TISD M1 Pro - Examen du 6 janvier 2009

Durée : 2 heures. Seul un formulaire sur feuille double est autorisé. Calculatrice autorisée.
Le barème est sur 30 points (renormalisation sur 20 en fonction de vos résultats) et est donné ***à titre indicatif***.

Exercice 1 (Maximum de vraisemblance et régression logistique, 19 points)
On considère un échantillon de couples *i.i.d.*$(Y_i, X_i)_{i\in[\![1,n]\!]}$, où $Y_i \in \{0,1\}$ et $X_i \in \mathbb{R}$ est une variable explicative. On suppose que :

$$Y_i = \mathbb{1}_{Y_i^*>0} \quad \text{avec} \quad Y_i^* = \alpha + \beta X_i + \varepsilon_i, \tag{1}$$

où les $(\varepsilon_i)_{i\in[\![1,n]\!]}$ sont des résidus *i.i.d.*, indépendants des X_i, centrés et à densité. On notera f leur densité (supposée continue) et F leur fonction de répartition. On supposera de plus que leur distribution est symétrique, ce qui se traduit par : f paire ou de façon équivalente $F(u) = 1 - F(-u)$.

1 (1.5 pts). Quelle est la loi de Y_i conditionnellement à $X_i = x_i$? Montrer que $\mathbb{P}(Y_i = 1 \,|\, X_i = x_i) = F(\alpha + \beta\, x_i)$. En déduire l'espérance et la variance conditionnelles de Y_i.

2 (0.5). On pose $p_i = F(\alpha + \beta x_i)$. Ecrire la vraisemblance des $Y_1, \cdots Y_n$ conditionnellement aux $X_1, \cdots X_n$.

3 (2). Montrer que :

$$\frac{\partial \ln\, p_i}{\partial \alpha} = \frac{f(\alpha + \beta x_i)}{F(\alpha + \beta x_i)}$$

faire de même pour la dérivée partielle par rapport à β et pour les dérivées partielles de $\ln(1-p_i)$.

4 (1). En déduire que l'estimateur du maximum de vraisemblance $(\widehat{\alpha}, \widehat{\beta})$ satisfait les équations suivantes :

$$\begin{aligned}
&\sum_{i=1}^{n} Y_i \frac{f(\widehat{\alpha} + \widehat{\beta} X_i)}{F(\widehat{\alpha} + \widehat{\beta} X_i)} - (1 - Y_i) \frac{f(\widehat{\alpha} + \widehat{\beta} X_i)}{1 - F(\widehat{\alpha} + \widehat{\beta} X_i)} = 0 \\
&\sum_{i=1}^{n} X_i Y_i \frac{f(\widehat{\alpha} + \widehat{\beta} X_i)}{F(\widehat{\alpha} + \widehat{\beta} X_i)} - X_i (1 - Y_i) \frac{f(\widehat{\alpha} + \widehat{\beta} X_i)}{1 - F(\widehat{\alpha} + \widehat{\beta} X_i)} = 0.
\end{aligned} \tag{2}$$

Dans la suite du problème, on s'intéresse à la régression logistique, obtenue pour des résidus ε_i tels que

$$F(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

5 (1). Montrer que cette distribution est symétrique.

6 (1.5). Montrer que F est strictement croissante et en déduire que les paramètres (α, β) sont identifiables, c'est à dire que si l'on connaît $\mathbb{P}(Y_i = 1 \,|\, X_i = x)$ pour tout x, alors α et β sont déterminés de façon unique.

7 (2). Réécrire dans le cas de la régression logistique les dérivées partielles de la question **3** en fonction de F la fonction de répartition des ε_i. Montrer que la log-vraisemblance dans le cas de la régression logistique est une fonction strictement concave de α et β. Pourquoi ce résultat est-il intéressant ?

8 (1). Citer un algorithme qu'on pourrait utiliser pour maximiser numériquement la log-vraisemblance. Pourquoi cet algorithme fonctionnerait-il bien ici ?

9 (1). Montrer que dans le cas de la régression logistique, le système (2) se ré-écrit :

$$\begin{aligned} &\sum_{i=1}^{n} \big(Y_i - F(\widehat{\alpha} + \widehat{\beta} X_i)\big) = 0 \\ &\sum_{i=1}^{n} \big(Y_i - F(\widehat{\alpha} + \widehat{\beta} X_i)\big) X_i = 0. \end{aligned} \tag{4}$$

10 (1). Donner l'information de Fisher du modèle sous forme d'une espérance et en faisant intervenir la densité f des résidus.

Jusqu'à la fin du problème, on se place dans le cas d'un modèle logistique où X est également une variable dichotomique, c'est-à-dire qui prend les valeurs 0 ou 1. On se donne le tableau de contingence suivant :

		Y	
		0	1
X	0	$n_{00} = 26$	$n_{01} = 26$
	1	$n_{10} = 16$	$n_{11} = 32$

TAB. 1 – Tableau de contingence de X et Y

11 (1). Réécrire les équations (4) avec uniquement n_{00}, n_{01}, n_{10}, n_{11} (sans les X_i ou Y_i).

12 (2). Résoudre les équations obtenues à la question **11** pour obtenir l'estimateur du maximum de vraisemblance (Indication : on pourra poser $a = F(\widehat{\alpha})$ et $b = F(\widehat{\alpha} + \widehat{\beta})$).

13 (3.5). On admet que les hypothèses pour appliquer les résultats de normalité asymptotique des estimateurs du maximum de vraisemblance sont satisfaits. Donner un intervalle de confiance à 95% pour β.

Exercice 2 (Pression et température du mercure, 11 points)
Nous analysons des données pour obtenir des relations entre la température en degrés Celsius et la pression en millimètre de mercure du mercure gazeux. La base de données, sous **R**, s'appelle

`pressure`. Elle contient deux variables, T (`temperature`) et P (`pressure`).

1 (0.5). On cherche graphiquement une relation entre la pression et la température. On trace successivement les graphiques de la Figure 1. Commenter.

(a) (b) (c)

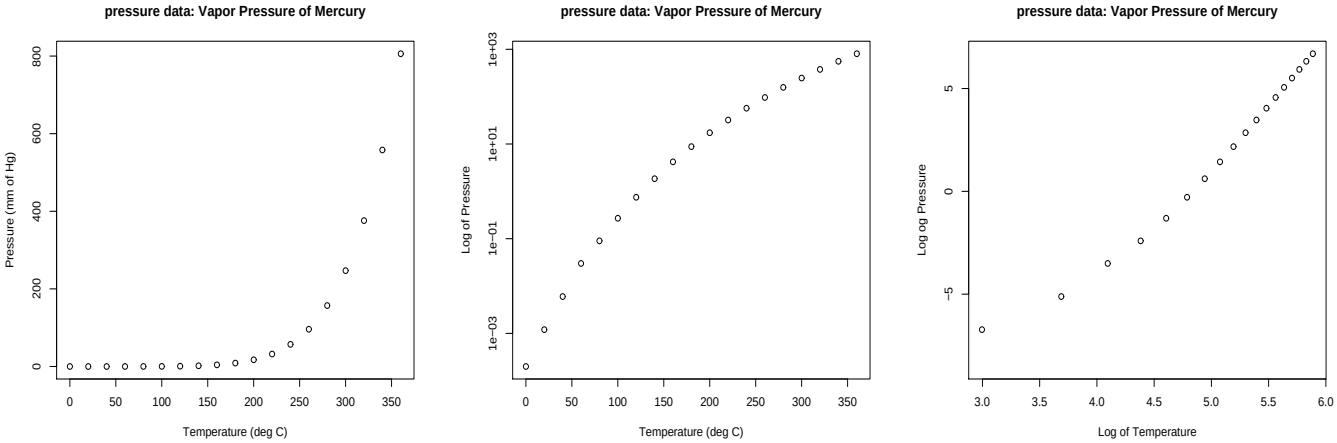

FIG. 1 – *(a) : Pression en fonction de la température. (b) : Log de la pression en fonction de la température. (c) : Log de la pression en fonction du log de la température.*

2 (0.5). Afin de quantifier la relation entre pression et température, on réalise les deux régressions linéaires `mco` et `mco2` (voir sorties ci-dessous) :

```
> mco<-lm(log(pressure$pressure) ~ pressure$temperature)
> summary(mco)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -6.068144   0.483831  -12.54 5.10e-10 ***
pressure$temperature  0.039792   0.002296   17.33 3.07e-12 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.096 on 17 degrees of freedom
Multiple R-Squared: 0.9464,     Adjusted R-squared: 0.9433
F-statistic: 300.3 on 1 and 17 DF,  p-value: 3.070e-12

> mco2<-lm(log(pressure$pressure[2:19]) ~ log(pressure$temperature[2:19]))
> summary.lm(mco2)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -23.5908     1.0742  -21.96 2.25e-13 ***
log(pressure$temperature[2:19])   5.0260     0.2115   23.76 6.62e-14 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6994 on 16 degrees of freedom
Multiple R-Squared: 0.9724,     Adjusted R-squared: 0.9707
F-statistic: 564.5 on 1 and 16 DF,  p-value: 6.624e-14
```

Dans la seconde régression, on n'utilise les données que de 2 à 19, car la donnée numéro 1 correspond à la température 0 (et son log n'est pas défini).

Comment peut-on comparer ces deux régressions ? Laquelle vous semble la meilleure ?

3 (2 pts). Nous obtenons les résultats suivants sous **R** :

```
mean(log(pressure$temperature[2:19]))=5.02
mean(log(pressure$temperature[2:19])^2)=25.78

mean(log(pressure$pressure[2:19]))=1.63
mean(log(pressure$pressure[2:19])^2)=18.43

mean(log(pressure$pressure[2:19])*log(pressure$temperature[2:19]))=11.22
```

Calculer les variances des variables `log(temperature)` et `log(pressure)`. Calculer la covariance de ces deux variables.

4 (2.5). On s'intéresse à la régression de $\log(P)$ sur $\log(T)$:

$$\log(P) = a\log(T) + b + \varepsilon \tag{5}$$

Donner les expressions littérales des estimateurs MCO $\widehat{a}$ et $\widehat{b}$, puis retrouver à l'aide de la question **3** les résultats de la question **2**. D'où viennent les différences à votre avis ?

5 (1). Les coefficients pour a et b sont-ils significatifs au seuil 5% ? Justifier.

6 (0.5). Montrer que (5) est équivalent au fait que la pression varie comme une puissance de la température, si l'on omet le bruit ε.

7 (0.5). On étudie les résidus de la régression (5).

```
res2=residuals(mco2)   mean(res2)=8.83e-17   var(res2)=0.46
```

(On rappelle que la commande `var` de **R** donne la variance corrigée (*i.e.* renormalisée par $n-1$)) Commentez ces résultats ainsi que la Figure 2(a). Tracer ce QQ-plot a-t-il un intérêt pour l'exploitation des résultats de la régression ? si oui, lequel ?

8 (3). Sur la Figure 2(b), on peut voir que la courbe présente une légère inflexion au niveau de la 5^{e}valeur. On réalise un test de Chow pour savoir s'il y a deux comportements différents.

(a) (b)

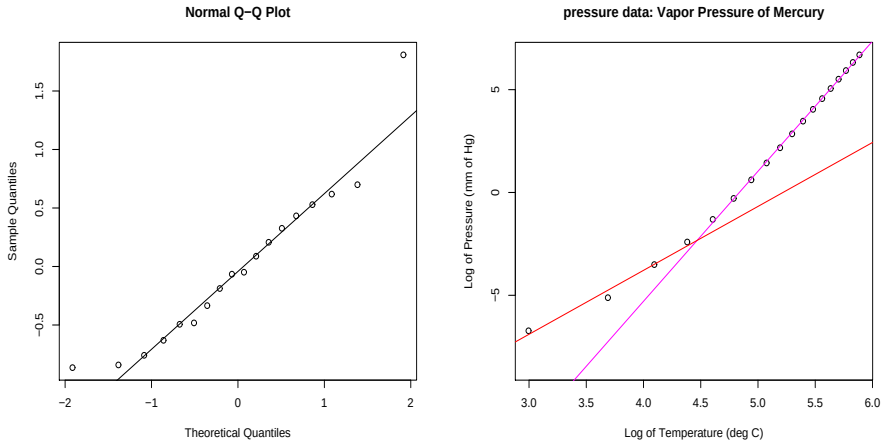

FIG. 2 – *(a) : QQ-plot des résidus de la régression (5) avec ceux de la loi normale. (b) : Droites de régression de la question 10.*

```
> mcop3<-lm(log(pressure$pressure[6:19]) ~ log(pressure$temperature[6:19]))
> summary.lm(mcop3)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -30.5862     0.2508  -122.0   <2e-16 ***
log(pressure$temperature[6:19])   6.3237     0.0466   135.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06748 on 12 degrees of freedom
Multiple R-Squared: 0.9993,     Adjusted R-squared: 0.9993
F-statistic: 1.842e+04 on 1 and 12 DF,  p-value: < 2.2e-16

> mcop4<-lm(log(pressure$pressure[2:5]) ~ log(pressure$temperature[2:5]))
> summary.lm(mcop4)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -16.2226     1.1660  -13.91  0.00513 **
log(pressure$temperature[2:5])   3.1089     0.3048   10.20  0.00947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3173 on 2 degrees of freedom
Multiple R-Squared: 0.9811,     Adjusted R-squared: 0.9717
F-statistic: 104.1 on 1 and 2 DF,  p-value: 0.009474

> resp3=residuals(mcop3)
```

```
> resp4=residuals(mcop4)
> var(resp3)=0.004
> var(resp4)=0.067
> qf(0.95,2,14)=3.739
> qf(0.025,12,2)=0.196
> qf(0.975,12,2)=39.415
```

Effectuer le test de Chow. Vérifier les hypothèses du test de Chow.

9 (0.5). Conclure sur le comportement de la pression en fonction de la température.