Exercices sur le modèle de régression linéaire simple

Exercice 1

Le tableau ci-dessous représente l'évolution du revenu disponible brut et de la consommation des ménages en euros pour un pays donné sur la période 1992-2001. [Pour les calculs, prendre 4 chiffres après la virgule].

Année	Revenu	Consommation
1992	8000	7389.99
1993	9000	8169.65
1994	9500	8831.71
1995	9500	8652.84
1996	9800	8788.08
1997	11000	9616.21
1998	12000	10593.45
1999	13000	11186.11
2000	15000	12758.09
2001	16000	13869.62

On cherche à expliquer la consommation des ménages (C) par le revenu (R), soit :

$$C_t = \alpha + \beta R_t + u_t$$

Travail à faire :

(i) Tracer le nuage de points et commenter.

(ii) Estimer la consommation autonome et la propension marginale à consommer $\hat{\alpha}$ et $\hat{\beta}$.

(iii) En déduire les valeurs estimées $\hat{C}_t$ de C_t.

(iv) Calculer les résidus et vérifier la propriété selon laquelle la moyenne des résidus est nulle.

(v) Calculer l'estimateur de la variance de l'erreur.

(vi) Tester la significativité de la pente.

(vii) Construire l'intervalle de confiance au niveau de confiance de 95% pour le paramètre β.

(viii) Calculer le coefficient de détermination et effectuer le test de Fisher permettant de déterminer si la régression est significative dans son ensemble.

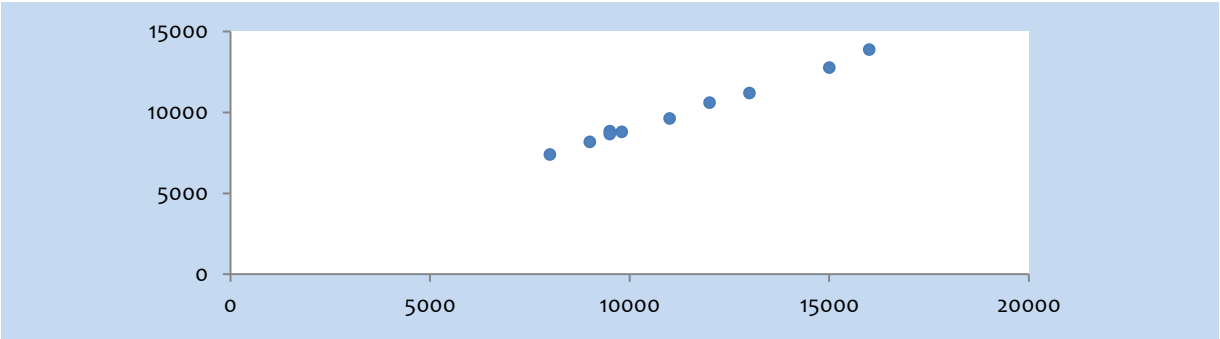
(ix) Ecrire et vérifier l'équation d'analyse de la variance. Interpréter.

(x) Après un travail minutieux, un étudiant de L1 FASE trouve le coefficient de corrélation linéaire entre C_t et R_t suivant r_{XY} = 0.99789619. Sans le moindre calcul, tester la significativité de ce coefficient. Argumenter.

(xi) En 2002 et 2003, on prévoit respectivement 16800 et 17000 euros pour la valeur du revenu. Déterminer les valeurs prévues de la consommation pour ces deux années, ainsi que l'intervalle de prévision au niveau de confiance de 95%.

Solution de l'exercice 1

(i) Le graphique nuage de points est donné ci-dessous :



Ce graphique témoigne de l'existence d'une association linéaire positive, presque parfaite, entre la consommation des ménages (C_t) par le revenu (R_t), ce qui autorise l'estimation de la relation les liant par la méthode des moindres ordinaires.

(ii) *Pour simplifier l'estimation de* la consommation autonome ($\hat{\alpha}$) et de la propension marginale à consommer $\hat{\beta}$, posons ce qui suit :

$Y_t = C_t$; $X_t = R_t$; $\alpha = \beta_0$ *et* $\beta = \beta_1$. Ce qui nous permet d'écrire le modèle donné dans l'exercice comme suit :

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

A partir des calculs effectués dans le tableau ci-dessous, on a :

Estimation de la propension marginale à consommer

$$\hat{\beta}_1 = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{50104729}{64156000} = 0{,}78098$$

Estimation de la consommation autonome

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9985{,}575 - 0{,}78098(11280) = 1176{,}0896$$

Le modèle estimé est par conséquent : $\hat{Y}_t = 1176{,}0896 + 0{,}78098X_t$.

(iii) *Voir tableau ci-dessous. Ces valeurs sont trouvées en remplaçant dans l'équation de la droite des moindres :* $\hat{Y}_t = 1176{,}0896 + 0{,}78098X_t$, pour chaque date, X_t par sa valeur.

(iv) *Voir tableau ci-dessous. Les résidus sont calculés d'après la formule* $e_t = Y_t - \hat{Y}_t$

(v) L'estimateur de la variance de l'erreur est donnée par $\hat{\sigma}^2_{\hat{u}_t} = \frac{\sum e_t^2}{n-2}$, *connaissant n =10 et* $\sum e_t^2$ *(voir tableau), on obtient :*

$$\hat{\sigma}^2_{\hat{u}_t} = \frac{\sum e_t^2}{n-2} = \frac{165169{,}3826}{10-2} = 20646{,}1728$$

(vi) La pente ici est la propension marginale à consommer, soit $\hat{\beta}_1$. *Le test de significativité de ce coefficient requiert son écart-type* $\hat{\sigma}_{\hat{\beta}_i}$. *Connaissant la variance de l'erreur, la variance de* $\hat{\beta}_1$ *est calculée comme suit :*

$$Var\,(\hat{\beta}_1) = \frac{\hat{\sigma}^2_{\hat{u}_t}}{\sum x_t^2} = \frac{20646{,}17282}{64156000} = 0{,}0003 \quad \rightarrow \quad \hat{\sigma}_{\hat{\beta}_i} = \sqrt{Var\,(\hat{\beta}_1)} = 0{,}0179$$

Par conséquent son ratio de Student est :

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0{,}78098}{0{,}0179} = 43{,}5352 \sim t_{0.025\,;\,8} = 2.306.$$

Puisque $t_{\hat{\beta}_1} > t_{table} \rightarrow$ *la pente* $\hat{\beta}_1$ *est statistiquement significative.*

(vii) *L'intervalle de confiance au niveau de confiance de 95% (au seuil de 5%) pour le paramètre θ1 est construire comme suit :*

$$I = [\hat{\beta}_1 - (t_{0.025\,;\,8})\hat{\sigma}_{\hat{\beta}_1} \;;\quad \hat{\beta}_1 + (t_{0.025\,;\,8})\hat{\sigma}_{\hat{\beta}_1}]$$

En faisant les remplacements nécessaires, on trouve : $I = [0{,}7396 \;;\quad 0{,}8224]$

(viii) *Le coefficient de détermination R^2 peut être calculé par la formule (les valeurs viennent du tableau ci-dessous) :*

$$R^2 = \frac{SCE}{SCT} = \frac{\sum \hat{y}_t^2}{\sum y_t^2} = \frac{39130928,8011}{39296098,1837} = 0,9958$$

Connaissant le t de Student de la pente, la statistique de Fisher peut se calculer comme suit :

$F = t_{\hat{\beta}_1}^2 = (43,5352)^2 = 1895,3136 \sim F_{[1;8]} = 5,32$. *Puisque* $F > F_{table} \rightarrow RH_0$, *la régression est significative dans son ensemble.*

(ix) *L'équation d'analyse de la variance est :*

$$SCT = SCE + SCR \rightarrow \sum y_t^2 = \sum \hat{y}_t^2 + \sum e_t^2 \rightarrow 39296098,1837 = 39130928,8011 + 165169,3826$$

(x) *Nous savons que dans un modèle linéaire simple, accepter la significativité de la pente revient à accepter celle du coefficient de corrélation linéaire. La pente* $\hat{\beta}_1$ *étant significative, le* r_{XY} *l'est aussi naturellement.*

(xi) *La prévision ponctuelle ne pose aucun problème. La prévision par intervalle requiert l'estimation de l'écart-type de l'erreur de prévision. Elle est donnée par :*

$$Y_{n+h} \in I = \hat{Y}_{n+h} \pm (t_{0.025;8})\hat{\sigma}_{e_{n+h}}$$

où

$$\hat{\sigma}_{e_{n+h}} = \left\{\hat{\sigma}_{\hat{u}_t}^2 \left[1 + \frac{1}{n} + \frac{(X_{n+h} - \bar{x})^2}{\sum x_t^2}\right]\right\}^{\frac{1}{2}}$$

En effectuant les remplacements nécessaires, au niveau de confiance de 95% (au seuil de 5%), on a les résultats suivants :

	Prévision ponctuelle	**Prévision par intervalle**
2002	$\hat{Y}_t = 1176,0896 + 0,78098(16800) = 14296,5998$	$Y_{n+h} \in [13949,0697 ; \ 14644,1299]$
2003	$\hat{Y}_t = 1176,0896 + 0,78098(17000) = 14452,7963$	$Y_{n+h} \in [14105,2657 ; \ 14800,3269]$

Le tableau récapitulant tous les calculs est repris ci-dessous.

Année	Y_t	X_t	y_t	x_t	$x_t y_t$	x_t^2	$\hat{Y}_t$	e_t	e_t^2	y_t^2	$\hat{y}_t$	$\hat{y}_t^2$
1992	7389,99	8000	-2595,585	-3280	8513518,8	10758400	7423,9516	-33,9615958	1153,389989	6737061,4922	-2561,6234	6561914,4650
1993	8169,65	9000	-1815,925	-2280	4140309	5198400	8204,93434	-35,28434098	1244,984718	3297583,6056	-1780,6407	3170681,1566
1994	8831,71	9500	-1153,865	-1780	2053879,7	3168400	8595,42571	236,2842864	55830,26401	1331404,4382	-1390,1493	1932515,0386
1995	8652,84	9500	-1332,735	-1780	2372268,3	3168400	8595,42571	57,41428643	3296,400286	1776182,5802	-1390,1493	1932515,0386
1996	8788,08	9800	-1197,495	-1480	1772292,6	2190400	8829,72054	-41,64053713	1733,934332	1433994,2750	-1155,8545	1335999,5393
1997	9616,21	11000	-369,365	-280	103422,2	78400	9766,89983	-150,6898313	22707,42527	136430,5032	-218,6752	47818,8294
1998	10593,5	12000	607,875	720	437670	518400	10547,8826	45,56742347	2076,390081	369512,0156	562,3076	316189,8106
1999	11186,1	13000	1200,535	1720	2064920,2	2958400	11328,8653	-142,7553217	20379,08188	1441284,2862	1343,2903	1804428,8884
2000	12758,1	15000	2772,515	3720	10313755,8	13838400	12890,8308	-132,7408121	17620,12319	7686839,4252	2905,2558	8440511,3336
2001	13869,6	16000	3884,045	4720	18332692,4	22278400	13671,8136	197,8064427	39127,38879	15085805,5620	3686,2386	13588354,7011
Σ			0	0	50104729	64156000		0	165169,3825	39296098,1837	0	39130928,8011

n=10 ; $\bar{y}$ =9985,575
$\bar{x}$ =11280

Exercice 2

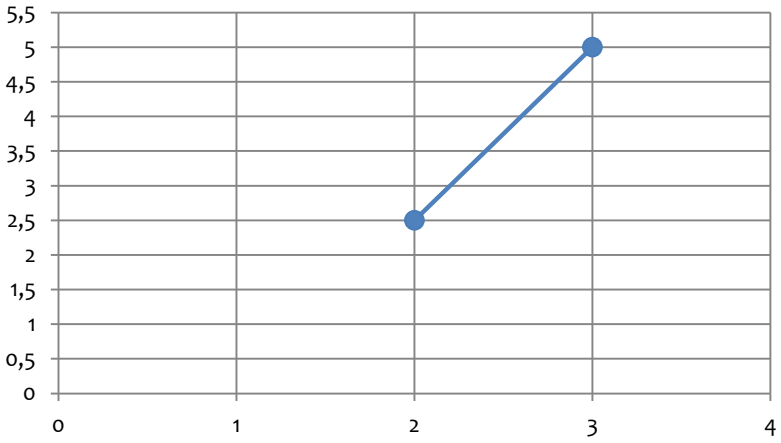
Soit le modèle linéaire $Y_t = \beta_0 + \beta_1 X_t + u_t$. Où Y_t représente la quantité offerte de pommes et X_t le prix.
On donne les informations suivantes : $\bar{Y} = 5$ et $\bar{X}=3$.

Après estimation, on a la droite de régression suivante : $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$. Connaissant le couple (Y=2.5 ; X=2) par lequel passe cette droite de régression, trouver $\hat{\beta}_0$ et $\hat{\beta}_1$.

Solution de l'exercice 2

Connaissant le couple (Y=2.5 ; X=2) et le centre de gravité du nuage de points ($\bar{Y} = 5$; $\bar{X}=3$), on peut reproduire la droite des moindres carrés de cette estimation comme ci-après :

Y	5	2.5
X	3	2



En mesurant la pente de cette droite, on trouve la pente $\hat{\beta}_1 = \frac{5-2,5}{3-2} = 2,5$.

Connaissant la pente $\hat{\beta}_1$ *et les deux moyennes* $\bar{Y}$ *et* $\bar{X}$, $\hat{\beta}_0$ *est calculé comme suit :*

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 5 - (2,5)3 = -2,5$

Et le modèle estimé (équation de la droite des MCO) est : $\hat{Y}_t = -2,5 + 2,5X_t$

Exercice 3

Soit un modèle linéaire simple : $Y_t = \beta_o + \beta_o X_t + u_t$

On donne les informations suivantes :

$\sum YX=184500$ $\sum Y^2=26350$ $\sum X^2=1400000$ $\bar{Y}=60$ $\bar{X}=400$ $n=7$

Travail demandé :

- Estimer les coefficients du modèle
- Evaluer la qualité de cet ajustement
- Tester la significativité globale du modèle

Solution de l'exercice 3

En fonction des données en présence, les formules suivantes seront utilisées pour répondre aux trois questions posées :

- $\hat{\beta}_1 = \frac{\sum X_t Y_t - n\bar{x}\bar{y}}{\sum X_t^2 - n\bar{x}^2}$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- $R^2 = \frac{\hat{\beta}_1^2[\sum X_t^2 - n\bar{x}^2]}{[\sum Y_t^2 - n\bar{y}^2]}$
- $F = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{(n-2)}}$

Après calcul, sachant que $\sum YX = \sum XY$, *on a les résultats suivants :*

$\hat{\beta}_1 = 0,0589$; $\hat{\beta}_0 = 36,44$; $R^2 = 0,8455$; $F = 27,3618$

Le R^2 *étant relativement élevé, environ 85%, l'ajustement effectué est de bonne qualité. Et puisque* $F > F_{[1;5]} = 6,61$, *on en conclut que le modèle est globalement bon.*

Exercice 4

Soit le modèle : $Y_t = \beta_0 + \beta_1 X_t + u_t$

Y_t : salaire moyen horaire par jour [en USD]
X_t : nombre d'années d'études

On donne par ailleurs les informations suivantes : $r_{XY} = 0.951916$; $\sigma_x = 3.894440$ et $\sigma_y = 2.945636$

Après estimation, sur base d'un échantillon de 13 observations, un étudiant de L1 FBA présente les résultats incomplets ci-après :

$$\hat{Y}_t = 0.030769 + \ldots\ldots\ldots\ldots\ldots X_t$$

Travail demandé :

(i) Compléter les pointillés.
(ii) Tester la significativité du r_{XY}.
(iii) Interpréter ces résultats. Semblent-ils logiques ?
(iv) Calculer le R^2.
(v) Tester la significativité de la pente et la significativité d'ensemble du modèle.

Solution de l'exercice 4

(i) *Connaissant r_{XY}, σ_x et σ_y, la pente $\hat{\beta}_1$ est estimée par la formule $\hat{\beta}_1 = \mathbf{r_{XY}} \frac{\sigma_Y}{\sigma_X}$, ce qui donne, en remplaçant :*

$\hat{\beta}_1 = 0,7200$. On a ainsi :

$$\hat{Y}_t = 0,030769 + 0,7200\, X_t$$

(ii) *Le t calculé pour $r_{XY} = 0,951916$ donne $t_{cal} = 10,3054$ et le $t_{0.025\,;\,11} = 2,201$. Puisque tcal > t_{table}, on conclut que le r_{XY} est statistiquement non nul.*

(iii) *Il y a lien fort et positif entre le salaire moyen horaire par jour et le nombre d'années d'études. En effet, ces résultats semblent logiques car il est tout à fait normal que ceux qui beaucoup étudié gagnent un peu plus que ceux qui ont étudié un peu moins.*

(iv) *On sait que, pour un modèle de régression linéaire simple avec terme constant, le R^2 n'est rien d'autre que le carré du coefficient de corrélation de Bravais – Pearson. Ainsi :*

$$R^2 = (0,951916)^2 = 0,9061$$

(v) *Connaissant le R^2, on a : $F = 106,2009 \sim F_{[1\,;\,11]} = 4,84$. On sait de plus que dans un modèle linéaire simple, le F n'est rien d'autre que le carré du t de Student associé à la pente. Le t de Student de la pente est donc obtenu en prenant la racine carré de F, soit :*

$$\left|t_{\hat{\beta}_1}\right| = 10,3054 > t_{0.025\,;\,11} = 2,201$$

En conclusion, la pente est statistiquement significative et le modèle est valable dans l'ensemble.

Exercice 5

Le tableau suivant donne l'âge et la tension artérielle Y de 12 femmes :

Individu	1	2	3	4	5	6	7	8	9	10	11	12
Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
Tension artérielle (Y)	136	132	136	130	138	132	136	130	142	134	136	140

Travail demandé :

(i) Déterminer l'équation de la droite de régression de Y sur X.
(ii) Tester la significativité de la pente. Quelle conclusion peut-on tirer ?
(iii) Estimer la tension artérielle d'une femme âgée de 50 ans.

Solution de l'exercice 5

L'équation de la droite de régression de Y sur X est :

$$\hat{Y}_t = 129{,}5193 + 0{,}1079X_t$$
$$(5{,}0449) \quad (0{,}0942)$$
(.) : écart-type

La statistique t de Student de la pente est $t_{cal} = 0{,}1079/0{,}0942 = 1{,}1455$. *Le Student théorique, au seuil de 5% et à 10 degrés de liberté est* $t_{table} = 2{,}228$. *D'où la pente est statistiquement nulle, ce qui signifie que l'âge n'explique en rien la tension artérielle.*

La tension artérielle d'une femme âgée de 50 ans est : $\hat{Y}_t = 129{,}5193 + 0{,}1079(50) = 134{,}9149$

Exercice 6

Les données statistiques ci-dessous portent sur les poids respectifs des pères et de leur fils aîné.

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

Travail demandé :

(i) Calculer la droite des moindres carrés du poids des fils en fonction du poids des pères.
(ii) Calculer la droite des moindres carrés du poids des pères en fonction du poids des fils.
(iii) Que vaut le produit des pentes des deux régressions ?
(iv) Juger de la qualité des ajustements faits en (i) et (ii).

Solution de l'exercice 6

Soient Y=Fils et X=Père.

La droite des moindres carrés du poids des fils en fonction en fonction des pères, après estimation est :

$$\hat{Y}_t = 35{,}8248031 + 0{,}47637795X_t$$

Et la droite des moindres carrés du poids des pères en fonction en fonction des fils, après estimation est :

$$\hat{Y}_t = -3{,}37687366 + 1{,}03640257X_t$$

Le produit de deux pentes donne le R^2 qui, comme le coefficient de corrélation linéaire, est un indicateur symétrique. On a ainsi :

$R^2 = 0{,}47637795 * 1{,}03640257 = 0{,}49371933$

Au regard de la valeur du R^2 faible, environ 49%, les ajustements effectués en (i) et (ii) ne sont de bonne qualité.

Exercice 7

Cocher la bonne la réponse.

1. La droite des MCO d'une régression linéaire simple avec constante passe-t-elle par le point $(\bar{x}, \bar{y})$?
   A. Toujours
   B. Jamais
   C. Parfois
2. Pour une régression linéaire simple, le R^2 est symétrique :
   A. Oui
   B. Non
   C. Parfois
3. Pour une régression linéaire simple, le R^2 correspond au carré du F de Fisher :
   A. Oui
   B. Non

Solution de l'exercice 7

1 A ; 2A ; 3B.

Exercice 8

Soient les données suivantes :

$\sum_1^6 Y_t = 114$ $\sum_1^6 X_t = 36$ $\sum_1^6 X_t^2 = 226$ $\sum_1^6 X_t Y_t = 702$

Estimer la relation $Y_t = \beta_0 + \beta_1 X_t + u_t$

Indication : n = 6.

Exercice 9

Soit le modèle suivant sans terme constant : $Y_t = \beta X_t + u_t$.

Trouver l'estimateur $\hat{\beta}$ des MCO.

Solution de l'exercice 9

En appliquant le critère des MCO, minimisation de la somme des erreurs quadratiques, à cette relation, on obtient :

$$\hat{\beta} = \frac{\sum X_t Y_t}{\sum X_t^2}$$