression très significative. De plus, le R^2 de 0,858 confirme que l'ajustement du modèle est assez bon.

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

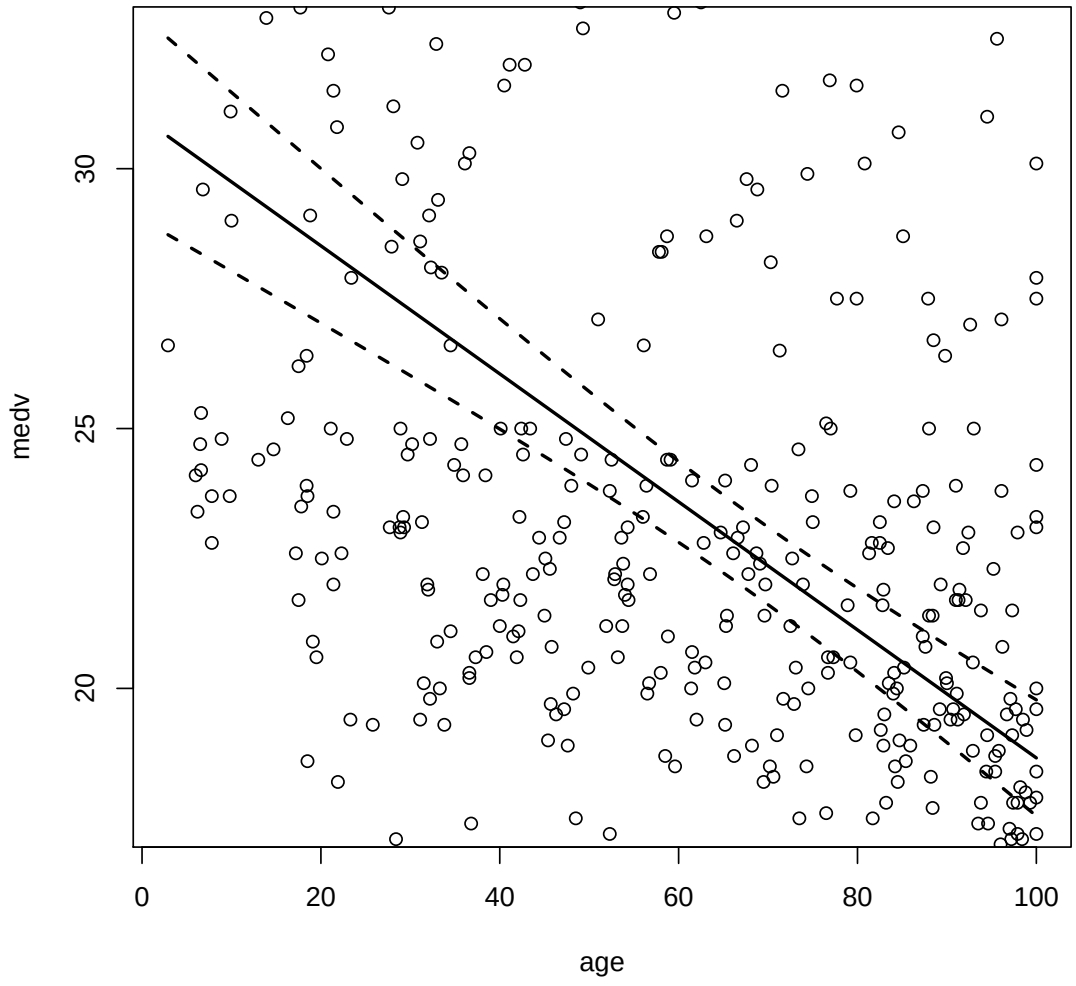

FIG. A.8 – Résultat de la régression de la variable `age` sur la variable `medv` des données `house.dat`

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

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

##       fit      lwr      upr
## 1 10.5669 8.432089 12.7017
```

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,

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

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)

##
## 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:
```

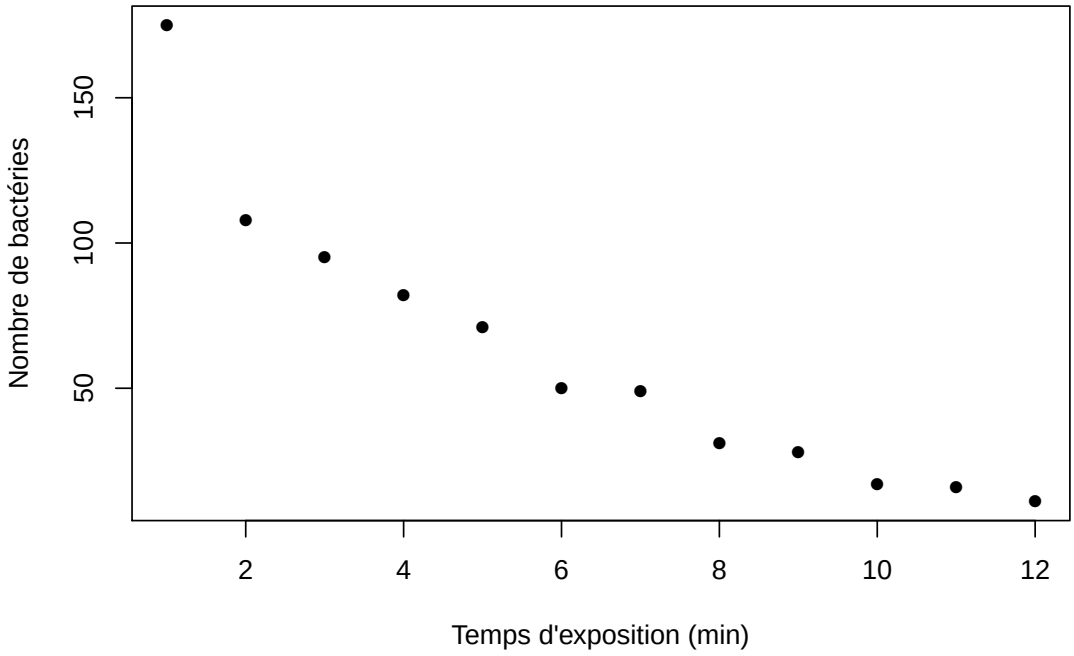

FIG. A.9 – Scatter Plot of the Number of Bacteria versus the Minutes of Exposure to 300^o F

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

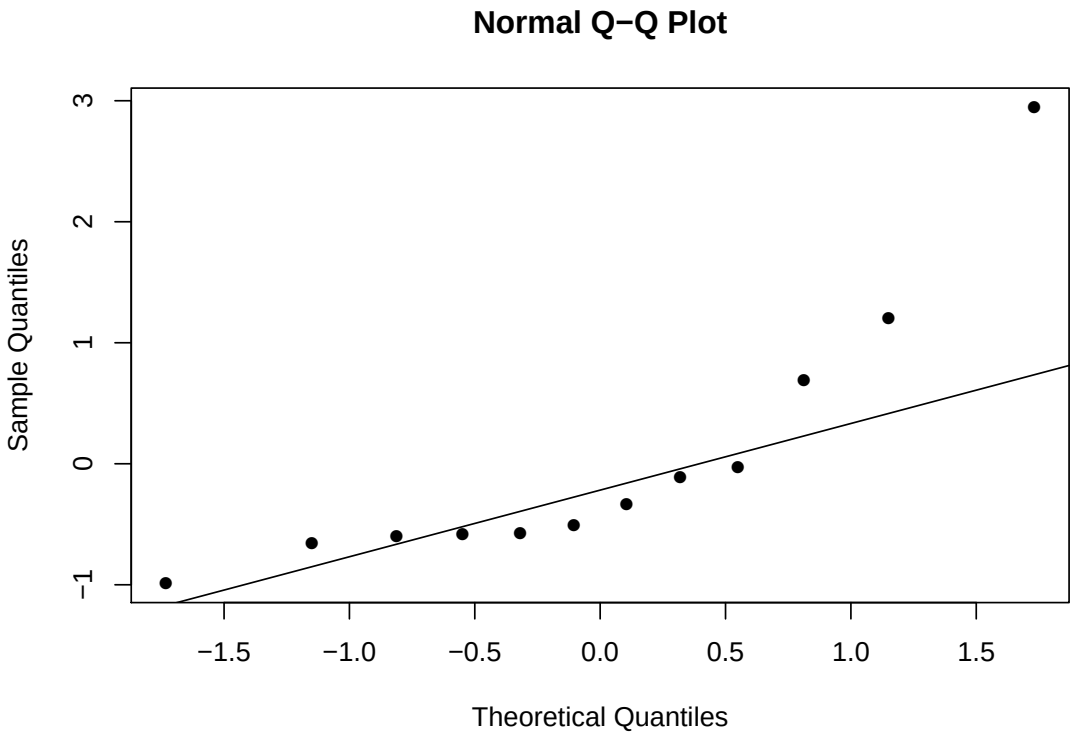

FIG. A.10 – Q-Q Plot for Simple Linear Model in Problem 5 b)

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 **??**. 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.

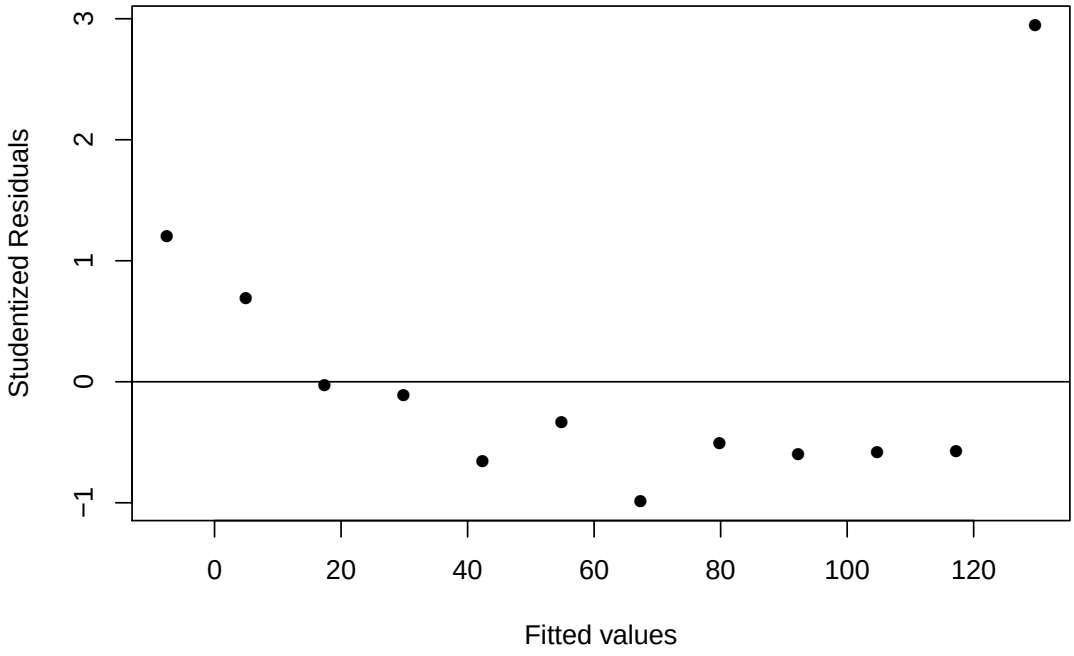

FIG. A.11 – Residuals versus the Fitted Values for Simple Linear Model in Problem 5 b)

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 :

```
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)

##
## Call:
## lm(formula = logbact ~ min)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.184303 -0.083994  0.001453  0.072825  0.206246
```

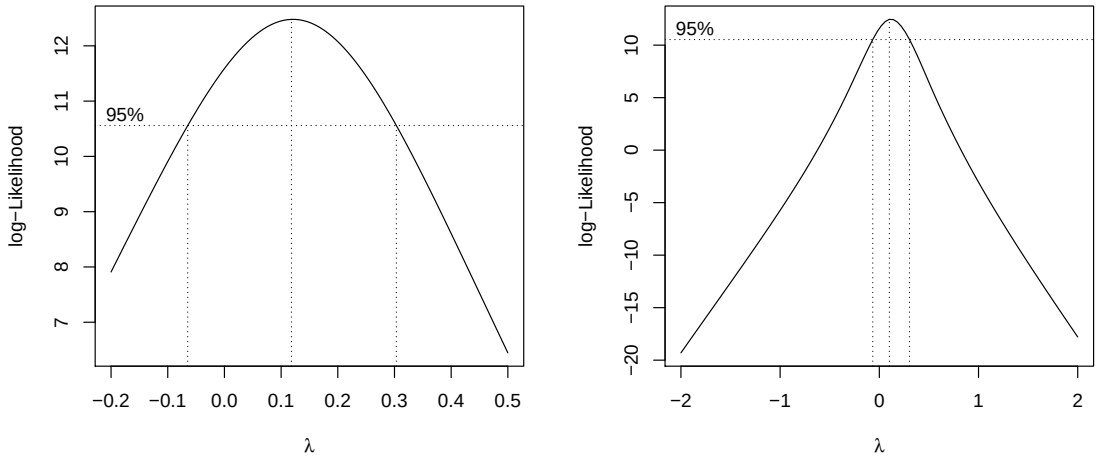

FIG. A.12 – Log-likelihood versus λ in the Box-Cox method for Problem 5 c)

```
##
## 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
##
## 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
## ---
```

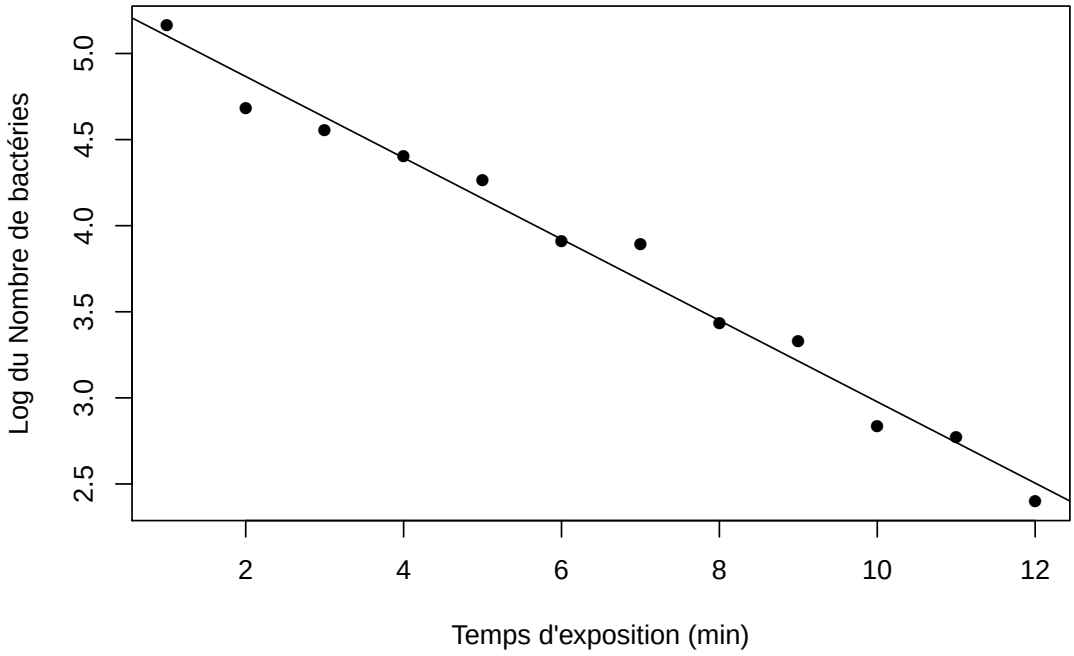

FIG. A.13 – Scatter Plot of the Logarithm of the Number of Bacteria versus the Minutes of Exposure to 300^o F

```
## 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.

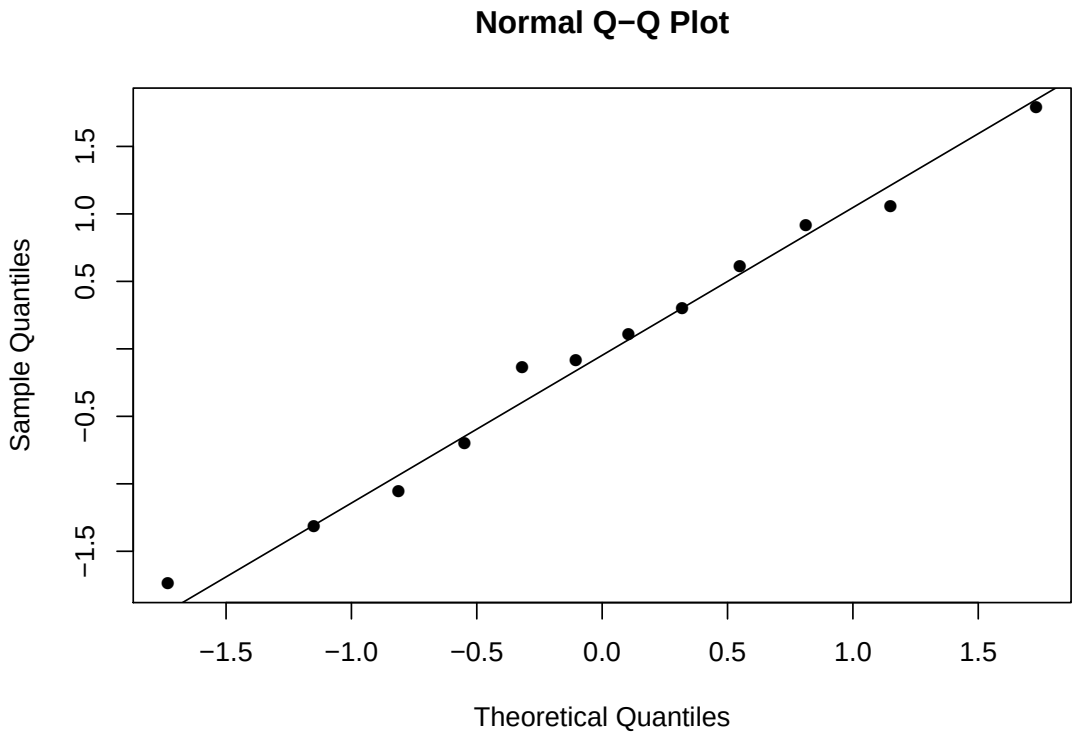

FIG. A.14 – Q-Q Plot of Model for the Logarithm of the Number of Bacteria in Problem 5 c)

Chapitre ??

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

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

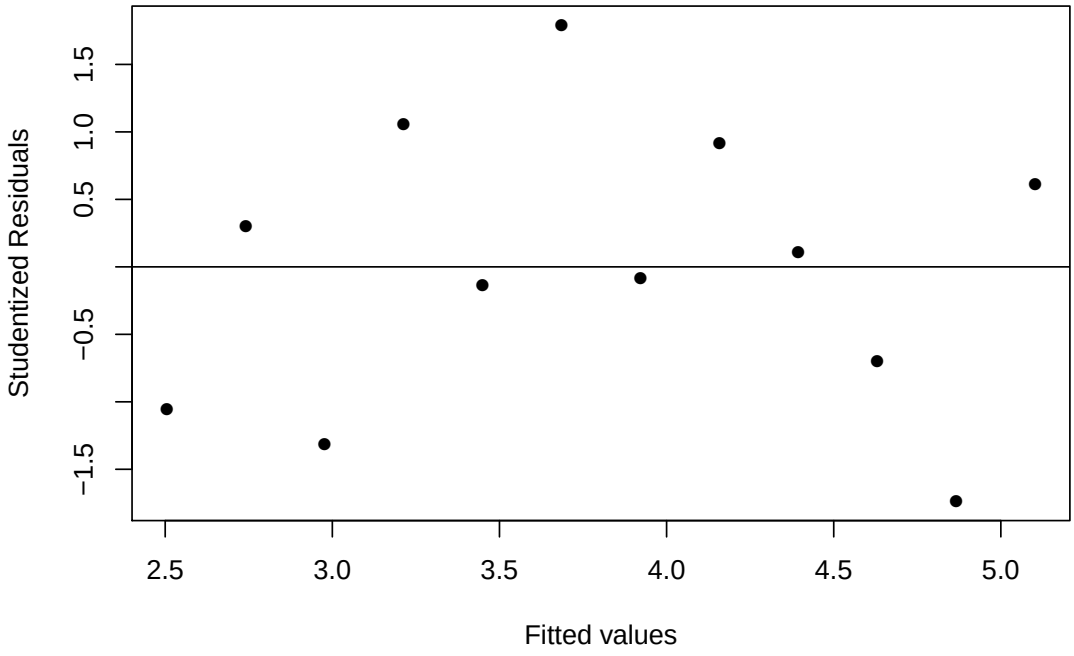

FIG. A.15 – Residuals versus the Fitted Values for Model for the Logarithm of the Number of Bacteria in Problem 5 c)

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 **??**.**??**.

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 **??**) de l'exercice **??**.**??**.

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 **??**. 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

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

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,

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

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

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

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

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

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

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

##
## Call:
## lm(formula = consommation ~ poids + cylindree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8799 -0.5595  0.1577  0.6051  1.7900
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.049304   1.098281  -2.776  0.00877 **
## poids        0.012677   0.001512   8.386 6.85e-10 ***
## cylindree   -1.122696   0.333479  -3.367  0.00186 **
## ---
## 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 **??**.**??** 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 ***
## 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)
```

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

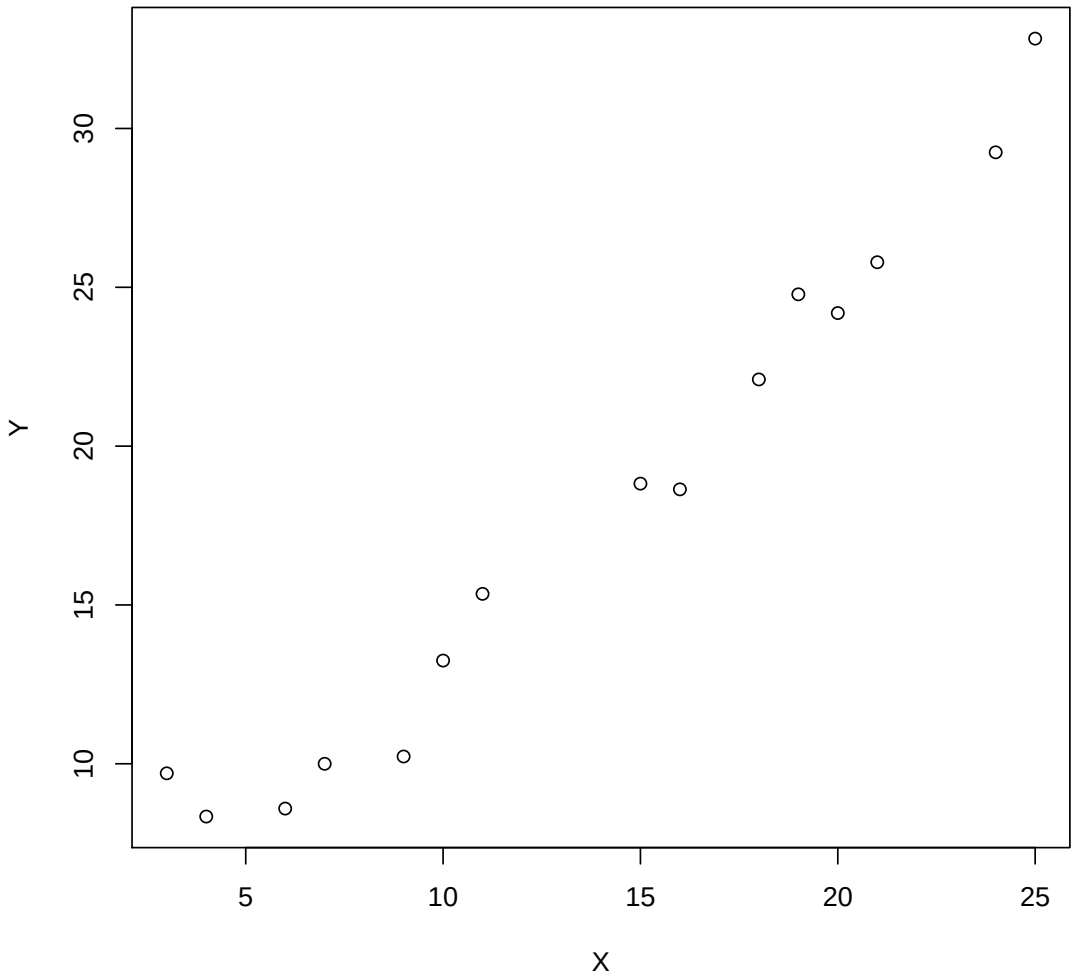

FIG. A.16 – Graphique des données de l'exercice **??.??**

```
## 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.

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))
```

```r
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))
```

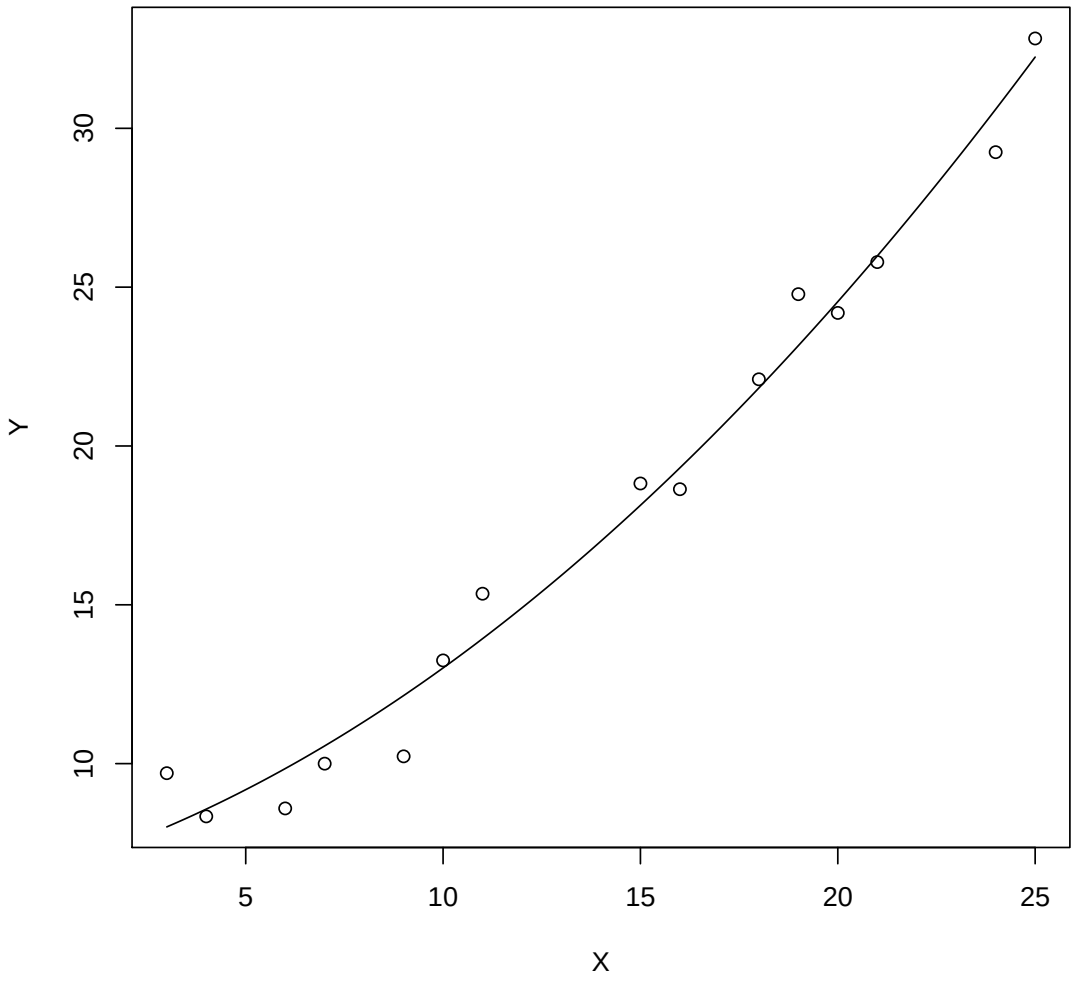

FIG. A.17 – Graphique des données de l'exercice **??.??** et courbe obtenue par régression

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

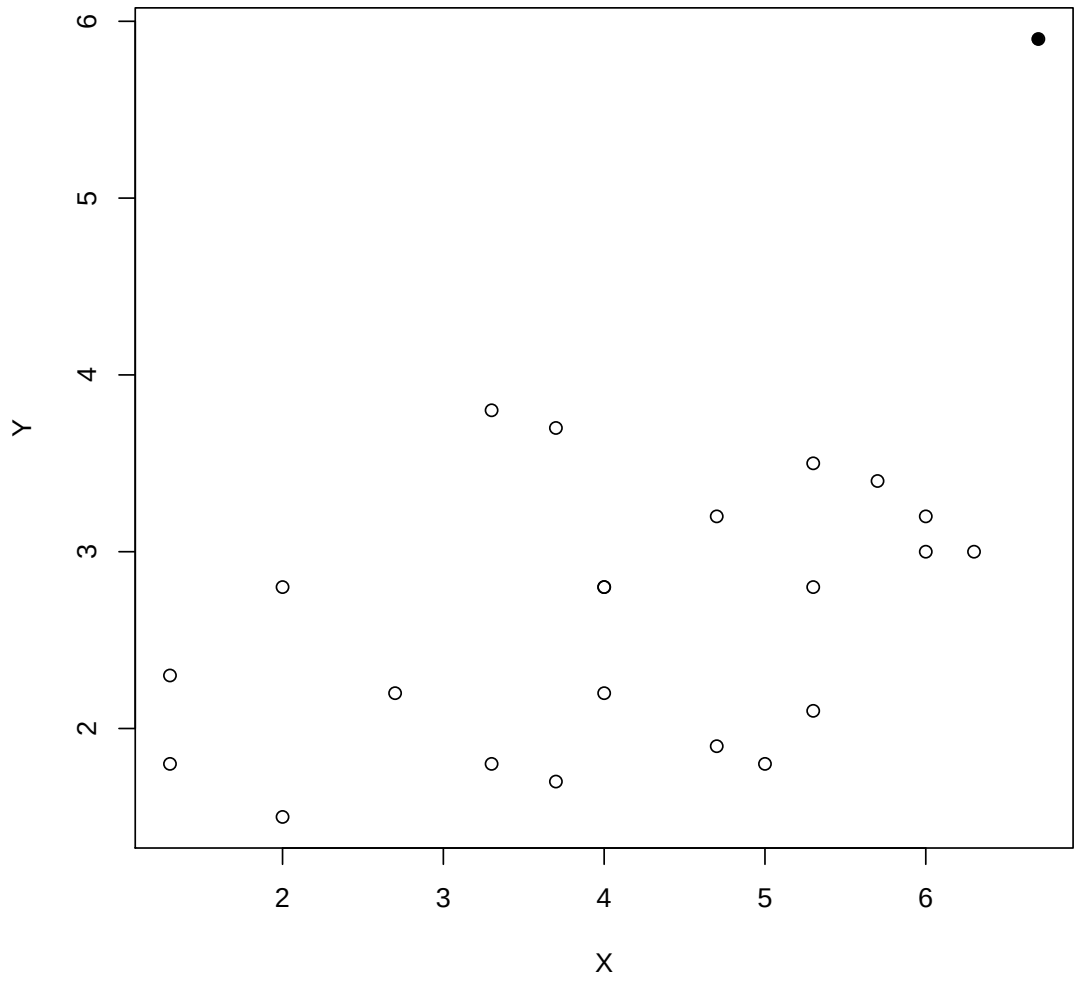

FIG. A.18 – Graphique des données de l'exercice **??.??**. Le cercle plein représente la donnée (X_{16}, Y_{16}).

```
##
## 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)
##
## 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.

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 ??

2.1 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.

2.2 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 2

2.1 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$(0,\beta)$: 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 $Y^* = Y/r$:

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

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

2.2 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.

2.3 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.

2.4 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

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

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

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

On maximise la log-vraisemblance :

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

Donc, les équations à résoudre sont

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

2.5 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

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

2.6 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 :

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

2.7 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.

2.8 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)
```

```
##
## 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.

2.9 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

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.

```r
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)
```

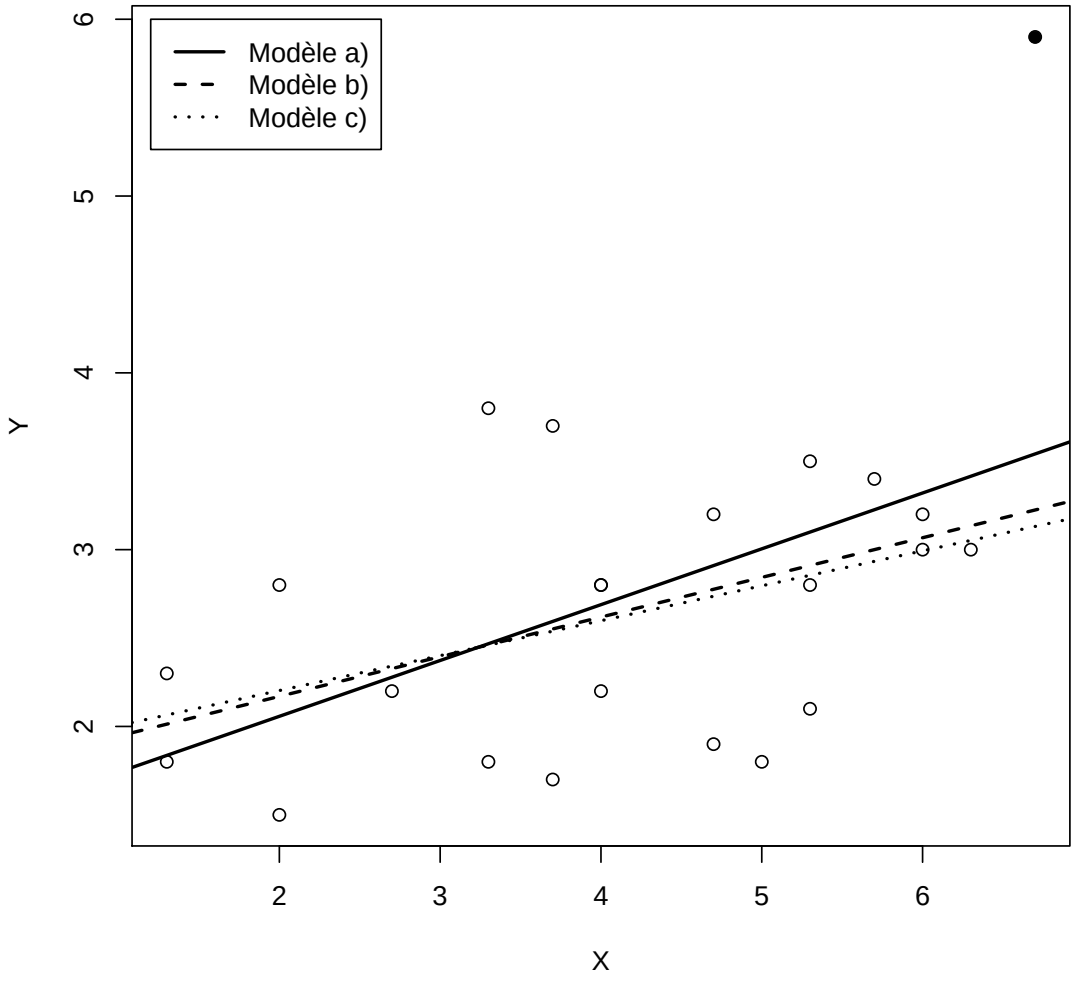

FIG. A.19 – Graphique des données de l'exercice **??.??** avec les droites de régression obtenues à l'aide des moindres carrés pondérés.

```r
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))
```

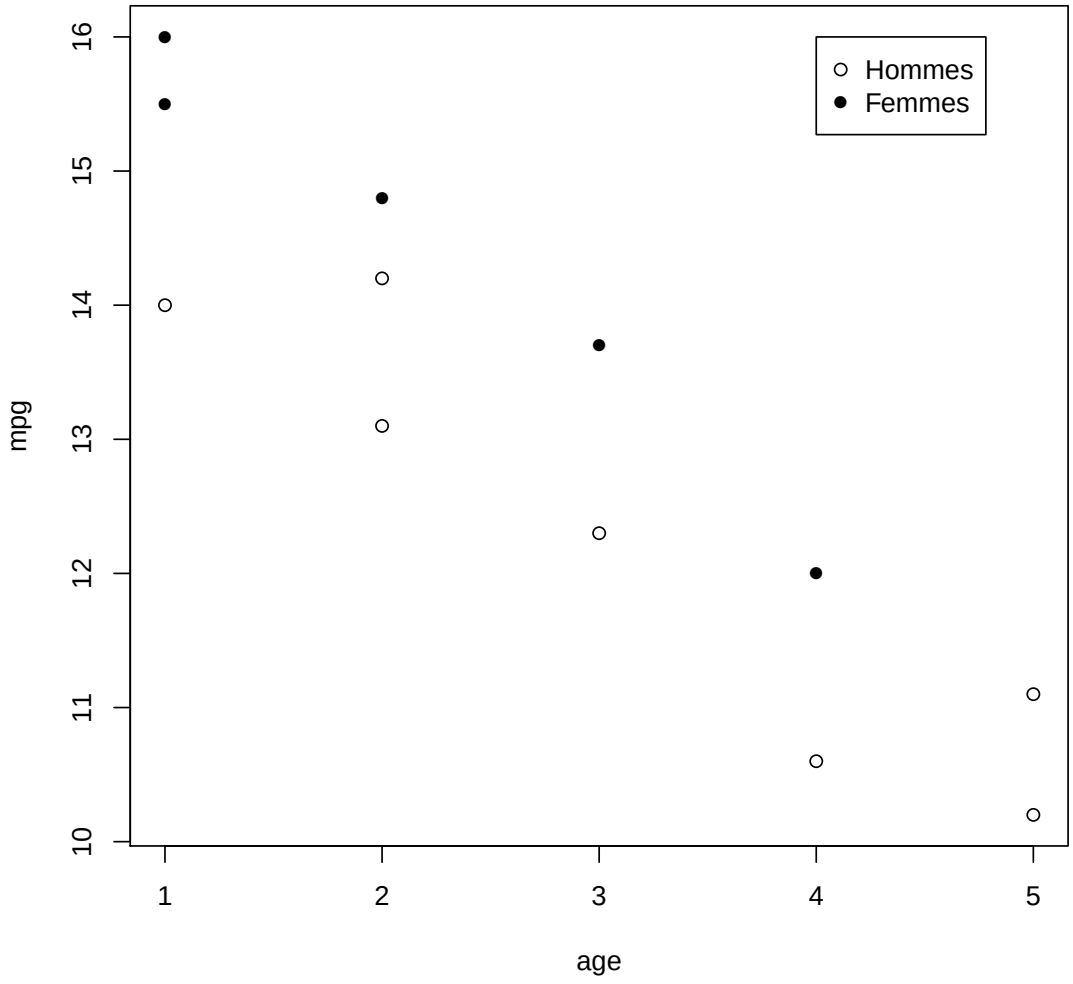

FIG. A.20 – Graphique des données de l'exercice **??.??**

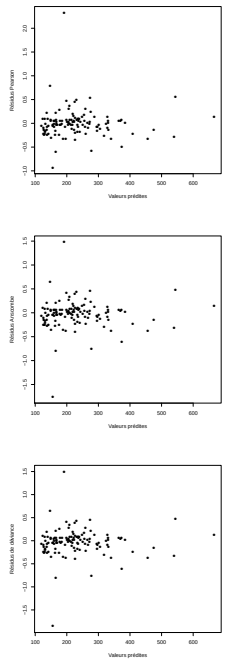

FIG. A.21 – Résidus pour GLM Gamma

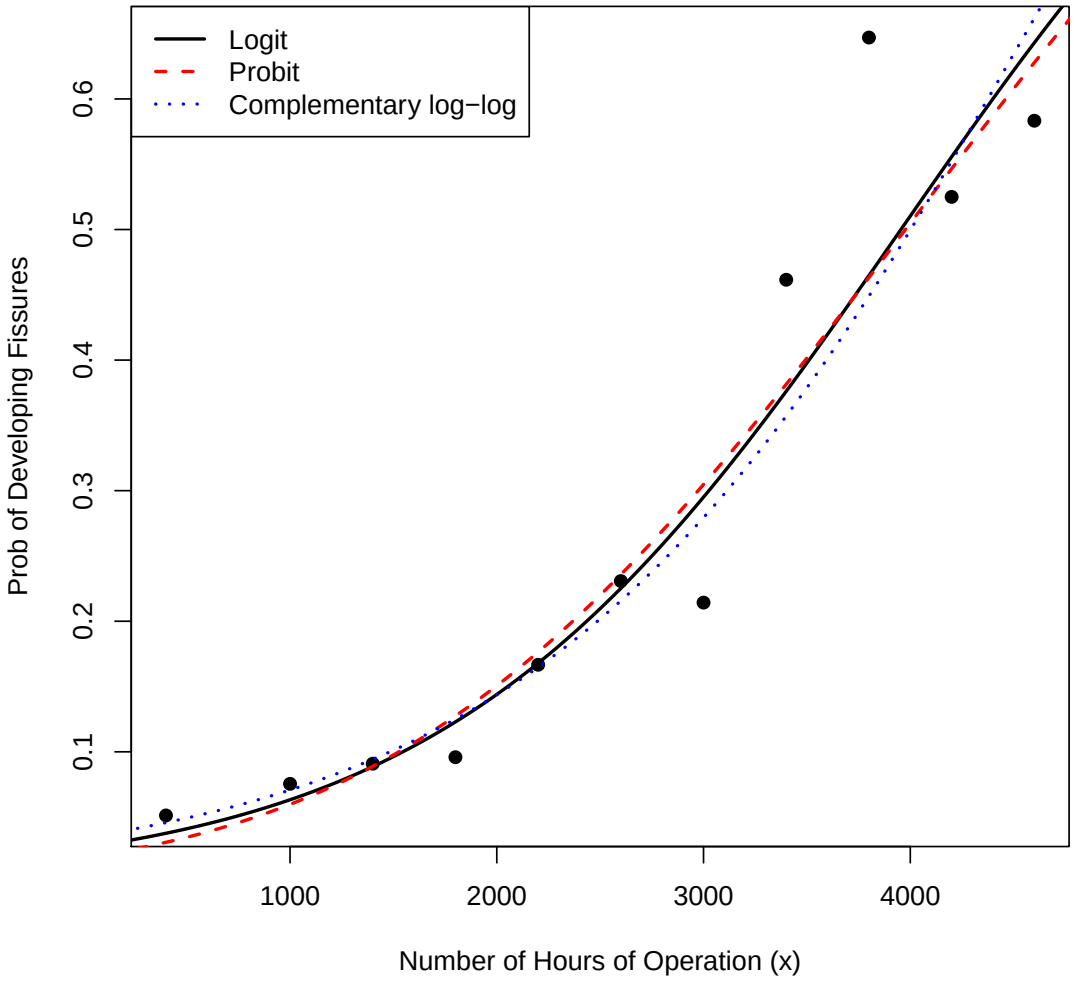

FIG. A.22 – Logistic, Probit and Complementary Log-Log Model Fit