TISD M1 Pro - Examen du 15 janvier 2008

Durée : 2 heures
Documents autorisés : sujets de TD, notes de cours manuscriptes
Documents NON autorisés : corrigés de DM, corrigé du partiel
La calculatrice est autorisée

Le nombre de décimales pour les résultats numériques est laissé à votre appréciation.

Dans tous les tests demandés, le seuil sera 5%. *Pour les tests lus dans les sorties SAS, on donnera la valeur numérique de la statistique de test et la p-valeur associée. Pour les tests réalisés "à la main", on donnera la statistique de test et le quantile auquel on doit la comparer.*

On rappelle que `qf(alpha,n1,n2)` *est la commande* **R** *qui donne le quantile d'ordre* α *d'une loi de Fisher* $\mathcal{F}(n_1, n_2)$.

`qf(0.95,7,518)=2.03, qf(0.025,281,237)=qf(0.025,237,281)=0.78,`
`qf(0.975,281,237)=qf(0.975,237,281)=1.28`
`qnorm(0.975,0,1)=-qnorm(0.025,0,1)=1.96`

Exercice 1 (Déterminants du salaire et discrimination salariale)
Nous analysons les données d'une enquête auprès des ménages effectuée par le U.S. Census Bureau en mai 1985. Le fichier étudié, `cps85.sas7bdat`, contient 532 individus, pour lesquels sont renseignées les variables suivantes : le **gain** qui est le log du salaire horaire en dollars (`LNWAGE`), l'âge (`AGE`), le nombre d'années d'étude (`ED`), le nombre d'années d'expérience (`EX=AGE-(ED+6)`, qui est le nombre d'années à partir de la fin des études en admettant que celles-ci commencent à 6 ans), le carré du nombre d'années d'expérience (`EXSQ=EX*EX`), le produit du nombre d'années d'expérience et du nombre d'années d'études (`EDEX=EX*ED`), les indicatrices `FE`, `NONWH`, `HISP` et `UNION` qui valent 1 si l'individu est respectivement une femme, ni Blanc ni Hispanique, Hispanique, syndiqué.

Nous allons nous intéresser à l'impact de l'expérience et de l'éducation sur le salaire dans une première partie, puis nous étudierons les discrimations salariales hommes/femmes. Nous considérons le modèle linéaire suivant :

$$\texttt{LNWAGE} = \alpha + \alpha_F\,\texttt{FE} + \alpha_U\,\texttt{UNION} + \alpha_N\,\texttt{NONWH} + \alpha_H\,\texttt{HISP} + \beta_1\,\texttt{ED} + \beta_2\,\texttt{EX} + \beta_3\,\texttt{EXSQ} + \varepsilon \quad (1)$$

où ε est un bruit. Les paramètres de ce modèle ont été estimés par MCO sous **SAS**, avec le code suivant. Les sorties **SAS** sont à l'Annexe 2.

```
proc reg data=malib.cps85;
model lnwage=fe union nonwh hisp ed ex exsq / clb;
output out=malib.cps85 r=epsilon p=predit;
run;quit;
```

1. Donner le coefficient de détermination R^2 de la régression.

2.1. A partir du code **SAS** donné ci-dessus, pouvez-vous dire ce que contient la variable `epsilon` une fois la procédure effectuée ?

2.2. Une étude de cette variable est donnée à l'Annexe 3. Peut-on considérer que cette variable est centrée ? Dans la rubrique `Test for normality`, plusieurs tests avec l'hypothèse nulle H_0 *: la loi de epsilon est une loi normale* sont réalisés. Que pouvez-vous dire des p-valeurs de ces tests ? Accepte-t-on la normalité de la variable `epsilon` ? Ceci est-il corroboré par l'histogramme et le QQ-plot de la Figure 1 ? (justifier).

2.3. Pourquoi tester la normalité de la variable `epsilon` est-il important ?

(a) (b)

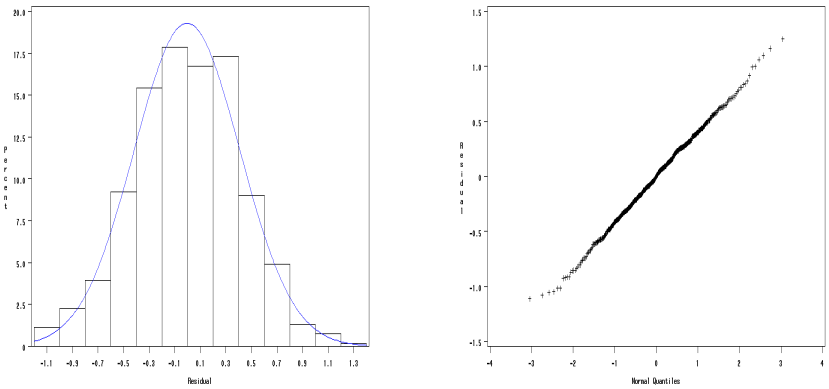

FIG. 1 – *(a) : Histogramme de la variable* epsilon *et superposition de la densité d'une loi normale. (b) : QQ-plot des quantiles de la variable* epsilon *et de ceux de la loi normale.*

3. A partir de l'Annexe 2, donner les estimateurs MCO de β_1, β_2 et β_3 ainsi que leurs intervalles de confiance à 95%. Ces coefficients sont-ils significatifs ? Commenter les signes de ces coefficients. Pour quel niveau d'expérience le salaire est-il maximisé ? Calculer le salaire horaire maximum prédit par le modèle, en dollars, pour un homme Blanc non syndiqué et ayant 14 années d'étude.

4. La théorie économique du Capital Humain nous dit que si les capacités individuelles sont corrélées avec le niveau d'étude d'une part et avec la rémunération d'autre part, alors le salaire devrait croître plus vite avec l'expérience pour les plus diplômés. Une manière d'incorporer un tel effet est de rajouter au modèle (1) la variable `EDEX` dont on appellera β_4 le coefficient. L'effet de l'expérience sur le log du salaire dépend alors de l'expérience *et* du niveau d'étude :

$$\partial \texttt{LNWAGE}/\partial \texttt{EX} = \beta_2 + 2\beta_3 \texttt{EX} + \beta_4 \texttt{ED}.$$

La régression avec la variable supplémentaire `EDEX` a été estimée à l'Annexe 4. Quel test peut-on faire pour tester la validité de cette théorie économique ? Quelle est la conclusion ?

Nous revenons au modèle (1) et nous intéressons à la différence de salaires entre hommes et femmes.

5. Interpréter le coefficient α_F et son exponentielle (dans l'Annexe 2). Tester l'hypothèse selon laquelle la différence de salaire entre un homme et une femme est nulle, toutes les autres variables étant égales.

6. A partir de l'Annexe 6, faire un test de Chow pour tester H_0 *: les paramètres du modèle (1) pour la sous-population des hommes et sur la sous-population des femmes sont les mêmes.* Vérifier la validité du test de Chow. Conclure.

Exercice 2 (Chiffre d'affaire)
Une société vend du matériel multimédia. Le Tableau 1 donne les chiffres d'affaires $(Y_t)_{t\in[\![1,12]\!]}$ du premier trimestre 2001 au dernier trimestre 2003. Cette série est représentée graphiquement à la Figure 2.

Trimestre	Année 1	Année 2	Année 3
1	53	62	70
2	70	78	85
3	41	50	63
4	82	91	102

TAB. 1 – Chiffre d'affaire par trimestre

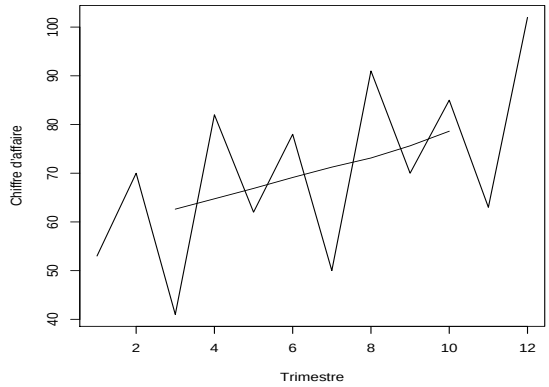

FIG. 2 – Séries (Y_t) et $(MM_{4,t})$ (Moyenne mobile).

1. Au vu de la Figure 2, le modèle additif suivant semble-t-il approprié :

$$Y_t = f_t + s_t + \varepsilon_t$$

où f_t est la tendance, s_t la saisonnalité et ε_t le bruit ? Pourquoi ?

2. On va estimer la tendance par une moyenne mobile centrée d'ordre 4. Rappeler la définition de cette moyenne mobile et calculer les valeurs manquantes dans le Tableau 2 (calcul de la moyenne mobile $MM_{4,t}$ et calcul de $Y_t - MM_{4,t}$).

3. Quelle contrainte doivent satisfaire les coefficients saisonniers ? Calculer les coefficients saisonniers relatifs à chaque trimestre. Commenter.

année	trimestre	série Y_t	série $MM_{4,t}$	série $Y_t - MM_{4,t}$
1	1	53		
1	2	70		
1	3	41	62.625	-21.625
1	4	82	64.750	17.250
2	1	62		
2	2	78		
2	3	50		
2	4	91		
3	1	70	75.625	-5.625
3	2	85	78.625	6.375
3	3	63		
3	4	102		

TAB. 2 – Tableau à compléter

Exercice 3 (Modèle de régression à variables qualitatives)
Soit Y une variable observable prenant les valeurs 1, avec probabilité p et 0, avec probabilité $1-p$. Nous observons un échantillon $(Y_i)_{i\in[\![1,n]\!]}$ de variables aléatoires indépendantes identiquement distribuées de même loi que Y.

1.1. Quelles sont les lois de Y et $\sum_{i=1}^n Y_i$?

1.2. Ecrire la vraisemblance des observations et calculer l'estimateur du maximum de vraisemblance $\widehat{p}$ de p.

1.3. $\widehat{p}$ est-il convergent ?

1.4. Enoncer un théorème de normalité asymptotique pour $\widehat{p}$.

1.5. Construire un intervalle de confiance asymptotique à 5% pour p.

Nous considérons maintenant un couple de variables aléatoire (Y, X) où X est une variable explicative prenant les valeurs 0 ou 1, et où les Y est une variable à expliquer, prenant elle aussi les valeurs 0 et 1, et pour laquelle on suppose qu'il existe une variable latente Y^* telle que
- conditionnellement à X, Y^* suit une loi $\mathcal{N}(aX+b, 1)$ avec a, $b \in \mathbb{R}$,
- $Y = \mathbf{1}_{Y^*>0}$.

2.1. Quelles sont les lois de Y et Y^* conditionnellement à X ? Montrer que $\mathbb{P}(Y=1 \,|\, X=1) = \Phi(a+b)$ où Φ, la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. Calculer les probabilités $\mathbb{P}(Y=1 \,|\, X=0)$, $\mathbb{P}(Y=0 \,|\, X=1)$ et $\mathbb{P}(Y=0 \,|\, X=0)$.

Nous disposons d'un échantillon de 100 observations, $(y_i, x_i)_{i\in[\![1,100]\!]}$ réalisations des 100 couples $(Y_i, X_i)_{i\in[\![1,100]\!]}$ indépendants deux à deux et de même loi que (Y, X) étudié à la question 2.1. Ces observations sont résumées dans le Tableau 3.
2.2. Ecrire la vraisemblance des observations $(Y_1, \cdots Y_{100})$ conditionnellement à $(X_1, \cdots X_{100})$

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

TAB. 3 – Tableau de contingence de X et Y

en utilisant les nouvelles variables $\alpha = \Phi(a + b)$ et $\beta = \Phi(b)$ ainsi que les effectifs n_{00}, n_{01}, n_{10} et n_{11} définis dans le Tableau 3.

2.3. Donner la log-vraisemblance. Puis, écrire les conditions du premier ordre du programme de maximisation de la log-vraisemblance par rapport à α et β.

2.4. En déduire les estimateurs du maximum de vraisemblance $\widehat{\alpha}$ et $\widehat{\beta}$ pour α et β. Faire l'application numérique avec les valeurs données dans le Tableau 3.

2.5. Donner la valeur de $\Phi(0)$. En déduire $\Phi^{-1}(1/2)$. Puis, en utilisant que $\Phi^{-1}(2/3) = 0.43$, déduire les valeurs numériques des EMV $\widehat{a}$ et $\widehat{b}$ pour les observations décrites au Tableau 3.

2.6. Les estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$ sont-ils convergents lorsque $n_{00} + n_{01} \to +\infty$ et $n_{10} + n_{11} \to +\infty$? Qu'en déduire pour $\widehat{a}$ et $\widehat{b}$?

2.7. Qu'en est-il de la normalité asymptotique ?