Méthodes d'analyse biostatistique projet 1

Li, Qingyue; Wen, Zehai

2023-10-16 12:28:29-04:00

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 $(\Omega, \mathcal{F}, P)$ 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 $(\Omega, \mathcal{F}, P)$ 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

Les modèles linéaires généralisés (GLM) sont formulés comme :

$$g(E[Y]) = \mathbf{X}\beta$$

où,

- $E[Y]$ est la valeur attendue de la variable de réponse Y.
- g est la fonction de lien.
- $\mathbf{X}$ est la matrice de conception.
- β est le vecteur de paramètres à estimer.

Considérons les composants principaux des GLM :

1. **Fonction de lien et moyenne** : Les GLM supposent qu'une fonction de la variable de réponse (souvent la moyenne) est une combinaison linéaire des variables prédictives. Par exemple, dans la régression logistique, la log-vraisemblance de la réponse est une fonction linéaire des variables prédictives :

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}\beta$$

où $p = E[Y]$ est la probabilité que la réponse soit 1.

2. **Fonction de vraisemblance** : Les paramètres du modèle sont estimés en maximisant la fonction de vraisemblance. Pour un ensemble de données observées, la fonction de vraisemblance est donnée par :

$$L(\beta) = f(y|\mathbf{X}, \beta)$$

où f est la fonction de densité de probabilité des données pour un ensemble donné de paramètres β. Notez que nous nous soucions uniquement de la valeur de la densité de probabilité à ce moment précis, et non de la distribution complète de Y.

3. **Pourquoi une distribution complète n'est-elle pas nécessaire ?** Notre objectif est de trouver les valeurs des paramètres qui maximisent la fonction de vraisemblance. Nous devons donc uniquement savoir quelle est la vraisemblance des données observées pour un ensemble donné de paramètres, sans nous soucier du reste de la distribution. Nous n'avons pas besoin de connaître la forme entière de la distribution, mais seulement comment la fonction de densité ou de masse de probabilité se comporte pour les données observées sous un ensemble donné de paramètres.

En conclusion, l'estimation des paramètres dans les GLM se concentre principalement sur la vraisemblance des données sous un ensemble donné de paramètres, et non sur la distribution complète de Y. ////

Exercice 1e

1. **Définition du modèle**:

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

où:

- Y_{ij} est la réponse observée pour l'individu i à l'occasion j.
- a_i est l'interception aléatoire pour l'individu i.
- b_i est la pente aléatoire pour l'individu i.
- β est la pente fixe associée à la variable explicative X.
- X_{ij} est la variable explicative pour l'individu i à l'occasion j.
- ϵ_{ij} est l'erreur résiduelle pour l'individu i à l'occasion j.

2. **Hypothèses sur les composants aléatoires**:

$$a_i \sim N(4, \tau^2)$$
$$b_i \sim N(0, \psi^2)$$
$$\text{Cov}(\epsilon_{ij}) = \sigma^2 V$$

où V est une matrice qui définit une corrélation de type AR(1) entre les erreurs pour un même individu i.

Solution Nous revoyons d'abord le modèle donné dans la question:

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

i) Pour séparer le modèle en deux composantes, considérons d'abord la partie de régression linéaire:

La partie de régression linéaire contient les éléments déterministes ou systématiques du modèle. Elle représente la tendance générale des données X_{ij}. Dans ce modèle, la partie linéaire est:

$$a_i + b_i X_{ij} + \beta X_{ij}$$

Nous pouvons combiner les termes associés à X_{ij}:

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

Ceci est la partie de régression linéaire.

Ensuite, considérons la partie d'erreur W_{ij}:

$$W_{ij} = \epsilon_{ij}$$

L'erreur W_{ij} reflète la variabilité que la partie linéaire ne peut pas expliquer. Elle est une combinaison des erreurs individuelles et globales, mais dans ce modèle, elle est seulement représentée par ϵ_{ij}.

En conclusion, le modèle peut être séparé comme:

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

où:

- $a_i + (b_i + \beta) X_{ij}$ est la partie de régression linéaire.
- W_{ij} est la partie d'erreur.

Ceci est cohérent avec la façon dont nous séparons les modèles linéaires en classe (une partie déterministe et une partie d'erreur).

ii) Pour la variance de Y_{ij}, nous considérons la somme des variances des trois effets aléatoires (intercept, pente, et l'erreur) :

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

En utilisant les informations fournies :

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

Cette variance sera l'élément diagonal de la matrice de covariance. La covariance entre deux observations de Y à deux moments différents j et k est donnée par :

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(a_i + b_i X_{ij} + \epsilon_{ij}, a_i + b_i X_{ik} + \epsilon_{ik})$$

En tenant compte des effets aléatoires a_i et b_i qui sont indépendants, et de la structure de covariance entre les erreurs ϵ_{ij} qui est définie par la matrice V :

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

où $\rho^{|j-k|}$ est défini par la structure AR(1) de la matrice V.

Nous pouvons maintenant construire la matrice de covariance 5×5 $\text{Cov}(Y_i)$ pour chaque individu i en utilisant les formules de variance et de covariance ci-dessus.

Exercice 2a

Dans cette section, nous analyserons les données à l'aide du logiciel R pour répondre à la question a.

Importation de données

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

```
library(rsq)

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

Modèle de régression de Poisson

Ensuite,nous avons utilisé un modèle de régression de Poisson pour analyser les données et prédire le nombre de satellites chez les araignées de mer mâles.

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

Le test de surdispersion

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.

Cela signifie que les variations du nombre de satellites mâles ne sont pas bien prises en compte dans le modèle de Poisson standard, et que des modèles de surdiscrétisation plus complexes peuvent être envisagés pour mieux s'adapter à ces données.

Exercice 2b

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. Lorsque la variance observée est supérieure à la moyenne prévue, nous parlons de **surdispersion**.

Signification de la surdispersion

Si le test de surdispersion rejette l'hypothèse nulle, cela signifie que la dispersion observée dans les données dépasse ce qui serait attendu sous une distribution de Poisson. Cela peut être dû à :

- Une hétérogénéité non observée.
- Une accumulation de comptages issus de plusieurs processus indépendants.

L'utilisation d'une régression de Poisson en présence de surdispersion peut entraîner une sous-estimation des erreurs standard des coefficients, conduisant ainsi à des p-valeurs sous-estimées et augmentant le risque d'erreurs de type I.

Méthode alternative

Face à la surdispersion, on se tourne généralement vers d'autres modèles pour les données de comptage. La **régression binomiale négative** est un choix courant car elle introduit un paramètre supplémentaire pour capturer la surdispersion. Plus précisément, la distribution binomiale négative peut être considérée comme la somme de plusieurs distributions géométriques indépendantes, où chaque distribution géométrique représente le nombre d'essais nécessaires avant la prochaine occurrence d'un événement.

Exercice 2c

Il y a $2 \times 4 - 1 = 15$ modèles possibles, il faut choisir une critière pour sélectionner le meilleur modèle Disons, 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))
```

On trouve que AIC = 7, le modèle color-weight est le meilleur modèle

Exercice 2d

Analyse des résidus pour le modèle color-weight choisi précédemment dans la question c).

L'analyse choisit d'utiliser 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))
```

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 3a

Approfondissons davantage la structure de corrélation du poids dans les données longitudinales.

Les données longitudinales font référence aux données collectées plusieurs fois sur le même échantillon ou entité. Par exemple, les données collectées sur une personne à différents moments peuvent être considérées comme des données longitudinales. Compte tenu de cette caractéristique des données longitudinales, nous pouvons nous attendre à la structure de corrélation suivante dans les données :

Autocorrélation : Il pourrait y avoir une autocorrélation entre les mesures de poids à des moments consécutifs. En raison des caractéristiques physiologiques du corps humain, le poids ne change pas facilement en peu de temps. Par exemple, si le poids d'une personne est de 70 kg lors de la première mesure, lors de la deux-

ième mesure, son poids pourrait toujours être proche de 70 kg et il est peu probable qu'il augmente ou diminue soudainement de 10 kg.

Continuité des habitudes de vie : Les habitudes de vie, comme l'alimentation, l'exercice et les routines quotidiennes, ont un impact direct sur le poids. À court terme, ces habitudes sont susceptibles de rester relativement stables. Par conséquent, si les habitudes alimentaires et d'exercice à un moment donné entraînent une prise de poids, nous pourrions également observer une prise de poids dans les mesures ultérieures.

Facteurs saisonniers : À différents moments de l'année, le poids peut être affecté par les jours fériés, les changements saisonniers et les changements d'habitudes alimentaires associés. Par exemple, en hiver, le poids pourrait augmenter en raison de moins d'activités de plein air et de repas festifs plus copieux.

Influence d'autres variables : Outre l'effet temporel, d'autres variables telles que l'âge, le sexe, la situation socio-économique (SES) et les habitudes de tabagisme peuvent également être liées aux variations de poids. Par exemple, le poids des adolescents pourrait augmenter avec l'âge. Le tabagisme peut affecter l'appétit et le métabolisme, influençant ainsi le poids.

En résumé, nous pouvons nous attendre à ce qu'il y ait une forte corrélation positive entre les mesures de poids à des moments consécutifs pour un individu particulier. De plus, cette corrélation pourrait diminuer progressivement à mesure que l'intervalle de temps entre les deux points augmente. Par exemple, par rapport à la mesure du poids de la veille, la mesure du poids d'un mois plus tôt pourrait avoir une corrélation plus faible avec la mesure du poids actuelle.

Exercice 3b

Objectifs de la question:

1. Utiliser trois modèles différents pour vérifier l'effet des variables explicatives sur le poids (`weight`).
2. Pour ces modèles, nous devons comparer leur corrélation résiduelle.

Trois modèles:

1. Un modèle avec seulement une constante.
2. Un modèle avec `age` et `sex` comme variables explicatives.
3. Un modèle avec `age`, `sex`, `SES` et `smoking` comme variables explicatives.

Analyse du code:

1. **Modèle 1** (seulement une constante):

```
 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. **Modèle 2** (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. **Modèle 3** (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))
}
```

4. **Comparaison des corrélations résiduelles des modèles** :

```
   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 (sans variables explicatives, seulement l'ordonnée à l'origine) 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 (avec *age* et *sex* comme variables explicatives) et ceux du troisième modèle (avec *age*, *sex*, *SES* et *smoking* comme variables explicatives). 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 (avec *age*, *sex*, *SES* et *smoking* comme variables explicatives) et ceux du premier modèle (sans variables explicatives, seulement l'ordonnée à l'origine). Cette corrélation positive signifie que lorsque les résidus du troisième modèle augmentent, ceux du premier modèle ont également tendance à augmenter, et vice versa.

En conclusion, ces valeurs de corrélation résiduelle reflètent les différences de performance prédictive et de structure des résidus entre les trois modèles. En comparant ces valeurs, on peut obtenir des insights sur l'efficacité des modèles et le choix des variables. Par exemple, la forte corrélation négative entre les deuxième et troisième modèles pourrait signifier qu'il y a un chevauchement dans la sélection des variables entre eux, tandis que le premier modèle pourrait avoir une performance prédictive différente en raison de l'absence de variables explicatives.

Exercice 3c

On utilise SES comme une variable explicative pour la variable de réponse weight.

3.1) Modèle linéaire en assumant l'indépendance entre toutes les observations:

```
modele_lineaire <- lm(weight ~ SES, data = donnee)
print(summary(modele_lineaire)) #std error for ses = 0.1328
```

On trouve que std error for SES est 0.1328.

3.2) 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)) #std error for ses = 0.1868304
```

On trouve que std error for SES est 0.1868304.

3.3) GEE avec une matrice de corrélation interchangeable:

```
require(geepack)
GEE <- summary(geese(weight ~ SES, id = ID, data = donnee, corstr =
   'exchangeable'))
print(GEE) #std error for ses est 0.1646921
```

On trouve que std error for SES est 0.1646921.

3.4) 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)) #ne converge pas
print(confint(modele_lineaire))
```

Lors de l'estimation des paramètres du modèle, le processus d'estimation n'a pas atteint une solution stable, de sorte que le modèle n'a pas convergé.

Conclusion:

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 modèle 3.1 présente l'erreur standard la plus faible (0.1328). On conclut que le modèle linéaire pour 3.1) est le meilleur modèle pour les données.