

Chapitre 2, exercice 1

Instructions pour employer CH02EX01.XLS

Le fichier CH02EX01.XLS comporte un exercice de base destiné à tous les apprenants et un exercice avancé réservé aux seuls apprenants de la version avancée.

Le répertoire CH02EX01 comporte un exercice de base destiné à tous les apprenants

Exercice de base (Pour tous les utilisateurs du cours)

Préalable



Le chapitre 2 du cours de base doit avoir été suivi jusqu'à la page 50, pour les parties 1, 2 et 3, jusqu'à la page 58 pour la partie 4, jusqu'à la page 65 pour la partie 5.

Objectif



Le but de l'exercice est d'introduire — ou de rappeler pour les apprenants qui ont déjà bénéficié d'un cours de statistique de base — les principaux éléments de la régression linéaire simple par la méthode des moindres carrés.

Données



Les données sont relatives à une petite enquête pilote sur le montant des dépenses en vacances en fonction du nombre des personnes du ménage. Ces données en très petit nombre de manière à permettre de vérifier les calculs et d'appliquer aisément plusieurs approches.

Structure de l'exercice

L'exercice comporte cinq parties :

- Dans la partie 1, le but de l'exercice est d'introduire la régression linéaire à partir d'un graphique des données dans Microsoft Excel et d'expérimenter à partir de modifications des données.
- Dans la partie 2, le but de l'exercice est de comparer les résultats de la partie 1 à ceux fournis par l'outil Regression de la boîte à outils d'analyse de Microsoft Excel et d'expérimenter le principe des moindres carrés et de définir les résidus, la variance résiduelle et l'écart-type résiduel.
- Dans la partie 3, le but de l'exercice est de commenter les fonctions de Microsoft Excel qui fournissent les principaux résultats de la régression linéaire simple au sens de la méthode des moindres carrés et de comparer ces résultats à ce qui résulte de l'application de formules.
- Dans la partie 4, le but de l'exercice est d'introduire la mesure de qualité de l'ajustement par le coefficient de corrélation et le coefficient de détermination.
- Dans la partie 5, le but de l'exercice est d'employer le logiciel Time Series Expert for Windows afin de réaliser une régression linéaire simple.

Partie 1 La variable x représente le nombre de personnes d'un ménage et la variable y est le montant des dépenses en vacances. Ces données sont en très petit nombre de manière à permettre d'appliquer aisément des approches voisines et, ultérieurement, de vérifier les calculs de la régression.

1.1 PRESENTATION DES DONNEES

Tableau Les données des dépenses en vacances.

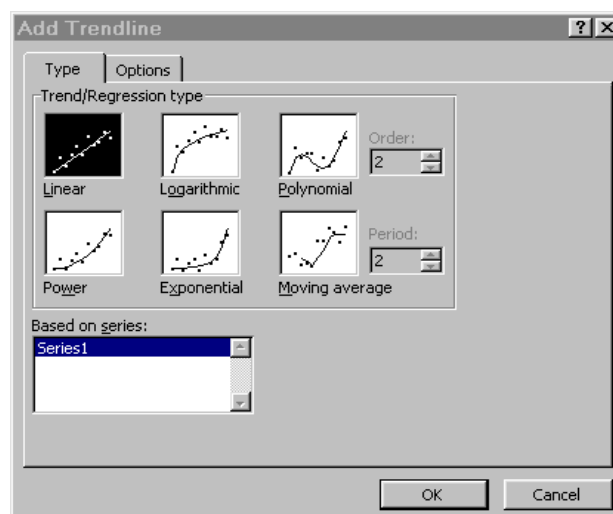
x (taille du ménage)	y (dépenses en vacances)
1	17
2	11
2	23
4	19
6	30

- ⇒ Dans la feuille Main, descendez vers la ligne 15 afin de voir le tableau des données
- ⇒ Cliquez ensuite sur l'onglet DATA afin de visualiser le graphique des données.

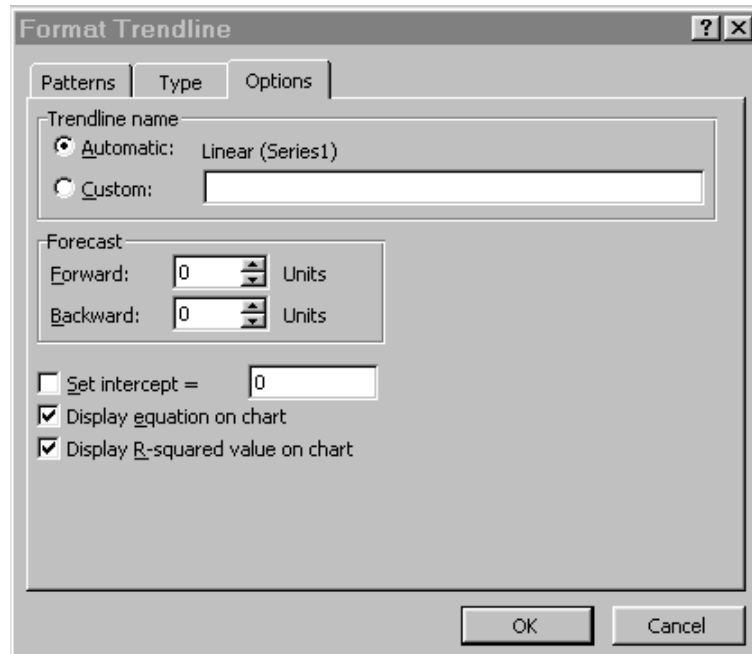
Ces données constituent ce qu'on appelle un nuage de points. Le but est de l'ajuster par une droite.

1.2 AJUSTEMENT PAR UNE DROITE DE REGRESSION

- ⇒ Cliquez sur le menu Chart ⇒ Add Trendline.



⇒ Cliquez sur l'onglet Options de la boîte de dialogue et cochez les options Display equation on chart et Display R-squared value on chart.



⇒ Cliquez sur l'onglet Type puis sur le type Linear.



Notez les informations qui se sont ajoutées au graphique. Il s'agit de l'équation de la droite de régression et de la valeur du coefficient de détermination R^2 (encore à définir).



1.2.1 Votre réponse

Remarques



1. Il est possible de réaliser un ajustement sans constante ou ajustement par l'origine. Il suffit pour cela de cocher la case Set intercept et de laisser la valeur 0 par défaut. Dans le présent cas, ce peut avoir un sens puisqu'on peut imaginer que les dépenses en vacances soient proportionnelles au nombre de personnes du ménage. Ce n'est toutefois que rarement recommandé.

2. La boîte de dialogue offre d'autres choix que l'ajustement linéaire. Nous verrons l'ajustement par une parabole et par une exponentielle dans le

chapitre 3. L'emploi des moyennes mobiles est décrit dans le chapitre 4 mais n'a de sens que pour des données chronologiques, donc sûrement pas ici.

1.3 MODIFICATION DES DONNEES

⇒ Cliquez de nouveau sur l'onglet Main. Changez une des données. Par exemple, changez la valeur 30 de y de la cinquième observation par 40. Cliquez sur l'onglet DATA afin de retrouver le graphique.



Notez les nouvelles informations.



1.3.1 Votre réponse

Nous avons vu qu'une modification d'une donnée dans le tableau se répercute sur le graphique. Inversement, on peut modifier un point du graphique et modifier la donnée correspondante dans le tableau.

- ⇒ Cliquez sur l'onglet DATA. Cliquez sur un des carrés représentant un des points. Cliquez de nouveau sur ce point. Le pointeur de souris se transforme en une quadruple flèche.
- ⇒ Déplacez le pointeur de souris, soit verticalement pour changer l'ordonnée y du point, soit horizontalement pour changer l'abscisse x du point, soit obliquement pour changer à la fois l'ordonnée y et l'abscisse x du point.



Remarquez-vous les changements dans l'équation de la droite et dans le coefficient de détermination (encore à définir) ?



1.3.2 Votre réponse

⇒ Cliquez sur l'onglet Main. Visualisez les modifications dans le

tableau.

Il est important pour la suite de revenir aux données de départ.

⇒ .Revenez aux données de départ en utilisant la macro-instruction CH02EX01Initial (ou la combinaison de touches CTRL SHIFT I si le classeur est le seul qui soit ouvert).

SYNTHESE

Nous avons vu — ou rappelé — comment obtenir un ajustement d'un nuage de points à deux dimensions par une droite à partir d'une représentation graphique de type diagramme x-y dans Excel. La méthode vue ici est limitée à une seule variable explicative. Dans la partie 2, nous voyons une méthode qui permettra, au chapitre 7, de réaliser de la régression linéaire multiple, et qui offre en outre des résultats supplémentaires.

Partie 2 Dans cette partie, le but de l'exercice est de comparer les résultats de la partie 1 à ceux fournis par l'outil Regression de la boîte à outils d'analyse de Microsoft Excel. On profite des possibilités pour expérimenter avec le principe des moindres carrés et définir les résidus, la variance résiduelle et l'écart-type résiduel, qui n'apparaissent pas parmi les sorties données dans les graphiques décrits dans la partie 1.

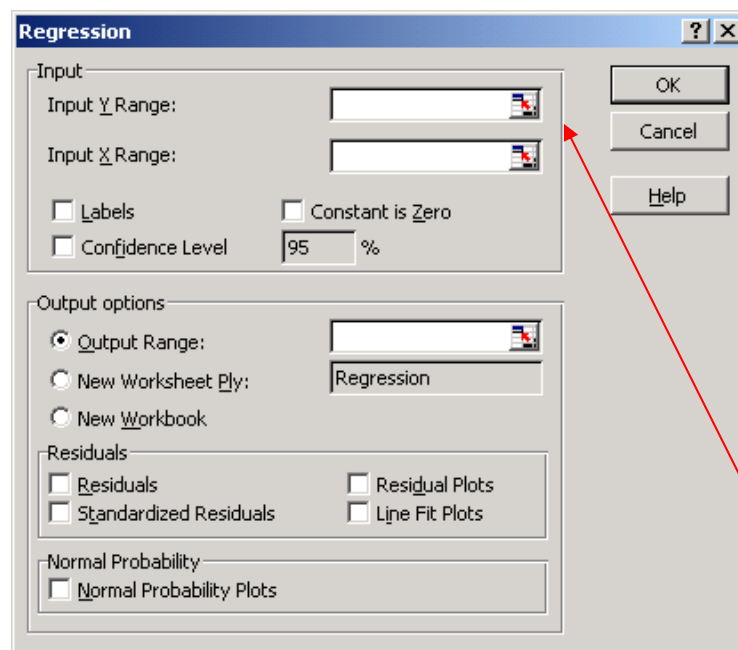
2.1 UTILISATION DE L'OUTIL DE REGRESSION

⇒ Cliquez sur le menu Tools ⇒ Data Analysis ⇒ Regression (descendez l'ascenseur à cet effet). Cliquez sur OK.



Remarque

Il est fréquent que les outils d'analyse ne soient pas installés parce qu'ils ne sont pas automatiquement au moyen du programme d'installation, à moins qu'une installation complète ait été demandée. S'il est nécessaire, cliquez sur le menu Tools ⇒ Add-ins et cochez la case Analysis Toolpak et Solver add-in qui servira un peu plus tard, par la même occasion. Recommencez l'étape précédente.

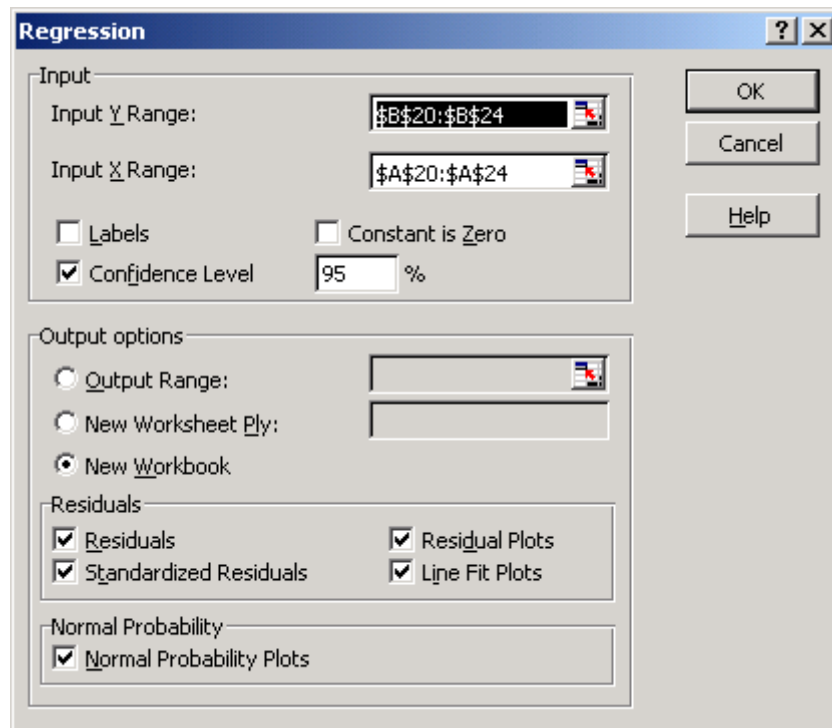


⇒ Cliquez sur le bouton d'accès à la feuille correspondant à Input Y Range et sélectionnez la plage de B20 à B24 qui contient les montants des dépenses en vacances.

⇒ Faites de même pour Input X Range et sélectionnez la plage de A20 à A24 qui contient les nombres de personnes du ménage.

Il faut évidemment veiller à sélectionner le même nombre de données pour la variable dépendante et pour la variable explicative. Comme les en-têtes ne sont pas sur la ligne avant les données, ne cochez pas la case Labels.

⇒ Pressez sur le bouton New Workbook (Output Options). Cochez les 5 cases relatives aux résidus et aux graphiques, pour les besoins futurs. Cliquez sur OK.



2.2 INTERPRETATION DES RESULTATS

⇒ Un classeur est créé. Sauvez-le, dans le dossier où vous mettez les travaux du chapitre 2, sous le nom CH02EX01Regression.XLS, par exemple. Notez les informations qui apparaissent dans le nouveau classeur ou mieux imprimez son contenu.

La sortie comporte six parties:

1. *Regression Statistics*.
2. ANOVA. Nous commenterons davantage ce tableau dans le chapitre 7.
3. *Coefficients*. Limitons-nous aux coefficients estimés — pour le restant, voir le chapitre 7 — la constante b_0 sur la ligne Intercept et le coefficient de régression b_1 sur la ligne X Variable.
4. *Residual Output*. Nous les emploierons ci-dessous.
5. *Probability Output*. Ne sort qu'en cochant Normal Probability Plot.
6. Les graphiques.

?

Comparez la valeur de R square à celle de R^2 dans la partie 1.

?

Comparez les coefficients estimés aux nombres obtenus dans l'équation à la partie 1.



2.2.1 Vos réponses

Remarques



1. Il est possible de réaliser un ajustement sans constante ou ajustement par l'origine. Il suffit pour cela de cocher la case Constant is zero. Dans le présent cas, ce peut avoir un sens puisqu'on peut imaginer que les dépenses en vacances soient proportionnelles au nombre de personnes du ménage. Ce n'est toutefois que rarement recommandé.

2. Excel n'est évidemment pas un logiciel conçu pour la régression. Une limitation que nous rencontrerons en régression multiple est la suivante : la plage des observations relatives aux variables explicatives doit être de forme rectangulaire.

2.3 ANALYSE DES VALEURS AJUSTEES ET DES RESIDUS



Revenez au classeur de nom CH02EX01.XLS, dans la feuille Main.



Entrez la valeur de b_0 dans la cellule C15 et la valeur de b_1 dans la cellule F15. Veillez à ne vous tromper et à entrer les nombres avec 4 chiffres après la virgule ou le point décimal.

Les *valeurs ajustées* sont calculées dans la colonne C. Ce sont les valeurs de $y = 12,3125 + 2,5625 x$ pour les différentes valeurs de x qui sont observées, c'est-à-dire les ordonnées des points observés (marqués par un carré dans la figure sous l'onglet DATA).

De même les *résidus* sont calculés dans la colonne D. Ce sont les valeurs de $y - (12,3125 + 2,5625 x)$, c'est-à-dire les différences d'ordonnées des points observés (marqués par un carré dans la figure sous l'onglet DATA) et des points correspondants sur la droite. Pour la commodité, nous

présentons ici le tableau obtenu.

Régression imposée		n =	5	
b0 = 12.3125		b1 =	2.5625	
x_i	Y_i	Y_i^*	e_i	e_i^2
1	17	14.8750	2.1250	4.5156
2	11	17.4375	-6.4375	41.4414
2	23	17.4375	5.5625	30.9414
4	19	22.5625	-3.5625	12.6914
6	30	27.6875	2.3125	5.3477
		Somme	0.0000	94.9375
MSE =		94.9375	0	2
Variance résiduelle		-----	(-----)	18.9875
		5	5	



Revenez au classeur de nom CH02EX01Regression.XLS.



Comparez quelques-uns des nombres de la colonne y_i^* ci-dessus avec les résultats de la colonne Predicted Y du nouveau classeur.



Comparez quelques-uns des nombres de la colonne e_i avec les résultats de la colonne Residuals du nouveau classeur.



Comparez la somme des carrés des résidus, e_i^2 , avec l'élément suivant du tableau ANOVA: ligne Residual, colonne SS. Notez-le ci-dessous.



2.3.1 Vos réponses



Comparez la quantité indiquée MSE ou variance résiduelle avec le carré de la quantité appelée Standard Error (erreur-type) sur la septième ligne. Notez-les.



?

On appelle *écart-type résiduel* la racine carrée de la variance résiduelle. Comparez-la à la quantité appelée Standard Error.

2.3.2 Vos réponses



Revenez au classeur CH02EX01.XLS, dans la feuille Main, et descendez sur la ligne 15.



Entrez d'autres valeurs des paramètres de la droite (cellules C15 et F15).

?

Comparez la somme des carrés des résidus dans la cellule F26 avec le nombre 94,9375 qui était obtenu précédemment. Le but du jeu est de battre le record, c'est-à-dire diminuer ce nombre.



2.3.3 Votre réponse



Recommencez à entrer d'autres valeurs des paramètres de la droite (cellules C15 et F15) jusqu'à ce que vous soyez convaincus qu'il n'est pas possible de battre le record.

?

Que peut-on en déduire comme propriété?



2.3.4 Votre réponse

2.4 EXAMEN DES GRAPHIQUES



Revenez au classeur CH02EX01Regression.XLS que vous avez sauvé. Cliquez une seule fois sur le graphique en haut à gauche (en

dessous de la pile) de manière à le faire apparaître. Il peut être utile d'agrandir le graphique en tirant son coin inférieur droit tout en enfonçant le bouton gauche de la souris. Approchez le pointeur de la souris afin de consulter les coordonnées des points.



?

De quels nombres s'agit-il? Justifiez le titre du graphique.

2.4.1 Votre réponse



Faites apparaître le graphique du milieu. Approchez le pointeur de la souris afin de consulter les coordonnées des points.



?

De quels nombres s'agit-il? Justifiez le titre du graphique.

2.4.2 Votre réponse

Nous n'allons pas examiner le troisième graphique maintenant.

SYNTHESE

Nous avons vu une deuxième manière de réaliser une régression linéaire en employant Microsoft Excel. Par rapport à la méthode vue dans la partie 1, celle-ci est à la fois plus générale et moins puissante. Plus générale, parce qu'elle fournit plus de résultats ainsi que des graphiques et qu'elle peut fonctionner dans le contexte de la régression linéaire multiple, donc quand il y a plusieurs variables explicatives. Moins puissante, parce qu'un changement des données ne se répercute pas automatiquement dans les résultats.

Nous avons expérimenté ici le fait que la somme des carrés des résidus est la plus petite possible. Nous avons donc minimisé le critère MSE, appelé ici la variance résiduelle.

Partie 3 Dans cette partie, le but de l'exercice est d'examiner les formules qui permettent de calculer les principaux éléments de la régression linéaire simple : la constante b_0 sur la ligne Intercept, le coefficient de régression b_1 ainsi que MSE ou la variance résiduelle. Certaines parties de cet exercice sont facultatives dans la mesure où nous n'aurons jamais besoin d'appliquer ces formules.

Le point important est d'être conscient de l'existence des formules qui fournissent une solution unique et presque toujours exacte (pour éclairer quelque peu cette restriction, voir la partie A. de l'exercice avancé). La justification de ces formules est fournie dans les parties B et C de l'exercice avancé.

3.1 UTILISATION DES FONCTIONS DE REGRESSION LINEAIRE SIMPLE

Ces fonctions sont illustrées dans un tableau de la feuille Main.

Remarque



La présentation n'est pas aussi soignée que celle d'Excel. Nous avons gardé celle du tableur Lotus 1-2-3, version 2, du début de la micro-informatique, qui a été un logiciel précurseur et dont les auteurs méritent bien cet hommage.

⇒ Pour atteindre ce tableau, pressez F5 et sélectionnez Regression_functions. Vous pouvez aussi cliquer sur le lien prévu en haut de la feuille principale Main « Fonctions Excel de régression linéaire simple ».

⇒ Placez le curseur sur un des nombres, par exemple 12,3125.

La formule d'Excel apparaît. Dans une version anglaise du logiciel que nous utilisons ici, la formule apparaît sous la forme INTERCEPT(Y ;X). Nous avons défini Y comme étant la plage B20:B24 et X comme étant la plage A20:A24. Nous supposons aussi que le « ; » est le séparateur (dans une configuration avec un point décimal, c'est souvent la virgule « , » qui est utilisée comme séparateur).

Sortie régression:

Constante	=INTERCEPT(Y;X)
Ecart-type résiduel	=STEYX(Y;X)
R carré	=RSQ (Y;X)
Nombre d'observations	=COUNT(Y)
Degrés de liberté	=COUNT(Y) – 2
Coefficient de régression	=SLOPE(Y;X)

**Remarques**

1. Un tableau à la fin de la partie 3 donne les équivalents dans les versions française et néerlandaise.
2. Il n'est pas possible de réaliser un ajustement sans constante avec ces fonctions.
3. Une autre solutions dans Excel consiste à employer la fonction "LINEST" ("DROITEREG" dans la version française, "LIJNSCH" dans la version néerlandaise) qui renvoie tous les résultats mentionnés ici (plus d'autres qui seront vus au chapitre 7). Cette solution est automatique mais est moins simple dans la mesure où les résultats sont présentés, presque en vrac dans une matrice à 4 lignes et qu'il faut de sérieuses connaissances pour récupérer les informations pertinentes. Nous l'emploierons au chapitre 7.



Comparez la valeur de R square à celle de R^2 dans les parties 1 et 2.



Comparez les coefficients estimés aux nombres obtenus dans l'équation dans les parties 1 et 2.



Comparez l'écart-type résiduel estimé fourni avec quantité appelée Standard Error dans la partie 2.

*3.1.1 Vos réponses*

Les deux paragraphes qui suivent sont facultatifs.

3.2 FORMULE POUR LE COEFFICIENT DE REGRESSION

facultatif

Considérons les écarts à la moyenne de chacune des deux variables et les carrés et produits de ces écarts. Ils sont présentés dans le tableau suivant.

x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-2	17	-3	4	9	6
2	-1	11	-9	1	81	9
2	-1	23	3	1	9	-3
4	1	19	-1	1	1	-1
6	3	30	10	9	100	30
\bar{x} : 3		\bar{y} : 20		16	200	41

Ces calculs sont détaillés dans un tableau de la feuille Main.

⇒ Pour atteindre ce tableau dans la feuille, pressez F5 et sélectionnez TAB2.3 (vous pouvez aussi cliquer sur le lien « Tableau 2.3 » prévu en haut de la feuille principale Main).

On introduit la variance des x_i

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^2]$$

?

Que vaut ici la variance de la variable explicative x ?



3.2.1 Votre réponse

La covariance entre les x_i et les y_i est donnée par l'expression

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]$$

?

Que vaut ici la covariance?



3.2.2 Votre réponse

?

Montrer que le rapport entre la covariance (réponse 3.2.2) et la variance de x (réponse 3.2.1) vaut 2,5625.



3.2.3 Votre réponse

On en déduit que le coefficient de régression b_1 peut s'obtenir comme

$$b_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

(Attention! Il s'agit bien, au dénominateur, de la variance de la variable explicative x , pas de celle de la variance dépendante y .)

facultatif

3.3 FORMULE POUR LA CONSTANTE

Nous avons vu que l'équation de la droite des moindres carrés est la suivante : $y = 12,3125 + 2,5625 x$.

Appliquez la formule suivante en utilisant la valeur de $b_1 = 2,5625$, en employant tous les chiffres et sans arrondir :

$$b_0 = \bar{y} - b_1 \bar{x}$$

?

Montrez que b_0 calculé par cette formule vaut 12,3125?



3.3.1 Votre réponse

On peut dire que la droite des moindres carrés passe par le *centre de gravité* des données, le point d'abscisse \bar{x} et d'ordonnée \bar{y} .

SYNTHESE

Nous avons vu une troisième manière de réaliser une régression linéaire simple en employant Microsoft Excel. Par rapport à la méthode vue dans la partie 2, celle-ci est à la fois moins générale et plus puissante. Elle est moins générale parce qu'elle fournit moins de résultats et pas de graphiques et qu'elle ne peut fonctionner dans le contexte de la régression linéaire multiple. Elle offre plus de résultats que la méthode de la partie 1 mais pas de graphique. Elle est aussi plus puissante que la méthode de la partie

2 dans la mesure où, comme pour la méthode de la partie 1, un changement des données se répercute automatiquement dans les résultats.

Annexe. Noms des fonctions pour la régression dans quelques versions d'Excel.

Terme	version française	Version néerlandaise	Version anglaise
Constante	=ORDONNEE.ORIGINE(Y;X)	=SNIJPUNT(Y;X)	=INTERCEPT(Y;X)
Ecart-type résiduel	=ERREUR.TYPE.XY(Y;X)	=STAND.FOUT.XY(Y;X)	=STEYX(Y;X)
R carré	=COEFFICIENT.DETERMINATION(Y;X)	=R.KWADRAAT(Y;X)	=RSQ(Y;X)
Nombre d'observations	=NB(Y)	=AANTAL(Y)	=COUNT(Y)
Degrés de liberté	=NB(Y) - 2	=AANTAL(Y) - 2	=COUNT(Y) - 2
Coefficient de régr.	=PENTE(Y;X)	=RICHTING(Y;X)	=SLOPE(Y;X)

Partie 4 Le but de cette partie de l'exercice est d'introduire la mesure de qualité de l'ajustement par le coefficient de corrélation et le coefficient de détermination.

On revient d'abord sur la variance résiduelle et l'écart-type résiduel avant d'introduire ou de rappeler le concept de coefficient de corrélation ainsi que celui de coefficient de détermination ou R carré.

A nouveau, nous présentons les formules pour les apprenants intéressés. Certaines parties de cet exercice sont facultatives dans la mesure où nous n'aurons jamais besoin d'appliquer ces formules. Celles-ci sont néanmoins intéressantes comme point de départ des interprétations. La justification de ces formules est fournie dans les parties B et C de l'exercice avancé.

4.1 LA VARIANCE RESIDUELLE

Nous avons vu que la qualité de l'ajustement peut être mesurée par la *variance résiduelle* notée $s_{y,x}^2$ (à gauche du point se trouve la variable dépendante, à droite, la variable explicative). Elle se confond ici avec le critère MSE.



Que vaut MSE dans l'exemple des dépenses en vacances?



4.1.1 Votre réponse

On peut aussi poser le problème de qualité de l'ajustement en termes de prédiction ou de prévision. Dans notre exemple, la prédiction des dépenses en vacances pour un client spécifié peut s'effectuer sans utiliser la variable explicative ou en l'utilisant.

a) Si l'on n'utilise pas la connaissance de x , la manière la plus raisonnable de prévoir les dépenses y est d'utiliser la moyenne des y_i dans l'échantillon.



Que vaut la meilleure prévision des dépenses, sans autre information?

4.1.2 Votre réponse



b) En employant la régression linéaire selon la méthode des moindres carrés, c'est-à-dire en employant la connaissance de x

?

Que vaut la meilleure prévision des dépenses, en employant l'information du nombre de personnes du ménage, égal à 5, par exemple?

4.1.3 Votre réponse



En général, on veut trouver *a priori* une mesure unique de la qualité des prédictions, en employant l'échantillon des données comme référence. A cette fin, on peut mesurer la dispersion des résidus par la variance.

a) Dans le premier cas, les résidus sont $y_i - \bar{y}$ et leur variance vaut s_y^2 .

b) Dans le second cas, la variance des résidus est la variance résiduelle $s_{y.x}^2$.

?

Que vaut la mesure de dispersion des résidus dans les deux cas:
a) sans employer d'information; b) en employant l'information du nombre de personnes du ménage, égal à 5, par exemple? Quel est, en pourcentage, l'amélioration réalisée?

4.1.4 Vos réponses



On peut simplifier les interprétations en introduisant deux notions : le coefficient de corrélation et le coefficient de détermination.

4.2 LE COEFFICIENT DE CORRELATION ET LE COEFFICIENT DE DETERMINATION

Le coefficient de corrélation entre x et y , r ou r_{xy} , se définit comme suit :

$$r_{xy} = \frac{S_{xy}}{S_x S_y} .$$



Avant de voir la suite, dites ce que vaut le coefficient de corrélation dans l'exemple?

4.2.1 Votre réponse



On définit le *coefficient de détermination* comme le carré du coefficient de corrélation (une autre définition sera donnée dans le cas de la régression multiple) :

$$R^2 = r_{xy}^2 = \left(\frac{s_{xy}}{s_x s_y} \right)^2$$

Le tableau suivant présente le calcul du coefficient de corrélation et du coefficient de détermination.

x_i	y_i	x_i^2	$x_i y_i$	y_i^2	
1	17	1	17	289	
2	11	4	22	121	
2	23	4	46	529	
4	19	16	76	361	
6	30	36	180	900	
$\bar{x}: 3$	$\bar{y}: 20$	61	341	2200	
$\bar{x} \cdot \bar{y}:$	60				
					$s_x^2 = \frac{61}{5} - (3)^2 = 3.2$
					$s_y^2 = \frac{2200}{5} - (20)^2 = 40$
					$s_{xy} = \frac{341}{5} - 60 = 8.2$
					$r = \frac{8}{(3.2 \cdot 40)^{1/2}} = 0.7248$
					$r^2 = R^2 = 0.5253$



➤ Pour atteindre ce tableau dans la feuille, pressez F5 et sélectionnez TAB2.8 (vous pouvez aussi cliquer sur le lien « Tableau 2.8 » prévu en haut de la feuille principale Main).



Avant de lire ce qui suit, comparez la valeur de R^2 à celle trouvée précédemment.

4.2.2 Vos réponses



4.3 LIEN ENTRE MSE ET LE COEFFICIENT DE DETERMINATION

Dans l'exemple, on a $R^2 = r^2 = s_{xy}^2 / s_x^2 s_y^2 = (8,2)^2 / (3,2 \times 40) = 0,5253$. Notons que le $s_{y.x}^2 / s_y^2 = 18,9875 / 40 = 0,4747$ et ceci vaut $1 - R^2$.

En effet, on peut montrer (voir cours avancé) qu'on a la relation très simple mais très importante:

$$\text{MSE} = s_{y.x}^2 = s_y^2 (1 - R^2).$$

Cela signifie que, contrairement à ce qui se passe en général avec les autres méthodes de prévision, une expression explicite est ici connue pour le critère MSE, sans devoir calculer aucun résidu (on retrouvera cette propriété pour la régression multiple).

Dans l'exemple, cela donne $\text{MSE} = 40 \times (1 - 0,5253) = 40 \times 0,4747 = 18,9875$. L'écart-type résiduel ("*residual standard deviation*") est évidemment égal à $\sqrt{\text{MSE}}$.

Nous avons constaté dans plusieurs des sorties que l'écart-type résiduel fourni par les logiciels n'est pas égal à la racine carrée de MSE. La raison est la suivante : on doit distinguer la variance de l'échantillon et l'estimation de la variance de la population. Nous avons déjà discuté ceci dans le chapitre 1, en introduisant l'estimation non biaisée de la variance. Au lieu de diviser la somme des carrés des écarts par le nombre d'observations n , on le divise par $n - 1$. Ce nombre $n - 1$ est appelé le nombre de degrés de liberté. Le principe est le même ici à la différence qu'au lieu d'estimer seulement la moyenne, donc un paramètre, on doit en estimer deux — les 2 coefficients de la régression — et le nombre de degré de liberté est alors $n - 2$. On divise donc par $n - 2$ au lieu de diviser par n .



Reprenez la comparaison entre les variances résiduelles et les écarts-types résiduels obtenus dans les parties 1, 2 et 3 de l'exercice.



4.3.1 Votre réponse

4.4 PRINCIPE DE DECOMPOSITION DE LA VARIANCE

Un premier aspect fondamental de cette expression de MSE est le principe de décomposition de la variance :

$$s_y^2 = s_y^2(1 - R^2) + s_y^2 R^2.$$

Il s'agit d'une simple identité algébrique dont on peut interpréter chaque terme. En effet s_y^2 mesure la dispersion des données relatives à la variable y qui nous intéresse. $MSE = s_y^2(1 - R^2)$ est la variance résiduelle $s_{y,x}^2$. Elle mesure la dispersion des résidus de la régression. C'est la partie de la variance de y qui *n'est pas* expliquée par la régression linéaire en fonction de x . Le facteur $1 - R^2$ représente donc la *proportion de la variance de y qui n'est pas expliquée par la régression*. En revanche, $s_y^2 R^2$ représente la partie de la variance de y qui *est* expliquée par la régression linéaire en fonction de x (on l'appelle parfois la *variance de régression*). Par conséquent, R^2 est la *proportion de la variance de y qui est expliquée par la régression linéaire en fonction de x* .



Décomposez la variance 40 des dépenses en vacances et déterminez la proportion de la variance des dépenses qui *n'est pas* expliquée par le nombre de personnes du ménage x et la proportion de la variance des dépenses qui *est* expliquée par x .

4.4.1 Votre réponse



4.5 SYNTHESSES DES CALCULS

Pour conclure, le tableau suivant présente, sur l'exemple, les calculs nécessaires pour l'application de la régression linéaire simple.

Partie 5 Le but de cette partie de l'exercice est d'employer un logiciel conçu pour le traitement de données temporelles. La seule raison pour l'employer ici est que nous l'emploierons dans plusieurs chapitres de la suite du cours. Ce n'est pas un logiciel conçu spécifiquement pour la régression multiple mais il offre néanmoins un peu plus de souplesse qu'Excel.

5.1 INTRODUCTION

Nous allons effectuer cette étude en employant Time Series Expert for Windows, en abrégé TSE.

- ⇒ Suivez les instructions rappelées en annexe du document introductif du cours afin de lancer le logiciel.
- ⇒ Choisissez le répertoire de données approprié sur votre disque (pas sur le CD-ROM): menu File ⇒ Open. Choisissez DATA puis CHAP02 puis CH02EX01.
- ⇒ Chargez le problème déjà préparé : HOLIDAY. Cliquez sur Open. Vous devez alors voir dans le bas de l'écran que la variable dépendante est Y, que l'échantillon d'estimation est 1 – 5 et que les prévisions seront calculées jusqu'en 6.
- ⇒ Pour visualiser le tableau des données: menu Data ⇒ Spreadsheet.
- ⇒ Pressez la touche fonction F3 (Load) pour charger une série dans la colonne A du tableau. Sélectionnez Y. Cliquez sur Open.
- ⇒ Cliquez dans la colonne B.
- ⇒ Pressez la touche fonction F3 pour charger une série dans la colonne B du tableau. Sélectionnez X. Cliquez sur Open.

Comparez avec les données dans Excel.

- ⇒ Sortez du tableur en cliquant le File ⇒ Exit TSE Spreadsheet.
- ⇒ Pour visualiser graphiquement les séries: menu Graphics ⇒ Series.
- ⇒ Cliquez sur Select. Sélectionnez Y (cliquez sur ce nom et

cliquez) puis X (enfoncez la touche Ctrl puis cliquez sur ce nom). Cliquez sur Open puis sur OK.

Comparez avec les données dans Excel.

- ⇒ Pour visualiser graphiquement les séries: menu Graphics ⇒ XY graphic/Scatter diagram
- ⇒ Cliquez sur Select Y. Sélectionnez Y. Cliquez sur Open.
- ⇒ Cliquez sur Select X. Sélectionnez X. Cliquez sur Open.
- ⇒ Cliquez sur OK pour obtenir le graphique.

Comparez avec les données dans Excel.

- ⇒ Fermez le graphique

5.2 ESTIMATION DE LA CONSTANCE

La régression linéaire simple nécessite de définir la variable explicative. Il n'y a pas de variable explicative qui soit définie dans le problème. Nous pouvons le voir de la manière suivante.

- ⇒ Pour estimer les paramètres d'une régression linéaire, menu Methods ⇒ Ordinary least squares ⇒ Expertise.
- ⇒ Dans la fenêtre de dialogue, sur la ligne Save residuals, vous devez voir "RES". Sur la ligne Save forecasts, vous devez voir "PRED". Cliquez OK.

Nous n'allons pas consulter la sortie de manière détaillée. Regardons surtout le début. La liste des variables explicatives ne contient que CONSTANT.



Quelle est la valeur estimée de la constante, située sur la ligne CONSTANT? Avez-vous déjà rencontré ce nombre dans les parties précédentes de l'exercice?





5.2.1 Votre réponse

?

Que vaut la valeur de R^2 sur la ligne Coefficient of determination (R square) ?

?

Que vaut la quantité sur la ligne Residual variance ? Est-ce un des résultats discutés dans le paragraphe 4.3 ? De même, que vaut la valeur sur la ligne Residual standard deviation ? Est-ce un des résultats discutés dans le paragraphe 4.3 ?

5.2.2 Vos réponses



Quittez la sortie.

Il est instructif de regarder la prévision pour la donnée supplémentaire pour laquelle $x = 5$. C'est la sixième donnée.



Pour obtenir le graphique des valeurs ajustées et prévisions en parallèle avec les données, procédez comme suit : menu Graphics
⇒ Predictions/Forecasts. Cliquez OK.

Comme pour tous les graphiques de TSE, il est possible de focaliser sur un point.



A cette fin, approchez le pointeur de la souris d'un point, par exemple la sixième prévision.



?

La prévision réalisée pour $x = 5$ (sur la courbe des prévisions, qui dans ce cas est une droite horizontale) est-elle bien celle qui a été discutée dans le paragraphe 4.1 quand aucune information n'est disponible?

5.2.3 Votre réponse



Pour quitter le mode graphique : menu File \Rightarrow Exit.

5.3 ESTIMATION DES DEUX PARAMETRES DE LA REGRESSION LINEAIRE

Nous allons ajouter X comme variable explicative en plus de la constante.



Pour définir X comme variable explicative : menu Data \Rightarrow Variables \Rightarrow Chosen. Une fenêtre de dialogue apparaît. Cliquez sur la ligne X et pressez la touche E. Ceci déclare que la variable X est explicative. « Exp » apparaît en face du nom. Cliquez OK pour valider.

Remarque

Plusieurs touches sont actives dans cette fenêtre. N'utilisez que la touche E. Ne tapez pas E quand le curseur est sur la ligne RES ou PRED. Si vous le faites, tapez à nouveau sur la touche E pour annuler la sélection.



Relancez la régression comme précédemment, c'est-à-dire par le menu Methods \Rightarrow Ordinary least squares \Rightarrow Expertise.



Dans la fenêtre de dialogue, sur la ligne Save residuals, vous devez voir "RES". Sur la ligne Save forecasts, vous devez voir "PRED". Laissez ces noms. Cliquez OK.

Nous n'allons pas consulter la sortie de manière détaillée. Regardons à nouveau le début. Cette fois X doit apparaître à côté de CONSTANT comme variable explicative. Vous devez voir la valeur de R^2 sur la ligne Coefficient of determination (R square), la variance résiduelle, l'écart-type résiduel et les estimations des coefficients, parmi d'autres résultats.



?

Comparez les coefficients estimés en face de CONSTANT et X aux nombres obtenus dans l'équation dans les parties 1, 2 et 3.

5.3.1 Votre réponse

?

Comparez la valeur de R square à celle de R^2 dans les parties 1 à 4.

?

Comparez la variance résiduelle fournie avec la variance résiduelle dans la partie 4.

?

Comparez l'écart-type résiduel estimé fourni avec quantité appelée Standard Error dans la partie 4.



5.3.2 Vos réponses



Pour visualiser les résidus dans un tableau: menu Data ⇒ Spreadsheet.



Pressez la touche fonction F3 pour charger une série dans la colonne A du tableau. Sélectionnez RES. Cliquez sur Open.

?

Les résidus sont-ils bien identiques à ceux trouvés dans les parties 2 et 4 ?



5.3.3 Votre réponse

Un graphique révélateur de la qualité de l'ajustement est constitué des résidus en fonction des valeurs ajustées.

- ⇒ Quittez le tableur de TSE par le menu Exit TSE Spreadsheet.
- ⇒ Pour l'obtenir, procédez comme suit : menu Graphics
XY graphic/Scatter diagram
- ⇒ Cliquez sur Select Y. Sélectionnez RES. Cliquez sur Open.
- ⇒ Cliquez sur Select X. Sélectionnez PRED. Cliquez sur Open.
- ⇒ Cliquez encore OK pour obtenir le graphique.

Il reste à regarder la prévision pour la donnée supplémentaire pour laquelle $x = 5$. C'est la sixième donnée.

- ⇒ Pour obtenir le graphique des valeurs ajustées et prévisions en parallèle avec les données, procédez comme suit : menu Graphics
⇒ Predictions/forecasts. Cliquez OK pour obtenir le graphique.
- ⇒ Focaliser sur le sixième point.



?

La prévision réalisée pour $x = 5$ est-elle bien celle qui a été obtenue dans le paragraphe 4.1 ?

5.3.4 Votre réponse

- ⇒ Pour quitter le mode graphique menu File ⇒ Exit.
- ⇒ Pour quitter TSE, menu File ⇒ Exit. Confirmez en pressant la touche Yes.

SYNTHESE

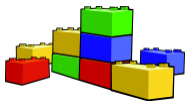
Notons d'abord que la régression linéaire sans variable explicative, c'est-à-dire seulement avec la constante fournit la moyenne comme valeur de cette constante. On peut en effet vérifier que la moyenne de l'échantillon est l'estimateur de la moyenne de la population, au sens des moindres carrés.

Nous avons vu une autre manière de réaliser une régression linéaire, cette fois en employant le logiciel Time Series Expert for Windows. Nous avons ainsi retrouvé la plupart des résultats précédents. Cette approche sera surtout intéressante dans les chapitres suivants et en particulier dans le chapitre 7 consacré à la régression linéaire multiple. Par rapport à Microsoft Excel, l'ajout ou la suppression d'une variable est en effet plus facile à réaliser.



Exercice avancé (Pour les utilisateurs de la version avancée du cours)

Préalable



Le chapitre 2 du cours de base et du cours avancé doit avoir été suivi jusqu'à la page 71.

Objectif



Le but est ici de discuter les aspects numériques de calcul ainsi que justifications empiriques et théoriques (facultatives).

Données



Les mêmes données que pour l'exercice de base.

Structure de l'exercice

L'exercice avancé comporte trois parties :

- Dans la partie A, le but de l'exercice est de discuter les aspects numériques de calcul.
- Dans la partie B, le but de l'exercice *facultatif* est de fournir les justifications *empiriques* de la méthode des moindres carrés et des formules utilisées dans les parties 3 et 4 de l'exercice de base. Ces justifications empiriques sont en grande partie expérimentales mais recourent néanmoins à la formule du minimum d'une fonction du second degré.
- Dans la partie C, le but de l'exercice *facultatif* est de fournir les justifications *théoriques* de la méthode des moindres carrés et des formules utilisées dans les parties 3 et 4 de l'exercice de base. Ces justifications théoriques reposent sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction.

Partie A Dans cette partie, le but de l'exercice est de discuter les aspects numériques de calcul.

Le tableau donné dans le paragraphe 4.2 a illustré le calcul de b_1 et de b_0 . Le procédé requiert deux passages des données : le premier passage sert à calculer les moyennes; le second passage permet alors de déterminer les sommes des carrés et des produits des écarts aux moyennes. Ce procédé présente un grave inconvénient: dans le cas du traitement sur ordinateur d'un très long fichier de données, il nécessite une double lecture de ces données. Lors de l'usage d'une calculatrice comportant des possibilités statistiques, le procédé est même impraticable car il faudrait une double saisie des données. La formule alternative de b_1 obtenue au paragraphe 3.2 n'a pas cet inconvénient. Elle permet d'utiliser l'algorithme suivant

Algorithme "calculatrice" en un passage pour le calcul des moyennes et variances.

```

{Initialisation}
  S1 ← 0
  S2 ← 0
{Corps}
  Pour i = 1 à n
    Lire  $x_i$ 
    S1 ← S1 +  $x_i$ 
    S2 ← S2 +  $x_i^2$ 
  fin pour
{Clôture}
   $\bar{x}$  ← S1/n
   $s_x^2$  ← S2/n -  $\bar{x}^2$ 

```



Le procédé en un seul passage a néanmoins un très gros inconvénient : il est peu précis. Pour s'en convaincre, il suffit de calculer la variance des nombres 1 000 000, 1 000 001 et 1 000 002 sur une calculatrice avec possibilités statistiques, si vous en avez une. Cette variance est la même que celle de 0, 1 et 2 et vaut donc $2/3$. Si la calculatrice affiche zéro, c'est que le nombre de décimales ne permet pas de distinguer $(1\,000\,000)^2$ de $(1\,000\,001)^2$. Pourtant, il n'est pas rare qu'une entreprise ou une administration ait à manipuler des nombres de cet ordre de grandeur. Pour une illustration, nous avons fait un calcul similaire dans Excel.

⇒ Cliquez sur la feuille Second. La deuxième ligne contient des nombres. On considère chaque fois trois données égales au nombre plus 1, 2 et 3. On en calcule la variance et l'écart-type. Elles sont correctes jusqu'à un certain ordre de grandeur puis deviennent fausses.

**Remarques**

1. Notons que la touche σ_n des calculatrices correspond à $\sqrt{s_x^2}$ tandis que σ_{n-1} correspond à $\sqrt{(ns_x^2)/(n-1)}$.

2. L'erreur mentionnée ci-dessus se produit dans Excel 97, 2000 et 2002 est due à un algorithme incorrect. Cette erreur a été corrigée dans Excel 2003.

Il existe également des procédés de calcul par récurrence qui n'emploient qu'un passage des données mais assurent une assez grande précision. Un tel algorithme est décrit comme suit.

Algorithme par récurrence en un passage pour le calcul des moyennes et variances.

```
{Initialisation}
  Lire  $x_1$ 
  S1  $\leftarrow x_1$ 
  S2  $\leftarrow 0$ 
{Corps}
  Pour  $i = 2$  à  $n$  {Attention! 2 et non 1}
    Lire  $x_i$ 
    S1  $\leftarrow S1 + x_i$ 
    S2  $\leftarrow S2 + ((i*x_i - S1)^2)/(i*(i - 1))$ 
  fin pour
{Clôture}
   $\bar{x} \leftarrow S1/n$ 
   $s_x^2 \leftarrow S2/n$ 
```

**Remarque**

Pour la covariance, il y a lieu de traiter les deux moyennes en utilisant deux accumulateurs S1x et S1y, disons, et de remplacer la ligne de mise à jour de S2 par :

$$S2 \leftarrow S2 + ((i*x_i - S1x)*(i*y_i - S1y))/(i*(i - 1)).$$

SYNTHESE

Nous avons montré qu'il n'est pas simple de calculer une variance efficacement (en un passage) et avec précision. Il faut employer un bon algorithme et effectuer les calculs avec une précision suffisante. Ceci montre a fortiori qu'il n'est pas recommandé de bricoler un programme de régression linéaire.

Partie B *Facultatif*

Dans cette partie, le but de l'exercice est de fournir les justifications *empiriques* de la méthode des moindres carrés et des formules utilisées dans les parties 3 et 4 de l'exercice de base. Cette démarche empirique n'est pas recommandée pour l'obtention des résultats numériques.

Cette partie ne suppose aucune connaissance mathématique, mais plutôt introduit les notions nécessaires.

B.a EXAMEN DU TABLEAU

Pour la commodité de la typographie, nous allons temporairement remplacer le coefficient de régression b_1 par b et la constante b_0 par a . Dans une première étape, il n'est pas nécessaire de consulter la feuille de calcul d'Excel, simplement le tableau ci-dessous. En parallèle, nous nous baserons néanmoins sur le classeur CH02EX01.XLS, feuille Main.

x_i	$x_i - \bar{x}$	y_i	$y_i + b(x_i - \bar{x})$	e_i	e_i^2
1	-2	17	$20 + (-2)b$	$-3 + (-2)b$	$9 + (-12)b + 4b^2$
2	-1	11	$20 + (-1)b$	$-9 + (-1)b$	$81 + (-18)b + 1b^2$
2	-1	23	$20 + (-1)b$	$3 + (-1)b$	$9 + (-6)b + 1b^2$
4	1	19	$20 + (1)b$	$-1 + (-1)b$	$1 + (-2)b + 1b^2$
6	3	30	$20 + (3)b$	$10 + (-3)b$	$100 + (-60)b + 9b^2$
$\bar{x}: 3$		$\bar{y}: 20$		$0 + (0)b$	$200 + (-82)b + 16b^2$
à minimiser					

?

Vérifiez le contenu des différentes colonnes.

?

Vérifiez la somme de la dernière colonne.

B.a.1 Votre réponse

B.b EXPERIMENTATION AVEC LA REGRESSION IMPOSEE

Afin de diminuer la somme des carrés des résidus, nous allons changer les deux paramètres, l'ordonnée à l'origine (cellule C15) et la pente (cellule F15) de la droite. Vous pouvez consulter la somme des résidus dans la cellule D26 et la somme des carrés dans la cellule F26.

B.c CENTRE DE GRAVITE

Notez que la moyenne arithmétique des x_i vaut $\bar{x} = 3$ et que celle des y_i vaut $\bar{y} = 20$. Le point d'abscisse \bar{x} et d'ordonnée \bar{y} est appelé le centre de gravité.

Le but est ici de montrer que la droite de régression au sens des moindres carrés passe par le centre de gravité.

Cela peut se faire en appliquant les étapes suivantes:

?

Essayez la droite d'équation $y = 13 + 2x$ et montrez que la somme des résidus (ou erreurs d'ajustement) vaut 5 et que la somme des carrés de ces mêmes résidus vaut 105.

B.c.1 Votre réponse

Afin de diminuer la somme des carrés des résidus, il faut veiller à ce que la somme de ces résidus soit nulle.

?

Justifiez pourquoi augmenter de 1 l'ordonnée à l'origine de la droite, donc remplacer 13 par $13 + 1 = 14$.

B.c.2 Votre réponse



?

Montrez que la droite obtenue a pour équation
 $y - \bar{y} = 2(x - \bar{x})$.

B.c.3 Votre réponse

Il faut montrer comment obtenir le coefficient de régression au sens des moindres carrés.

B.d COEFFICIENT DE REGRESSION OPTIMAL

Il reste à trouver le coefficient optimal à mettre à la place de 2, qui sera appelé le coefficient de régression (au sens des moindres carrés). A ce stade on peut dire que la droite des moindres carrés passe par le *centre de gravité* des données, le point d'abscisse \bar{x} et d'ordonnée \bar{y} et a donc pour équation $y - \bar{y} = b_1 (x - \bar{x})$.

?

Afin de diminuer la somme des carrés, remplacez la pente 2 de la droite par b et montrez, en vous basant sur le tableau du paragraphe B.a, que la valeur de b qui rend minimum la somme des carrés des résidus $e_i = y_i - [20 + b(x_i - 3)]$ vaut $b_1 = b = 82/32 = 2,5625$.



Remarque

Voici une indication. Ceci se base sur le fait que la fonction du second degré en x , $ax^2 + bx + c$ atteint un maximum ou un minimum pour $x = -b/(2a)$. Il s'agit d'un minimum si $a > 0$.



B.d.1 Votre réponse

B.e VERIFICATION

Montrez que la droite d'équation $y = 12,3125 + 2,5625 x$ passe par le centre de gravité.

?

Trouvez pour cela la moyenne des x et la moyenne des y et montrez qu'en remplaçant x par la moyenne des x dans le membre de droite on trouve la moyenne des y dans le membre de gauche. En outre on voit que la somme des erreurs dans la cellule D26 contient 0.



B.e.1 Votre réponse

SYNTHESE

Il semble évident que s'il fallait expérimenter aussi longtemps pour réaliser un ajustement de régression linéaire simple, la méthode ne serait jamais utilisée. L'avantage de cette approche est de fournir l'intuition nécessaire pour justifier les formules.

Partie C *Facultatif*

Dans cette partie, le but de l'exercice est de fournir les justifications *théoriques* de la méthode des moindres carrés et des formules utilisées dans les parties 3 et 4 de l'exercice de base.

C.a OBJECTIF

Montrons que les constatations relatives à cet exemple unique sont bien fondées et que l'application de la méthode des moindres carrés conduit aux résultats suivants:

$$b_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Le principe des moindres carrés consiste à minimiser la somme des carrés des résidus en fonction de b_0 et de b_1 . Pour la commodité de la typographie, nous allons temporairement remplacer b_1 par b et b_0 par a . La somme des carrés des résidus est une fonction Q de a et de b :

$$Q(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

C.b OBTENTION DES DERIVEES PREMIERES

Pour obtenir le minimum de cette fonction, on exprime que les *dérivées partielles premières* de $Q(a, b)$ par rapport à a , d'une part, et par rapport à b , d'autre part, sont nulles.

Rappelons que la dérivée partielle de $Q(a, b)$ par rapport à a est la dérivée ordinaire par rapport à a en considérant que b est une constante. Cette dérivée partielle s'écrit :

$$\begin{aligned} \frac{\partial Q(a, b)}{\partial a} &= \sum_{i=1}^n \frac{\partial}{\partial a} [(y_i - a - bx_i)^2] \\ &= \sum_{i=1}^n \left[2(y_i - a - bx_i) \frac{\partial (y_i - a - bx_i)}{\partial a} \right] \\ &= \sum_{i=1}^n [2(y_i - a - bx_i)(-1)], \end{aligned}$$

parce que y_i , x_i et b sont considérées comme des constantes dans la dérivée partielle de $(y_i - a - bx_i)$ par rapport à a . Par conséquent, imposer que la dérivée partielle de $Q(a, b)$ par rapport à a soit nulle équivaut à écrire

$$\sum_{i=1}^n (y_i - a - bx_i) = 0$$

Cette dernière relation exprime aussi que la moyenne \bar{e} des résidus $e_i = y_i - (a + b x_i)$ est nulle. On peut encore écrire

$$\sum_{i=1}^n y_i - \sum_{i=1}^n a - \sum_{i=1}^n b x_i = 0,$$

ou, en notant que $\sum_{i=1}^n a = na$ et en divisant les deux membres de l'équation par n : $\bar{y} - a - b\bar{x} = 0$. Ceci démontre la relation cherchée

$$a = \bar{y} - b\bar{x}.$$

En procédant de même pour la dérivée partielle de $Q(a,b)$ par rapport à b , on obtient :

$$\begin{aligned} \frac{\partial Q(a,b)}{\partial b} &= \sum_{i=1}^n \left[2(y_i - a - b x_i) \frac{\partial (y_i - a - b x_i)}{\partial b} \right] \\ &= \sum_{i=1}^n [2(y_i - a - b x_i)(-x_i)] \\ &= -2 \sum_{i=1}^n (x_i y_i - a x_i - b x_i^2). \end{aligned}$$

Ceci fournit une seconde équation

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n a x_i - \sum_{i=1}^n b x_i^2 = 0,$$

ou

$$\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0.$$

C.c RESOLUTION DU SYSTEME D'EQUATIONS

Nous avons trouvé deux équations liant a et b . Remplaçons dans la deuxième équation a par $\bar{y} - b\bar{x}$, qui vient de la première équation. Ceci donne l'équation en b suivante :

$$\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - b \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = 0,$$

ou, en divisant par n ,

$$\frac{1}{n} \left(\sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y} - b \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right) = 0.$$

Le lien entre l'expression trouvée pour b

$$b = \frac{\frac{1}{n} \left(\sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y}}{\frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) - \bar{x}^2}$$

et l'expression recherchée

$$b = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

est établi grâce aux relations suivantes :

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^2] = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

et

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y}.$$

La première relation correspond aux deux expressions équivalentes de la variance, rappelées dans la partie A. Démontrons la dernière relation relative à la covariance:

$$\begin{aligned} s_{xy} &= \frac{1}{n} \sum_{i=1}^n [x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}] \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \bar{y} - \bar{x} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) + \frac{1}{n} \sum_{i=1}^n \bar{x} \bar{y} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y}. \end{aligned}$$

Nous avons donc établi les formules donnant la pente b_1 et l'ordonnée à l'origine b_0 de la droite de régression au sens des moindres carrés. Il y a toutefois une condition : le dénominateur de b_1 doit être différent de zéro. Or nous venons de montrer que c'est la variance des valeurs de la variable explicative. Cette variance ne peut être nulle que quand tous les x_i sont égaux à \bar{x} . Or, dans ce cas, les x_i sont tous égaux entre eux et le problème posé au départ n'a pas de sens. On remarque également que la solution du problème est unique.



Remarque

Notons que nous n'avons pas montré que l'extremum de la fonction $Q(a,b)$ est bien un minimum; c'est toutefois évident sur base de la partie B.

SYNTHESE

Ces justifications théoriques reposent sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction. Non seulement elle fournit une justification des formules employées explicitement ou au travers de logiciels mais aussi elle montre le caractère unique de la solution obtenue.

[Retour au chapitre 2](#)

$$b = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

est établi grâce aux relations suivantes :

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^2] = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

et

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y}.$$

La première relation correspond aux deux expressions équivalentes de la variance, rappelées dans la partie A. Démontrons la dernière relation relative à la covariance:

$$\begin{aligned} s_{xy} &= \frac{1}{n} \sum_{i=1}^n [x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}] \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \bar{y} - \bar{x} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) + \frac{1}{n} \sum_{i=1}^n \bar{x} \bar{y} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y}. \end{aligned}$$

Nous avons donc établi les formules donnant la pente b_1 et l'ordonnée à l'origine b_0 de la droite de régression au sens des moindres carrés. Il y a toutefois une condition : le dénominateur de b_1 doit être différent de zéro. Or nous venons de montrer que c'est la variance des valeurs de la variable explicative. Cette variance ne peut être nulle que quand tous les x_i sont égaux à \bar{x} . Or, dans ce cas, les x_i sont tous égaux entre eux et le problème posé au départ n'a pas de sens. On remarque également que la solution du problème est unique.

Remarque

Notons que nous n'avons pas montré que l'extremum de la fonction $Q(a,b)$ est bien un minimum; c'est toutefois évident sur base de la partie B.



SYNTHÈSE

Ces justifications théoriques reposent sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction. Non seulement elle fournit une justification des formules employées explicitement ou au travers de logiciels mais aussi elle montre le caractère unique de la solution obtenue.

[Retour au chapitre 2](#)