Devoir 2

Partie théorique

T-1

#1

On veut montrer que $\bar{e} = \sum_{i=1}^{n} \frac{e_i}{n} = 0$. On peut donc développer la formule de $\bar{e}$:

$$\begin{aligned}
\bar{e} &= \sum_{i=1}^{n} \frac{e_i}{n} = \sum_{i=1}^{n} \frac{Y_i - \hat{Y}_i}{n} \\
&= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'} \hat{\beta}}{n} \\
\sum_{i=1}^{n} x_i^{'} \hat{\beta} &= \hat{\beta}_0 \sum_{i=1}^{n} 1 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + ... + \hat{\beta}_{p'} \sum_{i=1}^{n} x_{i,p'} \\
\frac{\sum_{i=1}^{n} x_i^{'} \hat{\beta}}{n} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'}
\end{aligned}$$

Tel que donné dans l'énoncé, on prend pour acquis que:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta_1} \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'}$$

Ainsi, on obtient:

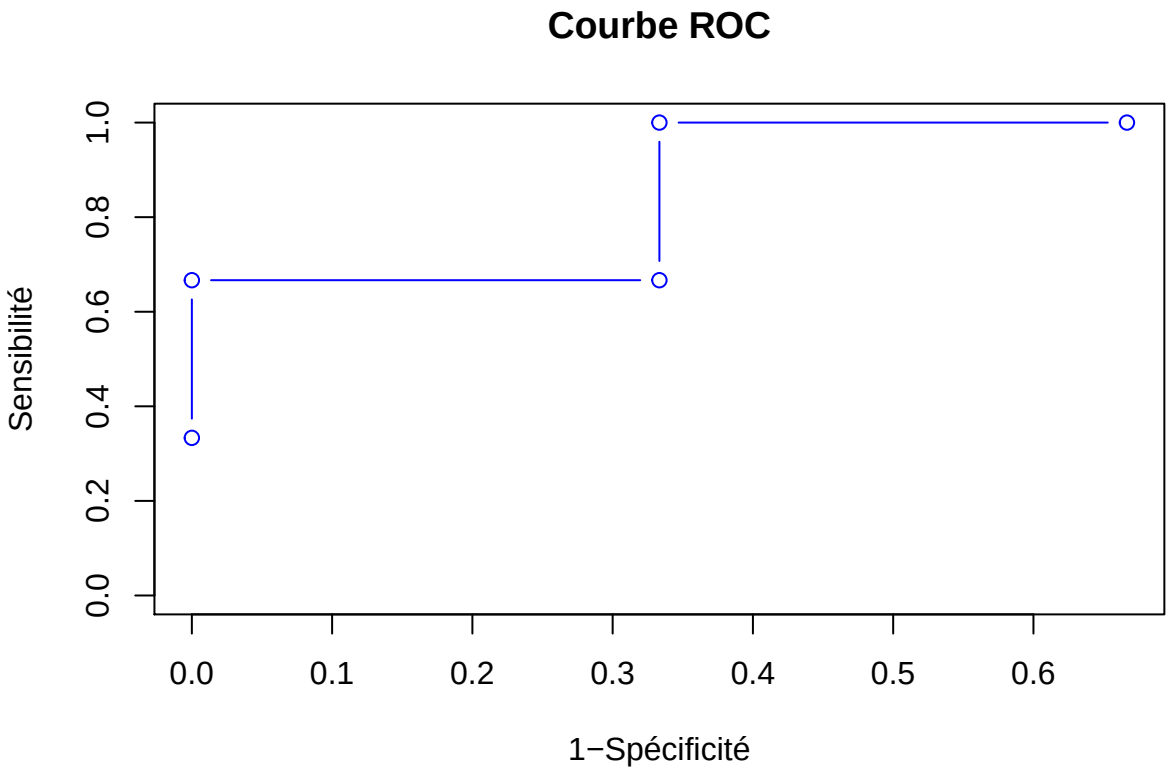
$$\begin{aligned}
\bar{e} &= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'} \hat{\beta}}{n} \\
&= \bar{Y} - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... \hat{\beta}_{p'} \bar{x}_{p'} \right) \\
&= \bar{Y} - \left(\left[\bar{Y} - \hat{\beta_1} \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'} \right] + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'} \right) \\
&= \bar{Y} - \bar{Y} \\
\bar{e} &= 0
\end{aligned}$$

#4

Pour faire la courbe ROC, il faut calculer la valeur de $p_i = P(Y_i = 1, x_i) = \frac{1}{1+exp(-(-4.2+1.2x_i))}$. Par la suite on calcule une prédiction de $\hat{Y}_i$ selon un certain seuil u_k : si $p_i > u_k$, alors $\hat{Y}_i = 1$, sinon $\hat{Y}_i = 0$. On calule alors la sensibilié et la spécificité pour différents seuils. Voici un tableau des résulats ($\hat{Y}_i$ est représenté comme étant Y^k selon la valeur de u_k), ainsi que le graphique de la courbe ROC:

Observation	x_i	p_i	Y_i	Y^1	Y^2	Y^3	Y^4	Y^5
u_k	NA	NA	NA	0.1	0.2	0.5	0.8	0.9
obs 1	1	0.0474258731775668	0	0.0	0.0	0.0	0.0	0.0
obs 2	2	0.141851064900488	0	1.0	0.0	0.0	0.0	0.0
obs 3	3	0.354343693774204	1	1.0	1.0	0.0	0.0	0.0
obs 4	4	0.645656306225795	0	1.0	1.0	1.0	0.0	0.0
obs 5	5	0.858148935099512	1	1.0	1.0	1.0	1.0	0.0
obs 6	6	0.952574126822433	1	1.0	1.0	1.0	1.0	1.0

Métriques	Y^1	Y^2	Y^3	Y^4	Y^5
VP	3.0000000	3.0000000	2.0000000	2.0000000	1.0000000
FN	0.0000000	0.0000000	1.0000000	1.0000000	2.0000000
VN	1.0000000	2.0000000	2.0000000	3.0000000	3.0000000
FP	2.0000000	1.0000000	1.0000000	0.0000000	0.0000000
Sensibilité	1.0000000	1.0000000	0.6666667	0.6666667	0.3333333
Spécificité	0.3333333	0.6666667	0.6666667	1.0000000	1.0000000



T-2

#4

Le modèle est défini de la façon suivante:

$$Y_{ij} = \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1})x_{ij} + \epsilon_{ij}$$

On peut définir la matrice L pour nos tests d'hypothèse qui prendront tous la même forme:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{11} \\ \gamma_{21} \end{bmatrix}$$

$$H_0 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{d}$$

$$H_1 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \neq \boldsymbol{d}$$

a)

$$\beta_0 + \gamma_{10} = 0 \text{ et } \beta_0 + \gamma_{20} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

b)

$$\beta_1 + \gamma_{11} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

c)

$$\beta_1 + \gamma_{11} = 0 \text{ et } \beta_1 + \gamma_{21} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

b)

$$\beta_1 + \gamma_{11} = \beta_1 + \gamma_{21} \leftrightarrow \gamma_{21} = \gamma_{11} \leftrightarrow \gamma_{11} - \gamma_{21} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

#5

a)

On peut définir le modèle de cette façon: $Y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 x_{ij} + \epsilon_{ij}$. Voici la notation matricielle:

$$\mathbf{Y}' = \begin{bmatrix} Y_{11} & Y_{12} & Y_{21} & Y_{22} & Y_{31} & Y_{32} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{31} \\ 1 & x_{32} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\epsilon' = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{21} & \epsilon_{22} & \epsilon_{31} & \epsilon_{32} \end{bmatrix}$$

On peut également remplacer les valeurs symboliques de $\mathbf{Y}$ et $\mathbf{X}$ par leurs valeurs numériques:

$$\mathbf{Y}' = \begin{bmatrix} 70 & 80 & 50 & 60 & 100 & 70 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

b)

On trouve les matrices de variance:

$$\mathbf{D} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \sigma_0^2 \end{bmatrix}$$

$$\mathbf{V} = Var(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = Var(\mathbf{Y}) = \mathbf{ZDZ}' + \mathbf{V}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2+\sigma_0^2 & \sigma_0^2 & 0 & 0 & 0 & 0 \\ \sigma_0^2 & \sigma^2+\sigma_0^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2+\sigma_0^2 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & \sigma^2+\sigma_0^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2+\sigma_0^2 & \sigma_0^2 \\ 0 & 0 & 0 & 0 & \sigma_0^2 & \sigma^2+\sigma_0^2 \end{bmatrix}$$

c)

Ces deux valeurs peuvent s'estimer avec les formules suivantes:

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

$$\hat{\boldsymbol{\gamma}} = \mathbf{D}\mathbf{Z}^{'}\mathbf{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{X\beta})$$

On peut utiliser R pour calculer les valeurs numériques de ces estimés:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 80 \\ -16.67 \end{bmatrix}$$

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} 2.67 \\ -13.33 \\ 10.67 \end{bmatrix}$$

d)

On peut calculer l'estimé à partir de cette équation:

$$\boldsymbol{\hat{Y}_i} = \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1}\boldsymbol{X_i^{'}}\hat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i * n_i} - \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1})\boldsymbol{Y_i}$$

Où:

$$\boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1} = \begin{bmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{bmatrix}$$

De cette façon, on obtient la valeur moyenne de $\boldsymbol{\hat{Y}_i} = 56.67$ pour notre estimé.

#6

a)

Dans cette situation, le paramètre d'intérêt est β_3 puisque celui-ci affectera la valeur de Y_{ij} au fil du temps lorsque $xi = 1$

b)

Il serait acceptable de choisir les struture AR(1) et UN(1) étant donné que cette paire a la plus petite valeur d'AIC peut importe la méthode utilisé (REML ou ML).

c)

On procède à un test d'hypothèse formel en testant un modèle complet et un modèle réduit:

$$H_0 : Y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 x_i + \beta_2 t_j + \beta_3 x_i t_j + \epsilon_{ij}$$

$$H_1 : Y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 x_i + (\beta_2 + \gamma_{2i}) t_j + \beta_3 x_i t_j + \epsilon_{ij}$$

On pose : $\epsilon = 2(\ell_1 - \ell_0) = 2(-88 + 89.5) = 3$. (On pourrait également utiliser les mesures REML pour des résultats semblables)

Nous rejeterons l'hypothèse si la p-value du test est trop élevée:

$$p = 0.5P[\chi^2_{m_1 - m_0 - 1} > \epsilon] + 0.5P[\chi^2_{m_1 - m_0} > \epsilon]$$

$m_0 = 2$(2 variances) et $m_1 = 3$(3 variances). Par conséquent:

$$p = 0.5P[\chi^2_0 > \epsilon] + 0.5P[\chi^2_1 > \epsilon] = 0 + 0.5 * 0.08326 = 0.04163226$$

La p-value n'atteint pas un seuil significatif (inférieur à 0.05), par conséquent on rejète l'hypothèse du modèle réduit. Il est toutefois à noter que la p-value est assez proche de 0.05 avec une valeur de 0.0416

Partie pratique

P1

```
#Modèle complet
modele_complet=lm(SOMA ~ WT2+HT2+WT9+HT9+LG9+ST9, data = data_tp1)

#Modèle final
#model_final=lm(SOMA ~ WT2+WT9+HT9+ST9, data = data_tp1)

Y<-data_tp1$SOMA
si <- studres(modele_complet) # residus studentises
hatYi <- modele_complet$fitted.values # valeurs ajustees
i <- 1:length(Y)

ols_plot_resid_fit(modele_complet)


# Résidus pour chaque observation
plot(i,si,xlab="i",ylab="si",main="Résidus de chaque observation")
abline(h=0,lty=2)

# QQ-plot
ols_plot_resid_qq(modele_complet)

# Tests de normalité
ols_test_normality(modele_complet)

# Transformation de Box-Cox
boxcox(modele_complet)
```

```r
################################################################################
#Influence
################################################################################
# Valeurs des h_ii
ols_leverage(modele_complet)


# DFBETAS
ols_plot_dfbetas(modele_complet)


# DFFITS
ols_plot_dffits(modele_complet)


# Distances de Cook
ols_plot_cooksd_chart(modele_complet)


# Residus vs h_ii
ols_plot_resid_lev(modele_complet)


# covratios
covratio(modele_complet)


# tableau résumé
influence.measures(modele_complet)
```