12

RÉGRESSION LINÉAIRE SIMPLE

STATISTIQUES APPLIQUÉES

*Alliance Data Systems**
Dallas, État du Texas

Alliance Data Systems (ADS) fournit des moyens de traitement des transactions, des services de crédit et des services marketing à ses clients dans le domaine de la gestion des relations client, aujourd'hui en croissance. Les clients de ADS sont concentrés dans quatre secteurs : le commerce de détail, les stations-service, les services publics et les transports. En 1983, Alliance a commencé à proposer des services de traitement des crédits aux entreprises appartenant aux secteurs du commerce de détail (y compris les stations-service) et de la restauration ; aujourd'hui cette société emploie plus de 6 500 personnes et offre ses services à des clients à travers le monde. Gérant plus de 140 000 points de vente aux États-Unis, ADS traite plus de 2,5 milliards de transactions par an. La société se place au deuxième rang des sociétés américaines privées de services de crédit, en gérant 49 programmes touchant près de 72 millions de détenteurs d'une carte de crédit. En 2001, ADS a fait une première offre publique d'achat et est maintenant cotée à la bourse de New York.

L'un des services marketing d'ADS consiste à élaborer des campagnes promotionnelles par courrier. Grâce à sa base de données contenant des informations sur les habitudes d'achat de plus de 100 millions de consommateurs, ADS peut cibler les consommateurs qui seront les plus sensibles à une campagne promotionnelle. Le bureau de développement analytique utilise l'analyse de la régression pour construire des modèles permettant de mesurer et de prévoir la sensibilité des consommateurs à des campagnes marketing ciblées. Certains modèles de régression prédisent la probabilité d'achat des individus recevant une réduction, d'autres prédisent le montant dépensé par les consommateurs qui effectuent un achat.

Lors d'une campagne promotionnelle particulière, une chaîne de magasins souhaitait attirer de nouveaux consommateurs. Pour prévoir l'effet de la campagne, les analystes de ADS ont sélectionné un échantillon de consommateurs dans leur base de données, ont envoyé à ces individus un bon d'achat et ont ensuite collecté des données sur les transactions de ces clients : le montant d'achat ainsi que plusieurs variables spécifiques à chaque consommateur susceptibles d'être utiles pour prévoir les ventes. La variable spécifique à chaque consommateur la plus pertinente pour prévoir le montant des achats, était le montant total des dépenses effectuées dans des magasins similaires au cours des 39 derniers mois. Les analystes de ADS ont effectué une régression entre le montant des achats et le montant dépensé dans des magasins similaires :

$$\hat{y} = 26,7 + 0,00205x$$

où $\hat{y}$ correspond au montant des achats et x au montant dépensé dans des magasins similaires.

En utilisant cette équation, nous pouvons prédire qu'une personne qui a dépensé 10 000 dollars au cours des 39 derniers mois dans des magasins similaires, dépensera 47,20 dollars en réponse à la campagne promotionnelle ciblée. Dans ce chapitre, vous apprendrez à effectuer ce type de régression.

* Les auteurs remercient Philip Clemance, directeur du développement analytique chez Alliance Data Systems, de leur avoir fourni ce Statistiques appliquées.

Le modèle final développé par les analystes de ADS incluait également plusieurs autres variables, augmentant ainsi le pouvoir prédictif de l'équation précédente, telles que la possession ou non d'une carte de crédit bancaire, le revenu estimé et le montant moyen dépensé par visite dans un magasin particulier. Dans le chapitre suivant, nous verrons comment de telles variables additionnelles peuvent être incorporées dans un modèle de régression multiple.

Les décisions prises par un responsable sont souvent basées sur la relation qui existe entre deux ou plusieurs variables. Par exemple, après avoir considéré la relation entre les dépenses publicitaires et les ventes, un responsable marketing peut essayer de prévoir les ventes pour un montant donné de dépenses publicitaires. Autre exemple, un fournisseur d'électricité peut se servir de la relation entre la température journalière maximale et la demande en électricité pour prévoir la demande en électricité, en se basant sur les températures maximales prévues le mois suivant. Parfois, un responsable peut se fier à son intuition pour déterminer le type de relation qui lie deux variables. Cependant, s'il est possible d'obtenir des données, une procédure statistique, appelée *analyse de la régression*, permet de construire une équation indiquant de quelle manière les variables sont liées.

Dans la terminologie utilisée dans le cadre d'une analyse de la régression, la variable que l'on cherche à prévoir est appelée **variable dépendante**. La variable ou les variables utilisées pour prévoir la valeur de la variable dépendante sont appelées **variables indépendantes**. Par exemple, en analysant les effets des dépenses publicitaires sur les ventes, le responsable marketing cherche à prévoir les ventes ; les ventes correspondent donc à la variable dépendante et les dépenses publicitaires correspondent à la variable indépendante, utilisée pour prévoir les ventes. Dans la notation statistique usuelle, la variable dépendante est notée y et la variable indépendante est notée x.

Dans ce chapitre, nous considérons l'analyse de la régression la plus simple impliquant une variable indépendante et une variable dépendante, dont la relation est estimée par une ligne droite. Il s'agit de la **régression linéaire simple**. L'analyse de la régression impliquant au moins deux variables indépendantes, appelée **analyse de la régression multiple**, sera étudiée au chapitre 13.

Les méthodes statistiques utilisées pour étudier la relation entre deux variables ont été employées pour la première fois par Sir Francis Galton (1822-1911). Galton s'intéressait à la relation entre la taille d'un père et celle de son fils. Le disciple de Galton, Karl Pearson (1857-1936), analysa la relation entre la taille d'un père et celle de son fils à partir d'un échantillon de 1 078 paires de sujets.

12.1 LE MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE

Les pizzerias Armand sont une chaîne de restaurants italiens, implantée dans cinq États américains. Les restaurants les plus fréquentés se situent près des campus universitaires. Les responsables pensent que les ventes trimestrielles de ces restaurants (notées y) sont positivement liées à la taille de la population étudiante (notée x) ; en d'autres termes, les restaurants situés près des campus universitaires de grande taille ont tendance à générer un plus gros chiffre d'affaires que ceux situés près des campus de plus petite taille. En utilisant l'analyse de la régression, nous pouvons construire une équation indiquant de quelle manière la variable dépendante y est liée à la variable indépendante x.

12.1.1 Modèle de régression et équation de la régression

Dans l'exemple des pizzerias Armand, la population étudiée correspond à l'ensemble des restaurants Armand. À chaque restaurant de la population sont associées une valeur x (la population étudiante) et une valeur y (les ventes trimestrielles). L'équation qui décrit la relation qui lie y à x et à un terme d'erreur, correspond à un **modèle de régression**. Le modèle de régression utilisé dans une régression linéaire simple s'écrit de la façon suivante :

► **Modèle de régression linéaire simple**

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (12.1)$$

β_0 et β_1 correspondent aux paramètres du modèle et ε (la lettre grecque epsilon) est une variable aléatoire appelée terme d'erreur. Le terme d'erreur prend en compte la variabilité de y qui n'est pas expliquée par la relation linéaire entre x et y.

La population de tous les restaurants Armand peut être vue comme un ensemble de sous-populations, une pour chaque valeur particulière de x. Par exemple, l'une des sous-populations est constituée de tous les restaurants Armand situés près de campus universitaires regroupant 8 000 étudiants ; une autre sous-population est constituée de tous les restaurants Armand situés près de campus universitaires regroupant 9 000 étudiants ; etc. Chaque sous-population a une distribution particulière des valeurs y. Ainsi, une distribution des valeurs y est associée aux restaurants situés près de campus regroupant 8 000 étudiants ; une distribution des valeurs y est associée aux restaurants situés près de campus regroupant 9 000 étudiants ; etc. Chaque distribution des valeurs y a sa propre moyenne ou espérance mathématique. L'équation qui décrit comment la moyenne ou l'espérance mathématique de y, notée $E(y)$, est liée à x, est appelée **équation de la régression**. L'équation de la régression dans le cadre d'une régression linéaire simple s'écrit :

► **Équation de la régression linéaire simple**

$$E(y) = \beta_0 + \beta_1 x \qquad (12.2)$$

L'équation de la régression linéaire simple est représentée graphiquement par une droite ; β_0 correspond à l'ordonnée à l'origine de la droite de régression, β_1 correspond à la pente et $E(y)$ est la moyenne ou espérance mathématique de y pour une valeur donnée de x.

La figure 12.1 regroupe quelques exemples de droites de régression possibles, dans le cadre d'une régression linéaire simple. Dans le cas A, la moyenne de y est positivement liée à x, de plus grandes valeurs de $E(y)$ étant associées à de plus grandes valeurs de x. Dans le cas B, la moyenne de y est négativement liée à x, de plus petites valeurs de $E(y)$ étant associées à de plus grandes valeurs de x. Dans le cas C, la moyenne de y n'est pas liée à x, la moyenne de y étant la même pour chaque valeur de x.

12.1.2 Équation estimée de la régression

Si la valeur des paramètres de la population β_0 et β_1 était connue, nous pourrions utiliser l'équation (12.2) pour calculer la moyenne de y pour une valeur donnée de x. En pratique, la valeur des paramètres n'est pas connue et doit être estimée en utilisant les données d'un échantillon. Les statistiques d'échantillon (notées b_0 et b_1) servent d'estimations des paramètres de la population β_0 et β_1. En substituant les valeurs de b_0 et b_1 à la place de β_0 et β_1 dans l'équation de la régression, nous obtenons **l'équation estimée de la régression**. L'équation estimée de la régression, dans le cadre d'une régression linéaire simple, s'écrit :

► **Équation estimée de la régression linéaire simple**

$$\hat{y} = b_0 + b_1 x \qquad (12.3)$$

La figure 12.2 résume le processus d'estimation dans le cadre d'une régression linéaire simple.

Le graphique de l'équation estimée de la régression linéaire simple est appelé *droite de régression estimée* ; b_0 correspond à l'ordonnée à l'origine et b_1 correspond à la pente. Dans la section suivante, nous montrerons comment appliquer la méthode des moindres carrés pour calculer les valeurs de b_0 et b_1 dans l'équation estimée de la régression.

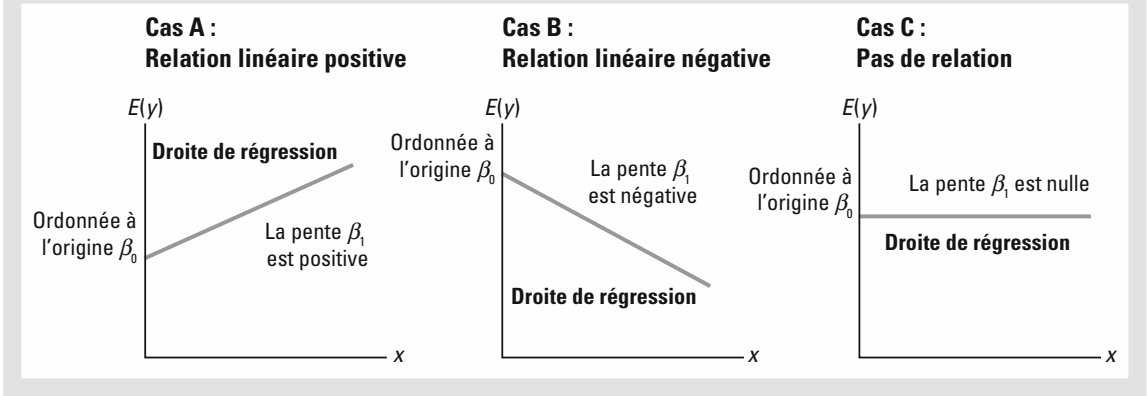

Figure 12.1 ***Droites de régression possibles dans une régression linéaire simple***

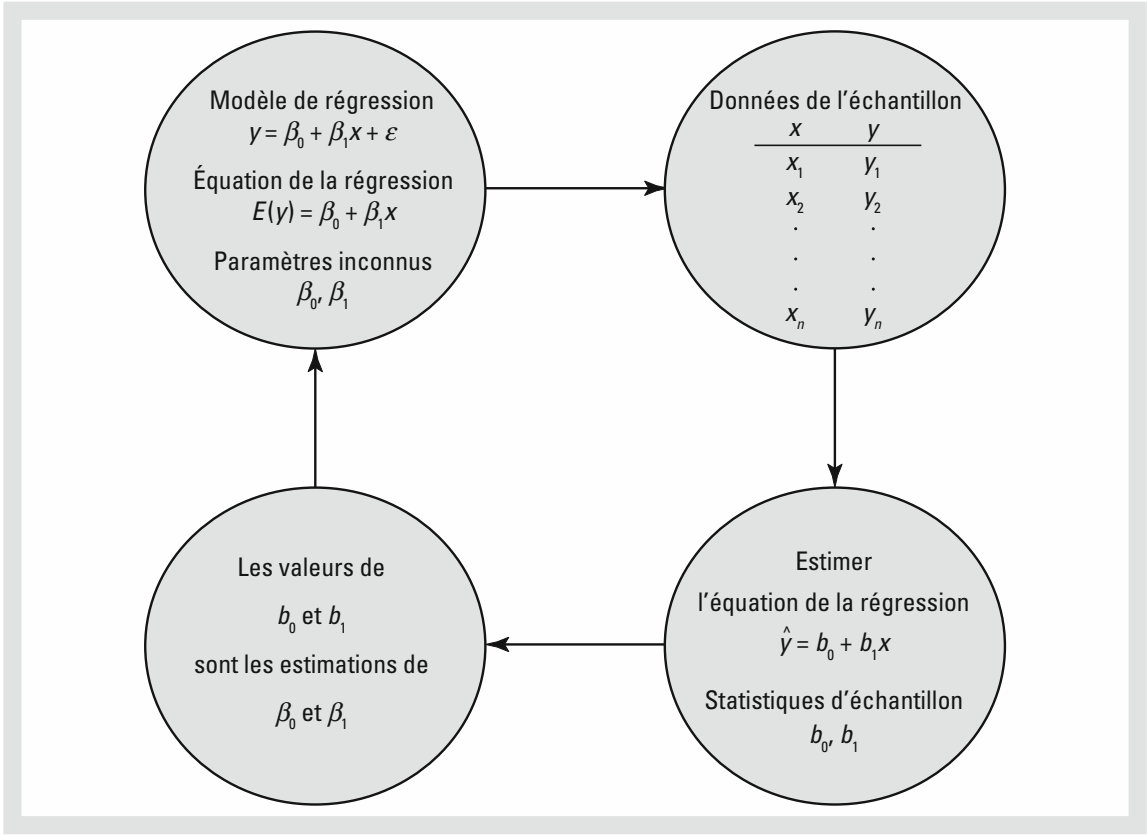

Figure 12.2 ***Processus d'estimation dans le cadre d'une régression linéaire simple***

L'estimation de β_0 et β_1 est une procédure statistique semblable à l'estimation de □ décrite dans le chapitre 7. β_0 et β_1 sont les paramètres inconnus qui nous intéressent et b_0 et b_1 sont les statistiques d'échantillon utilisées pour estimer les paramètres.

En général, $\hat{y}$ correspond à l'estimateur ponctuel de $E(y)$, la valeur moyenne de y pour une valeur particulière de x. Ainsi, pour estimer la moyenne des ventes trimestrielles des restaurants situés près de campus universitaires regroupant 10 000 étudiants, il faut substituer 10 000 à x dans l'équation (12.3). Dans certains cas, cependant, les restaurants Armand seront davantage intéressés par les prévisions de ventes dans un restaurant particulier. Par exemple, supposez qu'Armand veuille prévoir les ventes trimestrielles du restaurant situé près de l'université Talbot, comptant 10 000 étudiants. La meilleure estimation de y pour une valeur donnée de x est également fournie par $\hat{y}$. Ainsi, pour prévoir les ventes trimestrielles du restaurant situé près de l'université Talbot, Armand substituera également la valeur 10 000 à x dans l'équation (12.3).

La valeur de $\hat{y}$ fournit à la fois une estimation ponctuelle de E(y) pour une valeur donnée de x et une prédiction d'une valeur individuelle y pour une valeur donnée de x.

REMARQUES

1. L'analyse de la régression ne peut pas être interprétée comme une procédure établissant une relation de cause à effet entre deux variables. Elle peut simplement indiquer comment ou dans quelle mesure les variables sont associées les unes avec les autres. Toute conclusion sur les causes et les effets doit être basée sur l'opinion des individus les plus à même de porter un tel jugement.
2. L'équation de la régression dans une régression linéaire simple est $E(y) = \beta_0 + \beta_1 x$. Des ouvrages plus avancés sur l'analyse de la régression écrivent souvent l'équation de la régression $E(y|x) = \beta_0 + \beta_1 x$ pour souligner le fait que l'équation de la régression fournit la moyenne de y pour une valeur donnée de x.

12.2 LA MÉTHODE DES MOINDRES CARRÉS

La **méthode des moindres carrés** est une procédure qui permet d'utiliser les données de l'échantillon pour estimer l'équation de la régression. Pour illustrer la méthode des moindres carrés, supposons que nous ayons collecté des données sur un échantillon de 10 restaurants Armand, situés près de campus universitaires. Pour le i^e restaurant de l'échantillon, x_i correspond à la taille de la population étudiante (en milliers) et y_i correspond aux ventes trimestrielles (en milliers de dollars). Les valeurs de x_i et y_i associées aux 10 restaurants de l'échantillon sont présentées dans le tableau 12.1 (cf. fichier en ligne Armand). Le restaurant 1, caractérisé par $x_1 = 2$ et $y_1 = 58$, est situé près d'un campus regroupant 2 000 étudiants et ses ventes trimestrielles s'élèvent à 58 000 dollars. Le restaurant 2, caractérisé par $x_2 = 6$ et $y_2 = 105$, est situé près d'un campus regroupant 6 000 étudiants et ses ventes trimestrielles s'élèvent à 105 000 dollars. Le restaurant 10, situé sur un campus de 26 000 étudiants, détient le montant des ventes le plus élevé, avec 202 000 dollars de ventes trimestrielles.

Tableau 12.1 ***Données sur la population étudiante et les ventes trimestrielles de dix restaurants Armand***

Restaurant i	x_i = Population étudiante (en milliers)	y_i = Ventes trimestrielles (en milliers de dollars)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Dans une régression linéaire simple, chaque observation est composée de deux valeurs : l'une est associée à la variable dépendante, l'autre à la variable indépendante.

La figure 12.3 correspond au nuage de points, obtenu avec les données du tableau 12.1. L'axe des abscisses représente la taille de la population étudiante et l'axe des ordonnées représente la valeur des ventes trimestrielles. Les **nuages de points** des analyses de la régression sont construits en plaçant les valeurs de la variable indépendante x sur l'axe des abscisses et les valeurs de la variable dépendante y sur l'axe des ordonnées. Les nuages de points nous permettent d'observer graphiquement les données et de tirer des conclusions préliminaires sur la relation éventuelle entre les variables.

Quelles conclusions préliminaires pouvez-vous tirer de la figure 12.3 ? Les ventes trimestrielles semblent être supérieures sur les campus regroupant plus d'étudiants. De plus, pour ces données, la relation entre la taille de la population étudiante et les ventes trimestrielles semble pouvoir être estimée par une droite ; il semble donc y avoir une relation linéaire positive entre x et y. Nous choisissons par conséquent un modèle de régression linéaire simple pour représenter la relation entre les ventes

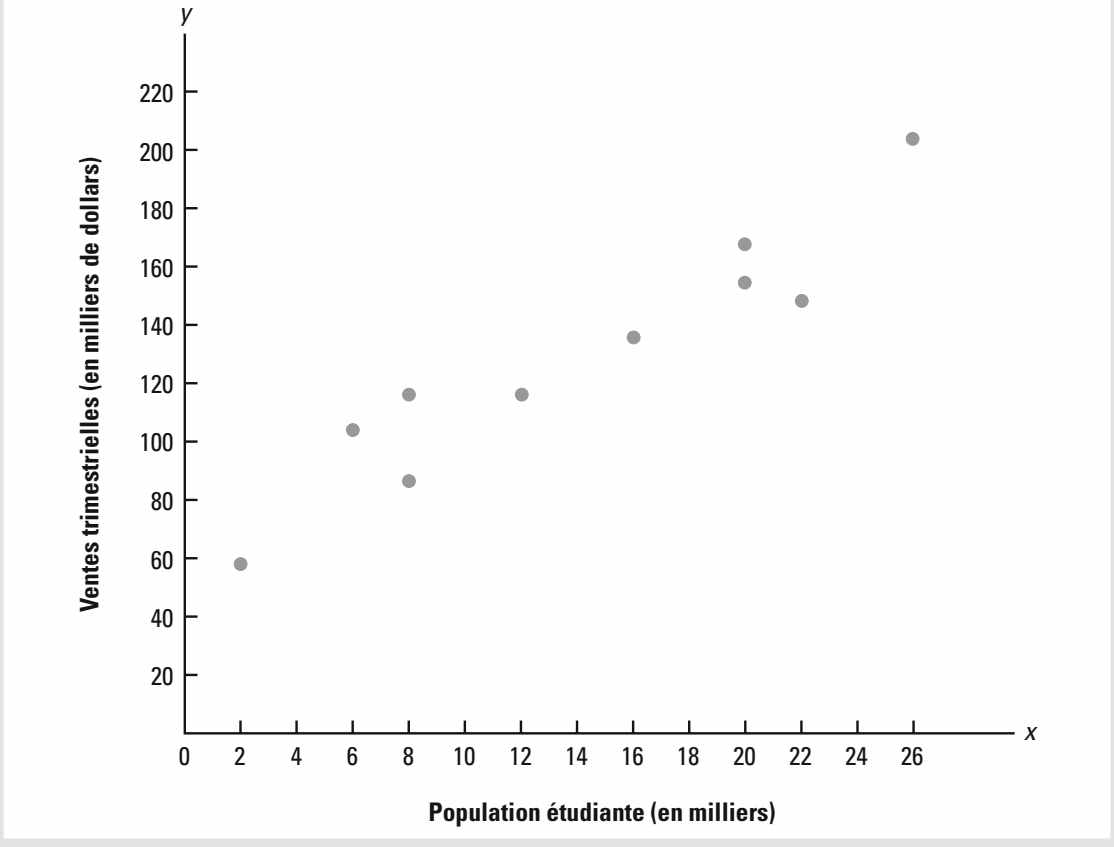

Figure 12.3 *Nuage de points de la population étudiante et des ventes trimestrielles pour les restaurants Armand*

trimestrielles et la population étudiante. L'étape suivante consiste à utiliser les données d'échantillon du tableau 12.1 pour déterminer les valeurs de b_0 et b_1 dans l'équation estimée de la régression linéaire simple. Pour le i^e restaurant, l'équation estimée de la régression s'écrit

$$\hat{y}_i = b_0 + b_1 x_i \tag{12.4}$$

où

$\hat{y}_i$ correspond à la valeur estimée des ventes trimestrielles (en milliers de dollars) du i^e restaurant

b_0 correspond à l'ordonnée à l'origine de la droite de régression estimée

b_1 correspond à la pente de la droite de régression estimée

x_i correspond à la taille de la population étudiante (en milliers) associée au i^e restaurant

Avec les ventes trimestrielles observées (réelles) du restaurant i notées y_i et $\hat{y}_i$ représentant la valeur estimée des ventes trimestrielles du i^e restaurant, chaque restaurant de l'échantillon est caractérisé par une valeur observée des ventes trimestrielles y_i et une valeur estimée des ventes trimestrielles $\hat{y}_i$. Si l'écart entre les valeurs observées et les valeurs estimées est faible, on peut considérer que la droite de régression estimée est bien adaptée aux données.

La méthode des moindres carrés utilise les données de l'échantillon pour fournir les valeurs de b_0 et b_1 qui minimisent la *somme des écarts au carré* entre les valeurs observées de la variable dépendante y_i et les valeurs estimées de cette dernière $\hat{y}_i$. L'expression (12.5) formule le critère de la méthode des moindres carrés.

► **Critère des moindres carrés**

$$\min \sum \left(y_i - \hat{y}_i\right)^2 \tag{12.5}$$

où

y_i correspond à la valeur observée de la i^e observation de la variable dépendante

$\hat{y}_i$ correspond à la valeur estimée de la i^e observation de la variable dépendante

La méthode des moindres carrés a été élaborée par Carl Friedrich Gauss (1777-1855).

Un calcul différentiel permet de démontrer que les valeurs de b_0 et b_1 qui minimisent l'expression (12.5), peuvent être obtenues en utilisant les expressions (12.6) et (12.7).

► Pente et ordonnée à l'origine de l'équation estimée de la régression[1]

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{12.6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{12.7}$$

où

x_i correspond à la valeur de la i^e observation de la variable indépendante
y_i correspond à la valeur de la i^e observation de la variable dépendante
$\bar{x}$ correspond à la moyenne de la variable indépendante
$\bar{y}$ correspond à la moyenne de la variable dépendante
n correspond au nombre total d'observations

Lors du calcul de b_1 avec une calculatrice, utilisez le plus grand nombre possible de chiffres décimaux dans les calculs intermédiaires. Nous recommandons d'utiliser au moins quatre chiffres après la virgule.

Le tableau 12.2 présente certains calculs nécessaires à l'obtention de l'équation estimée de la régression des moindres carrés dans le cadre des restaurants Armand. Avec un échantillon de 10 restaurants, nous avons 10 observations ($n = 10$). Nous commençons par calculer $\bar{x}$ et $\bar{y}$, nécessaires à l'application des équations (12.6) et (12.7).

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1\,300}{10} = 130$$

En utilisant les expressions (12.6) et (12.7), et les informations contenues dans le tableau 12.2, nous pouvons calculer la pente et l'ordonnée à l'origine de l'équation estimée de la régression dans le cadre des restaurants Armand. Les calculs de la pente (b_1) suivent.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{2\,840}{568} = 5$$

1 Une formule alternative pour b_1 est $b_1 = \dfrac{\sum x_i y_i - \left(\sum x_i \sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n}$. Cette forme de l'équation (12.6) est souvent recommandée lorsqu'une calculatrice est utilisée pour obtenir b_1.

Tableau 12.2 ***Calculs associés à l'estimation par les moindres carrés de l'équation de la régression pour les restaurants Armand***

Restaurant *i*	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totaux	140 $\sum x_i$	1 300 $\sum y_i$			2 840 $\sum (x_i - \bar{x})(y_i - \bar{y})$	568 $\sum (x_i - \bar{x})^2$

Les calculs de l'ordonnée à l'origine (b_0) suivent.

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= 130 - 5(14) \\ &= 60 \end{aligned}$$

Ainsi, l'équation estimée de la régression s'écrit :

$$\hat{y} = 60 + 5x$$

Le graphique 12.4 représente cette équation au milieu du nuage de points.

La pente de l'équation estimée de la régression ($b_1 = 5$) est positive, impliquant que lorsque la taille de la population étudiante augmente, les ventes trimestrielles augmentent. En fait, nous pouvons conclure qu'une augmentation de la taille de la population de 1 000 étudiants entraînera une augmentation des ventes trimestrielles de 5 000 dollars ; en d'autres termes, les ventes trimestrielles devraient augmenter de 5 dollars par étudiant.

Si nous pensons que l'équation estimée par la méthode des moindres carrés décrit correctement la relation entre x et y, il est raisonnable d'utiliser l'équation estimée de la régression pour prévoir la valeur de y pour une valeur donnée de x. Par exemple, si nous voulions prévoir les ventes d'un restaurant situé près d'un campus de 16 000 étudiants, nous calculerions

$$\hat{y} = 60 + 5(16) = 140$$

Par conséquent, nous prévoirions des ventes trimestrielles d'un montant de 140 000 dollars dans ce restaurant. Dans les sections suivantes, nous discuterons des méthodes qui permettent de juger de la pertinence de l'utilisation de l'équation estimée de la régression pour effectuer des prévisions.

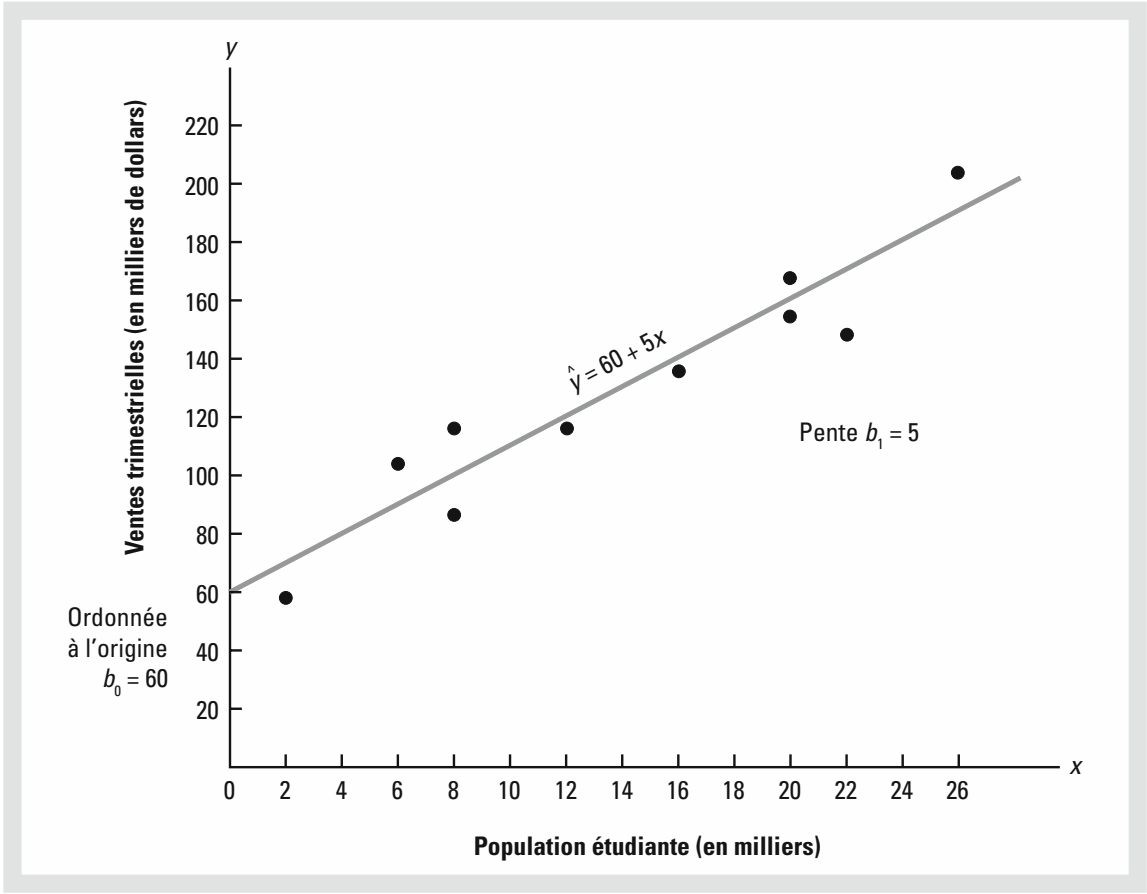

Figure 12.4 ***Graphique de l'équation estimée de la régression pour les restaurants Armand : $\hat{y}_i = 60 + 5x$***

Il faut être prudent lorsqu'on utilise l'équation estimée de la régression pour effectuer des prévisions pour des valeurs de la variable indépendante qui sortent de l'intervalle étudié, car il n'est pas certain que la relation reste valable pour de telles valeurs de la variable indépendante.

REMARQUES

La méthode des moindres carrés fournit une équation estimée de la régression qui minimise la somme des écarts au carré entre les valeurs observées de la variable dépendante, y_i et les valeurs estimées de la variable dépendante, $\hat{y}_i$. Le critère des moindres carrés permet d'obtenir l'équation la mieux adaptée aux données. Si on utilise d'autres critères, tels que la minimisation de la somme des écarts en valeur absolue entre y_i et $\hat{y}_i$, on obtiendra une équation différente. En pratique, la méthode des moindres carrés est la plus répandue.

EXERCICES

Méthode

1. Ci-dessous sont présentées les données concernant cinq observations de deux variables, x et y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Représenter le nuage de points associé à ces données.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Essayer de décrire la relation entre x et y en traçant une droite à travers le nuage de points.

d) Développer l'équation estimée de la régression en calculant les valeurs de b_0 et b_1 grâce aux expressions (12.6) et (12.7).

e) Utiliser l'équation estimée de la régression pour prévoir la valeur de y lorsque $x = 4$.

2. Ci-dessous sont présentées les données concernant cinq observations de deux variables, x et y.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Représenter le nuage de points associé à ces données.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Essayer de décrire la relation entre x et y en traçant une droite à travers le nuage de points.

d) Développer l'équation estimée de la régression en calculant les valeurs de b_0 et b_1 grâce aux expressions (12.6) et (12.7).

e) Utiliser l'équation estimée de la régression pour prévoir la valeur de y lorsque $x = 10$.

3. Ci-dessous sont présentées les observations collectées lors d'une analyse de la régression avec deux variables.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a) Représenter le nuage de points associé à ces variables.

b) Développer l'équation estimée de la régression correspondant à ces données.

c) Utiliser l'équation estimée de la régression pour prévoir la valeur de y lorsque $x = 6$.

Applications

4. Les données suivantes correspondent au pourcentage de femmes employées dans cinq entreprises dans le secteur du commerce de détail. Le pourcentage de postes à responsabilité confiés à des femmes dans chaque entreprise est également indiqué.

% de femmes employées	67	45	73	54	61
% de femmes responsables	49	21	65	47	33

a) Représenter le nuage de points associé à ces données en utilisant le pourcentage de femmes travaillant dans l'entreprise comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Essayer de décrire la relation entre le pourcentage de femmes travaillant dans l'entreprise et le pourcentage de postes à responsabilité confiés à des femmes.

d) Développer l'équation estimée de la régression en calculant les valeurs de b_0 et b_1.

e) Prédire le pourcentage de postes à responsabilité confiés à des femmes dans une entreprise employant 60 % de femmes.

5. La société Brawdy Plastics fabrique des ceintures de sécurité pour General Motors dans son usine de Buffalo, dans l'État de New York. Une fois assemblées et peintes, les pièces sont placées sur une chaîne de montage qui les entraînent jusqu'au poste d'inspection finale. La rapidité à laquelle les pièces passent devant le poste d'inspection finale dépend de la vitesse de la chaîne de montage (mesurée en pied par minute). Bien que des vitesses accrues soient désirables, la direction s'inquiète du fait qu'une très forte augmentation de la vitesse de la chaîne de montage ne fournisse pas suffisamment de temps aux inspecteurs pour identifier les pièces défectueuses. Pour tester cette théorie, Brawdy Plastics a mené une expérimentation dans laquelle le même ensemble de pièces, dont le nombre de pièces défectueuses était connu, a été inspecté à différentes vitesses de la chaîne de montage. Les données suivantes ont été collectées.

Vitesse de la chaîne de montage	Nombre de pièces défectueuses trouvées
20	23
20	21
30	19
30	16
40	15
40	17
50	14
50	11

a) Représenter le nuage de points associé à ces données en considérant la vitesse de la chaîne de montage comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Utiliser la méthode des moindres carrés pour estimer l'équation de la régression.

d) Prédire le nombre de pièces défectueuses trouvées pour une chaîne de montage avançant à la vitesse de 25 pieds par minute.

6. La ligue nationale de football (NFL) enregistre différentes données sur les performances des individus et des équipes. Pour déterminer l'importance des passes dans le pourcentage de parties gagnées par une équipe, des données (cf. fichier en ligne NFL Passes) sur le nombre moyen de yards parcourus en faisant des passes (yards) et le pourcentage de parties gagnées (% parties gagnées) ont été collectées à partir d'un échantillon aléatoire de 10 équipes de la NFL au cours de la saison 2011 (site Internet de la NFL, 12 février 2012).

Équipe	Yards	% parties gagnées
Arizona Cardinals	6,5	50
Atlanta Falcons	7,1	63
Carolina Panthers	7,4	38
Chicago Bears	6,4	50
Dallas Cowboys	7,4	50
New England Patriots	8,3	81
Philadelphia Eagles	7,4	50
Seattle Seahawks	6,1	44
St. Louis Rams	5,2	13
Tampa Bay Buccaneers	6,2	25

a) Représenter le nuage de points associé à ces données, avec le nombre de yards parcourus en faisant des passes sur l'axe horizontal et le pourcentage de parties gagnées sur l'axe vertical.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Développer l'équation de régression estimée qui pourrait être utilisée pour prédire le pourcentage de parties gagnées étant donné le nombre moyen de yards parcourus en faisant des passes.

d) Interpréter la pente de l'équation de la régression estimée.

e) Au cours de la saison 2011, le nombre moyen de yards parcourus en faisant des passes par les Kansas City Chiefs fut de 6,2. Utiliser l'équation de la régression estimée pour prédire le pourcentage de parties gagnées par cette équipe. (Remarque : au cours de la saison 2011, les Kansas City Chiefs ont gagné 9 parties et en ont perdu 7). Comparer votre prédiction au pourcentage réel de parties gagnées par les Kansas City Chiefs.

7. Un responsable des ventes a collecté les données suivantes sur les années d'expérience et le montant des ventes annuelles de différents vendeurs (cf. fichier en ligne Ventes).

Vendeur	Années d'expérience	Ventes annuelles (milliers de dollars)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111

Vendeur	Années d'expérience	Ventes annuelles (milliers de dollars)
7	10	119
8	10	123
9	11	117
10	13	136

a) Représenter le nuage de points associé à ces données, en utilisant le nombre d'années d'expérience comme variable indépendante.

b) Estimer l'équation de la régression qui peut être utilisée pour prévoir les ventes annuelles sachant le nombre d'années d'expérience du vendeur.

c) Utiliser l'équation estimée de la régression pour prévoir les ventes annuelles d'un vendeur qui a neuf années d'expérience.

8. L'enquête en ligne sur les courtiers de l'Association Américaine des Investisseurs Individuels (AAII) sonde les membres de l'association sur leurs expériences avec des courtiers. On demande notamment aux membres d'évaluer la qualité de la rapidité d'exécution des ordres et de fournir une note de satisfaction globale des transactions électroniques (cf. fichier en ligne Notation Courtiers). Les réponses possibles (notes) étaient : sans opinion (0), insatisfait (1), assez satisfait (2), satisfait (3) et très satisfait (4). Pour chaque courtier, une note résumant son appréciation a été établie sur la base de la moyenne pondérée des notes fournies par chaque membre interrogé. Une partie des résultats de l'enquête est fournie ci-dessous (site Internet de l'AAII, 7 février 2012).

Courtier	Rapidité d'exécution	Satisfaction
Scottrade, Inc.	3,4	3,5
Charles Schwab	3,3	3,4
Fidelity Brokerage Services	3,4	3,9
TD Ameritrade	3,6	3,7
E*Trade Financial	3,2	2,9
Vanguard Brokerage Services	3,8	2,8
USAA Brokerage Services	3,8	3,6
Thinkorswim	2,6	2,6
Wells Fargo Investments	2,7	2,3
Interactive Brokers	4,0	4,0
Zecco.com	2,5	2,5

Notation Courtiers

a) Représenter le nuage de points associé à ces données en utilisant la rapidité d'exécution comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Estimer par les moindres carrés l'équation de la régression.

d) Interpréter la pente de l'équation estimée de la régression.

e) Supposez que Zecco.com ait développé un nouveau logiciel pour augmenter la note qui lui est attribuée au regard de la rapidité d'exécution des ordres. Si le nouveau logiciel est capable d'accroître sa note de la valeur actuelle de 2,5 à la note moyenne des 10 autres courtiers étudiés, quelle serait la note de satisfaction globale selon vous ?

9. Les sociétés de location de voiture américaines varient fortement au regard de la taille de leur flotte, de leur nombre d'agences et de leur revenu annuel. En 2011, Hertz avait 320 000 véhicules de location en service et un revenu annuel d'environ 4,2 milliards de dollars. Les données suivantes indiquent le nombre de véhicules en service (en milliers) et le revenu annuel (en millions de dollars) pour six sociétés de location de voiture plus petites (site Internet de *Auto Rental News*, 7 août 2012).

Société	Véhicules (milliers)	Revenu (millions de dollars)
U-Save Auto Rental System, Inc.	11,5	118
Payless Car Rental System, Inc.	10,0	135
ACE Rent A Car	9,0	100
Rent-A-Wreck of America	5,5	37
Triangle Rent-A-Car	4,2	40
Affordable/Sensible	3,3	32

a) Représenter le nuage de points associé à ces données en utilisant le nombre de véhicules de location en service comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Estimer par les moindres carrés l'équation de la régression.

d) Pour chaque véhicule de location en service supplémentaire, estimer la variation du revenu annuel.

e) Fox Rent-A-Car possède une flotte de 11 000 voitures en service. Utiliser l'équation estimée de la régression obtenue à la question (c) pour prédire le revenu annuel de Fox Rent-A-Car.

10. Le 31 mars 2009, les actions de la société Ford Motor s'échangeaient à 2,63 dollars, le plus bas niveau depuis 26 ans. Le directoire de Ford avait alors octroyé au PDG des options sur les actions d'une valeur estimée à 16 millions de dollars. Le 26 avril 2011, le prix de l'action Ford avait augmenté à 15,58 dollars et les actions du PDG valaient alors 202,8 millions de dollars, soit un gain de 186,8 millions de dollars. Le tableau suivant indique le cours de l'action en 2009 et 2011 de 10 sociétés, ainsi que la valeur des options accordées à leur PDG en 2009 et 2011. Les augmentations en pourcentage du prix de l'action et des gains engrangés par les PDG sont également fournies (*The Wall Street Journal*, 27 avril 2011).

Société	Cours de l'action en 2009 ($)	Cours de l'action en 2011 ($)	% d'augmentation du cours de l'action	Valeur des options en 2009 (millions de dollars)	Valeur des options en 2011 (millions de dollars)	% de gain des options
Ford Motor	2,63	15,58	492	16,0	202,8	1168
Abercrombie & Fitch	23,80	70,47	196	46,2	196,1	324
Nabors Industries	9,99	32,06	221	37,2	132,2	255
Starbucks	9,99	32,06	221	12,4	75,9	512
Salesforce.com	32,73	137,61	320	7,8	67,0	759
Starwood Hotels	12,70	60,28	375	5,8	57,1	884
Caterpillar	27,96	111,94	300	4,0	47,5	1088
Oracle	18,07	34,97	94	61,9	97,5	58
Capital One	12,24	54,61	346	6,0	40,6	577
Dow Chemical	8,43	39,97	374	5,0	38,8	676

Actions PDG

a) Représenter le nuage de points associé à ces données avec le pourcentage d'augmentation du cours de l'action comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Estimer par les moindres carrés l'équation de la régression.

d) Interpréter la pente de l'équation estimée de la régression.

e) Les rémunérations des PDG semblent-elles basées sur les performances, mesurées par le cours de l'action ?

11. Pour aider les consommateurs dans leur achat d'un ordinateur portable, *Consumer Reports* attribue une note globale à chaque ordinateur testé sur la base d'une évaluation de différents éléments comme l'ergonomie, la portabilité, la performance, l'affichage et la durée de vie de la batterie. Une note élevée indique une qualité élevée. Les données suivantes (cf. fichier en ligne Ordinateur) correspondent au prix de vente moyen et à la note globale de dix modèles de 13 pouces (site Internet de *Consumer Reports*, 25 octobre 2012).

Marque et modèle	Prix ($)	Note globale
Samsung Ultrabook NP900X3C-A01US	1250	83
Apple MacBook Air MC965LL/A	1300	83
Apple MacBook Air MC231LL/A	1200	82
HP Envy 13-2050nr Spectre XT	950	79
Sony VAIO SVS13112FXB	800	77
Acer Aspire S5-391-9880 Ultrabook	1200	74
Apple MacBook Pro MD101LL/A	1200	74
Apple MacBook Pro MD313LL/A	1000	73
Dell Inspiron I13Z-6591SLV	700	67
Samsung NP535U3C-A01US	600	63

Ordinateur

a) Représenter le nuage de points associé à ces données avec le prix comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Estimer par la méthode des moindres carrés l'équation de la régression.

d) Interpréter la pente de l'équation estimée de la régression.

e) Un autre ordinateur portable testé par *Consumer Reports* est l'Acer Aspire S3-951-6646 Ultrabook ; le prix de cet ordinateur était de 700 dollars. Prédire la note globale de cet ordinateur en utilisant l'équation estimée de la régression.

12. La société Concur Technologies est une importante société de gestion des dépenses située à Redmond, dans l'État de Washington. Le *Wall Street Journal* a demandé à Concur d'examiner les données issues de 8,3 millions de rapports afin d'en tirer des enseignements sur les dépenses en matière de voyages d'affaires. Leur analyse des données a révélé que New York était la ville la plus chère, avec un tarif moyen pour une nuit d'hôtel de 198 dollars et une dépense moyenne en divertissement (incluant les repas de groupe et les tickets pour des spectacles ou d'autres évènements) de 172 dollars. En comparaison, les moyennes américaines pour ces deux catégories de dépenses s'élevaient à 89 dollars pour une chambre d'hôtel et 99 dollars pour un divertissement. Le tableau suivant (cf. fichier en ligne Voyage d'affaires) fournit le prix moyen d'une nuit d'hôtel et la dépense moyenne pour un divertissement pour un échantillon aléatoire de 9 des 25 villes américaines les plus visitées (*The Wall Street Journal*, 18 août 2011).

Ville	Tarif d'une chambre ($)	Divertissement ($)
Boston	148	161
Denver	96	105
Nashville	91	101
Nouvelle Orléans	110	142
Phoenix	90	100
San Diego	102	120
San Francisco	136	167
San José	90	140
Tampa	82	98

a) Représenter le nuage de points associé à ces données, en considérant le prix d'une chambre d'hôtel comme variable indépendante.

b) Quelle relation le nuage de points indique-t-il entre le tarif d'une chambre et celui d'un divertissement ?

c) Utiliser la méthode des moindres carrés pour estimer l'équation de la régression.

d) Interpréter la pente de l'équation estimée de la régression.

e) Le prix moyen d'une chambre à Chicago est de 128 dollars, bien supérieur à la moyenne américaine. Prédire le prix d'un divertissement à Chicago.

13. Un grand hôpital a mené une étude pour mieux cerner la relation entre le nombre de jours d'absence non autorisée des employés par an et la distance (en miles) entre leur domicile et leur lieu de travail. Un échantillon de 10 employés a été sélectionné et les données suivantes ont été collectées.

Distance au travail (miles)	Nombre de jours d'absence
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a) Représenter le nuage de points associé à ces données. Une relation linéaire semble-t-elle raisonnable ? Expliquer.

b) Utiliser la méthode des moindres carrés pour estimer l'équation de la régression qui lie la distance au travail au nombre de jours d'absence.

c) Prédire le nombre de jours d'absence pour un employé qui vit à 5 miles de l'hôpital.

14. Lorsque vous utilisez un système de navigation GPS dans votre voiture, vous entrez une destination et le système détermine une route, vous indique oralement les directions à suivre et indique votre progression au fur et à mesure du trajet. Aujourd'hui, même les systèmes les moins chers incluent des fonctionnalités que l'on ne trouvait que sur les modèles les plus chers. *Consumer Reports* a effectué une série de tests sur des GPS et leur a attribué une note globale sur la base de critères comme la facilité d'utilisation, l'information fournie, l'affichage et la durée d'autonomie de la batterie. Les données suivantes (cf. fichier en ligne GPS) indiquent le prix et la note d'un échantillon de 20 GPS ayant un écran de 4,3 pouces testés par *Consumer Reports* (site Internet de *Consumer Reports*, 17 avril 2012).

Marque et modèle	Prix ($)	Note globale
Garmin Nuvi 3490 LMT	400	82
Garmin Nuvi 3450	330	80
Garmin Nuvi 3790T	350	77
Garmin Nuvi3790 LMT	400	77
Garmin Nuvi 3750	250	74
Garmin Nuvi 2475 LT	230	74
Garmin Nuvi 2455LT	160	73
Garmin Nuvi 2370LT	270	71
Garmin Nuvi 2360 LT	250	71
Garmin Nuvi 2360 LMT	220	71
Garmin Nuvi 755 T	260	70
Motorola Motonab TN565t	200	68
Motorola Motonab TN555	200	67

Marque et modèle	Prix ($)	Note globale
Garmin Nuvi 1350T	150	65
Garmin Nuvi 1350 LMT	180	65
Garmin Nuvi 2300	160	65
Garmin Nuvi 1350	130	64
Tom Tom VAI 1435T	200	62
Garmin Nuvi 1300	140	62
Garmin Nuvi 1300LM	180	62

a) Représenter le nuage de points associé à ces données en utilisant le prix comme variable indépendante.

b) Quelle relation entre les deux variables le nuage de points indique-t-il ?

c) Utiliser la méthode des moindres carrés pour estimer l'équation de la régression.

d) Prédire la note globale d'un GPS de 4,3 pouces dont le prix serait de 200 dollars.

12.3 LE COEFFICIENT DE DÉTERMINATION

Dans le cadre des restaurants Armand, nous avons estimé l'équation de la régression $\hat{y} = 60 + 5x$ pour déterminer la relation linéaire entre la taille de la population étudiante x et les ventes trimestrielles y. À présent la question est : Dans quelle mesure l'équation estimée de la régression s'ajuste-t-elle aux données ? Dans cette section, nous montrerons que le **coefficient de détermination** fournit une mesure de l'adéquation de l'équation estimée de la régression aux données.

Pour la i^e observation, l'écart entre la valeur observée de la variable dépendante, y_i, et la valeur estimée de la variable dépendante, $\hat{y}_i$, est appelé le ***i*^e^ résidu**. Le i^e résidu représente l'erreur commise en utilisant $\hat{y}_i$ pour estimer y_i. Ainsi, pour la i^e observation, le résidu est égal à $y_i - \hat{y}_i$. La somme de ces résidus, ou erreurs, au carré correspond à la quantité minimisée par la méthode des moindres carrés. Cette quantité, aussi appelée *somme des carrés des résidus*, est notée *SCres*.

► **Somme des carrés des résidus**

$$SCres = \sum \left(y_i - \hat{y}_i\right)^2 \tag{12.8}$$

La valeur de *SCres* est une mesure de l'erreur commise en utilisant l'équation estimée de la régression pour estimer les valeurs de la variable dépendante dans l'échantillon.

Dans le tableau 12.3, nous détaillons les calculs nécessaires pour obtenir la somme des carrés des résidus dans le cadre de l'exemple des restaurants Armand. Par exemple, pour le restaurant 1, la valeur de la variable indépendante et celle de la variable dépendante sont respectivement 2 et 58. En utilisant l'équation estimée de la

régression, nous trouvons que la valeur estimée des ventes trimestrielles du restaurant 1 est égale à 70 ($\hat{y}_1 = 60 + 5(2) = 70$). Ainsi, l'erreur commise en utilisant $\hat{y}_1$ pour estimer y_1 pour le restaurant 1 est égale à $y_1 - \hat{y}_1 = 58 - 70 = -12$. L'erreur élevée au carré, $(-12)^2 = 144$, est notée dans la dernière colonne du tableau 12.3. Après avoir calculé et élevé au carré les résidus pour chaque restaurant de l'échantillon, la somme nous donne une *SCres* égale à 1 530. Ainsi, cette quantité mesure l'erreur commise en utilisant l'équation estimée de la régression $\hat{y} = 60 + 5x$ pour prévoir les ventes trimestrielles.

Supposons maintenant que nous voulions estimer les ventes trimestrielles sans connaître la taille de la population étudiante. Dans ce cas, nous utilisons la moyenne d'échantillon comme estimation des ventes trimestrielles d'un restaurant donné. D'après le tableau 12.2, $\sum y_i = 1\,300$. Par conséquent, la valeur moyenne des ventes trimestrielles pour l'échantillon des 10 restaurants Armand est $\bar{y} = \sum y_i / n = 1\,300/10 = 130$. Dans le tableau 12.4, nous indiquons la valeur de la somme des écarts au carré obtenue en utilisant la moyenne d'échantillon $\bar{y} = 130$ pour estimer les ventes trimestrielles pour chaque restaurant de l'échantillon. Pour le i^e restaurant de l'échantillon, l'écart $y_i - \bar{y}$ fournit une mesure de l'erreur commise en utilisant $\bar{y}$ pour estimer les ventes. La somme des carrés correspondante, appelée *somme des carrés totale*, est notée *SCT*.

► **Somme des carrés totale**

$$SCT = \sum (y_i - \bar{y})^2 \tag{12.9}$$

La somme en bas de la dernière colonne du tableau 12.4 correspond à la somme des carrés totale pour les restaurants Armand ; elle est égale à 15 730.

Tableau 12.3 *Calculs de SCres pour les restaurants Armand*

Restaurant *i*	x_i = Population étudiante (en milliers)	y_i = Ventes trimestrielles (en milliers de dollars)	Ventes prévues $\hat{y}_i = 60 + 5x_i$	Erreur $y_i - \hat{y}_i$	Erreur au carré $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SCres = 1530

Tableau 12.4 ***Calculs de la somme des carrés totale pour les restaurants Armand***

Restaurant *i*	x_i = Population étudiante (en milliers)	y_i = Ventes trimestrielles (en milliers de dollars)	Écart $y_i - \bar{y}$	Écart au carré $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				SCT = 15 730

La figure 12.5 représente la droite de régression estimée $\hat{y} = 60 + 5x$ et la droite correspondant à $\bar{y} = 130$. Notez que les points sont plus regroupés autour de la droite de régression estimée qu'autour de la droite $\bar{y} = 130$. Par exemple, pour le 10e restaurant de l'échantillon, l'erreur est beaucoup plus importante lorsqu'on utilise $\bar{y} = 130$ pour estimer y_{10} que lorsqu'on utilise $\hat{y}_{10} = 60 + 5(26) = 190$. Nous pouvons interpréter *SCT* comme une mesure de l'ajustement des observations autour de la droite $\bar{y}$ et *SCres* comme une mesure de l'ajustement des observations autour de la droite $\hat{y}$.

Avec $SCT = 15\,730$ et $SCres = 1\,530$, la droite de régression estimée est mieux ajustée aux données que la droite $y = \bar{y}$.

Pour déterminer dans quelle mesure les valeurs $\hat{y}$ de la droite de la régression estimée dévient de $\bar{y}$, une autre somme des carrés est calculée. Cette somme des carrées, appelée *somme des carrés de la régression*, est notée *SCreg*.

► **Somme des carrés de la régression**

$$SCreg = \sum (\hat{y}_i - \bar{y})^2 \qquad (12.10)$$

De par les précédentes discussions, on s'attend à ce que *SCT*, *SCreg* et *SCres* soient liées. De fait, la relation entre ces trois sommes des carrés fournit l'un des plus importants résultats en statistique.

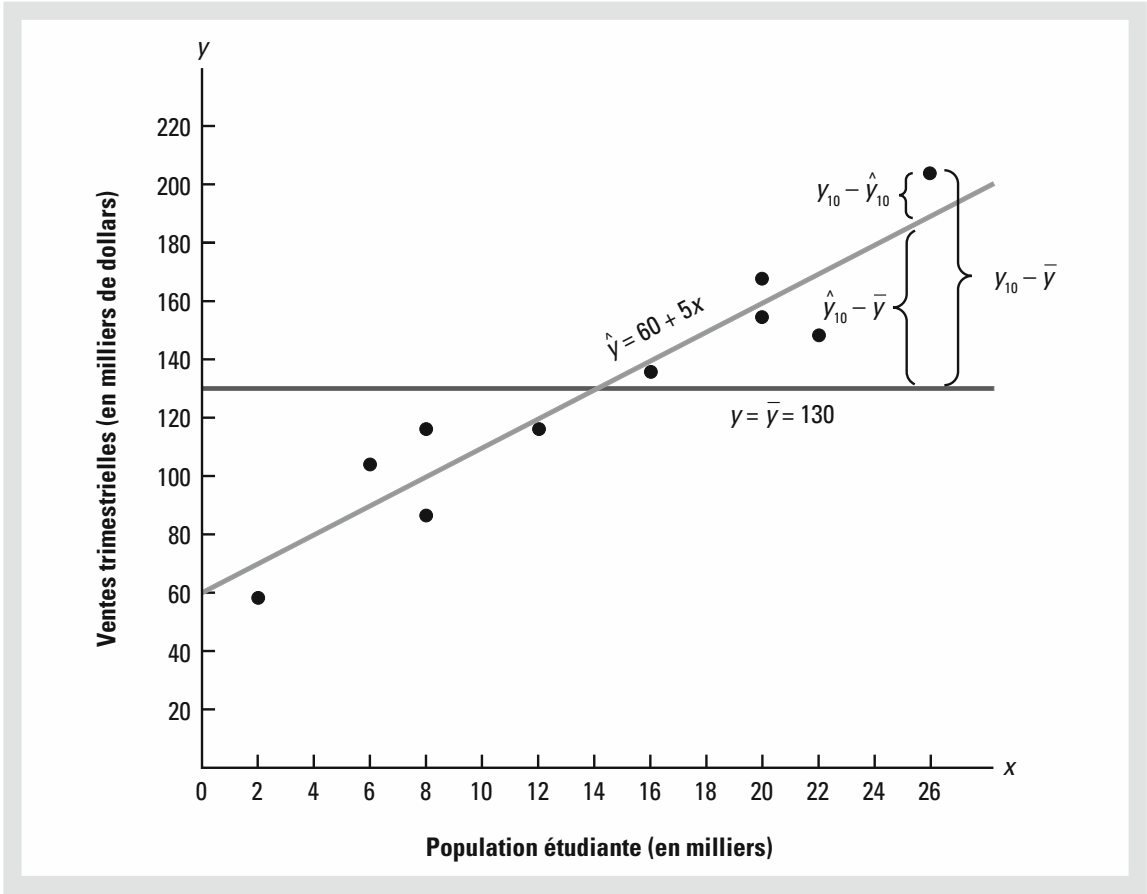

Figure 12.5 *Écarts par rapport à la droite de régression estimée et à la droite $y = \bar{y}$ dans le cadre des restaurants Armand*

► Relation entre SCT, SCreg et SCres

$$SCT = SCreg + SCres \tag{12.11}$$

où

SCT correspond à la somme des carrés totale
SCreg correspond à la somme des carrés de la régression
SCres correspond à la somme des carrés des résidus

SCreg peut être considérée comme la partie expliquée de SCT, et SCres comme la partie inexpliquée de SCT.

L'équation (12.11) indique que la somme des carrés totale peut être divisée en deux parties, la somme des carrés de la régression et la somme des carrés des résidus. Par conséquent, si les valeurs de ces deux sommes des carrés sont connues, la troisième somme des carrés peut être facilement calculée. Par exemple, dans le cadre de l'exemple des restaurants Armand, nous savons déjà que *SCres* est égale à 1 530 et *SCT* est égale à 15 730. La somme des carrés de la régression est donc égale à

$$SCreg = SCT - SCres = 15\,730 - 1\,530 = 14\,200$$

Voyons maintenant comment ces trois sommes, *SCT*, *SCreg* et *SCres*, peuvent fournir une mesure de l'adéquation de l'équation estimée de la régression. L'équation estimée de la régression s'ajusterait parfaitement aux données si toutes les valeurs de la variable dépendante y_i se trouvaient sur la droite de régression estimée. Dans ce cas, $y_i - \hat{y}_i$ serait nul pour chaque observation, et par conséquent *SCres* serait égale à zéro. Puisque $SCT = SCreg + SCres$, un parfait ajustement implique que *SCreg* soit égal à *SCT* et que le ratio $\left(SCreg/SCT\right)$ soit égal à un. Plus l'ajustement est imparfait, plus la valeur de *SCres* sera grande. Or, d'après l'équation (12.11), $SCres = SCT - SCreg$. Par conséquent, la plus grande valeur de *SCres* (et l'ajustement le plus imparfait) intervient lorsque $SCreg = 0$ et $SCres = SCT$.

Le ratio $\left(SCreg/SCT\right)$, compris entre zéro et un, est utilisé pour évaluer l'adéquation de l'équation estimée de la régression aux données. Ce ratio est appelé *coefficient de détermination* et est noté r^2.

► **Coefficient de détermination**

$$r^2 = \frac{SCreg}{SCT} \tag{12.12}$$

Dans l'exemple des restaurants Armand, le coefficient de détermination est égal à

$$r^2 = \frac{SCreg}{SCT} = \frac{14\ 200}{15\ 730} = 0,9027$$

Lorsqu'on exprime le coefficient de détermination en termes de pourcentage, on peut l'interpréter comme le pourcentage de la somme des carrés totale expliquée par l'équation estimée de la régression. Dans le cadre de l'exemple des restaurants Armand, nous concluons que 90,27 % de la somme des carrés totale peut être expliquée en utilisant l'équation estimée de la régression $\hat{y} = 60 + 5x$ pour prévoir les ventes trimestrielles. En d'autres termes, 90,27 % de la variation des ventes trimestrielles peut s'expliquer par la relation linéaire entre la taille de la population étudiante et les ventes trimestrielles. Une telle adéquation de l'équation estimée de la régression est satisfaisante.

12.3.1 Coefficient de corrélation

Au chapitre 3, nous avons introduit le **coefficient de corrélation** en tant que mesure descriptive de la robustesse de l'association linéaire entre deux variables, *x* et *y*. Le coefficient de corrélation est toujours compris entre –1 et +1. Une valeur égale à +1 indique que les deux variables *x* et *y* sont parfaitement liées de façon positive. En d'autres termes, tous les points sont sur une droite de pente positive. Une valeur égale à –1 indique que *x* et *y* sont parfaitement liés de façon négative, tous les points étant sur une droite de pente négative. Des valeurs proches de zéro indiquent que *x* et *y* ne sont pas linéairement liés.

Dans la section 3.5, nous avons présenté la formule de calcul du coefficient de corrélation d'un échantillon. Si une analyse de la régression a déjà été faite et si le coefficient de détermination r^2 a déjà été calculé, le coefficient de corrélation de l'échantillon peut être calculé de la façon suivante :

► **Coefficient de corrélation d'un échantillon**

$$r_{xy} = (\text{signe de } b_1)\sqrt{\text{Coefficient de détermination}}$$
$$= (\text{signe de } b_1)\sqrt{r^2} \qquad (12.13)$$

où b_1 correspond à la pente de l'équation estimée de la régression $\hat{y} = b_0 + b_1 x$.

Le signe du coefficient de corrélation d'un échantillon est positif si l'équation estimée de la régression est de pente positive ($b_1 > 0$) et négatif si l'équation estimée de la régression est de pente négative ($b_1 < 0$).

Pour l'exemple des restaurants Armand, le coefficient de détermination correspondant à l'équation estimée de la régression $\hat{y} = 60 + 5x$ est égal à 0,9027. Puisque la pente de l'équation estimée de la régression est positive, la formule (12.13) indique que le coefficient de corrélation est égal à $+\sqrt{0,9027} = +0,9501$. Avec un coefficient de corrélation égal à $r_{xy} = +0,9501$, on peut conclure qu'il existe une forte relation linéaire positive entre x et y.

Dans le cas d'une relation linéaire entre deux variables, à la fois le coefficient de détermination et le coefficient de corrélation fournissent une mesure de la robustesse de la relation. Le coefficient de détermination fournit une mesure entre zéro et un, alors que le coefficient de corrélation fournit une mesure entre –1 et +1. Alors que le coefficient de corrélation est restreint à des relations linéaires entre deux variables, le coefficient de détermination peut être utilisé dans le cas de relations non-linéaires et de relations comprenant plus de deux variables indépendantes. Le coefficient de détermination a donc un champ d'application plus large.

REMARQUES

1. En estimant l'équation de la régression par les moindres carrés et en calculant le coefficient de détermination, nous n'avons fait aucune hypothèse probabiliste sur le terme d'erreur ε et aucun test statistique relatif à la significativité de la relation entre x et y. Plus la valeur du coefficient de détermination est élevée, meilleure est l'adéquation de la droite des moindres carrés aux données ; c'est-à-dire, les observations sont bien regroupées autour de la droite des moindres carrés. Mais, en utilisant le coefficient de détermination seul, nous ne pouvons pas dire si la relation entre x et y est statistiquement significative. Une telle conclusion doit être fondée sur des considérations qui impliquent la taille de l'échantillon et les propriétés des distributions d'échantillonnage des estimateurs des moindres carrés.

2. D'un point de vue empirique, en sciences sociales, des valeurs du coefficient de détermination aussi petites que 0,25 sont souvent considérées comme utiles. Pour des données en sciences physiques ou naturelles, on trouve souvent des valeurs supérieures ou égales à 0,60 ; en fait, dans certains cas, on peut trouver des valeurs supérieures à 0,90. Dans les applications commerciales, les valeurs du coefficient de détermination varient beaucoup, en fonction des caractéristiques particulières de chaque exemple.

EXERCICES

Méthode

15. Reprendre les données de l'exercice 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

L'équation estimée de la régression associée à ces données est $\hat{y} = 0,20 + 2,60x$.

a) Calculer *SCres*, *SCT* et *SCreg* en utilisant les expressions (12.8), (12.9) et (12.10).

b) Calculer le coefficient de détermination r^2. Commenter l'adéquation de la régression aux données.

c) Calculer le coefficient de corrélation de l'échantillon.

16. Reprendre les données de l'exercice 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

L'équation estimée de la régression associée à ces données est $\hat{y} = 68 - 3x$.

a) Calculer *SCres*, *SCT* et *SCreg*.

b) Calculer le coefficient de détermination r^2. Commenter l'adéquation de la régression aux données.

c) Calculer le coefficient de corrélation de l'échantillon.

17. Reprendre les données de l'exercice 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

L'équation estimée de la régression, associée à ces données, est $\hat{y} = 7,6 + 0,9x$. Quel est le pourcentage de la somme des carrés totale attribuable à l'équation estimée de la régression ? Quelle est la valeur du coefficient de corrélation de l'échantillon ?

Applications

18. Les données suivantes fournissent la marque, le prix (en dollars) et la note globale de six écouteurs stéréo testés par *Consumer Reports* (site Internet de *Consumer Reports*, 5 mars 2012). La note globale est basée sur la qualité sonore et l'efficacité des écouteurs à réduire le bruit ambiant. Les notes vont de 0 (la plus faible) à 100 (la plus élevée). L'équation estimée de la régression associée à ces données est $\hat{y} = 23,194 + 0,318x$ avec x le prix et y la note globale.

Marque	Prix ($)	Note
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O'Neill	70	56
Denon	70	40
JVC	35	26

a) Calculer *SCT*, *SCreg* et *SCres*.

b) Calculer le coefficient de détermination r^2. Commenter l'adéquation de la régression aux données.

c) Quelle est la valeur du coefficient de corrélation de l'échantillon ?

19. Dans l'exercice 7, un responsable des ventes a collecté les données suivantes (cf. fichier en ligne Ventes) sur les ventes annuelles (x) et les années d'expérience (y). L'équation estimée de la régression pour ces données est $\hat{y} = 80 + 4x$.

Vendeur	Années d'expérience	Ventes annuelles (milliers de dollars)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a) Calculer *SCT*, *SCreg* et *SCres*.

b) Calculer le coefficient de détermination r^2. Commenter l'adéquation de la régression aux données.

c) Quelle est la valeur du coefficient de corrélation de l'échantillon ?

20. *Bicycling*, le magazine de cyclisme leader sur le marché mondial, teste des centaines de vélos toute l'année. La rubrique « Rade-Race » du magazine contient des tests de vélos utilisés principalement pour les courses. L'un des plus importants facteurs de choix d'un vélo pour une course est son poids. Les données suivantes (cf. fichier en ligne Vélos de course) correspondent aux poids (en livres) et au prix (en dollars) de 10 vélos de course testés par le magazine (site Internet de *Bicycling*, 8 mars 2012).

Marque	Poids	Prix ($)
FELT F5	17,8	2 100
PINARELLO Paris	16,1	6 250
ORBEA Orca GDR	14,9	8 370
EDDY MERCKX EMX-7	15,9	6 200
BH RC1 Ultegra	17,2	4 000
BH Ultralight 386	13,1	8 600
CERVELO S5 Team	16,2	6 000
GIANT TCR Advanced 2	17,1	2 580
WILIER TRIESTINA Gran Turismo	17,6	3 400
SPECIALIZED S-Works Amira SL4	14,1	8 000

a) Utiliser ces données pour estimer l'équation de la régression qui pourrait être utilisée pour estimer le prix d'un vélo en fonction de son poids.

b) Calculer le coefficient de détermination. L'équation de la régression estimée est-elle bien ajustée aux données ?

c) Prédire le prix d'un vélo qui pèse 15 livres.

21. Une application importante de l'analyse de la régression en comptabilité concerne l'estimation des coûts. En collectant des données sur les quantités et sur les coûts et en utilisant la méthode des moindres carrés pour estimer l'équation de la relation entre ces deux variables, un comptable peut estimer le coût associé à un volume de production particulier. Considérez l'échantillon suivant de quantités produites et de coûts de production.

Volume de la production (unités)	Coût total ($)
400	4 000
450	5 000
550	5 400
600	5 900
700	6 400
750	7 000

a) Utiliser ces données pour estimer l'équation de la régression qui peut servir à prévoir le coût total d'un volume de production donné.

b) Quel est le coût variable par unité produite ?

c) Calculer le coefficient de détermination. Quel est le pourcentage de la variation du coût total expliqué par le volume produit ?

d) La société prévoit de produire 500 unités le mois prochain. Quel est le coût estimé de cette opération ?

22. Référez-vous à l'exercice 9, dans lequel les données suivantes ont été utilisées pour identifier la relation entre le nombre de véhicules en service (en milliers) et le revenu annuel (en millions de dollars) de six petites sociétés de location de voitures (site Internet de *Auto Rental News*, 7 août 2012).

Société	Véhicules (milliers)	Revenu (millions de dollars)
U-Save Auto Rental System, Inc.	11,5	118
Payless Car Rental System, Inc.	10,0	135
ACE Rent A Car	9,0	100
Rent-A-Wreck of America	5,5	37
Triangle Rent-A-Car	4,2	40
Affordable/Sensible	3,3	32

Avec x le nombre de véhicules en service (en milliers) et y le revenu annuel (en millions de dollars), l'équation estimée de la régression est $\hat{y} = -17,005 + 12,966x$. Pour ces données, $SCres = 1\,043,03$.

a) Calculer le coefficient de détermination.

b) L'équation estimée de la régression est-elle bien ajustée aux données ? Expliquer.

c) Quel est le coefficient de corrélation de l'échantillon ? Reflète-t-il une relation forte ou faible entre le prix et la note ?

12.4 LES HYPOTHÈSES DU MODÈLE

Dans le cadre de l'analyse de la régression linéaire simple, nous avons fait une hypothèse sur le modèle approprié pour estimer la relation entre la variable dépendante et la variable indépendante. Le modèle de la régression estimé est

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Ensuite, nous avons utilisé la méthode des moindres carrés pour estimer les paramètres du modèle β_0 et β_1. L'équation de la régression estimée qui en résulte s'écrit

$$\hat{y} = b_0 + b_1 x$$

Nous avons vu que la valeur du coefficient de détermination est une mesure de l'adéquation de l'équation estimée de la régression. Cependant, même avec une valeur élevée de r^2, l'équation estimée de la régression ne devrait pas être utilisée tant qu'une analyse plus approfondie de la robustesse du modèle n'a pas été faite. Une étape importante dans la détermination de la robustesse du modèle consiste à effectuer un test de signification de la relation. Les tests de signification dans l'analyse de la régression sont basés sur les hypothèses suivantes concernant le terme d'erreur ε.

► **Hypothèses sur le terme d'erreur ε dans le modèle de la régression** $y = \beta_0 + \beta_1 x + \varepsilon$

1. Le terme d'erreur ε est une variable aléatoire de moyenne nulle ; c'est-à-dire, $E(\varepsilon) = 0$.

Conséquences : Puisque β_0 et β_1 sont des constantes, $E(\beta_0) = \beta_0$ et $E(\beta_1) = \beta_1$; ainsi, pour une valeur donnée de x, l'espérance mathématique de y est égale à

$$E(y) = \beta_0 + \beta_1 x \tag{12.14}$$

Comme indiqué précédemment, l'expression (12.14) correspond à l'équation de la régression.

2. La variance de ε, notée σ^2, est la même pour toutes les valeurs de x.
Conséquences : La variance de y pour une valeur donnée de x est égale à σ^2 et est la même pour toutes les valeurs de x.

3. Les valeurs de ε sont indépendantes.
Conséquences : La valeur de ε associée à une valeur particulière de x n'est pas liée à la valeur de ε associée à une autre valeur de x ; ainsi, la valeur de y associée à une valeur particulière de x n'est pas liée à la valeur de y associée à une autre valeur de x.

4. Le terme d'erreur ε est une variable aléatoire normalement distribuée.
Conséquences : Puisque y est une fonction linéaire de ε, y est également une variable aléatoire normalement distribuée.

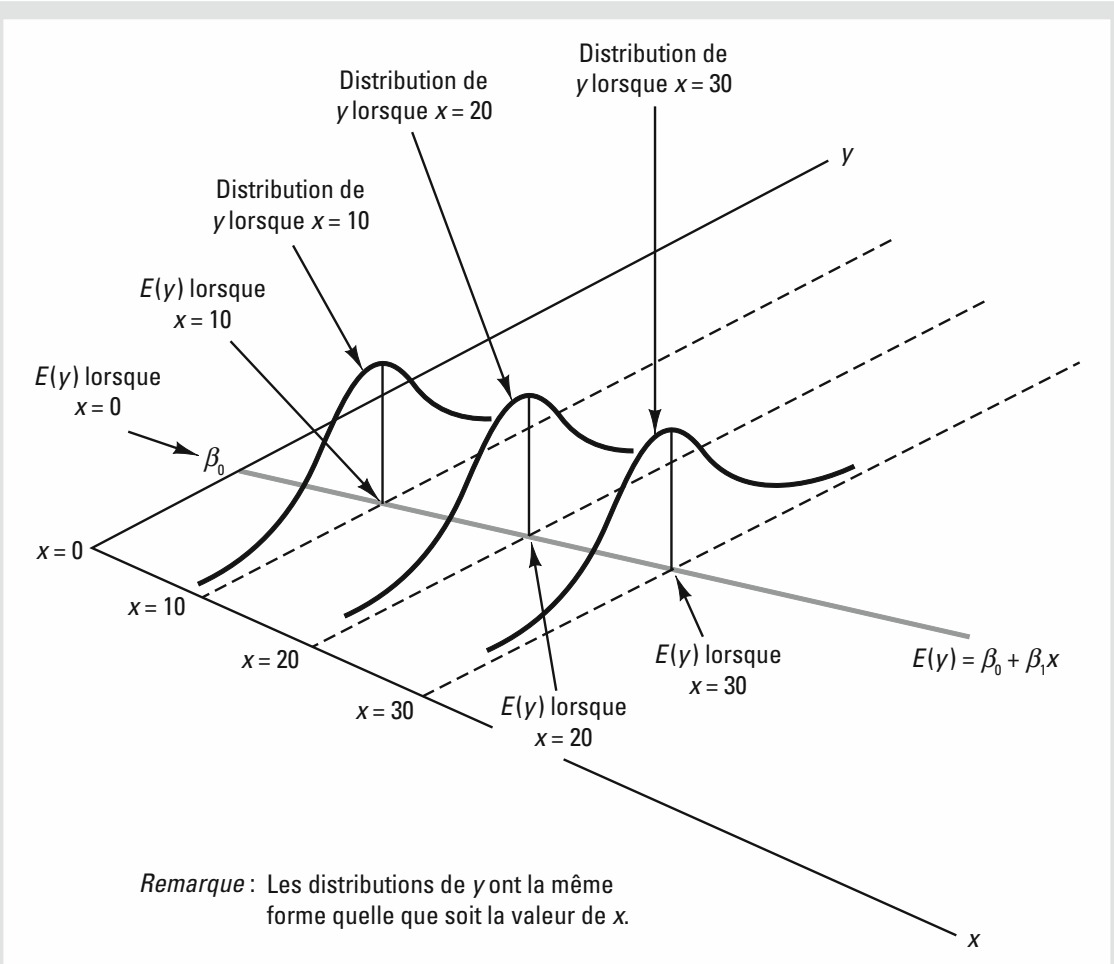

Figure 12.6 ***Hypothèses du modèle de régression***

La figure 12.6 est une illustration des hypothèses du modèle et de leurs conséquences ; notez que dans cette interprétation graphique, la valeur de $E(y)$ varie selon la valeur de x considérée. Cependant, sans tenir compte de la valeur de x, la distribution de probabilité de ε et donc la distribution de probabilité de y sont normales, chacune avec la même variance. La valeur spécifique du terme d'erreur ε dépend du fait que la valeur réelle de y soit supérieure ou inférieure à $E(y)$.

À ce point de la discussion, nous devons garder en mémoire le fait que nous avons également fait une hypothèse sur la forme de la relation entre x et y. En effet, nous avons supposé que la relation entre ces deux variables est linéaire, plus précisément de la forme $\beta_0 + \beta_1 x$. Nous ne devons pas oublier que d'autres modèles, par exemple $y = \beta_0 + \beta_1 x^2 + \varepsilon$, peuvent être plus appropriés pour décrire la relation qui lie x et y.

12.5 LES TESTS DE SIGNIFICATION

Dans une équation de régression linéaire simple, la moyenne ou l'espérance mathématique de y est une fonction linéaire de x : $E(y) = \beta_0 + \beta_1 x$. Si la valeur de β_1 est égale à zéro, $E(y) = \beta_0 + (0)x = \beta_0$. Dans ce cas, la moyenne de y ne dépend pas de la valeur de x ; nous pouvons donc en conclure que x et y ne sont pas linéairement liés. Par contre, si β_1 n'est pas égal à zéro, nous pouvons en conclure que les deux variables sont liées. Ainsi, pour tester si la relation est significative, nous devons effectuer un test d'hypothèses pour déterminer si β_1 est égal à zéro. Deux tests sont habituellement utilisés. Les deux requièrent une estimation de σ^2, la variance de ε.

12.5.1 Estimation de σ^2

À partir des hypothèses du modèle de régression, nous pouvons conclure que σ^2, la variance de ε, représente également la variance de y le long de la droite de régression. Rappelons que les écarts de y par rapport à la droite de régression estimée sont appelés les résidus. Ainsi, *SCres*, la somme des carrés des résidus, est une mesure de la variabilité de y le long de la droite de régression estimée. La **moyenne des carrés des résidus** (*MCres*) fournit une estimation de σ^2 ; cette moyenne des carrés des résidus correspond à la somme des carrés des résidus divisée par le nombre de ses degrés de liberté.

Avec $\hat{y}_i = b_0 + b_1 x_i$, la somme des carrés des résidus s'écrit :

$$SCres = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

À chaque somme des carrés est associé un nombre, appelé degrés de liberté. Des statisticiens ont démontré que la somme des carrés des résidus a $n - 2$ degrés de liberté, puisque deux paramètres (β_0 et β_1) doivent être estimés pour calculer cette somme des carrés des résidus. Ainsi, la moyenne des carrés des résidus est calculée en divisant *SCres* par $n - 2$. *MCres* fournit une estimation sans biais de σ^2. Puisque la valeur de la moyenne des carrés des résidus fournit une estimation de σ^2, la notation s^2 est aussi utilisée.

► **Moyenne des carrés des résidus (estimation de σ^2)**

$$s^2 = MCres = \frac{SCres}{n-2} \tag{12.15}$$

Dans la section 12.3, nous avons montré que la somme des carrés des résidus, dans le cadre de l'exemple des restaurants Armand, est égale à 1 530 ; par conséquent,

$$s^2 = MCres = \frac{1\,530}{8} = 191,25$$

fournit une estimation sans biais de σ^2.

Pour estimer σ, nous prenons la racine carrée de s^2. La valeur correspondante, s, est appelée **erreur type de l'estimation**.

► **Erreur type de l'estimation**

$$s = \sqrt{MCres} = \sqrt{\frac{SCres}{n-2}} \tag{12.16}$$

Dans l'exemple des restaurants Armand, $s = \sqrt{MCres} = \sqrt{191,25} = 13,829$. Dans la discussion qui suit, nous utiliserons l'erreur type de l'estimation pour effectuer des tests de signification de la relation entre x et y.

12.5.2 Le test t de Student

Le modèle de régression linéaire simple s'écrit $y = \beta_0 + \beta_1 x + \varepsilon$. Si x et y sont linéairement liés, nous devons avoir $\beta_1 \neq 0$. Le but du test de Student est d'utiliser les données de l'échantillon pour conclure si $\beta_1 \neq 0$. On teste les hypothèses suivantes concernant β_1 :

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Si on rejette H_0, on en conclut que $\beta_1 \neq 0$ et qu'une relation statistiquement significative existe entre les deux variables. Cependant, si on ne peut pas rejeter H_0, les preuves statistiques sont insuffisantes pour conclure qu'une relation significative existe. Les propriétés d'échantillonnage de b_1, l'estimateur des moindres carrés de β_1, fournissent les bases du test d'hypothèses.

Tout d'abord, considérons ce qui se serait passé si nous avions utilisé un autre échantillon pour effectuer la même analyse de la régression. Par exemple, supposons que nous ayons collecté des données sur les ventes trimestrielles d'un échantillon de dix autres restaurants Armand. Une analyse de la régression de ce nouvel échantillon devrait fournir une équation similaire à celle obtenue précédemment, $\hat{y} = 60 + 5x$. Cependant, il est très peu probable que nous obtenions exactement la même équation avec une ordonnée à l'origine égale à 60 et une pente égale à 5. En fait, b_0 et b_1, les estimateurs des moindres carrés, sont des statistiques d'échantillon qui ont leur propre distribution d'échantillonnage. Les propriétés de la distribution d'échantillonnage de b_1 sont décrites ci-dessous.

► **Distribution d'échantillonnage de b_1**

Espérance mathématique :

$$E(b_1) = \beta_1$$

Écart type :

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad (12.17)$$

Forme de la distribution :

Normale

Notez que l'espérance mathématique de b_1 est égale à β_1 ; b_1 est donc un estimateur sans biais de β_1.

Puisque que nous ne connaissons pas la valeur de σ, nous estimons σ_{b_1} en remplaçant σ par s dans l'équation (12.17). Nous obtenons ainsi l'estimateur suivant de σ_{b_1}.

► **Écart type estimé de b_1**

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad (12.18)$$

L'écart type de b_1 est également appelé erreur type de b_1. Ainsi, s_{b_1} fournit une estimation de l'erreur type de b_1.

Dans l'exemple des restaurants Armand, $s = 13{,}829$. Par conséquent, en utilisant les informations contenues dans le tableau 12.2, à savoir que $\sum (x_i - \bar{x})^2 = 568$, nous obtenons

$$s_{b_1} = \frac{13{,}829}{\sqrt{568}} = 0{,}5803$$

comme écart type estimé de b_1.

Le test de signification de Student est basé sur le fait que la statistique de test

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

suit une loi de Student à $n - 2$ degrés de liberté. Si l'hypothèse nulle est vraie, alors $\beta_1 = 0$ et $t = b_1 / s_{b_1}$.

Appliquons ce test de signification à l'exemple des restaurants Armand au seuil de signification $\alpha = 0{,}01$. La statistique de test est égale à

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0{,}5803} = 8{,}62$$

D'après la table de la distribution de Student (table 2 de l'annexe D), avec $n - 2 = 10 - 2 = 8$ degrés de liberté, $t = 3,355$ fournit une aire égale à 0,005 dans la queue supérieure de la distribution. Ainsi, l'aire dans la queue supérieure de la distribution de Student correspondant à la statistique de test $t = 8,62$ doit être inférieure à 0,005. Puisque le test est bilatéral, nous multiplions cette valeur par deux pour conclure que la valeur p associée à $t = 8,62$ est inférieure à 0,01. Minitab ou Excel indiquent que la valeur p est égale à 0,000. Puisque la valeur p est inférieure à $\alpha = 0,01$, nous rejetons H_0 et concluons que β_1 n'est pas égal à zéro. Les preuves statistiques sont suffisantes pour conclure qu'il existe une relation significative entre la population étudiante et les ventes trimestrielles. Un résumé du test de signification de Student dans le cadre d'une régression linéaire simple suit.

Les annexes 12.1 et 12.2 montrent comment utiliser Minitab et Excel pour calculer la valeur *p*.

► **Test de signification de Student dans le cadre d'une régression linéaire simple**

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

► **Statistique de test**

$$t = \frac{b_1}{s_{b_1}} \tag{12.19}$$

► **Règle de rejet**

Approche par la valeur *p* : Rejet de H_0 si la valeur $p \leq \alpha$

Approche par la valeur critique : Rejet de H_0 si $t \leq -t_{\alpha/2}$ ou si $t \geq t_{\alpha/2}$ où $t_{\alpha/2}$ est basé sur la distribution de Student à $n - 2$ degrés de liberté.

12.5.3 Intervalle de confiance pour β_1

La forme de l'intervalle de confiance pour β_1 est :

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

L'estimateur ponctuel est b_1 et la marge d'erreur est $t_{\alpha/2} s_{b_1}$. Le coefficient de confiance associé à cet intervalle est $1 - \alpha$ et $t_{\alpha/2}$ correspond à la valeur t fournissant une aire égale à $\alpha/2$ dans la queue supérieure de la distribution de Student à $n - 2$ degrés de liberté. Par exemple, supposez que nous voulions construire un intervalle de confiance à 99 % pour β_1 dans le cadre des restaurants Armand. D'après la table 2 de l'annexe B, la valeur t associée à $\alpha = 0,01$ et $n - 2 = 10 - 2 = 8$ degrés de liberté est égale à $t_{0,005} = 3,355$. Ainsi, l'intervalle de confiance à 99 % pour β_1 est

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3,355(0,5803) = 5 \pm 1,95$$

soit de 3,05 à 6,95.

En utilisant le test de signification de Student, les hypothèses testées étaient

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Au seuil de signification $\alpha = 0{,}01$, l'intervalle de confiance à 99 % nous offre une solution alternative pour effectuer le test d'hypothèses dans le cadre des restaurants Armand. Puisque 0, la valeur hypothétique de β_1, n'appartient pas à l'intervalle de confiance (de 3,05 à 6,95), nous pouvons rejeter H_0 et conclure qu'une relation statistiquement significative existe entre la taille de la population étudiante et les ventes trimestrielles. En général, un intervalle de confiance peut être utilisé pour tester tous les jeux d'hypothèses bilatérales concernant β_1. Si la valeur hypothétique de β_1 appartient à l'intervalle de confiance, ne pas rejeter H_0. Sinon, rejeter H_0.

12.5.4 Le test F de Fisher

Un test de Fisher, basé sur la distribution de Fisher, peut également être utilisé pour tester si une relation est significative. Avec une seule variable indépendante, le test de Fisher conduit à la même conclusion que le test de Student ; c'est-à-dire, si le test de Student conclut que $\beta_1 \neq 0$ et qu'il existe une relation significative entre les variables, le test de Fisher conclura également à l'existence d'une relation significative. Par contre, avec plus d'une variable indépendante, seul le test de Fisher peut être utilisé pour tester la signification globale d'une relation.

La logique qui sous-tend l'utilisation du test de Fisher pour déterminer si la relation est statistiquement significative, est basée sur la construction de deux estimations indépendantes de σ^2. Nous avons vu que la moyenne des carrés des résidus, *MCres*, fournit une estimation de σ^2. Si l'hypothèse nulle $H_0 : \beta_1 = 0$ est vraie, la somme des carrés de la régression, *SCreg*, divisée par le nombre de ses degrés de liberté, fournit une autre estimation indépendante de σ^2. Cette estimation est appelée *moyenne des carrés de la régression* et est notée *MCreg*. De façon générale,

$$MCreg = \frac{SCreg}{\text{Nombre de degrés de liberté}}$$

Pour les modèles de régression que nous considérons ici, le nombre de degrés de liberté est toujours égal au nombre de variables indépendantes ; ainsi,

$$MCreg = \frac{SCreg}{\text{Nombre de variables indépendantes}} \qquad (12.20)$$

Puisque nous ne considérons dans ce chapitre que les modèles de régression à une seule variable indépendante, $MCreg = SCreg/1 = SCreg$. Dans le cadre de l'exemple des restaurants Armand, $MCreg = SCreg = 14\,200$.

Si l'hypothèse nulle ($H_0 : \beta_1 = 0$) est vraie, *MCreg* et *MCres* sont deux estimations indépendantes de σ^2 et la distribution d'échantillonnage de $MCreg/MCres$ suit une loi de Fisher avec un degré de liberté au numérateur et $n - 2$ degrés de liberté au

dénominateur. Par conséquent, lorsque $\beta_1 = 0$, la valeur de $MCreg/MCres$ doit être proche de un. Par contre, si l'hypothèse nulle est fausse ($\beta_1 \neq 0$), $MCreg$ surestime σ^2 et la valeur de $MCreg/MCres$ augmente ; ainsi, des valeurs élevées de $MCreg/MCres$ conduisent au rejet de H_0 et à la conclusion selon laquelle la relation entre x et y est statistiquement significative.

Appliquons le test de Fisher à l'exemple des restaurants Armand. La statistique de test est

$$F = \frac{MCreg}{MCres} = \frac{14\ 200}{191,25} = 74,25$$

D'après la table 4 de l'annexe B, avec un degré de liberté au numérateur et 8 degrés de liberté au dénominateur, la valeur $F = 11,26$ fournit une aire égale à 0,01 dans la queue supérieure de la distribution de Fisher. Ainsi, l'aire dans la queue supérieure de la distribution de Fisher correspondant à la statistique de test $F = 74,25$ doit être inférieure à 0,01. Nous concluons par conséquent que la valeur p associée à cette statistique de test est inférieure à 0,01. Minitab ou Excel indiquent que la valeur p est égale à 0,000. Puisque la valeur p est inférieure à $\alpha = 0,01$, nous rejetons H_0 et concluons que β_1 n'est pas égal à zéro. Les preuves statistiques sont suffisantes pour conclure qu'il existe une relation significative entre la population étudiante et les ventes trimestrielles. Un résumé du test de Fisher dans le cadre d'une régression linéaire simple suit.

Le test de Fisher et le test de Student fournissent des résultats identiques dans le cadre d'une régression linéaire simple.

► **Test de signification de Fisher**

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

► **Statistique de test**

$$F = \frac{MCreg}{MCres} \qquad (12.21)$$

► **Règle de rejet**

Approche par la valeur p : Rejet de H_0 si la valeur $p \leq \alpha$

Approche par la valeur critique : Rejet de H_0 si $F \geq F_\alpha$

où F_α est basé sur la distribution de Fisher à un degré de liberté au numérateur et $n - 2$ degrés de liberté au dénominateur.

Si H_0 est fausse, MCres reste un estimateur sans biais de σ^2 et MCreg surestime σ^2.
Si H_0 est vraie, à la fois MCres et MCreg sont des estimateurs sans biais de σ^2 ; dans ce cas, la valeur de $MCreg/MCres$ sera proche de un.

Dans le chapitre 10, nous avons discuté de l'analyse de la variance (ANOVA) et montré comment utiliser un **tableau ANOVA** pour résumer les calculs de l'analyse de la variance. Un tableau ANOVA similaire peut être utilisé pour résumer les résultats du test de signification de Fisher. Le tableau 12.5 présente la forme générale d'un tableau ANOVA dans le cadre d'une étude de la régression impliquant une seule variable indépendante. Le tableau 12.6 présente le tableau ANOVA avec les calculs du test de Fisher effectué dans le cadre de l'exemple des restaurants Armand. Régression, résidus et totale sont les trois sources de variation, avec *SCreg*, *SCres* et *SCT* apparaissant dans la deuxième colonne. Les degrés de liberté, 1 pour Régression, $n - 2$ pour Résidus et $n - 1$ pour Totale, sont notés dans la troisième colonne. La quatrième colonne contient les valeurs de *MCreg* et *MCres* et la cinquième colonne, la valeur de $F = MCreg / MCres$. La sixième et dernière colonne contient la valeur *p* correspondante à la valeur *F* obtenue dans la colonne 5. Presque tous les logiciels fournissent un résumé de l'analyse de la régression sous forme d'un tableau ANOVA.

Dans chaque tableau d'analyse de la variance, la somme des carrés totale est égale à la somme de la somme des carrés de la régression et de la somme des carrés des résidus ; de plus, le nombre total de degrés de liberté est égal à la somme des degrés de liberté associés à la régression et des degrés de liberté associés aux résidus.

Tableau 12.5 ***Forme générale d'un tableau ANOVA dans le cadre d'une régression linéaire simple***

Source de la variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	*F*	Valeur *p*
Régression	SCreg	1	$MCreg = \frac{SCreg}{1}$	$F = \frac{MCreg}{MCres}$	
Résidu	SCres	$n-2$	$MCres = \frac{SCres}{n-2}$		
Totale	SCT	$n-1$			

Tableau 12.6 ***Tableau ANOVA pour le problème des restaurants Armand***

Source de la variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	*F*	Valeur *p*
Régression	14 200	1	$\frac{14\,200}{1} = 14\,200$	$\frac{14\,200}{191,25} = 74,25$	0,000
Résidu	1 530	8	$\frac{1530}{8} = 191,25$		
Totale	15 730	9			

12.5.5 Quelques précautions à prendre dans l'interprétation des tests de signification

Rejeter l'hypothèse nulle $H_0 : \beta_1 = 0$ et conclure que la relation entre x et y est statistiquement significative ne nous permet pas de conclure qu'une relation de cause à effet lie x et y. Un analyste ne peut conclure à une relation de cause à effet que s'il dispose d'une justification théorique attestant de la causalité de la relation. Dans l'exemple des restaurants Armand, nous pouvons conclure qu'une relation significative existe entre la taille de la population étudiante x et les ventes trimestrielles y ; de plus, l'équation estimée de la régression $\hat{y} = 60 + 5x$ correspond à l'estimation par les moindres carrés de la relation. Nous ne pouvons, cependant, pas conclure que des changements dans la population étudiante x *causent* des changements dans les ventes trimestrielles y, uniquement parce que nous avons identifié une relation statistiquement significative entre ces deux variables. La justesse d'une telle conclusion de causalité est laissée au jugement de l'analyste, étayé par une justification théorique. Les responsables des restaurants Armand pensaient que des augmentations de la population étudiante entraîneraient des augmentations des ventes trimestrielles. Ainsi, le résultat du test de signification leur permet de conclure qu'une relation de cause à effet existe.

L'analyse de la régression, utilisée pour identifier l'existence d'une relation entre deux variables, ne prouve pas l'existence d'une quelconque relation de causalité.

De plus, le fait de rejeter $H_0 : \beta_1 = 0$ et de conclure à l'existence d'une relation significative ne nous permet pas de conclure que la relation entre x et y est linéaire. Nous pouvons seulement affirmer que x et y sont liés et qu'une relation linéaire explique une partie significative de la variabilité de y par rapport aux valeurs de x observées dans l'échantillon. La figure 12.7 illustre cette situation. Le test de signification a conduit au rejet de l'hypothèse nulle $H_0 : \beta_1 = 0$ et à la conclusion que x et y sont significativement liés, mais la figure prouve que la relation effective entre x et y n'est pas linéaire. Bien qu'une approximation linéaire fournie par $\hat{y} = b_0 + b_1 x$ soit correcte au regard des valeurs de x observées dans l'échantillon, elle devient plus mauvaise pour les valeurs de x qui n'appartiennent pas à l'échantillon.

Dans la mesure où la relation est significative, nous pouvons utiliser, avec confiance, l'équation estimée de la régression pour effectuer des prévisions pour des valeurs de x appartenant à l'intervalle des valeurs observées dans l'échantillon. Dans le cadre de l'exemple des restaurants Armand, cet intervalle correspond aux valeurs de x comprises entre 2 et 26. Par contre, à moins que certains éléments indiquent que le modèle reste valable pour des valeurs de x situées hors de cet intervalle, les prévisions pour des valeurs de la variable indépendante qui n'appartiennent pas à l'intervalle observé, sont sujettes à caution. Dans l'exemple des restaurants Armand, puisque la relation de la régression est significative au seuil de 0,01, nous pouvons l'utiliser avec confiance pour prévoir les ventes trimestrielles des restaurants situés sur des campus dont la population étudiante varie entre 2 000 et 26 000 personnes.

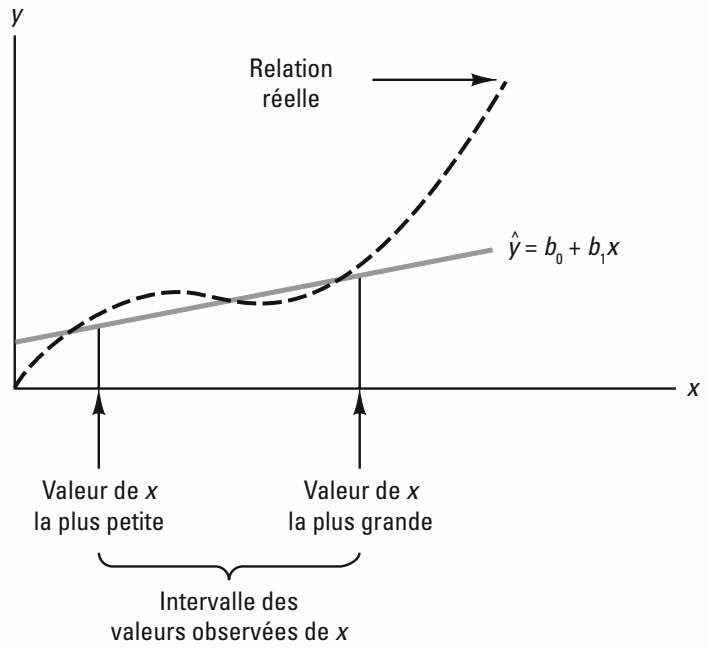

Figure 12.7 ***Exemple d'approximation linéaire d'une relation non-linéaire***

REMARQUES

1. Les hypothèses faites à propos du terme d'erreur (section 12.4) rendent légitimes les tests de signification effectués dans cette section. Les propriétés de la distribution d'échantillonnage de b_1 et les tests de Student et de Fisher découlent directement de ces hypothèses.

2. Ne confondez pas la signification statistique avec la signification pratique. Avec de très grands échantillons, des résultats statistiquement significatifs peuvent être obtenus pour de petites valeurs de b_1 ; dans de tels cas, il faut être prudent en concluant que la relation est significative d'un point de vue pratique.

3. Un test de signification d'une relation linéaire entre x et y peut également être effectué en utilisant le coefficient de corrélation de l'échantillon r_{xy}. Avec ρ_{xy} correspondant au coefficient de corrélation de la population, les hypothèses sont les suivantes.

$$H_0 : \rho_{xy} = 0$$
$$H_a : \rho_{xy} \neq 0$$

Si H_0 est rejetée, on peut conclure à l'existence d'une relation significative. Le détail de ce test est fourni dans des ouvrages plus avancés. Cependant, les tests de Student et de Fisher présentés précédemment fournissent le même résultat que le test de signification effectué avec le coefficient de corrélation. Effectuer un test de signification avec le coefficient de corrélation est donc inutile si un test de Student ou de Fisher a déjà été effectué.

EXERCICES

Méthode

23. Reprendre les données de l'exercice 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Calculer la moyenne des carrés des résidus en utilisant l'expression (12.15).

b) Calculer l'erreur type de l'estimation en utilisant l'expression (12.16).

c) Calculer l'écart type estimé de b_1 en utilisant l'expression (12.18).

d) Utiliser le test de Student pour tester les hypothèses suivantes ($\alpha = 0{,}05$) :

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

e) Utiliser le test de Fisher pour tester les hypothèses de la question (d) au seuil de 0,05. Présenter les résultats sous forme d'un tableau d'analyse de la variance.

24. Reprendre les données de l'exercice 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Calculer la moyenne des carrés des résidus en utilisant l'expression (12.15).

b) Calculer l'erreur type de l'estimation en utilisant l'expression (12.16).

c) Calculer l'écart type estimé de b_1 en utilisant l'expression (12.18).

d) Utiliser le test de Student pour tester les hypothèses suivantes ($\alpha = 0{,}05$) :

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

e) Utiliser le test de Fisher pour tester les hypothèses de la question (d) au seuil de 0,05. Présenter les résultats sous forme d'un tableau d'analyse de la variance.

25. Reprendre les données de l'exercice 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a) Quelle est la valeur de l'erreur type de l'estimation ?

b) Tester l'existence d'une relation significative en utilisant le test de Student au seuil $\alpha = 0{,}05$.

c) Utiliser le test de Fisher pour tester l'existence d'une relation significative au seuil $\alpha = 0{,}05$. Quelle est votre conclusion ?

Applications

26. À l'exercice 18, nous avons présenté les données suivantes concernant le prix et la note globale de six écouteurs stéréo testés par *Consumer Reports* (site Internet de *Consumer Reports*, 5 mars 2012).

Marque	Prix ($)	Note
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O'Neill	70	56
Denon	70	40
JVC	35	26

a) Est-ce que le test de Student révèle l'existence d'une relation significative entre la note moyenne et le salaire mensuel ? Quelle est votre conclusion ? Utiliser un seuil de signification $\alpha = 0{,}05$.

b) Tester l'existence d'une relation significative en utilisant le test de Fisher. Quelle est votre conclusion ? Utiliser un seuil de signification de 0,05.

c) Construire le tableau ANOVA.

27. Le nombre de pixels d'un appareil photo numérique est l'un des plus importants facteurs déterminant la qualité de l'image. Mais les appareils photo ayant le plus grand nombre de pixels coûtent-ils plus chers ? Les données suivantes (cf. fichier en ligne Appareils photo numériques) indiquent le nombre de pixels (en millions) et le prix (en dollars) de 10 appareils photo numériques (*Consumer Reports*, mars 2009).

Marque et modèle	Pixels (en millions)	Prix ($)
Canon PowerShot SD110 IS	8	180
Casio Exilim Card EX-510	10	200
Sony Cyber-shot DSC-T70	7	230
Pentax Optio M50	8	120
Canon PowerShot G10	15	470
Canon PowerShot A590 IS	8	140
Canon PowerShot E1	10	180
Fujifilm FinePi F00FD	12	310
Sony Cyber-shot DSC-W170	10	250
Canon PowerShot A470	7	110

a) Utiliser ces données pour développer l'équation estimée de la régression, permettant d'estimer le prix d'un appareil photo numérique en fonction du nombre de pixels.

b) Au seuil de signification de 0,05, déterminer si le nombre de pixels et le prix sont liés. Expliquer.

c) Pensez-vous que l'équation estimée de la régression est suffisamment robuste pour prévoir le prix d'un appareil photo numérique étant donné le nombre de pixels ? Expliquer.

d) L'appareil photo numérique Kodak EasyShare Z1012 IS a 10 millions de pixels. Prévoir le prix de cet appareil en utilisant l'équation estimée de la régression obtenue à la question (a).

28. Dans l'exercice 8, des données (cf. fichier en ligne Notation Courtiers) sur la rapidité d'exécution des ordres (x) et la note de satisfaction globale des transactions électroniques (y) ont fourni l'équation de régression estimée $\hat{y} = 0,2046 + 0,9077x$ (site Internet de l'AAII, 7 février 2012). Tester, au seuil de signification de 0,05, l'existence d'une relation significative entre la rapidité d'exécution des ordres et la satisfaction globale. Construire un tableau ANOVA. Quelle est votre conclusion ?

29. Reprendre l'exercice 21, dans lequel des données sur le volume et les coûts de production ont permis d'estimer une équation de la régression liant le volume de la production et son coût pour une opération de fabrication particulière. Tester, au seuil de signification de 0,05, l'existence d'une relation significative entre le volume de production et les coûts totaux. Construire le tableau ANOVA. Quelle est votre conclusion ?

30. Reprendre l'exercice 9, dans lequel les données suivantes ont été utilisées pour étudier la relation entre le nombre de véhicules en service (en milliers) et le revenu annuel (en millions de dollars) de six petites sociétés de location de voitures (site Internet de *Auto Rental News*, 7 août 2012).

Société	Véhicules (milliers)	Revenu (millions de dollars)
U-Save Auto Rental System, Inc.	11,5	118
Payless Car Rental System, Inc.	10,0	135
ACE Rent A Car	9,0	100
Rent-A-Wreck of America	5,5	37
Triangle Rent-A-Car	4,2	40
Affordable/Sensible	3,3	32

Avec x le nombre de véhicules en service (en milliers) et y le revenu annuel (en millions de dollars), l'équation estimée de la régression est $\hat{y} = -17,005 + 12,966x$. Pour ces données, $SCres = 1\,043,03$ et $SCT = 10\,568$. Existe-t-il une relation significative entre le nombre de véhicules en service et le revenu annuel ?

31. Dans l'exercice 20, des données (cf. fichier en ligne Vélos de course) sur le poids en livres (x) et le prix en dollars (y) de 10 vélos de courses ont fourni l'équation estimée de la régression suivante : $\hat{y} = 28,574 - 1\,439x$ (site Internet de *Bicycling*, 8 mars 2012). Pour ces données, $SCres = 7\,102\,922,54$ et $SCT = 52\,120\,800$. Utiliser le test de Fisher pour déterminer si le poids d'un vélo et son prix sont liés au seuil de signification égal à 0,05.

12.6 UTILISER L'ÉQUATION ESTIMÉE DE LA RÉGRESSION POUR ESTIMER ET PRÉVOIR

Lorsqu'on utilise un modèle de régression linéaire simple, on fait une hypothèse sur la relation entre x et y. En utilisant la méthode des moindres carrés, on obtient l'équation estimée de la régression linéaire simple. Si les résultats prouvent l'existence d'une relation statistiquement significative entre x et y, et si le coefficient de détermination indique que l'équation estimée de la régression semble bien adaptée aux données, l'équation estimée de la régression peut servir à faire des estimations et des prévisions.

Dans l'exemple des restaurants Armand, l'équation estimée de la régression s'écrit $\hat{y} = 60 + 5x$. À la fin de la section 12.1, nous avons affirmé que $\hat{y}$ pouvait être utilisé comme un estimateur ponctuel de $E(y)$, la moyenne ou valeur espérée de y pour une valeur donnée de x. Par exemple, supposez que les responsables des restaurants Armand veuillent effectuer une estimation ponctuelle de la moyenne des ventes trimestrielles pour *tous* les restaurants situés près de campus universitaires regroupant 10 000 étudiants. En utilisant l'équation estimée de la régression $\hat{y} = 60 + 5x$, nous voyons que pour $x = 10$ (soit 10 000 étudiants), $\hat{y} = 60 + 5(10) = 110$. Ainsi, une *estimation ponctuelle* de la moyenne des ventes trimestrielles pour tous les restaurants situés près de campus comptant 10 000 étudiants est 110 000 dollars. Dans ce cas, nous avons utilisé $\hat{y}$ comme estimateur ponctuel de la valeur moyenne de y lorsque x est égal à 10.

Nous pouvons également utiliser l'équation estimée de la régression pour *prédire* une valeur individuelle de y pour une valeur donnée de x. Par exemple, pour prévoir les ventes trimestrielles d'un nouveau restaurant situé près du collège Talbot, une école comptant 10 000 étudiants, nous calculons $\hat{y} = 60 + 5(10) = 110$. Par conséquent, nous pouvons utiliser $\hat{y}$ comme *prévision* de y pour une nouvelle observation lorsque $x = 10$.

Lorsque nous utilisons l'équation estimée de la régression pour estimer la valeur moyenne de y ou prédire une valeur individuelle de y, il est clair que l'estimation ou la prévision dépendent de la valeur de x considérée. Pour cette raison, lors de nos discussions sur les questions relatives à l'estimation et à la prévision, nous adopterons la notation suivante pour clarifier les choses.

x^* = la valeur considérée de la variable indépendante x

y^* = la variable aléatoire correspondant aux valeurs possibles de la variable dépendante y lorsque $x = x^*$

$E(y^*)$ = la moyenne ou l'espérance mathématique de la variable dépendante y lorsque $x = x^*$

$\hat{y}^* = b_0 + b_1 x^*$ = l'estimateur ponctuel de $E(y^*)$ et le prédicteur d'une valeur individuelle de y^* lorsque $x = x^*$

Pour illustrer l'usage de cette notation, supposez que nous souhaitions estimer la valeur moyenne des ventes trimestrielles de tous les restaurants Armand situés près d'un campus de 10 000 étudiants. Dans ce cas $x^* = 10$ et $E(y^*)$ correspond à la valeur moyenne

inconnue des ventes trimestrielles pour tous les restaurants où $x^* = 10$. Ainsi, l'estimation ponctuelle de $E(y^*)$ est fournie par $\hat{y}^* = 60 + 5(10) = 110$, soit 110 000 dollars. Mais, en utilisant cette notation, $\hat{y}^* = 110$ correspond aussi à la prévision des ventes trimestrielles pour le nouveau restaurant situé près du collège Talbot, une école de 10 000 étudiants.

12.6.1 Estimation par intervalle

Les estimations ponctuelles et les prévisions ne fournissent aucune information sur la précision de l'estimation et/ou de la prévision. Pour cela, il faut développer des intervalles de confiance et des intervalles de prévision. Un **intervalle de confiance** est une estimation par intervalle de la *valeur moyenne de y* pour une valeur donnée de *x*. Un **intervalle de prévision** est utilisé lorsqu'on souhaite *prédire une valeur individuelle de y* pour une nouvelle observation correspondant à une valeur donnée de *x*. Bien que la prévision de *y* pour une valeur donnée de *x* soit identique à l'estimation ponctuelle de la valeur moyenne de *y* pour une valeur donnée de *x*, les estimations par intervalle que nous obtenons dans les deux cas, sont différentes. Comme nous le verrons, la marge d'erreur est plus importante dans le cas d'intervalles de prévision. Nous commençons par montrer comment construire une estimation par intervalle de la valeur moyenne de *y*.

Les intervalles de confiance et les intervalles de prévision indiquent la précision des résultats de la régression. Plus les intervalles sont petits, plus le degré de précision est élevé.

12.6.2 Intervalle de confiance de la valeur moyenne de y

En général, $\hat{y}^*$ n'est pas exactement égal à $E(y^*)$. Si l'on souhaite faire de l'inférence sur l'écart entre $\hat{y}^*$ et la vraie moyenne $E(y^*)$, il faut estimer la variance de $\hat{y}^*$. La formule pour estimer la variance de $\hat{y}^*$ sachant x^*, notée $s^2_{\hat{y}^*}$ correspond à

$$s^2_{\hat{y}^*} = s^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{12.22}$$

L'estimation de l'écart type de $\hat{y}^*$ correspond à la racine carrée de l'expression (12.22).

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{12.23}$$

D'après les résultats obtenus dans le cadre de l'exemple des restaurants Armand dans la section 12.5, $s = 13{,}829$. Avec $x_p = 10$, $\bar{x} = 14$ et $\sum(x_i - \bar{x})^2 = 568$, on peut utiliser l'expression (12.23) pour obtenir

$$s_{\hat{y}_p} = 13{,}829\sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}}$$
$$= 13{,}829\sqrt{0{,}1282} = 4{,}95$$

L'expression générale pour un intervalle de confiance s'écrit de la façon suivante.

► **Intervalle de confiance pour $E(y_p)$**

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{12.24}$$

où le coefficient de confiance est égal à $1-\alpha$ et $t_{\alpha/2}$ est basé sur la distribution de Student à $n-2$ degrés de liberté

La marge d'erreur associée à cette estimation par intervalle est $t_{\alpha/2} s_{\hat{y}_p}$.

Pour pouvoir utiliser l'expression (12.24) pour construire un intervalle de confiance à 95 % de la moyenne des ventes trimestrielles pour tous les restaurants Armand situés près de campus regroupant 10 000 étudiants, il nous faut connaître la valeur de t pour $\alpha/2 = 0{,}025$ et $n - 2 = 10 - 2 = 8$ degrés de liberté. D'après la table 2 de l'annexe B, $t_{0{,}025} = 2{,}306$. Ainsi, avec $\hat{y}^* = 110$ et une marge d'erreur égale à $t_{\alpha/2} s_{\hat{y}^*} = 2{,}306(4{,}95) = 11{,}415$ l'estimation par intervalle de confiance à 95 % est

$$110 \pm 11{,}415$$

En dollars, l'intervalle de confiance à 95 % de la moyenne des ventes trimestrielles de tous les restaurants situés près des campus de 10 000 étudiants est 110 000 ± 11 415 dollars. Par conséquent, l'intervalle de confiance à 95 % de la moyenne des ventes trimestrielles lorsque la population étudiante compte 10 000 individus va de 98 585 dollars à 121 415 dollars.

Notez que l'écart type estimé de $\hat{y}^*$ donné par l'expression (12.23) est le plus faible lorsque $x^* - \overline{x} = 0$. Dans ce cas, l'écart type estimé de $\hat{y}^*$ devient

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(\overline{x} - \overline{x})^2}{\sum (x_i - \overline{x})^2}} = s\sqrt{\frac{1}{n}}$$

Ce résultat implique que la meilleure estimation ou l'estimation la plus précise de la moyenne de y est obtenue lorsque $x^* = \overline{x}$. En fait, plus x^* est loin de $\overline{x}$, plus $x^* - \overline{x}$ s'accroît. Par conséquent, les intervalles de confiance pour la moyenne de y deviennent plus larges lorsque x^* s'écarte de $\overline{x}$. La figure 12.8 illustre graphiquement ce résultat.

12.6.3 Intervalle de prévision d'une valeur individuelle de y

Supposez que plutôt qu'estimer la moyenne des ventes trimestrielles des restaurants Armand situés près des campus de 10 000 étudiants, nous voulions estimer les ventes trimestrielles d'un nouveau restaurant qu'Armand envisage de construire près du collège Talbot qui compte 10 000 étudiants. Comme souligné précédemment, la prévision de y^*, la valeur de y associée à x^*, correspond à $\hat{y}^* = b_0 + b_1 x^*$. Pour un nouveau restaurant situé près du collège Talbot, $x^* = 10$ et les ventes trimestrielles correspondantes sont estimées à $\hat{y}^* = 60 + 5(10) = 110$ soit 110 000 dollars. Notez que cette valeur est identique à l'estimation ponctuelle de la moyenne des ventes trimestrielles pour tous les restaurants situés près de campus de 10 000 étudiants.

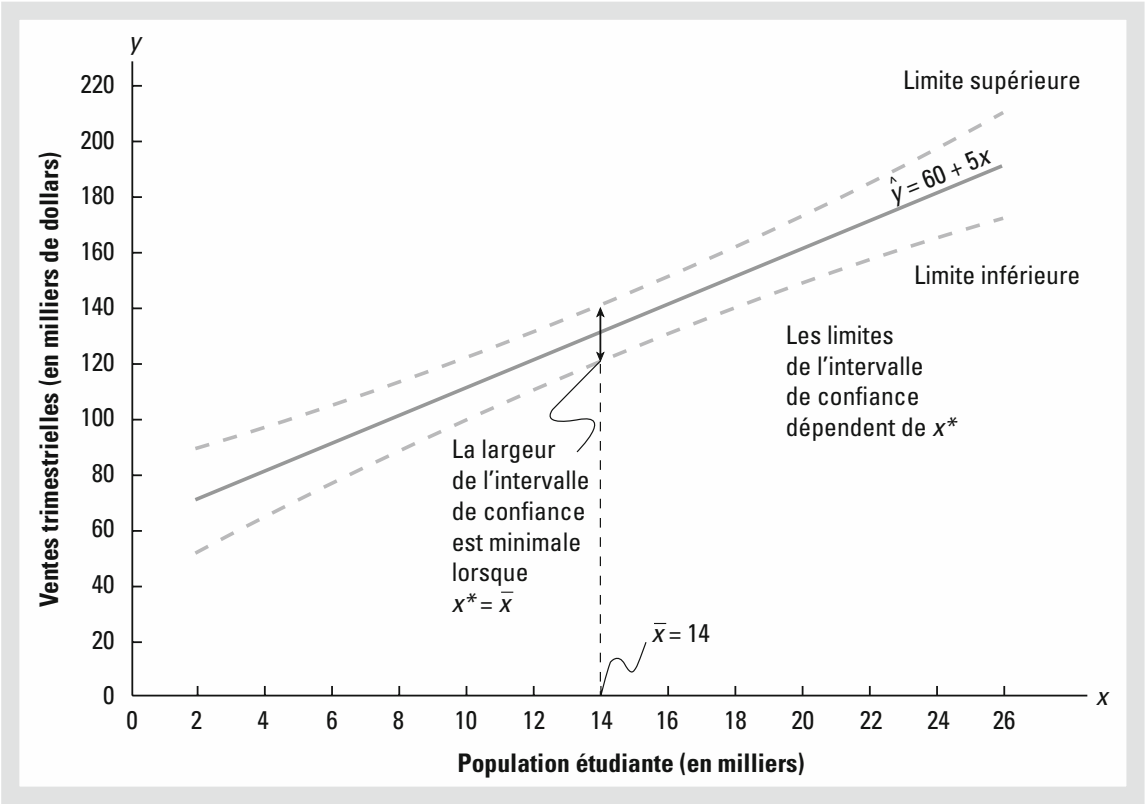

Figure 12.8 ***Intervalles de confiance de la moyenne des ventes trimestrielles y pour des valeurs données de la population étudiante x***

Pour développer un intervalle de prévision, nous devons tout d'abord estimer la variance associée à l'utilisation de $\hat{y}^*$ comme estimateur de y lorsque $x = x^*$. Cette variance est composée de la somme des deux éléments suivants :

1. La variance des valeurs de y^*, par rapport à la moyenne $E(y^*)$, estimée par s^2 ;

2. La variance associée à l'utilisation de $\hat{y}_p$ pour estimer $E(y^*)$, estimée par $s^2_{\hat{y}^*}$.

La formule pour estimer la variance associée à la prévision d'une valeur de y lorsque $x = x^*$, notée s^2_{prev}, est

$$\begin{aligned} s^2_{prev} &= s^2 + s^2_{\hat{y}^*} \\ &= s^2 + s^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \\ &= s^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \end{aligned} \tag{12.25}$$

Par conséquent, une estimation de l'écart type associé à la prévision d'une valeur de y^* est donnée par

$$s_{prev} = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]} \tag{12.26}$$

Dans le cadre de l'exemple des restaurants Armand, l'écart type estimé correspondant à la prévision des ventes trimestrielles d'un nouveau restaurant situé près du collège Talbot, un campus de 10 000 étudiants, est calculé de la façon suivante.

$$\begin{aligned} s_{prev} &= 13{,}829\sqrt{\left[1 + \frac{1}{10} + \frac{(10-14)^2}{568}\right]} \\ &= 13{,}829\sqrt{1{,}282} \\ &= 14{,}69 \end{aligned}$$

L'expression générale d'un intervalle de prévision est la suivante.

► **Intervalle de prévision de y_p**

$$\hat{y}_p \pm t_{\alpha/2} s_{prev} \tag{12.27}$$

où le coefficient de confiance est égal à $1-\alpha$ et $t_{\alpha/2}$ est basé sur la distribution de Student à $n-2$ degrés de liberté

La marge d'erreur associée à cette estimation par intervalle est $t_{\alpha/2} s_{prev}$.

L'intervalle de prévision à 95 % pour les ventes trimestrielles d'un nouveau restaurant situé près du collège Talbot peut être trouvé en utilisant $t_{0,025} = 2{,}306$ et $s_{prev} = 14{,}69$. Ainsi, avec $\hat{y}^* = 110$ et une marge d'erreur égale à $t_{0,025} s_{prev} = 2{,}306(14{,}69) = 33{,}875$, l'intervalle de prévision à 95 % est le suivant

$$110 \pm 33{,}875$$

En dollars, l'intervalle de prévision est le suivant : 110 000 ± 33 875 dollars, soit de 76 125 dollars à 143 875 dollars. Notez que l'intervalle de prévision pour le nouveau restaurant situé près du collège Talbot, un campus de 10 000 étudiants, est plus large que l'intervalle de confiance pour la moyenne des ventes de tous les restaurants situés près de campus de 10 000 étudiants. La différence reflète le fait que nous sommes capables d'estimer la valeur moyenne de y de façon plus précise qu'une valeur individuelle de y.

À la fois les estimations par intervalle de confiance et par intervalle de prévision sont plus précises lorsque la valeur de la variable indépendante x^* est proche de $\bar{x}$. Les formes générales des intervalles de confiance et des intervalles de prévision, plus larges, sont représentées à la figure 12.9.

En général, les courbes représentant les limites des intervalles de confiance et de prévision ont la même forme.

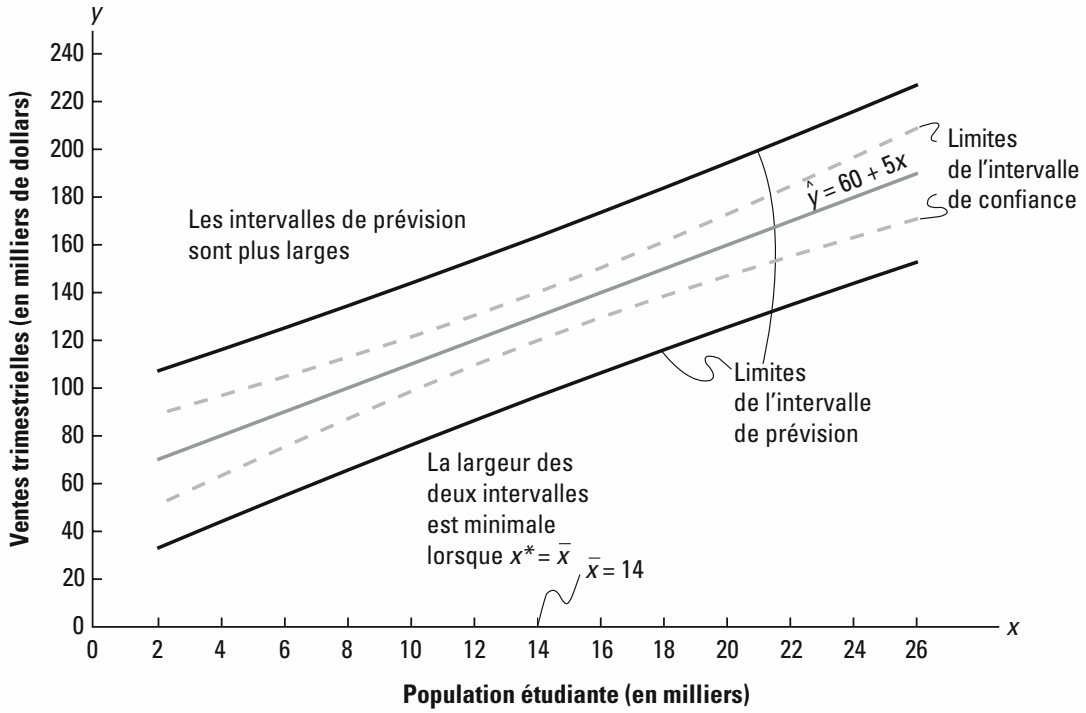

Figure 12.9 ***Intervalles de confiance et de prévision des ventes trimestrielles y pour des valeurs données de la population étudiante x***

REMARQUES

Un intervalle de prévision est utilisé pour prévoir la valeur de la variable dépendante *y* pour une *nouvelle observation*. À titre d'illustration, nous avons montré comment construire un intervalle de prévision des ventes trimestrielles d'un nouveau restaurant qu'Armand envisage de construire près du collège Talbot, un campus de 10 000 étudiants. Le fait que la valeur de $x = 10$ ne soit pas une des valeurs de la population d'étudiants appartenant à l'échantillon de données du tableau 12.1, n'implique pas que les intervalles de prévision ne peuvent pas être construits pour des valeurs de *x* appartenant aux données d'échantillon. Mais, pour les 10 restaurants qui constituent l'échantillon du tableau 12.1, construire un intervalle de prévision pour les ventes trimestrielles pour l'un de ces restaurants ne fait pas sens puisque nous connaissons déjà la valeur des ventes trimestrielles de chacun de ces restaurants. En d'autres termes, un intervalle de prévision n'a de sens que pour quelque chose de nouveau, dans ce cas, une nouvelle observation correspondant à une valeur particulière de *x* qui peut ou peut ne pas être égale à une des valeurs de *x* contenues dans l'échantillon.

EXERCICES

Méthode

32. Reprendre les données de l'exercice 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a) Utiliser l'expression (12.23) pour estimer l'écart type de $\hat{y}^*$ lorsque $x = 4$.
b) Utiliser l'expression (12.24) pour construire un intervalle de confiance à 95 % pour la valeur attendue de y lorsque $x = 4$.
c) Utiliser l'expression (12.26) pour estimer l'écart type d'une valeur individuelle de y lorsque $x = 4$.
d) Utiliser l'expression (12.27) pour construire un intervalle de prévision à 95 % pour $x = 4$.

33. Reprendre les données de l'exercice 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a) Estimer l'écart type de $\hat{y}^*$ lorsque $x = 8$.
b) Construire l'intervalle de confiance à 95 % pour la valeur attendue de y lorsque $x = 8$.
c) Estimer l'écart type d'une valeur individuelle de y lorsque $x = 8$.
d) Construire l'intervalle de prévision à 95 % pour y lorsque $x = 8$.

34. Reprendre les données de l'exercice 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

Construire les intervalles de confiance et de prévision à 95 % lorsque $x = 12$. Expliquer pourquoi ces deux intervalles sont différents.

Applications

35. Les données suivantes correspondent aux salaires mensuels y et à la note moyenne x des étudiants diplômés d'une licence en école de commerce.

Note moyenne	Salaire mensuel ($)
2,6	3 600
3,4	3 900
3,6	4 300
3,2	3 800
3,5	4 200
2,9	3 900

L'équation estimée de la régression associée à ces données est $\hat{y} = 2\,090{,}5 + 581{,}1x$ et $MCres = 21\,284$.

a) Quelle est l'estimation ponctuelle du salaire mensuel de base d'un étudiant qui a eu une note moyenne de 3 ?

b) Construire un intervalle de confiance à 95 % pour le salaire moyen de base de tous les étudiants qui ont obtenu une note moyenne égale à 3.

c) Construire un intervalle de prévision à 95 % pour Ryan Dailey, un étudiant qui a obtenu une note moyenne de 3.

d) Discuter des différences entre vos réponses aux questions (b) et (c).

36. Dans l'exercice 7, les données (cf. fichier en ligne Ventes) sur les ventes annuelles (en milliers de dollars) (x) et le nombre d'années d'expériences (y) d'un échantillon de 10 vendeurs ont fourni l'équation de régression estimée $\hat{y} = 80 + 4x$. Pour ces données, $\overline{x} = 7$, $\sum (x_i - \overline{x})^2 = 142$ et $s = 4{,}6098$.

a) Construire un intervalle de confiance à 95 % pour les ventes annuelles moyennes de tous les vendeurs qui ont neuf ans d'expérience professionnelle.

b) La société envisage d'embaucher Tom Smart, un vendeur qui a neuf années d'expérience professionnelle. Construire l'intervalle de prévision à 95 % des ventes annuelles que pourrait réaliser Tom Smart.

c) Discuter des différences entre vos réponses aux questions (b) et (c).

37. Dans l'exercice 5, les données suivantes sur le nombre de pièces défectueuses (x) et la vitesse (en pied par minute) de la chaîne de montage (y) dans le processus de production de Brawdy Plastics ont fourni l'équation estimée de la régression $\hat{y} = 27{,}5 - 0{,}3x$.

Vitesse de la chaîne de montage	Nombre de pièces défectueuses trouvées
20	23
20	21
30	19
30	16
40	15
40	17
50	14
50	11

Pour ces données, $SCres = 16$. Construire un intervalle de confiance à 95 % pour le nombre moyen de pièces défectueuses sur une chaîne de production avançant à 25 pieds par minute.

38. Référez-vous à l'exercice 21, dans lequel des données sur le volume de la production x et le coût total y d'une opération de fabrication particulière, ont permis d'estimer l'équation de la régression $\hat{y} = 1\,246{,}67 + 7{,}6x$.

a) D'après le planning de production de la société, 500 unités devraient être produites le mois prochain. Quelle est l'estimation ponctuelle du coût total pour le mois prochain ?

b) Construire un intervalle de prévision à 99 % pour le coût total du mois prochain.

c) Si un rapport comptable sur les coûts, écrit à la fin du mois suivant, indique que le coût réel de la production au cours du mois était de 6 000 dollars, les responsables devraient-ils s'inquiéter d'avoir supporté un coût total aussi élevé ? Discuter.

39. Dans l'exercice 12, les données suivantes sur le prix moyen d'une chambre d'hôtel (x) et le montant dépensé en divertissement (y) (*The Wall Street Journal*, 18 août 2011) a fourni l'équation estimée de la régression $\hat{y} = 17,49 + 1,0334x$ (cf. fichier en ligne Voyage d'affaires). Pour ces données, $SCres = 1\,541,4$.

Ville	Tarif d'une chambre ($)	Divertissement ($)
Boston	148	161
Denver	96	105
Nashville	91	101
Nouvelle Orléans	110	142
Phoenix	90	100
San Diego	102	120
San Francisco	136	167
San José	90	140
Tampa	82	98

a) Prévoir le montant dépensé en divertissement pour une ville particulière dans laquelle le tarif d'une chambre d'hôtel s'élève à 89 dollars.

b) Construire un intervalle de confiance à 95 % pour le montant moyen dépensé en divertissement dans toutes les villes dans lesquelles le tarif d'une chambre d'hôtel s'élève à 89 dollars.

c) Le tarif moyen d'une chambre à Chicago s'élève à 128 dollars. Construire un intervalle de prévision à 95 % pour le montant dépensé en divertissement à Chicago.

12.7 SOLUTION INFORMATIQUE

Faire une analyse de la régression sans l'aide d'un ordinateur peut être chronophage. Dans cette section, nous verrons comment minimiser les calculs en utilisant un logiciel comme Minitab.

Nous avons enregistré les données relatives à la population étudiante et aux ventes trimestrielles des restaurants Armand, dans une feuille de calcul Minitab. Nous avons nommé la variable indépendante POP et la variable dépendante SALES pour faciliter l'interprétation du résultat de la programmation, illustré à la figure 12.10.[2] L'interprétation de ce résultat suit.

2 Les étapes de la programmation nécessaires à l'obtention de l'output sont décrites dans l'annexe 12.1.

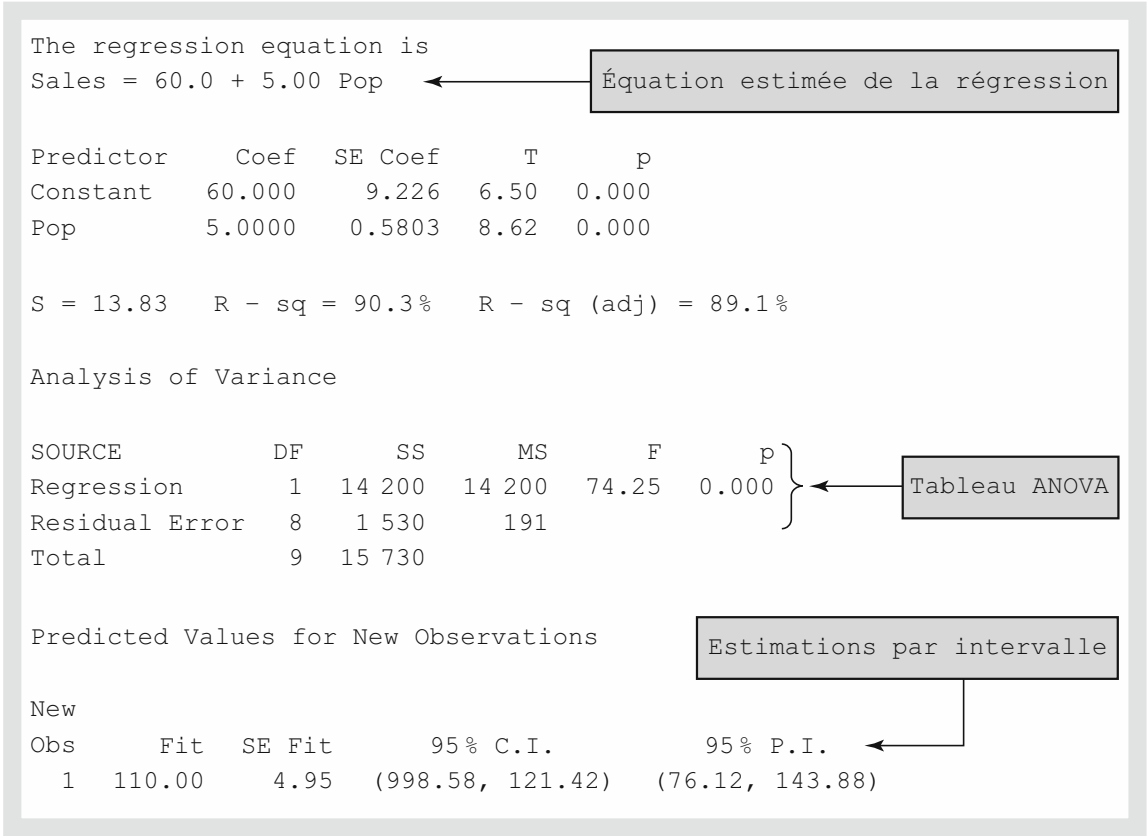

Figure 12.10 ***Feuille de résultats Minitab dans le cadre du problème des restaurants Armand***

1. Minitab affiche l'équation estimée de la régression de la façon suivante : $SALES = 60.0 + 5.00\,POP$.
2. Minitab affiche un tableau dans lequel apparaissent les valeurs des coefficients b_0 et b_1, l'écart type de chaque coefficient, la valeur t obtenue en divisant la valeur du coefficient par son écart type, et la valeur p associée au test de Student. Puisque la valeur p est égale à zéro (avec trois chiffres après la virgule), les résultats d'échantillon indiquent que l'hypothèse nulle ($H_0: \beta_1 = 0$) doit être rejetée. De manière alternative, on peut comparer 8,62 (situé dans la colonne T) à la valeur critique appropriée. Cette procédure a été décrite pour le test de Student dans la section 12.5.
3. Minitab affiche l'erreur type de l'estimation, $s = 13{,}8293$, ainsi que des informations sur l'adéquation du modèle aux données. Notez que « R – sq = 90,3 % » correspond au coefficient de détermination exprimé en pourcentage. La valeur « R-Sq(adj) = 89.1 % » sera discutée au chapitre 13.
4. Le tableau ANOVA est affiché en dessous du titre « Analysis of variance ». Minitab utilise le titre « Residual Error » pour exprimer la source de variation que sont les erreurs. Notez que DF est une abréviation de degrés de liberté (« degrees of freedom ») et que la moyenne des carrés de la régression (*MCreg*) est égale à 14 200 et la moyenne des carrés des résidus (*MCres*) est

égale à 191. Le rapport de ces deux valeurs fournit la valeur F, égale à 74,25 et la valeur p qui lui est associée, égale à 0. Puisque la valeur p est nulle (avec trois chiffres après la virgule), la relation entre *Sales* et *Pop* est jugée statistiquement significative.

5. L'estimation par intervalle de confiance à 95 % des ventes trimestrielles attendues et l'estimation par intervalle de prévision à 95 % des ventes trimestrielles d'un restaurant situé près d'un campus de 10 000 étudiants sont affichées sous le tableau ANOVA. L'intervalle de confiance est [98,58 ; 121,42] et l'intervalle de prévision est [76,12 ; 143,87] comme nous l'avons vu dans la section 12.6.

EXERCICES

Applications

40. Le département commercial d'une agence immobilière a effectué une analyse de la régression de la relation entre x, les loyers bruts annuels (en milliers de dollars) et y, le prix de vente (en milliers de dollars) d'un immeuble. Les données collectées concernent plusieurs propriétés récemment vendues, et les résultats informatiques suivants ont été obtenus.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor        Coef     SE Coef        T
Constant       20.000      3.2213     6.21
X               7.210      1.3626     5.29

Analysis of Variance

SOURCE              DF          SS
Regression           1     41587.3
Residual Error       7
Total                8     51984.1
```

a) Combien d'immeubles l'échantillon comprend-t-il ?

b) Écrire l'équation estimée de la régression.

c) Quelle est la valeur de s_{b_1} ?

d) Utiliser la statistique de Fisher pour tester l'existence d'une relation significative au seuil de 0,05.

e) Prédire le prix de vente d'un immeuble dont le loyer brut annuel s'élève à 50 000 dollars.

41. Ci-dessous est présentée une partie du résultat de la programmation d'une analyse de la régression reliant les dépenses de maintenance (en dollars par mois), y, et l'usage (en heures par semaine) d'une marque particulière d'un terminal informatique, x.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor        Coef     SE Coef
Constant       6.1092      0.9361
X              0.8951      0.1490

Analysis of Variance

SOURCE              DF          SS          MS
Regression           1     1575.76     1575.76
Residual Error       8      349.14       43.64
Total                9     1924.90
```

a) Écrire l'équation estimée de la régression.

b) Utiliser un test de Student pour déterminer si les dépenses mensuelles de maintenance du terminal sont liées à son utilisation, au seuil de signification de 0,05.

c) Utiliser l'équation estimée de la régression pour prévoir les dépenses mensuelles de maintenance pour tout terminal utilisé 25 heures par semaine.

42. Un modèle de régression reliant x, le nombre de vendeurs d'une succursale, à y, les ventes annuelles de la succursale (en milliers de dollars), a été développé. Le résultat de la programmation de l'analyse de la régression est présenté ci-dessous.

```
The regression equation is
Y = 80.0 + 50.0 X

Predictor        Coef     SE Coef         T
Constant         80.0      11.333      7.06
X                50.0       5.482      9.12

Analysis of Variance

SOURCE              DF          SS          MS
Regression           1      6828.6      6828.6
Residual Error      28      2298.8        82.1
Total               29      9127.4
```

a) Écrire l'équation estimée de la régression.

b) Combien de succursales l'étude comprend-t-elle ?

c) Calculer la statistique de Fisher et tester l'existence d'une relation significative au seuil de 0,05.

d) Prévoir les ventes annuelles de la succursale de Memphis. Cette succursale emploie 12 vendeurs.

43. Les frais d'inscription dans des écoles de commerce peuvent être très élevés mais le salaire de base et les bonus auxquels peuvent prétendre les diplômés de ces écoles peuvent s'avérer également substantiels. Les données suivantes (cf. fichier en ligne Écoles de commerce) indiquent les frais d'inscription (arrondis au millier de dollars le plus proche)

et la rémunération (salaire de base plus bonus) de récents diplômés de ces écoles (arrondis au millier de dollars le plus proche) pour un échantillon de 20 écoles de commerce (*U.S. News & World Report 2009 Edition America's Best Graduate Schools*).

École	Frais d'inscription (en milliers de dollars)	Rémunération (en milliers de dollars)
Université d'État d'Arizona	28	98
Babson College	35	94
Université de Cornell	44	119
Université de Georgetown	40	109
Institut technologique de Géorgie	30	88
Université de l'Indiana – Bloomington	35	105
Université d'État du Michigan	26	99
Université Northwestern	44	123
Université d'État de l'Ohio	35	97
Université de Purdue – West Lafayette	33	96
Université de Rice	36	102
Université de Stanford	46	135
Université de Californie – Davis	35	89
Université de Floride	23	71
Université de l'Iowa	25	78
Université du Minnesota – Twin Cities	37	100
Université de Notre Dame	36	95
Université de Rochester	38	99
Université de Washington	30	94
Université du Wisconsin – Madison	27	93

a) Représenter un nuage de points avec la rémunération comme variable dépendante.

b) Une relation apparaît-elle entre ces variables ? Expliquer.

c) Estimer l'équation de la régression qui pourrait être utilisée pour prévoir la rémunération des jeunes diplômés étant donnés les frais d'inscription à l'école.

d) Tester l'existence d'une relation significative au seuil de 0,05. Quelle est votre conclusion ?

e) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

f) Supposez que nous sélectionnions aléatoirement un jeune diplômé de l'Université de Virginie. Les frais d'inscription s'élèvent à 43 000 dollars. Estimer la rémunération de ce diplômé.

44. Les courses automobiles, les écoles de conduite de haut niveau et les programmes d'éducation des automobilistes proposés par les clubs automobiles voient leur popularité s'accroître. Toutes ces activités imposent aux participants de porter un casque certifié par la fondation Snell Memorial, une organisation à but non lucratif dédiée à la recherche, au test et au développement des casques de sécurité. Les casques professionnels évalués par Snell « SA » (Sports Application) sont conçus pour les courses automobiles et offrent une protection

optimale contre le feu et une bonne résistance aux impacts extrêmes. L'un des facteurs clés dans le choix d'un casque est le poids, puisque des casques plus légers minimisent l'impact sur la nuque. Les données suivantes (cf. fichier en ligne Casques de course) indiquent le poids et le prix de 18 casques SA (site Internet de SoloRacer, 20 avril 2008).

Casques de course

Casque	Poids (onces)	Prix ($)
Pyrotect Pro Airflow	64	248
Pyrotect Pro Airflow Graphics	64	278
RCi Full Race	64	200
RaceQuip RidgeLine	64	200
HJC AR-10	58	300
HJC Si-12	47	700
HJC HX-10	49	900
Impact Racing Super Sport	59	340
Zamp FSA-1	66	199
Zamp RZ-2	58	299
Zamp RZ-2 Ferrari	58	299
Zamp RZ-3 Sport	52	479
Zamp RZ-3 Sport Painted	52	479
Bell M2	63	369
Bell M4	62	369
Bell M4 Pro	54	559
G Force Pro Force 1	63	250
G Force Pro Force 1 Grafx	63	280

a) Représenter le nuage de points avec le poids comme variable indépendante.

b) Une relation apparaît-elle entre les deux variables ?

c) Estimer l'équation de la régression qui peut servir à prévoir le prix en fonction du poids.

d) Tester l'existence d'une relation significative au seuil de 0,05.

e) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

12.8 L'ANALYSE DES RÉSIDUS : VALIDER LES HYPOTHÈSES DU MODÈLE

Comme nous l'avons noté précédemment, le *résidu* de l'observation i est la différence entre la valeur observée de la variable dépendante (y_i) et la valeur estimée de la variable dépendante ($\hat{y}_i$).

L'analyse des résidus est le principal outil pour déterminer si le modèle de régression utilisé est approprié.

► **Résidu de l'observation *i***

$$y_i - \hat{y}_i \tag{12.28}$$

où

y_i correspond à la valeur observée de la variable dépendante

$\hat{y}_i$ correspond à la valeur estimée de la variable dépendante

En d'autres termes, le i^e résidu est l'erreur qui résulte de l'utilisation de l'équation estimée de la régression pour prévoir la valeur de la variable dépendante y_i. Le calcul des résidus associés à l'exemple des restaurants Armand est présenté dans le tableau 12.7. Les valeurs observées de la variable dépendante sont notées dans la deuxième colonne et les valeurs estimées de la variable dépendante, obtenues en utilisant l'équation estimée de la régression $\hat{y} = 60 + 5x$, dans la troisième colonne. Les résidus correspondants sont inscrits dans la quatrième colonne. Une analyse de ces résidus permet de déterminer si les hypothèses qui ont été faites sur le modèle de régression sont appropriées.

Revoyons maintenant les hypothèses faites dans le cadre de l'exemple des restaurants Armand. Un modèle de régression linéaire simple a été utilisé :

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{12.29}$$

Par ce modèle, nous avons supposé que les ventes trimestrielles (y) dépendaient linéairement de la taille de la population étudiante (x) et d'un terme d'erreur ε. Dans la section 12.4, nous avons fait les hypothèses suivantes sur le terme d'erreur ε.

1. $E(\varepsilon) = 0$.

2. La variance de ε, notée σ^2, est la même pour toutes les valeurs de x.

3. Les valeurs de ε sont indépendantes.

4. Le terme d'erreur ε est normalement distribué.

Tableau 12.7 ***Résidus obtenus pour le problème des restaurants Armand***

Population étudiante x_i	Ventes trimestrielles y_i	Ventes estimées $\hat{y}_i = 60 + 5x_i$	Résidus $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

Ces hypothèses forment la base théorique des tests de Student et de Fisher, utilisés pour déterminer si la relation entre x et y est significative, ainsi que des estimations par intervalle de confiance et de prévision, présentées à la section 12.6. Si les hypothèses sur le terme d'erreur ε sont remises en question, les tests de signification de la relation de régression et les estimations par intervalle peuvent ne pas être corrects.

Les résidus fournissent la meilleure information sur ε ; par conséquent, une analyse des résidus est une étape importante pour déterminer si les hypothèses sur ε sont appropriées. La plus grande part de l'analyse des résidus est basée sur un examen graphique. Dans cette section, nous introduirons les graphiques des résidus suivants.

1. Un graphique des résidus en fonction de la variable indépendante x
2. Un graphique des résidus en fonction des valeurs estimées de la variable dépendante y

12.8.1 Graphique des résidus en fonction de x

Un **graphique des résidus** en fonction de la variable indépendante x est un graphique dont l'axe des abscisses représente les valeurs de la variable indépendante et l'axe des ordonnées les valeurs des résidus. Chaque résidu est représenté par un point. La première coordonnée de chaque point correspond à la valeur de x_i et la seconde coordonnée correspond à la valeur du résidu $y_i - \hat{y}_i$. Les coordonnées du premier point du graphique des résidus, associé à l'exemple des restaurants Armand (cf. tableau 12.7) sont (2, –12) : $x_1 = 2$ et $y_1 - \hat{y}_1 = -12$. Les coordonnées du second point sont (6, 15) : $x_2 = 6$ et $y_2 - \hat{y}_2 = 15$. Et ainsi de suite. La figure 12.11 présente le graphique des résidus obtenu avec les données de l'exemple des restaurants Armand.

Avant d'interpréter ce graphique, considérons les différentes formes de graphique des résidus qui peuvent être observées. Trois formes typiques sont représentées à la figure 12.12. Si l'hypothèse selon laquelle la variance de ε est la même pour toutes les valeurs de x est correcte et si le modèle de régression est une représentation adéquate de la relation entre les variables, le graphique des résidus devrait former une bande de points, comme représenté dans la partie A de la figure 12.12. Par contre, si la variance de ε n'est pas la même pour toutes les valeurs de x – par exemple, si la variabilité de la droite de régression est plus importante pour les plus grandes valeurs de x – on peut observer une forme similaire à celle dessinée dans la partie B de la figure 12.12. Dans ce cas, l'hypothèse d'une variance constante de ε est violée. Une autre forme possible d'un graphique des résidus est présentée dans la partie C. Dans ce cas, on peut conclure que le modèle de régression envisagé n'est pas approprié pour représenter la relation entre les variables. Un modèle de régression curviligne ou un modèle de régression multiple devraient être envisagés.

Revenons au graphique des résidus obtenu dans le cadre de l'exemple des restaurants Armand, figure 12.11. Les résidus semblent avoir la forme horizontale de la partie A de la figure 12.12. Par conséquent, nous en concluons que le graphique des résidus ne fournit pas de preuve remettant en question les hypothèses considérées lors de la constitution du modèle de régression pour l'exemple des restaurants Armand. À ce point de l'analyse, le modèle de régression linéaire simple semble valide.

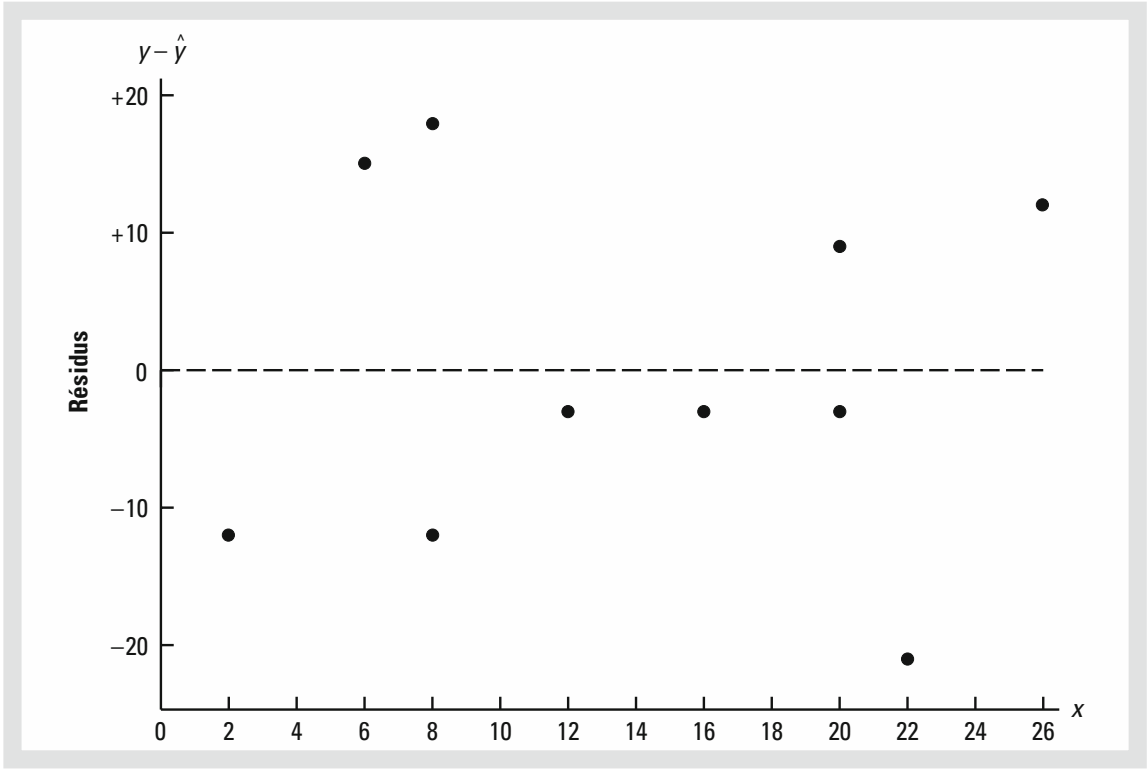

Figure 12.11 ***Graphique des résidus par rapport à la variable indépendante x pour le problème des restaurants Armand***

L'expérience et le bon sens sont des facteurs importants dans l'interprétation des graphiques des résidus. Rarement, un graphique des résidus a l'une des formes présentées à la figure 12.12. Toutefois, les analystes qui effectuent régulièrement des études de la régression et qui analysent des graphiques des résidus, sont à même de pouvoir déterminer les différences entre les formes qui sont raisonnables et celles qui remettent en question les hypothèses du modèle. Un graphique des résidus est l'une des techniques utilisées pour garantir la validité des hypothèses d'un modèle de régression.

12.8.2 Graphique des résidus en fonction de $\hat{y}$

Un autre graphique des résidus représente les valeurs estimées de la variable dépendante $\hat{y}$ sur l'axe des abscisses et les valeurs des résidus sur l'axe des ordonnées. Chaque résidu est représenté par un point. La première coordonnée de chaque point correspond à la valeur de $\hat{y}_i$ et la seconde coordonnée correspond à la valeur du résidu $y_i - \hat{y}_i$. Les coordonnées du premier point du graphique des résidus, associé à l'exemple des restaurants Armand (cf. tableau 12.7) sont (70, –12) : $\hat{y}_1 = 70$ et $y_1 - \hat{y}_1 = -12$. Les coordonnées du second point sont (90, 15) : $\hat{y}_2 = 90$ et $y_2 - \hat{y}_2 = 15$. Et ainsi de suite. La figure 12.13 présente ce graphique des résidus. Notez que la forme de ce graphique des résidus est identique à celle du graphique des résidus en fonction de la variable indépendante x. Il ne s'agit pas d'une forme entraînant la remise en question des hypothèses du modèle. Dans le cadre d'une régression linéaire simple, le graphique des résidus en fonction de x et le graphique

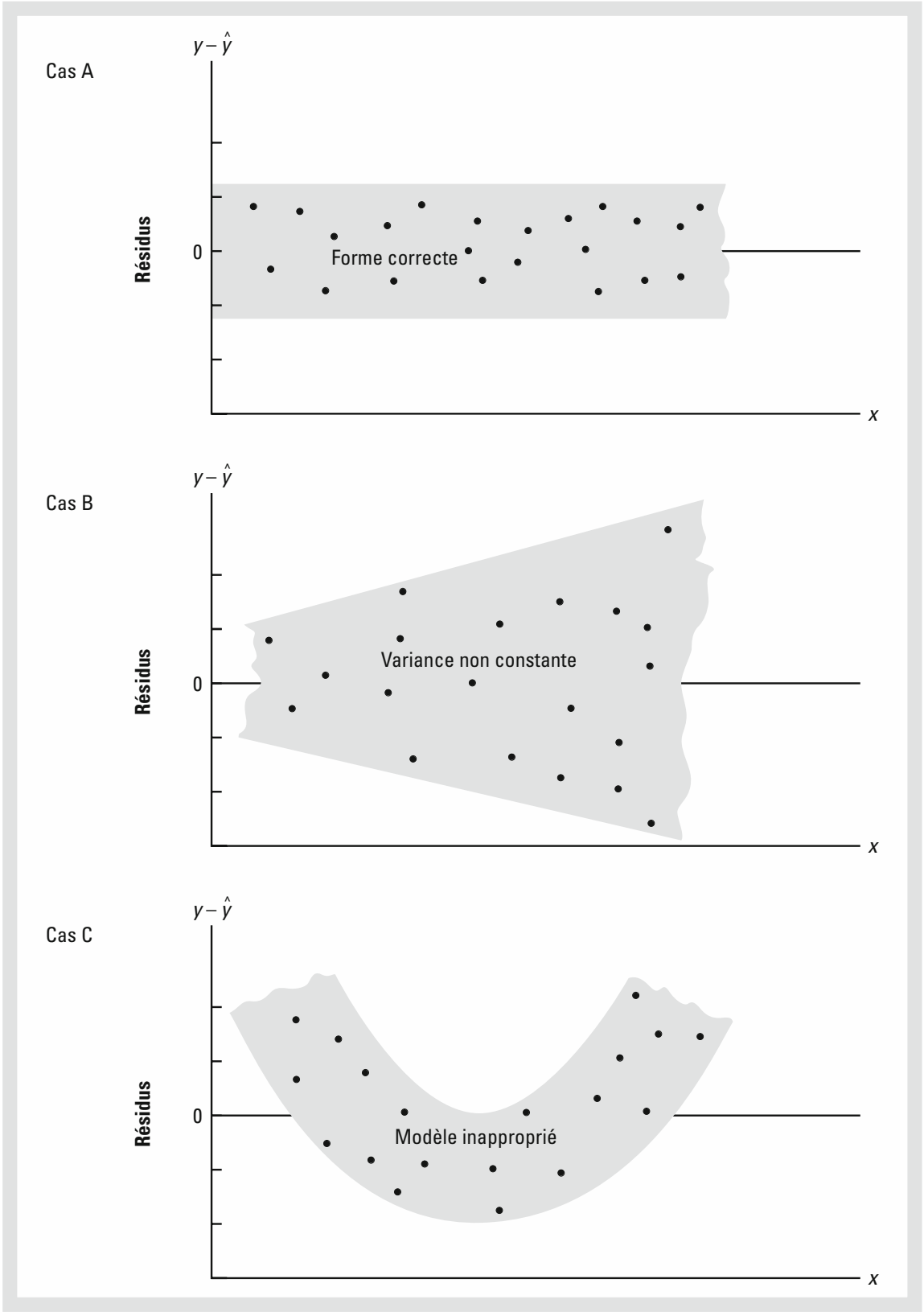

Figure 12.12 ***Graphique des résidus pour trois études de la régression***

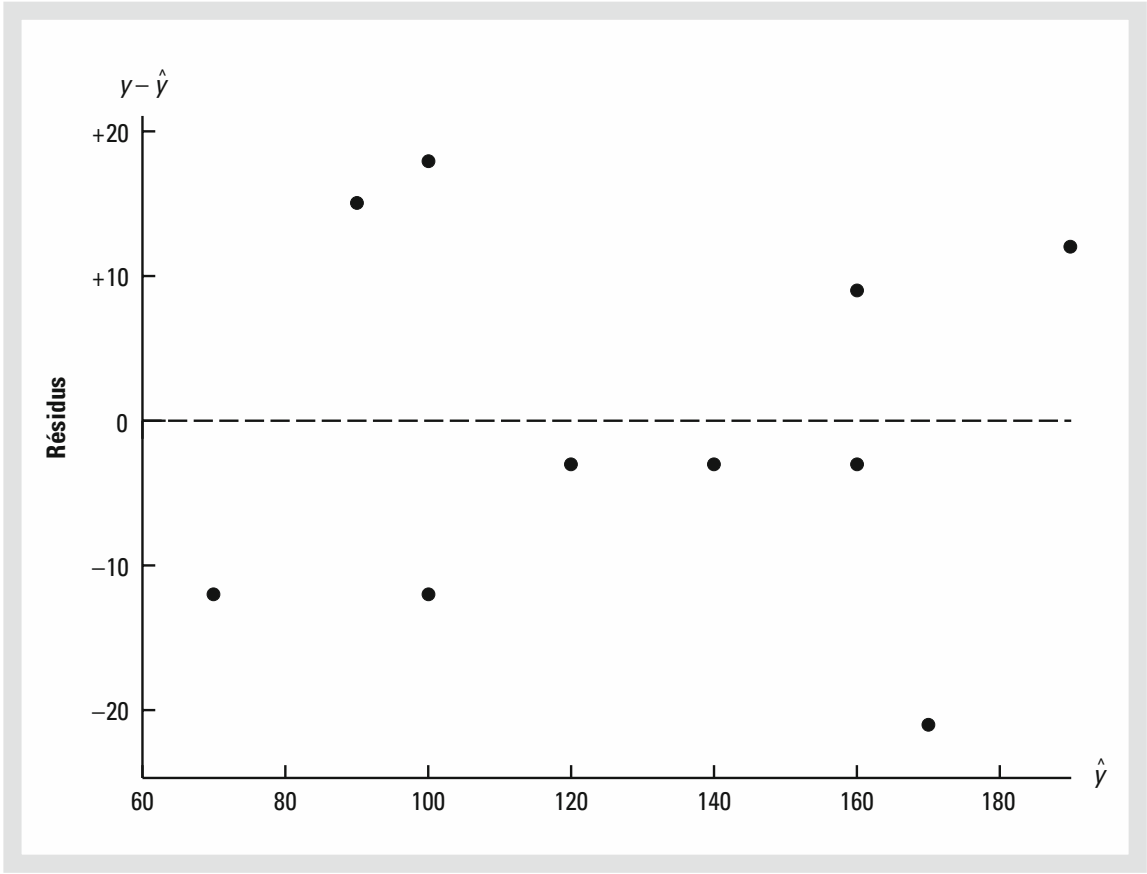

Figure 12.13 ***Graphique des résidus en fonction des valeurs estimées ŷ pour le problème des restaurants Armand***

des résidus en fonction de $\hat{y}$ ont la même forme. Dans le cadre d'une régression multiple, le graphique des résidus en fonction de $\hat{y}$ est plus souvent utilisé, en raison de la présence de plusieurs variables indépendantes.

REMARQUES

1. Nous utilisons les graphiques des résidus pour valider les hypothèses d'un modèle de régression. Si l'analyse des résidus indique qu'une ou plusieurs hypothèses sont contestables, un modèle de régression différent ou une transformation des données doivent être considérés. Les mesures prises lorsque certaines hypothèses ne sont pas vérifiées doivent être basées sur le bon sens ; les recommandations d'un statisticien expérimenté peuvent, à ce titre, être utiles.
2. L'analyse des résidus est la principale méthode que les statisticiens utilisent pour valider les hypothèses associées à un modèle de régression. Même si aucune violation n'est trouvée, il n'est pas certain que le modèle fournisse de bonnes prévisions. Cependant, si les tests statistiques permettent de conclure que les paramètres du modèle sont significatifs et si le coefficient de détermination est important, il devrait être possible de développer de bonnes estimations en utilisant l'équation estimée de la régression.

EXERCICES

Méthode

45. Ci-dessous sont présentées les données de deux variables, x et y.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a) Estimer l'équation de la régression associée à ces données.

b) Calculer les résidus.

c) Dessiner le graphique des résidus par rapport à la variable indépendante x. Les hypothèses concernant les termes d'erreur semblent-elles satisfaites ?

46. Les données suivantes ont été utilisées dans une étude de la régression.

Observation	x_i	y_i
1	2	4
2	3	5
3	4	4
4	5	6
5	7	4

Observation	x_i	y_i
6	7	6
7	7	9
8	8	5
9	9	11

a) Estimer l'équation de la régression associée à ces données.

b) Dessiner le graphique des résidus. Les hypothèses sur le terme d'erreur semblent-elles être satisfaites ?

Applications

47. Dans le tableau suivant sont regroupées des données sur les dépenses publicitaires et le chiffre d'affaires (en milliers de dollars) du restaurant Les Quatre Saisons.

Dépenses publicitaires	Chiffre d'affaires
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a) Soit x les dépenses publicitaires et y le chiffre d'affaires. Utiliser la méthode des moindres carrés pour développer une approximation linéaire de la relation entre les deux variables.

b) Tester l'existence d'une relation significative entre le chiffre d'affaires et les dépenses publicitaires, au seuil de 0,05.

c) Dessiner le graphique des résidus en fonction de la variable dépendante (en fonction de $\hat{y}$).

d) Quelle conclusion pouvez-vous tirer de l'analyse des résidus ? Devrait-on utiliser ce modèle ou en chercher un meilleur ?

48. Reprendre l'exercice 7, dans lequel on a estimé une équation de la régression liant les années d'expérience aux ventes annuelles.

a) Calculer les résidus et dessiner un graphique des résidus pour ce problème.

b) Les hypothèses sur le terme d'erreur semblent-elles raisonnables au regard du graphique des résidus ?

49. En 2011, le prix des maisons et les taux d'emprunt étaient tellement bas que dans un certain nombre de villes, il était moins coûteux d'acheter une maison que de louer un logement. Les données suivantes (cf. fichier en ligne Location-Emprunt) indiquent le loyer moyen demandé sur 10 marchés et le montant mensuel à rembourser suite à l'achat d'une maison au prix médian du marché (incluant les taxes et les assurances) dans 10 villes dans lesquelles le remboursement mensuel moyen d'un emprunt était inférieur au montant moyen des loyers (*The Wall Street Journal*, 26-27 novembre 2011).

Ville	Loyer (en dollars)	Emprunt (en dollars)
Atlanta	840	539
Chicago	1 062	1 002
Detroit	823	626
Jacksonville	779	711
Las Vegas	796	655
Miami	1 071	977
Minneapolis	953	776
Orlando	851	695
Phoenix	762	651
Saint Louis	723	654

a) Estimer l'équation de la régression qui pourrait être utilisée pour prévoir le montant mensuel de remboursement des emprunts étant donné le loyer moyen.

b) Dessiner le graphique des résidus en fonction de la variable indépendante.

c) Les hypothèses sur le terme d'erreur et la forme du modèle semblent-elles raisonnables au regard du graphique des résidus ?

RÉSUMÉ

Dans ce chapitre, nous avons tout d'abord montré comment utiliser l'analyse de la régression pour déterminer la relation entre une variable dépendante *y* et une variable indépendante *x*. Dans une régression linéaire simple, le modèle de régression est

$y = \beta_0 + \beta_1 x + \varepsilon$. L'équation de la régression linéaire simple $E(y) = \beta_0 + \beta_1 x$ décrit la façon dont la moyenne ou l'espérance mathématique de *y* est liée à *x*. Nous avons utilisé les données d'un échantillon et la méthode des moindres carrés pour estimer l'équation de la régression $\hat{y} = b_0 + b_1 x$ où b_0 et b_1 sont les statistiques d'échantillon utilisées pour estimer les paramètres inconnus du modèle β_0 et β_1.

Le coefficient de détermination a été présenté comme une mesure de l'adéquation de l'équation estimée de la régression ; on peut l'interpréter comme la proportion de la variation de la variable dépendante *y* expliquée par l'équation estimée de la régression. Nous avons revu le coefficient de corrélation en tant que mesure de la robustesse d'une relation linéaire entre deux variables.

Les hypothèses concernant le modèle de régression et son terme d'erreur ε ont été examinées et les tests de Student et de Fisher, basés sur ces hypothèses, ont été présentés comme moyens de déterminer si la relation entre deux variables est statistiquement significative. Nous avons montré comment utiliser l'équation estimée de la régression pour construire des intervalles de confiance pour la moyenne de *y* et des intervalles de prévision pour des valeurs individuelles de *y*.

Nous avons finalement montré que les logiciels peuvent faciliter les calculs associés à l'analyse d'une régression linéaire simple et comment l'analyse des résidus permet de valider les hypothèses du modèle.

GLOSSAIRE

Variable dépendante. Variable qui est prédite ou expliquée. Elle est notée *y*.

Variable indépendante. Variable qui permet de prévoir ou d'expliquer la variable dépendante. Elle est notée *x*.

Régression linéaire simple. Analyse de la régression impliquant une variable indépendante et une variable dépendante dont la relation est décrite par une droite.

Modèle de régression. Équation qui décrit comment *y* est lié à *x* et à un terme d'erreur ε ; dans le cadre d'une régression linéaire simple, le modèle de régression est $y = \beta_0 + \beta_1 x + \varepsilon$.

Équation de la régression. Équation qui décrit comment la moyenne ou l'espérance mathématique de la variable dépendante est liée à la variable indépendante ; dans le cadre d'une régression linéaire simple, l'équation de la régression correspond à $E(y) = \beta_0 + \beta_1 x$.

Équation estimée de la régression. Estimation de l'équation de la régression faite à partir des données d'un échantillon en utilisant la méthode des moindres carrés. Dans le cadre d'une régression linéaire simple, l'équation estimée de la régression s'écrit $\hat{y} = b_0 + b_1 x$.

Méthode des moindres carrés. Procédure utilisée pour estimer l'équation de la régression. L'objectif est de minimiser $\sum (y_i - \hat{y}_i)^2$.

Nuage de points. Graphique sur lequel les valeurs de la variable indépendante sont représentées sur l'axe des abscisses et les valeurs de la variable dépendante sur l'axe des ordonnées.

Coefficient de détermination. Mesure de l'adéquation de l'équation estimée de la régression

aux données. Il peut être interprété comme la proportion de la variation de la variable dépendante y, expliquée par l'équation estimée de la régression.

i^e RÉSIDU. Écart entre la valeur observée de la variable dépendante et la valeur obtenue en utilisant l'équation estimée de la régression ; pour la i^e observation, le résidu correspond à $y_i - \hat{y}_i$.

COEFFICIENT DE CORRÉLATION. Mesure de la robustesse de la relation linéaire entre deux variables (cf. chapitre 3).

MOYENNE DES CARRÉS DES RÉSIDUS. Estimation sans biais de σ^2, la variance du terme d'erreur ε. Elle est notée *MCres* ou s^2.

ERREUR TYPE DE L'ESTIMATION. Racine carrée de la moyenne des carrés des résidus, notée s. Il s'agit de l'estimation de σ, l'écart type du terme d'erreur ε.

TABLEAU ANOVA. Tableau d'analyse de la variance utilisé pour résumer les calculs associés au test de signification de Fisher.

INTERVALLE DE CONFIANCE. Estimation par intervalle de la moyenne de y pour une valeur donnée de x.

INTERVALLE DE PRÉVISION. Estimation par intervalle d'une valeur individuelle de y pour une valeur donnée de x.

ANALYSE DES RÉSIDUS. Outil permettant de déterminer si les hypothèses faites sur le modèle de régression sont appropriées. L'analyse des résidus est également utilisée pour identifier les valeurs extrêmes.

GRAPHIQUE DES RÉSIDUS. Représentation graphique des résidus qui peut servir à déterminer si les hypothèses concernant le modèle de régression sont valables.

FORMULES CLÉ

Modèle de régression linéaire simple

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{12.1}$$

Équation de la régression linéaire simple

$$E(y) = \beta_0 + \beta_1 x \tag{12.2}$$

Équation estimée de la régression linéaire simple

$$\hat{y} = b_0 + b_1 x \tag{12.3}$$

Critère des moindres carrés

$$\min_y \sum (y_i - \hat{y}_i)^2 \tag{12.5}$$

Pente et ordonnée à l'origine de l'équation estimée de la régression

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \tag{12.6}$$

$$b_0 = \overline{y} - b_1 \overline{x} \tag{12.7}$$

Somme des carrés des résidus

$$SCres = \sum (y_i - \hat{y}_i)^2 \tag{12.8}$$

Somme des carrés totale

$$SCT = \sum (y_i - \overline{y})^2 \tag{12.9}$$

Somme des carrés de la régression

$$SCreg = \sum (\hat{y}_i - \overline{y})^2 \tag{12.10}$$

Relation entre SCT, SCreg et SCres

$$SCT = SCreg + SCres \tag{12.11}$$

Coefficient de détermination

$$r^2 = \frac{SCreg}{SCT} \tag{12.12}$$

Coefficient de corrélation d'un échantillon

$$\begin{aligned} r_{xy} &= (\text{signe de } b_1)\sqrt{\text{Coefficient de détermination}} \\ &= (\text{signe de } b_1)\sqrt{r^2} \end{aligned} \tag{12.13}$$

Moyenne des carrés des résidus (estimation de σ^2)

$$s^2 = MCres = \frac{SCres}{n-2} \tag{12.15}$$

Erreur type de l'estimation

$$s = \sqrt{MCres} = \sqrt{\frac{SCres}{n-2}} \tag{12.16}$$

Écart type de b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}} \tag{12.17}$$

Écart type estimé de b_1

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \overline{x})^2}} \tag{12.18}$$

Statistique de test de Student

$$t = \frac{b_1}{s_{b_1}} \tag{12.19}$$

Moyenne des carrés de la régression

$$MCreg = \frac{SCreg}{\text{Nombre de variables indépendantes}} \tag{12.20}$$

Statistique de test de Fisher

$$F = \frac{MCreg}{MCres} \tag{12.21}$$

Écart type estimé de $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}} \tag{12.23}$$

Intervalle de confiance de $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{12.24}$$

Écart type estimé d'une valeur individuelle

$$s_{prev} = s\sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}} \tag{12.26}$$

Intervalle de prévision de y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{prev} \tag{12.27}$$

Résidu de l'observation *i*

$$y_i - \hat{y}_i \tag{12.28}$$

EXERCICES SUPPLÉMENTAIRES

50. Les indices Dow Jones Industriel (DJIA) et Standard & Poor's 500 (S&P500) sont des indicateurs des mouvements sur le marché boursier. Le DJIA est basé sur les variations de prix des 30 plus grandes sociétés ; le S&P500 est un indice composé de 500 actions. Certains disent que le S&P500 est un meilleur indicateur des performances du marché boursier dans la mesure où il est plus large. Les prix de clôture des indices DJIA et S&P500 durant 15 semaines, à partir du 6 janvier 2012 (site Internet de *Barron's*, 17 avril 2012) sont fournis ci-dessous (cf. fichier en ligne DJIAS&P500).

DJIAS&P500

Date	DJIA	S&P
6 janvier	12 360	1 278
13 janvier	12 422	1 289
20 janvier	12 720	1 315
27 janvier	12 660	1 316
3 février	12 862	1 345
10 février	12 801	1 343
17 février	12 950	1 362
24 février	12 983	1 366
2 mars	12 978	1 370
9 mars	12 922	1 371
16 mars	13 233	1 404
23 mars	13 081	1 397
30 mars	13 212	1 408
5 avril	13 060	1 398
13 avril	12 850	1 370

a) Représenter un nuage de points avec l'indice DJIA comme variable indépendante.

b) Déterminer l'équation estimée de la régression.

c) Au seuil de signification de 0,05, existe-t-il une relation significative entre les deux variables ?

d) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

e) Supposez que le prix de clôture pour le DJIA soit de 13 500. Prédire le prix de clôture du S&P500.

f) Doit-on s'inquiéter du fait que la valeur de 13 500 associée au DJIA utilisée pour prévoir la valeur de l'indice S&P500 à la question (e) soit hors du champ des données utilisées pour estimer l'équation de la régression ?

51. Les données suivantes (cf. fichier en ligne Stocks500) indiquent l'estimation faite par Morningstar de la valeur des actions et le prix de l'action pour 28 sociétés. La valeur attribuée par Morningstar est une estimation de la valeur des actions de la société qui tient compte des prévisions de croissance de la société au cours des cinq années suivantes, de sa rentabilité, de son niveau de risque et d'autres facteurs (*Morningstar Stocks 500*, édition 2008).

Société	Valeur Morningstar (en dollars)	Prix des actions (en dollars)
Air Products and Chemical	80	98,63
Allied Waste Industries	17	11,02
America Mobile	83	61,39
AT&T	35	41,56
Bank of America	70	41,26
Barclays PLC	68	40,37
Citigroup	53	29,44
Costco Wholesale Corp.	75	69,76
Covidien, Ltd.	58	44,29
Darden Restaurants	52	27,71
Dun & Bradstreet	87	88,63
Equifax	42	36,36
Gannett Co.	38	39,00
Guine Parts	48	46,30
GloxoSmithKline PLC	57	50,39
Iron Mountain	33	37,02
ITT Corporation	83	66,04
Johnson & Johnson	80	66,70
Las Vegas Sands	98	103,05
Macrovision	23	18,33
Marrott International	39	34,18
Nalco Holding Company	29	24,18
National Interstate	25	33,10
Portugal Telecom	15	13,02
Qualcomm	48	39,35
Royal Dutch Shell Ltd.	87	84,20
SanDisk	60	33,17
Time Warner	42	27,60

a) Déterminer l'équation estimée de la régression qui peut être utilisée pour estimer le prix des actions en fonction de leur valeur.

b) Au seuil de signification de 0,05, existe-t-il une relation significative entre les deux variables ?

c) Utiliser l'équation estimée de la régression pour estimer le prix des actions d'une société dont la valeur est estimée à 50 dollars par Morningstar.

d) Pensez-vous que l'équation estimée de la régression fournit une bonne prévision du prix des actions ? Utiliser le coefficient de détermination pour étayer votre réponse.

52. Un des principaux changements dans l'éducation supérieure intervenus ces dernières années est l'apparition d'un nombre croissant d'universités en ligne. « Online Education Database » est une organisation indépendante dont la mission est de constituer une liste exhaustive des écoles et universités en ligne agréées. Le tableau suivant (cf. fichier en ligne Éducation en ligne) indique le taux de redoublement (%) et le taux de diplômés (%) pour 29 écoles en ligne (site Internet de Online Education Database, janvier 2009).

École	Taux de redoublement (%)	Taux de diplômés (%)
Université internationale de l'Ouest	7	25
Université du Sud	51	25
Université de Phoenix	4	28
Université intercontinentale américaine	29	32
Université de Franklin	33	33
Université de Devry	47	32
Université de Tiffin	63	34
Université de Post	45	36
Pierce College	60	36
Université Everest	62	36
Université de l'Iowa	67	36
Université d'État Dickinson	65	37
Université des gouverneurs de l'Ouest	78	37
Université Kaplan	75	38
Université internationale de Salem	54	39
Université Ashford	45	41
Institut technologique ITT	38	44
Berkeley College	51	45
Université du Grand Canyon	69	46
Université Nova	60	47
Westwood College	37	48
Université des Everglades	63	50
Université Liberty	73	51
Université LeTourneau	78	52
Rasmussen College	48	53
Université Keiser	95	55
Herzing College	68	56
Université nationale	100	57
Collège national de Floride	100	61

a) Représenter le nuage de points de cet ensemble de données, en prenant pour variable indépendante le taux de redoublement. Qu'indique le nuage de points à propos de la relation entre les deux variables ?

b) Estimer l'équation de la régression.

c) Tester l'existence d'une relation significative au seuil de 0,05.

d) L'équation estimée de la régression est-elle bien adaptée aux données ?

e) Supposez que vous soyez le doyen de l'Université du Sud. Après avoir revu les résultats, devriez-vous être inquiet de la performance de votre université comparée à celle des autres universités en ligne ?

f) Supposez que vous soyez le doyen de l'Université de Phoenix. Après avoir revu les résultats, devriez-vous être inquiet de la performance de votre université comparée à celle des autres universités en ligne ?

53. Jensen Tire & Auto s'interroge sur l'opportunité de signer un contrat de maintenance pour son nouvel appareil d'alignement et d'équilibrage des pneus. Les dirigeants pensent que le coût de la maintenance de cet appareil est lié à l'usage qui en ait fait et ont collecté des informations (cf. fichier en ligne Jensen) sur l'usage hebdomadaire (en heures) et le coût annuel de maintenance (en milliers de dollars).

Usage hebdomadaire (en heures)	Coût annuel de maintenance
13	17,0
10	22,0
20	30,0
28	37,0
32	47,0
17	30,5
24	32,5
31	39,0
40	51,5
38	40,0

a) Estimer l'équation de la régression qui relie le coût annuel de maintenance à l'usage hebdomadaire.

b) Tester la significativité de la relation obtenue à la question (a) au seuil de 0,05.

c) Jensen pense utiliser la nouvelle machine 30 heures par semaine. Construire un intervalle de prévision à 95 % du coût annuel de maintenance pour la société.

d) Si le coût du contrat de maintenance s'élève à 3 000 dollars par an, recommanderiez-vous de le signer ? Pourquoi ?

54. L'autorité de transport régional d'une grande métropole souhaite déterminer s'il existe une relation entre l'âge d'un bus et son coût annuel de maintenance. Un échantillon de 10 bus fournit les données suivantes (cf. fichier en ligne Âge-Coût).

Âge du bus (années)	Coût de maintenance ($)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

Âge-Coût

a) Déterminer l'équation estimée de la régression.

b) Au seuil de 0,05, déterminer si les deux variables sont significativement liées.

c) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

d) Construire un intervalle de prévision à 95 % du coût de maintenance d'un bus particulier âgé de 4 ans.

55. Reuters rapportait que la valeur bêta du marché de la société Xerox était égale à 1,22 (site Internet de Reuters, 30 janvier 2009). Les valeurs bêta du marché pour des titres individuels sont déterminées par une régression linéaire simple. Pour chaque action, la variable dépendante correspond à son rendement trimestriel, en pourcentage (accroissement du capital plus les dividendes) moins le rendement en pourcentage obtenu d'un investissement sans risque (le taux des bons du trésor est utilisé comme taux sans risque). La variable indépendante correspond à la rentabilité de l'ensemble du marché. Une équation de la régression est estimée avec les données trimestrielles : la valeur bêta du marché pour l'action considérée correspond à la pente de l'équation estimée de la régression (b_1). La valeur bêta du marché est souvent interprétée comme une mesure du risque associé à l'action. Les valeurs bêta supérieures à 1 indiquent que l'action est plus volatile que la moyenne du marché ; les valeurs inférieures à 1 indiquent que l'action est moins volatile que la moyenne du marché. Les écarts entre le rendement en pourcentage et le rendement sans risque, au cours de 10 trimestres, pour les actions S&P500 et Horizon Technology sont présentés ci-dessous (cf. fichier en ligne Bêta du marché).

S&P500	Horizon
1,2	-0,7
-2,5	-2,0
-3,0	-5,5
2,0	4,7
5,0	1,8
1,2	4,1
3,0	2,6
-1,0	2,0
0,5	-1,3
2,5	5,5

Bêta du marché

a) Déterminer l'équation estimée de la régression qui peut être utilisée pour calculer la valeur bêta pour Horizon Technology. Quelle est la valeur bêta pour Horizon Technology ?

b) Tester l'existence d'une relation significative au seuil de 0,05.

c) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

d) Utiliser les valeurs bêta de Xerox et Horizon Technology pour comparer les risques associés à ces deux actions.

56. La Toyota Camry est l'une des voitures les plus vendues aux États-Unis. Le prix de revente d'une Camry d'occasion dépend d'un certain nombre de facteurs, comme l'année du modèle, le kilométrage et son état général. Dans le but d'étudier la relation entre le kilométrage d'un modèle de 2007 et son prix de revente, les données suivantes sur le kilométrage et le prix de revente de 19 Camry d'occasion (cf. fichier en ligne Camry) ont été collectées (site Internet de PriceHub, 24 février 2012).

Kilométrage (en milliers de miles)	Prix (en milliers de dollars)
22	16,2
29	16,0
36	13,8
47	11,5
63	12,5
77	12,9
73	11,2
87	13,0
92	11,8
101	10,8
110	8,3
28	12,5
59	11,1
68	15,0
68	12,2
91	13,0
42	15,6
65	12,7
110	8,3

a) Représenter un nuage de points avec le kilométrage sur l'axe horizontal et le prix sur l'axe vertical.

b) Qu'indique le nuage de points sur la relation entre les deux variables ?

c) Déterminer l'équation estimée de la régression qui peut être utilisée pour prévoir le prix en fonction du kilométrage.

d) Au seuil de 0,05, déterminer s'il existe une relation significative entre les deux variables.

e) L'équation estimée de la régression est-elle bien adaptée aux données ? Expliquer.

f) Interpréter la pente de l'équation estimée de la régression.

g) Supposez que vous envisagiez l'achat d'une Camry de 2007 d'occasion qui a 60 000 miles au compteur. Utiliser l'équation estimée de la régression déterminée à la question (c) pour prédire le prix de cette voiture. Est-ce le prix que vous souhaitez offrir au vendeur ?

57. Une enquête menée en 2012 par IdeaWorks a fourni des données indiquant le pourcentage de sièges disponibles lorsque les consommateurs souhaitent échanger des points ou des miles contre un voyage gratuit (cf. fichier en ligne Sièges Compagnies aériennes). Pour chaque compagnie aérienne listée, la colonne intitulée Pourcentage 2011 indique le pourcentage de sièges disponibles en 2011 et la colonne intitulée Pourcentage 2012 fournit les pourcentages correspondants en 2012 (*The Wall Street Journal*, 17 mai 2012).

Compagnie	Pourcentage 2011	Pourcentage 2012
Air Berlin	96,4	100,0
Air Canada	82,1	78,6
Air France KLM	65,0	55,7
AirTran Airways	47,1	87,1
Alaska Airlines	64,3	59,3
American Airlines	62,9	45,7
British Airways	61,4	79,3
Cathay Pacific	66,4	70,7
Delta Air Lines	27,1	27,1
Emirates	35,7	32,9
GOL Airlines (Brésil)	100,0	97,1
Iberia	70,7	63,6
JetBlue	79,3	86,4
LAN (Chili)	75,7	78,6
Lufthansa, Suisse, Autriche	85,0	92,1
Qantas	75,0	78,6
SAS Scandinavian	52,9	57,9
Singapore Airlines	90,7	90,7
Southwest	99,3	100,0
Turkish Airways	49,3	38,6
United Airlines	71,4	87,1
US Airways	25,7	33,6
Virgin Australia	91,4	90,0

a) Représenter le nuage de points de cet ensemble de données en prenant le pourcentage 2011 comme variable indépendante.

b) Qu'indique le nuage de points de la question (a) quant à la relation entre les deux variables ?

c) Estimer l'équation de la régression.

d) Tester l'existence d'une relation significative au seuil de 0,05.

e) L'équation estimée de la régression est-elle bien adaptée aux données ?

f) Représenter un graphique des résidus. Commenter la forme du graphique ainsi que tout point qui vous semble inhabituel.

PROBLÈME 1 *Mesurer le risque sur le marché boursier*

L'écart type du rendement global (appréciation du capital plus dividendes) sur plusieurs périodes constitue une mesure du risque ou de la volatilité d'une action. Bien que l'écart type soit facile à calculer, il ne prend pas en compte l'ampleur à laquelle le prix d'une action varie en fonction d'un indice du marché, tel que le S&P 500. En conséquence, beaucoup d'analystes financiers préfèrent utiliser une autre mesure du risque appelée *bêta*.

Les valeurs bêta des actions sont déterminées par une simple régression linéaire. La variable dépendante correspond au rendement total d'une action et la variable indépendante correspond au rendement total du marché boursier[3]. Dans le cadre de ce problème, nous utiliserons l'indice S&P 500 comme mesure du rendement total du marché boursier et une équation estimée de la régression sera déduite de données mensuelles. La valeur bêta d'une action correspond à la pente de l'équation estimée de la régression (b_1). Le fichier en ligne Bêta fournit le rendement total (appréciation du capital plus dividendes) sur 36 mois de huit actions fréquemment échangées et de l'indice S&P 500.

La valeur bêta du marché boursier est toujours égale à 1 ; ainsi, les actions qui ont tendance à varier de façon similaire au marché boursier auront également un bêta proche de 1. Les bêtas supérieurs à 1 indiquent que l'action est plus volatile que le marché. Par exemple, si une action a un bêta de 1,4, elle est 40 % plus volatile que le marché, et si une action a un bêta de 0,4, elle est 60 % moins volatile que le marché.

Rapport

Vous êtes chargé d'analyser les caractéristiques de risque de ces actions. Préparez un rapport qui inclut mais ne se limite pas aux éléments suivants.

1. Calculez les statistiques descriptives pour chaque action et l'indice S&P 500. Commentez vos résultats. Quelles actions sont les plus volatiles ?

2. Calculez la valeur bêta de chaque action. Lesquelles sont les plus performantes sur un marché en croissance, selon vous ? Lesquelles seraient les plus performantes sur un marché en décroissance, selon vous ?

3. Discutez de la part du rendement des actions individuelles expliquée par le marché.

3 Des sources différentes utilisent des approches différentes pour calculer les valeurs bêta. Par exemple, certaines sources soustraient le rendement qui peut être obtenu d'un investissement sans risque (par exemple, les bons du Trésor) à la variable dépendante et à la variable indépendante avant de calculer l'équation estimée de la régression. D'autres sources utilisent différents indices du rendement total du marché boursier ; par exemple, *Value Line* calcule les valeurs bêta en utilisant l'indice composite de la bourse de New York.

PROBLÈME 2 *Le ministère américain des transports*

Dans le cadre d'une étude sur la sécurité des transports, le ministère américain des transports a collecté des données sur la proportion d'accidents mortels sur 1 000 permis de conduire et le pourcentage de conducteurs, détenteurs d'un permis, âgés de moins de 21 ans dans un échantillon de 42 villes. Les données collectées sur une période d'un an sont présentées ci-dessous. Ces données sont disponibles en ligne dans le fichier Sécurité.

Pourcentage de conducteurs âgés de moins de 21 ans	Accidents mortels sur 1 000 permis de conduire	Pourcentage de conducteurs âgés de moins de 21 ans	Accidents mortels sur 1 000 permis de conduire
13	2,962	17	4,100
12	0,708	8	2,190
8	0,885	16	3,623
12	1,652	15	2,623
11	2,091	9	0,835
17	2,627	8	0,820
18	3,830	14	2,890
8	0,368	8	1,267
13	1,142	15	3,224
8	0,645	10	1,014
9	1,028	10	0,493
16	2,801	14	1,443
12	1,405	18	3,614
9	1,433	10	1,926
10	0,039	14	1,643
9	0,338	16	2,943
11	1,849	12	1,913
12	2,246	15	2,814
14	2,855	13	2,634
14	2,352	9	0,926
11	1,294	17	3,256

Rapport

1. Résumez sous forme numérique et graphique les données.
2. Utilisez l'analyse de la régression pour étudier la relation entre le nombre d'accidents mortels et le pourcentage de conducteurs âgés de moins de 21 ans. Commentez vos résultats.
3. Quelles conclusions ou recommandations pouvez-vous tirer de votre analyse ?

PROBLÈME 3 *Choisir un appareil photo numérique*

Consumer Reports a testé 166 appareils photo numériques. Sur la base de facteurs tels que le nombre de pixels, le poids (onces), la qualité d'image et la facilité d'utilisation, ils ont attribué une note à chaque appareil testé. Les notes vont de 0 à 100, des notes élevées indiquant de meilleurs résultats aux tests. Choisir un appareil peut être difficile et le prix est certainement un critère de choix pour la plupart des consommateurs. En dépensant plus, un consommateur acquière-t-il un appareil de meilleure qualité ? Les appareils qui ont plus de pixels, un facteur souvent considérés comme une bonne mesure de la qualité de l'image, coûtent-ils plus cher que les appareils qui en ont moins ? Le tableau 12.8 (cf. fichier en ligne Appareils photo) indique la marque, le prix de vente moyen (en dollars), le nombre de pixels, le poids (en onces) et la note de 13 appareils photo Canon et 15 appareils Nikon testés par *Consumer Reports* (site Internet de *Consumer Reports*, 7 février 2012).

Tableau 12.8 ***Données pour 28 appareils photo numériques***

Observations	Marque	Prix ($)	Nombre de pixels	Poids (onces)	Note
1	Canon	330	10	7	66
2	Canon	200	12	5	66
3	Canon	300	12	7	65
4	Canon	200	10	6	62
5	Canon	180	12	5	62
6	Canon	200	12	7	61
7	Canon	200	14	5	60
8	Canon	130	10	7	60
9	Canon	130	12	5	59
10	Canon	110	16	5	55
11	Canon	90	14	5	52
12	Canon	100	10	6	51
13	Canon	90	12	7	46
14	Nikon	270	16	5	65
15	Nikon	300	16	7	63
16	Nikon	200	14	6	61
17	Nikon	400	14	7	59
18	Nikon	120	14	5	57
19	Nikon	170	16	6	56
20	Nikon	150	12	5	56
21	Nikon	230	14	6	55
22	Nikon	180	12	6	53
23	Nikon	130	12	6	53
24	Nikon	80	12	7	52

(*suite*)

Observations	Marque	Prix ($)	Nombre de pixels	Poids (onces)	Note
25	Nikon	80	14	7	50
26	Nikon	100	12	4	46
27	Nikon	110	12	5	45
28	Nikon	130	14	4	42

Rapport

1. Résumez sous forme numérique les données.
2. En utilisant la note comme variable dépendante, représentez trois diagrammes de points, l'un en utilisant le prix comme variable indépendante, l'un en utilisant le nombre de pixels comme variable indépendante et le dernier, en utilisant le poids comme variable indépendante. Laquelle de ces trois variables indépendantes semble être le meilleur inducteur de la note ?
3. En utilisant la régression linéaire simple, estimez l'équation de la régression qui permettrait de prévoir la note en fonction du prix de l'appareil photo. Pour cette équation estimée de la régression, analysez les résidus et discutez de vos résultats.
4. Analysez les données en utilisant uniquement les observations relatives aux appareils Canon. Discutez de la pertinence d'utiliser une régression linéaire simple. Quelles sont vos recommandations au regard des prévisions que l'on peut faire de la note à partir simplement du prix de l'appareil photo ?

PROBLÈME 4 *Trouver la meilleure offre pour une voiture*

Lorsque vous devez choisir quelle voiture acheter, la valeur réelle ne correspond pas nécessairement au coût d'achat. En effet, les voitures qui sont fiables et qui ne coûtent pas trop chères à l'entretien, représentent souvent les meilleures affaires. Mais, quels que soient son degré de fiabilité et son coût d'entretien, elle doit bien fonctionner.

Pour mesurer la valeur, *Consumer Reports* a construit une statistique appelée score de valeur. Le score de valeur est basé sur les coûts d'entretien sur cinq ans, les notes attribuées lors des tests sur route et les évaluations quant à la fiabilité du véhicule. Les coûts d'entretien sur cinq ans sont basés sur les dépenses supportées la première année, dont la dépréciation du véhicule, la consommation de carburant, les réparations, etc. En utilisant une moyenne nationale de 12 000 kilomètres parcourus par an, un coût moyen au kilomètre est utilisé pour mesurer les coûts d'entretien sur cinq ans. Les notes attribuées lors des tests sur route sont le résultat de plus de 50 tests et les notes vont de 0 à 100, les notes les plus élevées indiquant une meilleure performance, un meilleur confort, une meilleure praticité et une moindre consommation de carburant. La note la plus élevée a

été attribuée à la Lexus LS 460L (une note de 99 sur 100). Les évaluations relatives à la fiabilité (1 = mauvaise, 2 = convenable, 3 = bonne, 4 = très bonne et 5 = excellente) sont basées sur les données issues de l'enquête « auto » annuelle de *Consumer Reports*.

Une voiture ayant un score de valeur de 1,0 est considérée comme une « valeur moyenne ». Une voiture dont le score de valeur est de 2,0 est considérée être deux fois meilleure qu'une voiture dont le score est de 1,0 ; une voiture dont le score est de 0,5 est considérée comme moitié moins bonne que la moyenne, et ainsi de suite. Les données pour 20 berlines familiale, incluant le prix (en dollars) de chaque voiture testée, sont fournies ci-dessous (cf. fichier en ligne Berlines familiales).

Voiture	Prix ($)	Coût au km	Test sur route	Fiabilité	Score de valeur
Nissan Altima 2.5 S (4 cylindres)	23 970	0,59	91	4	1,75
Kia Optima LX (2.4)	21 885	0,58	81	4	1,73
Subaru Legacy 2.5i Premium	23 830	0,59	83	4	1,73
Ford Fusion Hybrid	32 360	0,63	84	5	1,70
Honda Accord LX-P (4 cylindres)	23 730	0,56	80	4	1,62
Mazda6 i Sport (4 cylindres)	22 035	0,58	73	4	1,60
Hyundai Sonata GLS (2.4)	21 800	0,56	89	3	1,58
Ford Fusion SE (4 cylindres)	23 625	0,57	76	4	1,55
Chevrolet Malibu LT (4 cylindres)	24 115	0,57	74	3	1,48
Kia Optima SK (2.0T)	29 050	0,72	84	4	1,43
Ford Fusion SEL (V6)	28 400	0,67	80	4	1,42
Nissan Altima 3.5 SR (V6)	30 335	0,69	93	4	1,42
Hyundai Sonata Limited (2.0T)	28 090	0,66	89	3	1,39
Honda Accord EX-L (V6)	28 695	0,67	90	3	1,36
Mazda6 s Grand Touring (V6)	30 790	0,74	81	4	1,34
Ford Fusion SEL (V6, AWD)	30 055	0,71	75	4	1,32
Subaru Legacy 3.6R Limited	30 094	0,71	88	3	1,29
Chevrolet Malibu LTZ (V6)	28 045	0,67	83	3	1,20
Chrysler 200 Limited (V6)	27 825	0,70	52	5	1,20
Chevrolet Impala LT (3.6)	28 995	0,67	63	3	1,05

Rapport

1. Résumez sous forme numérique les données.
2. Utilisez l'analyse de la régression pour estimer l'équation de la régression qui pourrait être utilisée pour prévoir le score de valeur étant donné le prix de la voiture.
3. Utilisez l'analyse de la régression pour estimer l'équation de la régression qui pourrait être utilisée pour prévoir le score de valeur étant donnés les coûts d'entretien sur cinq ans (coût au km).

4. Utilisez l'analyse de la régression pour estimer l'équation de la régression qui pourrait être utilisée pour prévoir le score de valeur étant donnée la note attribuée lors des tests sur route.
5. Utilisez l'analyse de la régression pour estimer l'équation de la régression qui pourrait être utilisée pour prévoir le score de valeur étant données les estimations en termes de fiabilité.
6. Quelles conclusions pouvez-vous tirer de votre analyse ?

ANNEXE 12.1 ANALYSE DE LA RÉGRESSION AVEC MINITAB

Dans la section 12.7, nous avons présenté le résultat du problème de régression associé aux restaurants Armand, obtenu avec Minitab (cf. fichier en ligne Armand). Dans cette annexe, nous décrirons les différentes étapes qui permettent d'obtenir ce résultat. Tout d'abord, on entre les données dans une feuille de calcul Minitab. Les données sur la population étudiante sont enregistrées dans la colonne C1 et les ventes trimestrielles dans la colonne C2. Les noms des variables POP et SALES correspondent au titre des colonnes. Dans les étapes suivantes, on utilise le nom des variables POP et SALES ou le numéro des colonnes C1 et C2 pour désigner les données. Les étapes suivantes décrivent la façon d'utiliser Minitab pour obtenir les résultats de la régression présentés dans la figure 12.10.

Étape 1. Sélectionner le menu **Stat**
Étape 2. Sélectionner le menu **Regression**
Étape 3. Choisir l'option **Regression**
Étape 4. Lorsque la boîte de dialogue Regression apparaît
Entrer SALES dans la boîte **Response**
Entrer POP dans la boîte **Predictors**
Cliquer sur le bouton **Options**
Lorsque la boîte de dialogue Regression-Options apparaît
Entrer 10 dans la boîte **Prediction intervals for new observations**
Cliquer sur **OK**
Lorsque la boîte de dialogue Regression apparaît
Sélectionner **OK**

La boîte de dialogue de régression Minitab fournit des informations supplémentaires, obtenues en sélectionnant les options désirées. Par exemple, pour obtenir un graphique des résidus qui indique la valeur prévue de la variable dépendante $\hat{y}$ sur l'axe horizontal et les résidus sur l'axe vertical, l'étape 4 devient :

Étape 4. Lorsque la boîte de dialogue Regression apparaît
Entrer SALES dans la boîte **Response**
Entrer POP dans la boîte **Predictors**
Cliquer sur le bouton **Graphs**

Lorsque la boîte de dialogue Regression-Graphs apparaît
Sélectionner **Regular** dans Residuals for Plots
Sélectionner **Residuals versus fits** dans Residual Plots
Cliquer sur **OK**
Lorsque la boîte de dialogue Regression apparaît
Sélectionner **OK**

ANNEXE 12.2 ANALYSE DE LA RÉGRESSION AVEC EXCEL

Décrivons l'analyse de la régression effectuée en utilisant Excel dans le cadre du problème des restaurants Armand (cf. fichier en ligne Armand). Référez-vous à la figure 12.14. Les noms Restaurant, Population et Ventes sont enregistrés dans les cellules A1:C1 d'une feuille de calcul Excel. Pour identifier chacune des dix observations, nous avons entré les chiffres 1 à 10 dans les cellules A2:A11. Les données d'échantillon sont entrées dans les cellules B2:C11. Les étapes suivantes décrivent comment utiliser Excel pour obtenir les résultats de la régression.

Étape 1. Cliquer sur le bouton **Data** dans la barre des tâches
Étape 2. Dans le groupe **Analysis**, cliquer sur **Data Analysis**
Étape 3. Choisir **Regression** dans la liste Analysis Tools
Étape 4. Cliquer sur **OK**
Étape 5. Lorsque la boîte de dialogue Regression apparaît
Entrer C1:C11 dans la boîte **Input Y Range**
Entrer B1:B11 dans la boîte **Input X Range**
Sélectionner **Labels**
Sélectionner **Confidence Level**
Entrer 99 dans la boîte **Confidence Level**
Sélectionner **Output Range**
Entrer A13 dans la boîte **Output Range**
(Cellule dans le coin gauche supérieur indiquant où commence l'affichage des résultats)
Cliquer sur **OK**

La première partie de la feuille de résultats, intitulée Statistiques de la régression, contient des statistiques descriptives telles que le coefficient de détermination (R^2). La deuxième partie, intitulée ANOVA, contient le tableau d'analyse de la variance. La dernière partie, qui n'a pas de titre, contient les coefficients estimés de la régression. Nous commençons notre discussion par l'interprétation des résultats de la régression en décrivant l'information contenue dans les cellules A28:I30.

	A	B	C	D	E	F	G	H	I	J
1	**Restaurant**	**Population**	**Ventes**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	RÉSULTATS									
14										
15	*Statistiques de la régression*									
16	Multiple R	0,9501								
17	R Square	0,9027								
18	Ajusted R square	0,8906								
19	Erreur type	13,8293								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Régression	1	14200	14200	74,2484	2,55E-05				
25	Résidus	8	1530	191,25						
26	Total	9	15730							
27										
28		*Coefficients*	*Erreur type*	*Statistique t*	*Valeur p*	*Inférieur 95 %*	*Supérieur 95 %*	*Inférieur 99 %*	*Supérieur 99 %*	
29	Constante	60	9,2260	6,5033	0,0002	38,7247	81,2753	29,0431	90,9569	
30	Population	5	0,5803	8,6167	2,55E-05	3,6619	6,3381	3,0530	6,9470	
31										

Figure 12.14 ***Résultat obtenu avec Excel dans le cadre du problème des restaurants Armand***

Interprétation des résultats de l'équation estimée de la régression

La valeur de la constante de la droite estimée de la régression, $b_0 = 60$, est indiquée dans la cellule B29 et la pente de la droite estimée de la régression, $b_1 = 5$, est reportée dans la cellule B30. Les noms « Constante » dans la cellule A29 et « Population » dans la cellule A30 identifient ces deux valeurs.

Dans la section 12.5 nous avons montré que l'écart type estimé de b_1 est $s_{b_1} = 0{,}5803$. Notez que la valeur de la cellule C30 est 0,5803. Le terme Erreur type dans la cellule C28 est la façon qu'a Excel d'indiquer que la valeur de la cellule C30 est l'erreur type ou l'écart type de b_1. Souvenez-vous que le test de Student d'une relation significative a nécessité le calcul de la statistique de test $t = b_1 / s_{b_1}$. Pour les données des restaurants Armand, la valeur t que nous avions calculée s'élevait à $t = 5/0{,}5803 = 8{,}62$. Le terme de la cellule D28, Statistique t, nous rappelle que la cellule D30 contient la valeur de la statistique de Student.

La valeur dans la cellule E30 est la valeur p associée au test de signification de Student. Excel a noté la valeur p dans la cellule E30 en utilisant la notation scientifique. Pour obtenir la valeur décimale, nous déplaçons la virgule décimale de 5 chiffres vers la gauche, obtenant ainsi la valeur 0,0000255. Puisque la valeur $p = 0{,}0000255 < \alpha = 0{,}01$, nous pouvons rejeter H_0 et conclure à l'existence d'une relation significative entre la population étudiante et les ventes trimestrielles.

L'information contenue dans les cellules F28:I30 peut être utilisée pour construire des intervalles de confiance des paramètres de l'équation estimée de la régression. Excel fournit toujours les limites inférieure et supérieure d'un intervalle de confiance à 95 %. Souvenez-vous que dans l'étape 4, nous avions choisi un niveau de confiance de 99 %. En conséquence, la feuille de résultats fournit également les limites inférieure et supérieure d'un intervalle à 99 %. La valeur dans la cellule H30 correspond à la limite inférieure de l'intervalle de confiance à 99 % pour β_1 et la valeur dans la cellule I30 correspond à la limite supérieure. Ainsi, en arrondissant, l'estimation par intervalle de confiance de β_1 est comprise entre 3,05 et 6,95. Les valeurs dans les cellules F30 et G30 fournissent les limites inférieure et supérieure de l'intervalle de confiance à 95 %, allant de 3,66 à 6,34.

Interprétation des résultats de l'analyse de la variance

L'information contenue dans les cellules A22:F26 est un résumé de l'analyse de la variance. Les trois sources de variation sont nommées Régression, Résidus et Totale. Le terme *df* dans la cellule B23 signifie degrés de liberté, le terme *SS* dans la cellule C23 somme au carré et le terme *MS* dans la cellule D23 moyenne des carrés.

Dans la section 12.5, nous avons établi que la moyenne des carrés des résidus, obtenue en divisant l'erreur ou la somme au carré des résidus par ses degrés de liberté, fournit une estimation de σ^2. La valeur dans la cellule D25, 191,25, est la moyenne des carrés des résidus dans le cadre du problème des restaurants Armand. Dans la section 12.5,

nous avons montré qu'un test de Fisher pouvait être utilisé pour tester la significativité d'une régression. La valeur dans la cellule F24, 0,0000255, est la valeur *p* associée au test de Fisher. Puisque la valeur $p = 0,0000255 < \alpha = 0,01$, nous pouvons rejeter H_0 et conclure à l'existence d'une relation significative entre la population étudiante et les ventes trimestrielles. Le terme qu'Excel utilise pour identifier la valeur *p* associée au test de Fisher est *Significance F*.

> Le terme Significance F a plus de sens si vous pensez à la valeur contenue dans la cellule F24 comme au seuil de signification observé pour le test de Fisher.

Interprétation des statistiques de la régression

Le coefficient de détermination, 0,9027, apparaît dans la cellule B17 ; le terme correspondant, R square, est contenu dans la cellule A17. La racine carré du coefficient de détermination fournit le coefficient de corrélation de l'échantillon, égal à 0,9501, contenu dans la cellule B16. Notez qu'Excel utilise le terme Multiple R (cellule A16) pour identifier cette valeur. Dans la cellule A19, le terme Erreur type est utilisé pour désigner la valeur de l'erreur type de l'estimation contenue dans la cellule B19. Ainsi, l'erreur type de l'estimation est égale à 13,8293. Attention : dans la feuille de résultats Excel, le terme Erreur type apparaît à deux endroits différents. Dans la partie Statistiques de la régression, le terme Erreur type fait référence à l'estimation de σ. Dans la partie sur l'équation estimée de la régression, le terme Erreur type fait référence à s_{b_1}, l'écart type de la distribution d'échantillonnage de b_1.

ANNEXE 12.3 ANALYSE DE LA RÉGRESSION AVEC STATTOOLS

Décrivons l'analyse de la régression effectuée en utilisant StatTools dans le cadre du problème des restaurants Armand (cf. fichier en ligne Armand). Commencez par utiliser Data Set Manager pour créer un ensemble de données StatTools en suivant la procédure décrite en annexe du chapitre 1. Les étapes suivantes décrivent comment utiliser StatTools pour obtenir les résultats de la régression.

Étape 1. Cliquer sur **StatTools** dans barre des tâches
Étape 2. Dans le groupe **Analyses**, cliquer sur **Regression and Classification**
Étape 3. Choisir l'option **Regression**
Étape 4. Lorsque la boîte de dialogue apparaît :
Sélectionner **Multiple** dans la boîte **Regression Type**
Dans la section **Variables**,
Cliquer sur le bouton **Format** et sélectionner **Unstacked**
Dans la colonne intitulée **I** sélectionner **Population**
Dans la colonne intitulée **D** sélectionner **Sales**
Cliquer sur **OK**

Les résultats de l'analyse de la régression apparaîtront.

Notez qu'à l'étape 4, nous avons sélectionné Multiple dans la boîte Regression Type. Avec StatTools, l'option Multiple est utilisée à la fois pour des régressions linéaires simples et des régressions multiples. La boîte de dialogue StatTools – Regression contient plusieurs options plus avancées pour effectuer des estimations par intervalle de prévision et représenter des graphiques des résidus. L'aide de StatTools fournit des informations sur l'utilisation de ces options.