CORRIGE PARTIEL B

M1 ECO

Mars 2012

1 EXERCICE-1

1. Regression simple
   a. $Y = a_1X_1 + a_0 + \varepsilon$; par la méthode MCO, on estime les paramètres avec les formules suivantes :
   $\widehat{a}_1 = \dfrac{\sum_{i=1}^{i=n} x_i y_i}{\sum_{i=1}^{i=n} x_i^2} = \dfrac{Cov(X;Y)}{V(X)}$ avec $\boxed{\widehat{a}_0 = \overline{Y} - \widehat{a}_1 * \overline{X}}$; on obtient avec la calculatrice :

a1	a0	R^2
-2,9566	25,7231	0.0871

   soit $\widehat{Y} = -2.9566X_1 + 25.7231$

   b. $\widehat{a}_1$ donne une estimation de la variation de Y suite à une augmentation d'une unité de X,ce qui indique une estimation d'une baisse de 2.96 forages suite à une augmentation d'un dollar du prix du baril. $\widehat{a}_0$ donne une estimation du nombre de forages pour un prix du baril nul, ce qui n'a pas de réalité ici.

   c. On a :

	Y	X1
Variance	4,73072653	0,04715153

   ; on a $SCR = (1 - R^2)SCT = (1 - R^2) * nV(Y) = (1-0.0871)*14*4.7307 = 60.4612$. On calcule l'erreur type de a_1 ; on a : $S^2_{\widehat{a}_1} = \dfrac{S^2_\varepsilon}{\sum_{i=1}^{i=n} x_i^2} = \frac{SCR/(n-2)}{nV(X)} = \frac{60.4612/12}{14*0.0472} =$ 7.6247 ce qui donne : $S_{\widehat{a}_1} = \sqrt{7.6247} = 2.7613$ et $t_{\widehat{a}_1} = \dfrac{\widehat{a}_1}{S_{\widehat{a}_1}} = \frac{-2.9566}{2.7613} = -1.0707$

   Nous allons donc tester si X contribue à expliquer Y en testant l'hypothèse nulle $a_1 = 0$ contre l'hypothèse alternative $a_1 \neq 0$:
   $\begin{cases} H_0 : a_1 = 0 \ (X \text{ n'influence pas } Y) \\ H_1 \quad\quad\quad : a_1 \neq 0 \end{cases}$. Sous l'hypothèse H_0, $t_{\widehat{a}_1} = \dfrac{\widehat{a}_1}{S_{\widehat{a}_1}}$, le ratio de Student, suit une distribution de Student avec $(n-2)$ degrés de liberté, soit ici un ddl de 12 ; ici $t_{\widehat{a}_1} = -1.0707$; le seuil de signification est $\alpha = 0.05$; il reste à comparer ce quotient avec la valeur lue dans la table de Student, de $t_{\alpha/2;n-2}$ soit ici : $t_{0.025\,;\,12} = 2.1788$; $|t_{\widehat{a}_1}| < t_{\alpha/2;n-2}$, on ne peut rejeter H_0, au seuil de $100\alpha\%$; on ne peut affirmer que a_1 est significativemen différent de zéro. On en conclut que X n'est pas significative et ne contribue pas à l'explication de Y.

   d.

ddl	Somme	
1	$SCE = SCT - SCR$	5.7686
$n-2=12$	SCR	60.4612
$n-1=13$	$SCT = nV(Y)$	$14 * 4.7307 = 66.2298$

   e. $S^2_\varepsilon = \dfrac{SCR}{(n-2)} = \frac{60.4612}{12} = 5.0384$; le terme d'erreur suit une loi normale de moyenne 0 et de variance inconnue constante σ^2_ε.Cette hypothèse de variance constante est l'hypothèse d'homoscédasticité ; on parle alors de série homoscédastique (par opposition à hétéroscédastique). On estime alors σ_ε par S_ε et $\dfrac{\epsilon}{S_\varepsilon}$ suit la loi de Student de ddl $n-2$ soit 13.

   f. $R^2 = \frac{SCE}{SCT} = \frac{5.7686}{66.2298} = 8.71 \times 10^{-2}$; il est très faible et représente le pourcentage de la variance de Y expliquée par le modèle, ici environ 8.71%.

2 EXERCICE-2

1. $M = 0$, donne $\widehat{Y} = 8.40 = \widehat{a_0}$; donne une estimation du salaire moyen des femmes et $M = 1$ donne $\widehat{Y} = 3.40 + 8.40 = 11.8$, estimation du salaire moyen des hommes ; $\widehat{a_1} = 3.40$ représente la différence entre le salaire moyen entre des hommes et celui des femmes.
2. On teste $\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$. Sous l'hypothèse H_0, $t_{\widehat{a}_1} = \dfrac{\widehat{a}_1}{S_{\widehat{a}_1}}$, le ratio de Student, suit une distribution de Student avec $(n-2)$ degrés de liberté, soit ici $70-2=68$; ici $t_{\widehat{a}_1} = \dfrac{\widehat{a}_1}{S_{\widehat{a}_1}} = \dfrac{3.40}{1.35} \simeq 2.5185$; il reste à comparer ce quotient avec la valeur lue dans la table de Student, soit ici $t_{0.005;68} \simeq 2.6479$. La valeur du quotient $\dfrac{\widehat{a}_1}{S_{\widehat{a}_1}}$ est inférieur au t de la table, on en déduit donc que l'on ne peut rejeter l'hypothèse H_0 et donc $\widehat{a}_1$ n'est pas significativement différent de zéro au niveau de risque de 1%. La différence de salaire n'est pas significativement différente de 0 au niveau de 1%.

3 EXERCICE-3

1. Question 1
   a. L'espérance de ε_i, $E(\varepsilon_i)$ est nulle pour tout i ($E(\varepsilon_i / X_i) = 0$)

      La variance $V(\varepsilon_i) = E\left((\varepsilon_i - E(\varepsilon_i))^2\right)$ est constante pour tout i, soit $V(\varepsilon_i) = \sigma_\varepsilon^2$. Cette hypothèse de variance constante est l'hypothèse d'homoscédasticité ; on parle alors de série homoscédastique (par opposition à hétéroscédastique).

      Abscence d'autocorrélation des erreurs : $Cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$. Le terme d'erreur n'est pas autocorrélé : la valeur du terme d'erreur ϵ_i n'est pas corrélé à celle de ϵ_j.

      Chaque ε_i suit une loi normale, les ε_i résultant de l'influence combinée d'un grand nombre de variables indépendantes non intégrées dans le modèle de régréssion.

      En conclusion : les erreurs suivent une loi normale : $\varepsilon_i \hookrightarrow \mathcal{N}(0; \sigma_\varepsilon)$ et sont indépendantes. Les erreurs sont **normalement et indépendamment distribuées** : $\epsilon_i \hookrightarrow \mathcal{N}\mathbf{id}(0; \sigma_\varepsilon)$.

   b. En cellule $C12$, $SCE = SCT - SCR = 66.2302 - 19.1465 = 47.0837$; $B7$, $S\epsilon = \sqrt{\frac{SCR}{n-k-1}} = \sqrt{\frac{19.1465}{10}} = 1.3837$; les degrés de liberté : $B12 : k = 3$, en $B13 : n-k-1 = 14-3-1 = 10$ et en $B14$, $n-1 = 13$; en $D12$ et $D13$, les moyennes, soit SCE/k et $SCR/(n-k-1)$, soit $D12 : 47.0837/3 = 15.6946$ et en $D13$: $\frac{19.1465}{10} = 1.9146$ et enfin en $E12$, $F = \frac{SCE/k}{SCR/(n-k-1)} = \frac{15.6946}{1.9146} = 8.1973$ En $D18$, $t_{\widehat{a}_1} = \dfrac{\widehat{a}_1}{S_{\widehat{a}_1}} = \dfrac{7.2808}{2.8547} = 2.5505$ Enfin en $F18$ et $G18$ les bornes de l'intervalle de confiance à 95%, soit $[\widehat{a}_1 - t_{0.025;10} S_{\widehat{a}_1}; \widehat{a}_1 + t_{0.025;10} S_{\widehat{a}_1}]$, soit : $[7.2808 - 2.2281 * 2.8547; 7.2808 + 2.2281 * 2.8547]$: soit $[0.9202 ; 13.6414]$

   c. $\widehat{a}_0 = -80.9175, \widehat{a}_1 = 7.2808, \widehat{a}_2 = 0.1174$ et $\widehat{a}_3 = -1.8149$.

   d. Théorème de Gauss-Markov : si les hypothèses de la MCO sont vérifiées, les estimateurs $\widehat{a}_1$ et $\widehat{a}_0$ **sont $\boldsymbol{BLUE}$** (Best Linear Unbiased Estimator), ce sont des estimateurs linéaires, sans biais, efficaces (variance minimale).

   e. Les estimations de $\widehat{a}_1, \widehat{a}_2$ et $\widehat{a}_3$ donnent une estimation des variations de Y, respectivement quand X_1 augmente d'une unité, les autres variables restant constantes.

   f. L'intervalle de confiance trouvé pour a_1 ne contient pas 0, donc on rejete l'hypothèse H_0 : $a_1 = 0$; donc on accepte l'hypothèse H_1que X_1 contribue à l'explication de Y. On a en $E18$, une $p - value$ de 2.88% ; c'est le plus petit niveau de risque à partir duquel on rejette H_0. Cette $p - value$ est inférieure à 5% ce qui est cohérent avec le rejet de H_0.

2. Le coefficient de détermination est $R^2 = 0.7109$ et le coefficient de détermination corrigé est $\overline{R^2} = 0.6242$;

L'introduction de nouvelles variables explicatives accroît la part expliquée ; pour une même variabilité totale, R^2 augmente quand on introduit des variables, mais cette augmentation tient au nombre de variables et non à leur pouvoir explicatif. Pour pallier à cet inconvénient, on introduit le coefficient de détermination corrigé qui tient compte de la baisse du ddl, consécutive à l'introduction de nouvelles variables explicatives indépendantes. On montre facilement que $\overline{R}^2 < R^2$.

3. $F = \dfrac{SCE/k}{SCR/(n-k-1)} = 8.1973$; sous l'hypothèse nulle $H_0 : a_1 = a_2 = a_3 = 0$, la variable aléatoire F suit la loi de Fisher avec pour degrés de liberté k et $n-k-1$, soit 3 et $10. F_{(3;10)} \simeq 3.71$

   Régle de décision : on prend un seuil de signification de 5%

   $F > F_{(k;n-k-1)}$, on rejete H_0, on en conclut qu'il existe au moins un des paramètres non nul, c'est-à dire au moins une variable qui contribue à expliquer Y.

Statistiques de la régression						
Coefficient de détermination multiple	0,8432					
Coefficient de détermination R^2	0,7109					
Coefficient de détermination R^2	0,6242					
Erreur-type	1,3837					
Observations	14					
ANALYSE DE VARIANCE						
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*	
Régression	3	47,0837	15,6945737	8,1971	0,0543	
Résidus	10	19,1465	1,91464505			
Total	13	66,2302				
	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*
Constante	-80,9175	24,6166	-3,2871	0,0082	-135,7667	-26,0683
Variable X 1	7,2808	2,8547	2,5505	0,0288	0,9202	13,6414
Variable X 2	0,1174	0,0301	3,8996	0,0030	0,0503	0,1844
Variable X 3	-1,8149	0,5651	-3,2115	0,0093	-3,0741	-0,5557