Méthodes d'analyse biostatistique projet 1

Wen, Zehai; Li, Qingyue

2023-10-26 12:57:40-04:00

Exercice 1a

Démontrez que la loi de Poisson appartient à la famille exponentielle sous la mesure de comptage.

Solution Soit ν la mesure de comptage supportée sur $\mathbb{N}$. La loi de Poisson avec paramètre $\lambda > 0$ est définie par $f(y)d\nu(y)$, où la densité $f : \mathbb{N} \to (0, \infty)$ est donnée par $f(y) = \frac{\lambda^y}{y!}e^{-\lambda}$ pour tout $y \in \mathbb{N}$. Écrire $\lambda = e^\theta$ pour un $\theta \in \mathbb{R}$. On a alors:

$$f(y) = \exp\{-\lambda + y\ln\lambda - \ln y!\} = \exp\big\{\theta y - e^\theta - \ln y!\big\}$$

Par conséquent, la loi de Poisson appartient à la famille exponentielle avec l'espace paramétrique naturelle $\mathbb{R}$. Sous la forme $f(y) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}$, on a que $a(\phi) \equiv 1$, $b(\theta) = e^\theta$ et $c(y,\phi) = -\ln y!$.////

Exercice 1b

Afin de faire la régression logistique, soit Y, X_1, X_2 et X_3 quatre variables aléatoires à valeur $\mathbb{R}$ définies sur l'espace de probabilité commune (Ω, F, P) telles que $Y \in \{0,1\}$, $X_1 \in \{0,1\}$ et X_2 sont variables binaires catégoriels et X_3 est continue. Supposons qu'il y a quatre constantes réelles β_0, β_1, β_2 et β_3 telles que :

$$P\{Y = 1 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3\} = \pi_1(x_1, x_2, x_3) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\}}$$

Assumer premièrement que X_2 et X_3 sont constantes presque partout, montrez que β_1 peut être interprété comme un log-rapport de cotes.

Dans le deuxième cas, trouvez la différence du log-rapport de cotes pour deux individus. Le premier individu a $X_1 = 1$ et $X_3 = 7$, le deuxième individu a $X_1 = 0$ et $X_3 = 5$, et les deux individus ont la même valeur de X_2.

Solution Si $f : \mathbb{R} \to (1, \infty)$ et $f(x) = \frac{1}{1+e^{-x}}$, alors f est une bijection parce que la fonction exponentielle est une bijection. La fonction inverse de f est $f^{-1} : (1, \infty) \to \mathbb{R}$ et $f^{-1}(y) = \ln\frac{y}{1-y}$.

Supposons que $X_2 = x_2$ et $X_3 = x_3$ presque partout. On a :

$$\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 x_3 = \ln\frac{\pi_1(X_1, x_2, x_3)}{1 - \pi_1(X_1, x_2, x_3)}$$

Si $X_1 = 1$, on a:

$$\beta_1 = \ln\frac{\pi_1(1, x_2, x_3)}{1 - \pi_1(1, x_2, x_3)} - \beta_2 x_2 - \beta_3 x_3 - \beta_0$$

Donc, β_1 peut être interprété comme un log-rapport de cotes. Dans le deuxième cas, supposons que $X_2 = x_2$ pour tous les deux individus, alors on a :

$$\ln\frac{\frac{\pi_1(1,x_2,7)}{1-\pi_1(1,x_2,7)}}{\frac{\pi_1(0,x_2,5)}{1-\pi_1(0,x_2,5)}} = \ln\frac{\pi_1(1, x_2, 7)}{1 - \pi_1(1, x_2, 7)} - \ln\frac{\pi_1(0, x_2, 5)}{1 - \pi_1(0, x_2, 5)} = \beta_1 + 2\beta_3$$

Le calcul est complet.////

Exercice 1c

Soit X, Y et Z trois variables aléatoires à valeur $\mathbb{R}$ définies sur l'espace de probabilité commune (Ω, F, P) telles que :

1. Pour tout $\omega \in \Omega$, la probabilité conditionnelle régulière $\mu_{Y|X,Z}(\omega, \cdot)$ est la loi de Poisson avec paramètre:

$$\exp(\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega))$$

   Où β_0, β_1 et β_2 sont des constantes réelles.

2. On a :

$$\mathbb{E}[Y \mid X] = \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]$$

Calculer Var$(Y \mid X)$.

Solution On a, pour chaque $\omega \in \Omega$:

$$\mathbb{E}[Y^2 \mid X, Z](\omega) = \int_{\mathbb{R}} y^2 d\mu_{Y|X=X(\omega),Z=Z(\omega)}(y) = \exp\{2(\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega))\} + \exp\{\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega)\}$$

Alors :

$$\mathbb{E}[Y^2 \mid X] = \mathbb{E}[\mathbb{E}[Y^2 \mid X, Z] \mid X] = \exp\{2(\beta_0 + \beta_1 X)\}\mathbb{E}[e^{2\beta_2 Z} \mid X] + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]$$

Donc, en utilisant la formule de la variance conditionnelle :

$$\begin{aligned}
&\text{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2 \\
&= \exp\{2(\beta_0 + \beta_1 X)\}\mathbb{E}[e^{2\beta_2 Z} \mid X] + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X] - \left(\exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]\right)^2 \\
&= \exp\{2(\beta_0 + \beta_1 X)\}\text{Var}(e^{\beta_2 Z} \mid X) + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]
\end{aligned}$$

Le calcul est complet. ////

Exercice 1d

Expliquez assez brièvement, pourquoi les modèles de régression linéaire généralisés ne nécessitent pas réellement la spécification complète de la distribution de la variable de réponse pour l'estimation des coefficients.

Solution Pour une mesure σ-finie ν définie sur une espace mesurable (Ω, F), une famille $\mathcal{P} = \{f_\theta d\nu : \theta \in \Theta \subseteq \mathbb{R}\}$ continue absolument est appelé une famille exponentiel si, dans une forme plus simple et comme dans les notes de cours, on a pour tout $x \in \Omega$ que :

$$f_\theta(x) = \exp\left\{\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right\}$$

Ici $\phi > 0$ est un paramètre de dispersion, $a : \mathbb{R} \to \mathbb{R}$, $b : \mathbb{R} \to \mathbb{R}$ et $c : \Omega \times \mathbb{R} \to \mathbb{R}$ sont des fonctions telles que $a \neq 0$, b est trois fois différentiables et $b'' > 0$. On obtient les modèles linéaires généralisés si on utilise

$\theta = g(\eta)$ pour une fonction de lien $g : \mathbb{R} \to \mathbb{R}$ et $\eta = \langle x, \beta \rangle_{\mathbb{R}^p}$ pour un vector constante $x \in \mathbb{R}^p$ et les paramètres $\beta \in \mathbb{R}^p$. Soit Y une réalisation de $f_\theta d\nu \in \mathcal{P}$. Comme dans le notes de cours, on a que :

$$\mathbb{E}[Y] = b'(\theta); \quad \text{Var}(Y) = b''(\theta)a(\phi)$$

On observe que la variance est une fonction de moyen, c'est-à-dire, on peut tout simpliment dériver le moyen afin d'obtenir la variance. En effet, si Y a plus de moments, les propriétés de la famille exponentiel ditent que tout les moments peuvent être dérivés en fonction du moyen. Si Y a une fonction génératrice des moments, alors sa fonction génératrice des moments est aussi contrôlé par le moyen. En conclusion, contrôler le moyen égale contrôler la distribution si Y est suffisament "régulière". Le pramètre β, dans un autre côté, ne dépend que le moment. Il s'agit toujours d'estimer le moment et puis utiliser le moment pour estimer β en faisant les inversions des fonctions ou matrices, sans spécifier la distribution de Y.////

Exercice 1e

Soit $\{Y_i\}_{i=1}^m$, $\{a_i\}_{i=1}^m$, $\{b_i\}_{i=1}^m$, $\{\epsilon_i\}_{i=1}^m$ trois suites de variables aléatoires à valeur $\mathbb{R}^5$, $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}^5$ respectivement tout définies sur une espace de probabilité commune (Ω, F, P). On suppose que :

1. Les quatres σ-algébres générées par les quatre suites sont indépendantes.
2. Il y existe une matrice X de taille 5×5. Pour tout $j \in \{1, \ldots, 5\}$, on a, presque partout :

$$Y_{ij} = a_i + b_i X_{ij} + \beta X_{ij} + \epsilon_{ij}$$

3. Il existe constantes $\tau > 0$, $\psi > 0$, $\sigma > 0$ et $\rho \in (0, 1)$ tels que, pour tout $i \in \{1, \cdots, m\}$, a_i est la loi normale avec moyen 4 et variance τ^2, b_i est la loi normale avec moyen 0 et variance ψ^2 et:

$$\text{Cov}(\epsilon_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Séparez et présentez le formule de Y_{ij} comme deux composantes: une composante de régression linéaire et une erreur. Calculer $\text{Cov}(Y_i)$ pour chaque $i \in \{1, \ldots, m\}$.

Solution Comme la partie de régression linéaire contient les termes déterministes ou systématiques du modèle, le modèle peut être séparé comme, pour tout $i \in \{1, \ldots, m\}$ et $j \in \{1, \ldots, 5\}$:

$$Y_{ij} = a_i + (b_i + \beta)X_{ij} + W_{ij}$$

Ici $a_i + (b_i + \beta)X_{ij}$ est la partie de régression linéaire et W_{ij} est la partie d'erreur. Ensuite, comme les quatre suites sont indépendants, on obtient, pour chaque $i \in \{1, \ldots, m\}$ et $j \in \{1, \ldots, 5\}$, que :

$$\text{Var}(Y_{ij}) = \text{Var}(a_i) + \text{Var}(b_i X_{ij}) + \text{Var}(\epsilon_{ij})$$

En utilisant les normalités assumées :

$$\text{Var}(Y_{ij}) = \tau^2 + \psi^2 X_{ij}^2 + \sigma^2$$

Si $j \neq k$, on a :

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(a_i + b_i X_{ij} + \epsilon_{ij}, a_i + b_i X_{ik} + \epsilon_{ik}) = \tau^2 + \psi^2 X_{ij} X_{ik} + \sigma^2 \rho^{|j-k|}$$

Le calcul est complet. ////

Exercice 2abcd

Reprendre le jeu de données sur les limules disponible à partir du progiciel `rsq` en R. La variable de réponse est toujours le nombre de `num.satellites`. Premièrement, ajustez un modèle de régression de Poisson utilisant les variables explicatives `color`, `spine`, et `width`. Évaluez la sur-dispersion potentielle en faisant un test de sur-dispersion et en évaluant le paramètre de sur-dispersion. Discutez des résultats. Supposons que le test de sur-dispersion rejète l'hypothèse nulle. Expliquez ce que rejetter cette hypothèse signifie, et suggérez une approche alternative sans analyse.

Ensuite, choisissez et réalisez une cretière pour sélectionner le meilleur modèle parmi les modèles possibles utilisant les varaibles explicatives `color`, `spine`, `width`, et `weight`. Présentez une analyse de résidus et discutez les observations.

Solution Tout d'abord, nous importons l'ensemble de données et examinons les informations de base sur les données.

```
library(rsq)

data("hcrabs")
attach (hcrabs)
```

Ensuite,nous réalisons un modèle de régression de Poisson avec la variable de réponse `num.satellites` et les variables explicatives `color`, `spine`, et `width`.

```
modele_poisson <-
glm ( num.satellites ~ width + spine + color, family = poisson ( link = log ) ,
data = hcrabs )
```

On fait puis le test de surdispersion et évaluer le paramèter de sur-dispersion.

```
library (AER)
print(dispersiontest(modele_poisson))
```

Le test donne que le paramètre de dispersion est 3.143975 avec p-valeur $4.07e-08$. Par conséquent, la sur-dispersion est significative. Alors la dispersion observée dans les données dépasse ce qui serait attendu sous une distribution de Poisson. Dans notre cas, les variations du nombre de satellites mâles ne sont pas bien prises en compte dans le modèle de Poisson standard et des modèles de surdiscrétisation plus complexes peuvent être envisagés pour mieux s'adapter à ces données. Lorsque nous effectuons une régression de Poisson, nous supposons que la variance est égale à la moyenne, une propriété inhérente à la distribution de Poisson. Cependant, les données réelles peuvent ne pas toujours respecter cette supposition.

Comme on a une surdispersion, il faudrait avoir une hétérogénéité non observée ou une accumulation de comptages issus de plusieurs processus indépendants. Face à la surdispersion, on se tourne généralement vers d'autres modèles plus complexes pour les données de comptage. Une choix est d'utilier un modèle mixte, qui s'appelle modèle binomial négatif. Il s'agit de commencer avec un modèle de Poisson avec le paramètre estimé par une distribution de Gamma.

Passons au sujet prochain, il y a $2 \times 4 - 1 = 15$ modèles possibles si on ne compte pas l'intercept. On choisit le critère AIC pour sélectionner le meilleur modèle. On rappelle que, si on a k modèles qui ont les valeurs d'AIC $\text{AIC}_1, \cdots, \text{AIC}_k$ respectivement, les valeurs $\exp\{-\frac{1}{2}(\text{AIC}_i - \text{AIC}_{min})\}$ sont les probabilités que le modèle i minimisant la perte d'informations.

```
modele_color <- glm ( num.satellites ~ color, family = poisson ( link = log )
   ,data = hcrabs )
```

```
modele_spine <- glm ( num.satellites ~ spine, family = poisson ( link = log )
   ,data = hcrabs )
modele_width <- glm ( num.satellites ~ width, family = poisson ( link = log )
   ,data = hcrabs )
modele_weight <- glm ( num.satellites ~ weight, family = poisson ( link = log )
   ,data = hcrabs )
modele_color_spine <- glm ( num.satellites ~ color + spine, family = poisson (
   link = log ) ,data = hcrabs )
modele_color_width <- glm ( num.satellites ~ color + width, family = poisson (
   link = log ) ,data = hcrabs )
modele_color_weight <- glm ( num.satellites ~ color + weight, family = poisson (
   link = log ) ,data = hcrabs )
modele_spine_width <- glm ( num.satellites ~ spine + width, family = poisson (
   link = log ) ,data = hcrabs )
modele_spine_weight <- glm ( num.satellites ~ spine + weight, family = poisson (
   link = log ) ,data = hcrabs )
modele_width_weight <- glm ( num.satellites ~ width + weight, family = poisson (
   link = log ) ,data = hcrabs )
modele_color_spine_width <- glm ( num.satellites ~ color + spine + width, family
   = poisson ( link = log ) ,data = hcrabs )
modele_color_spine_weight <- glm ( num.satellites ~ color + spine + weight,
   family = poisson ( link = log ) ,data = hcrabs )
modele_color_width_weight <- glm ( num.satellites ~ color + width + weight,
   family = poisson ( link = log ) ,data = hcrabs )
modele_spine_width_weight <- glm ( num.satellites ~ spine + width + weight,
   family = poisson ( link = log ) ,data = hcrabs )
modele_color_spine_width_weight <- glm ( num.satellites ~ color + spine + width +
   weight, family = poisson ( link = log ) ,data = hcrabs )

modeles <- list(modele_color, modele_spine, modele_width, modele_weight,
modele_color_spine, modele_color_width, modele_color_weight, modele_spine_width,
modele_spine_weight, modele_width_weight, modele_color_spine_width,
modele_color_spine_weight, modele_color_width_weight,
modele_spine_width_weight, modele_color_spine_width_weight)

AICs <- rep(0, 15)
for (i in 1:15) {
  AICs[i] <- AIC(modeles[[i]])
}

print(AICs)

min_AIC <- min(AICs)
proba <- rep(0, 15)
for (i in 1:15) {
  proba[i] <- exp(0.5*(min_AIC- AICs[i]))
}

print(proba)

print(which.max(proba))
```

À la fin, on trouve que le modèle color-weight est le meilleur modèle. Pour le résidus, on choisit les

résidus d'Anscombe. Les résidus d'Anscombe sont spécifiquement conçus pour les modèles linéaires généralisés et offrent de bonnes propriétés pour identifier des valeurs atypiques ou des observations influentes. Ce lien `https://www.sfu.ca/sasdoc/sashtml/insight/chap39/sect57.htm` fournit des informations supplémentaires sur les résidus d'Anscombe.

```
library(surveillance)
plot(anscombe.residuals(modele_color_weight, phi =1))
```

Toutes les modélisations sont complètes.////

Exercice 2e

Soit $\{X^{(i)} = (X_1^{(i)}, X_2^{(i)})\}_{i=1}^n$ et $\{Y_i\}_{i=1}^n$ deux suites de variables aléatoires indépendantes et identiquement distribuées à valeur $\mathbb{R}^2$ et $\mathbb{R}$ respectivement définies sur un espace de probabilité commun $(\Omega, \mathcal{F}, P)$ tels qu'il existe trois constantes β_0, β_1 et β_2 réels avec, pour chaque $i \in \{1, \cdots, n\}$:

$$\mathbb{E}\left[Y_i \mid X^{(i)}\right] = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)}$$

Avec les hypothèses appropriées, postulez une équation d'estimation pour β_0, β_1 et β_2. Calculez une expression pour β_0, β_1 et β_2 et ses variances étant donné les X. Réalisez tout utilisant le donnée `addhealth` avec $Y_i = \texttt{Weight}_i$, $X_1^{(i)} = \texttt{age}_i$ et $X_2^{(i)} = \texttt{SES}_i$.

Solution On fait les hypothèses suivantes:

1. La règle de décision est moindre carrés.
2. Seulement pour le calcul des variances, on suppose que Y_1 a distribution $N(0, \sigma^2)$.

Pour simplicité, écrire :

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(n)} & X_2^{(n)} \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}; \quad S(X, Y, \beta) = (Y - X\beta)^T (Y - X\beta)$$

L'espace de paramètre est $\mathbb{R}^3$. Selon la définition d'éaquation d'estimation, il s'agit de trouver, pour chaque $i \in \{1, \cdots, n\}$, une fonction $\psi_i : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ telle que, pour tout $\beta \in \mathbb{R}^3$, on a :

$$\mathbb{E}\left[\sum_{i=1}^n \psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta)\right] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

On écrit $\hat{\beta}$ pour la solution de l'équation d'estimation, qui est une variable aléatoire dépendant Y et X, telle que $\sum_{i=1}^n \psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \hat{\beta}) = 0$ presque partout. Selon la première hypothèse, la règle de décision est :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^3}{\arg\min}\, S(X, Y, \beta)$$

Donc on utilise :

$$\frac{\partial}{\partial\beta}S(X,Y,\beta) = -2X^T(Y - X\beta) = \sum_{i=1}^{n}\begin{bmatrix}(-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)})\\(-2Y_i X_1^{(i)} + 2\beta_0 X_1^{(i)} + 2\beta_1 (X_1^{(i)})^2 + 2\beta_2 X_1^{(i)} X_2^{(i)})\\(-2Y_i X_2^{(i)} + 2\beta_0 X_2^{(i)} + 2\beta_1 X_1^{(i)} X_2^{(i)} + 2\beta_2 (X_2^{(i)})^2)\end{bmatrix}$$

$$= \sum_{i=1}^{n}\psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta) = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

Comme :

$$\mathbb{E}[\psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta)] = \begin{bmatrix}\mathbb{E}\left[(-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)})\right]\\\mathbb{E}\left[(-2Y_i X_1^{(i)} + 2\beta_0 X_1^{(i)} + 2\beta_1 (X_1^{(i)})^2 + 2\beta_2 X_1^{(i)} X_2^{(i)})\right]\\\mathbb{E}\left[(-2Y_i X_2^{(i)} + 2\beta_0 X_2^{(i)} + 2\beta_1 X_1^{(i)} X_2^{(i)} + 2\beta_2 (X_2^{(i)})^2)\right]\end{bmatrix}$$

$$= \begin{bmatrix}\mathbb{E}\left[\mathbb{E}\left[(-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)}) \mid X^{(i)}\right]\right]\\\mathbb{E}\left[X_1^{(i)}\mathbb{E}\left[-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)} \mid X^{(i)}\right]\right]\\\mathbb{E}\left[X_2^{(i)}\mathbb{E}\left[-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)} \mid X^{(i)}\right]\right]\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

On conclut que l'équation définie est une équation d'estimation. On a donc :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

L'inverse peut être un inverse généralisé si $X^T X$ n'est pas inversible. Selon la deuxième hypothèse, on sait immédiatement que $\text{Var}(\hat{\beta} \mid X) = \sigma^2 (X^T X)^{-1}$. Maintenant, il s'agit de réaliser tout utilisant le donnée `addhealth` avec $Y_i = \texttt{Weight}_i$, $X_1^{(i)} = \texttt{age}_i$ et $X_2^{(i)} = \texttt{SES}_i$. Ce travail ne consiste que l'utilisation de la fonction `lm` en R :

```
set.seed(1234)

donnee <-
   read.delim("Chapters\\biostat_projet_1\\resource_content_1_addhealth.txt",
   header = TRUE)
#donnee <- read.delim("resource_content_1_addhealth.txt", header = TRUE)

attach(donnee)

#On commence par le netoyage
donnee$feeling_depressed <- as.factor(donnee$feeling_depressed)
donnee$feeling_depressed[is.na(donnee$feeling_depressed)] <- as.factor(
   floor(runif(sum(is.na(donnee$feeling_depressed)), min = 1, max = 4.9999)))

donnee$smoking <- as.factor(donnee$smoking)

donnee$weight <- as.numeric(donnee$weight)
donnee$weight[is.na(donnee$weight)] <- mean(donnee$weight, na.rm = TRUE)

donnee$time <- as.factor(donnee$time) #identiquement 1

donnee$age <- as.numeric(donnee$age)
```

```
donnee$age[is.na(donnee$age)] <- as.factor(mean(donnee$age, na.rm = TRUE))

donnee$sex <- as.factor(donnee$sex)
donnee$sex[is.na(donnee$sex)] <- as.factor(floor(runif(sum(is.na(donnee$sex)),
   min = 1, max = 2.9999)))

donnee$SES <- as.numeric(donnee$SES)
donnee$SES[is.na(donnee$SES)] <- as.numeric(floor(mean(donnee$SES, na.rm =
   TRUE)))

attach(donnee)

modele_moindre_carre <- lm(weight ~ SES + age, data = donnee)
print(summary(modele_moindre_carre))
```

Le code donne que, selon les donné, on a $\beta_0 = 59.5309$, $\beta_1 = 5.5819$ et $\beta_2 = -0.3637$. Les variances sont respectivement $3.7017^2 = 13.70258289$, $0.2223^2 = 0.04941729$ et $0.2149^2 = 0.4618201$. Le calcul et la démonstration sont complets. ////

Exercice 3

On s'intéresse à l'ensemble de variables suivantes: `age`, `sex`, `SES`, `smoking` et `weight`, notant que les variables `ID` et `time discrete` représentent respectivement l'identifiant de l'individu et le temps (il y a quatre temps: vague 1, 2, 3 ou 4) dans le jeu de données `addhealth long`.

Proposez une structure de corrélations entre les mesures de poids de chaque individu. Ensuite, réalisez et comparez les trois modèles suivants utilisant la corrélation résiduelle :

1. Un modèle de régression linéaire pour la variable dépendente `weight` sans aucune variable explicative, c'est-à-dire just une ordonnée à l'origine.
2. Un modèle de régression linéaire pour la variable dépendente `weight` et les variables explicatives `age` et `sex`.
3. Un modèle de régression linéaire pour la variable dépendente `weight` et les variables explicatives `age`, `sex`, `SES` et `smoking`.

Finalement, afin d'étudier la rélation entre la variable dépendent `weight` et la variable explicative `SES`, on propose quatre modèles suivants:

1. Un modèle linéaire en assumant l'indépendance entre toutes les observations.
2. Un modèle linéaire généralisé (normale) où on assume une matrice de corrélation interchangeable.
3. GEE avec une matrice de corrélation interchangeable.
4. Un modèle mixte avec une ordonnée à l'origine et une pente aléatoire pour chaque individu.

Comparez les quatre modèles utilisant la variance de la coefficient de la variable continue `SES`, pour laquelle faisez un intervalle de confiance.

Solution Parmi les quatres stuctures mentionnés en classe, on décide que la matrice de corrélation est celui de Toepliz. On ne choisit pas les trois autres parce que les poids d'un individu entre les points différents de temps ne sont pas indépendents ou interchangeable et que l'on croit pas les poids forment une chaîne de Markov. On choisit Toepliz parce qu'il y a une « délai » entre les observations et que nous n'avons pas trouvé une autre matrice de corrélation en ligne qui semble plus appropriée.

D'abord, on réalise les trois modèles linéaires :

1. Le modèle avec seulement une ordonnée à l'origine :

```
modeles_lineaires_aucune <- list()
residues_aucune <- rep(1,4)
for (i in 1:4){
   modele_i <- lm(donnee_separe[[i]]$weight ~ 1)
   modeles_lineaires_aucune <- append(modeles_lineaires_aucune, modele_i)
   residues_aucune[i] <- mean(residuals(modele_i))
}
```

2. Le modèle avec `age` et `sex` comme variables explicatives :

```
modeles_lineaires_age_sex <- list()
residues_age_sex <- rep(1,4)
for (i in 1:4){
   modele_i <- lm(donnee_separe[[i]]$weight ~ donnee_separe[[i]]$age +
      donnee_separe[[i]]$sex)
   modeles_lineaires_age_sex <- append(modeles_lineaires_age_sex, modele_i)
   residues_age_sex[i] <- mean(residuals(modele_i))
}
```

3. Le modèle avec `age`, `sex`, `SES` et `smoking` comme variables explicatives :

```
modeles_lineaires_age_SES_sex_smoking <- list()
residues_age_SES_sex_smoking <- rep(1,4)
for (i in 1:4){
   modele_i <- lm(donnee_separe[[i]]$weight ~ donnee_separe[[i]]$age +
      donnee_separe[[i]]$sex +
donnee_separe[[i]]$SES + donnee_separe[[i]]$smoking)
   modeles_lineaires_age_SES_sex_smoking <- append(
modeles_lineaires_age_SES_sex_smoking, modele_i)
   residues_age_SES_sex_smoking[i] <- mean(residuals(modele_i))
}
```

On imprime puis les corrélations résiduelles :

```
print("Les correlations residuelles sont :")
print(cor(residues_aucune, residues_age_sex)) # -0.5644551
print(cor(residues_age_sex, residues_age_SES_sex_smoking)) # -0.9982123
print(cor(residues_age_SES_sex_smoking, residues_aucune)) # 0.5159088
```

- -0.5644551: Cette valeur représente la corrélation entre les résidus du premier modèle et ceux du deuxième modèle (avec *age* et *sex* comme variables explicatives). Cette corrélation négative signifie que lorsque les résidus du premier modèle augmentent, ceux du deuxième modèle ont tendance à diminuer, et vice versa.

- -0.9982123: Cette valeur représente la corrélation entre les résidus du deuxième modèle et ceux du troisième modèle. Cette corrélation, très proche de -1, indique une forte relation négative entre les résidus des deux modèles. Cela signifie que lorsque les résidus du deuxième modèle augmentent, ceux du troisième modèle ont tendance à diminuer, et vice versa.
- 0.5159088: Cette valeur représente la corrélation entre les résidus du troisième modèle et ceux du premier modèle. Cette corrélation positive signifie que lorsque les résidus du troisième modèle augmentent, ceux du premier modèle ont également tendance à augmenter, et vice versa.

À partir de maintenant, on utilise `SES` comme une variable explicative pour la variable de réponse `weight`. Les modèles démandés sont réalisés ci-dessous :

1. Modèle linéaire en assumant l'indépendance entre toutes les observations:

```
modele_lineaire <- lm(weight ~ SES, data = donnee)
print(summary(modele_lineaire))
```

   On trouve que l'écat type d'erreur pour `SES` est 0.1328.

2. Modèle linéaire généralisé (normale) où on assume une matrice de corrélation interchangeable:

```
library(nlme)
glm_normal <- gls(weight ~ SES, data = donnee, correlation =
   corCompSymm(form = ~ 1 | SES))
print(summary(glm_normal))
```

   On trouve que l'écat type d'erreur pour `SES` est 0.1868304.

3. GEE avec une matrice de corrélation interchangeable:

```
require(geepack)
GEE <- summary(geese(weight ~ SES, id = ID, data = donnee, corstr =
   'exchangeable'))
print(GEE)
```

   On trouve que l'écat type d'erreur pour `SES` est 0.1646921.

4. Modèle mixte avec une ordonnée à l'origine et une pente aléatoire pour chaque individu:

```
library(lme4)
modele_mixte <- lmer(weight ~ (1 + SES | ID), data = donnee)
print(summary(modele_mixte))
print(confint(modele_lineaire))
```

   Nous avons manipulé beaucoup d'options. Ce modèle ne converge jamais.

On sait que des erreurs standard plus petites impliquent des estimations plus précises des paramètres du modèle, tandis que des erreurs standard plus grandes indiquent une plus grande incertitude. Ainsi, parmi les quatre modèles mentionnés ci-dessus, le permier modèle présente l'erreur standard la plus faible (0.1328). Par conséquent, le premier modèle est le meilleur modèle pour les données. On puis construire le 95% intervalle de confiance :

```
print(confint(modele_lineaire))
```

L'intervalle pour l'ordonnée à l'origine est $[161.6002042, 165.2151458]$ et l'intervalle pour la pente est $[-0.6829711, -$
Donc, en général, on peut dire que le poids diminue lorsque `SES` augmente. ////